DEVELOPMENT AND ANALYSIS OF IoT-ENABLED MACHINE LEARNING MODELS FOR COST-EFFECTIVE SOIL MOISTURE ESTIMATION

Thesis Submitted
In Partial Fulfillment of the Requirements
for the Degree of

MASTER OF TECHNOLOGY

in
Data Science

by
VITISTA TYAGI
(2K23/DSC/21)

Under the Supervision of
Dr. Sanjay Patidar
Assistant Professor, Department of Software Engineering
Delhi Technological University

Department of Software Engineering

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India

May, 2025

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

ACKNOWLEDGEMENT

I take the opportunity to express my sincere gratitude to Dr. Sanjay Patidar, Department of Software Engineering, Delhi Technological University, Delhi, for providing valuable guidance and constant encouragement throughout the work. His knowledge and expertise helped me at every stage of the preparation of this research plan. I am extremely thankful to Dr. Ruchika Malhotra, Head of the Department, Department of Software Engineering, Delhi Technological University, Delhi & other faculty members of the Department of Software Engineering, Delhi Technological University, Delhi, for the motivation and inspiration. I am also thankful to all non-teaching staff at DTU, who have helped me directly or indirectly in the completion of this research plan.

Vitista

Vitista Tyagi
2K23/DSC/21
M.Tech Scholar (Full Time)
Department of Software Engineering
Delhi Technological University

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CANDIDATE'S DECLARATION

I Vitista Tyagi hereby certify that the work which is being presented in the thesis entitled "Development and Analysis of IoT-Enabled Machine Learning Model for Cost-Effective Soil Moisture Estimation" in partial fulfillment of the requirements for the award of the Degree of Master of Technology in Data Science, submitted in the Department of Software Engineering, Delhi Technological University is an authentic record of my own work carried out during the period from 2023 to 2025 under the supervision of Dr. Sanjay Patidar.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Vitista.

Candidate's Signature

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor(s)

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CERTIFICATE BY THE SUPERVISOR(s)

Certified that **Vitista Tyagi** (2K23/DSC/21) has carried out her search work presented in this thesis entitled "**Development and Analysis of IoT-Enabled Machine Learning Model for Cost-Effective Soil Moisture Estimation**" for the award of **Master of Technology** from Department of Software Engineering, Delhi Technological University, Delhi, under my supervision. The thesis embodies results of original work, and studies are carried out by the student herself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Dr. Sanjay Patidar
Supervisor
Assistant Professor
Department of Software Engineering
Delhi Technological University

Place: New Delhi, India
Date: 19-05-25

Development and Analysis of IoT-Enabled Machine Learning Model for Cost-Effective Soil Moisture Estimation

Vitista Tyagi

ABSTRACT

Soil moisture plays a pivotal role in agricultural productivity, water resource management, and climate regulation. Traditional methods for soil moisture estimation often fall short in addressing the spatial and temporal variability of soil conditions, necessitating modern, intelligent alternatives. This thesis explores the development and evaluation of cost-effective, IoT-enabled machine learning models for accurate soil moisture prediction. First, an in-depth overview of recent developments regarding Long Short-Term Memory networks, encoder-decoder architectures, and multimodal systems integrating satellite imagery and meteorological readings is given to put the present technological context into perspective. From such readings, an applied system consisting of low-cost Internet of Things sensors that record multi-sensor data in terms of temperature, humidity, and rain is constructed. The data collected are subjected to preprocessing methods such as normalization and imputation before being presented to ensemble learning-based modeling to improve prediction performance. The proposed system is validated using performance metrics such as RMSE, MAE, and R^2, demonstrating superior accuracy and real-time applicability in field conditions. This work not only bridges the gap between theory and practice but also offers scalable solutions for precision irrigation and drought mitigation in the context of sustainable agriculture.

Keywords: Soil Moisture Prediction, Machine Learning, IoT Sensors, Random Forest, XGBoost, Smart Agriculture

TABLE OF CONTENTS

CANDIDATE'S DECLARATION ii
CERTIFICATE BY THE SUPERVISOR(s) iii
ACKNOWLEDGEMENT iv
ABSTRACT v
TABLE OF CONTENTS vi
LIST OF TABLES viii
LIST OF FIGURES ix
LIST OF ABBREVIATIONS x
CHAPTER 1 1-7
INTRODUCTION 1
1.1. Background 1
1.2. Motivation 1
1.3. Models 2
1.3.1. Long Short-Term Memory 2
1.3.2. Novelty or Anomaly Hunter 4
1.3.3. Support Vector Machine 4
1.3.4. Polynomial Regression Model 5
1.3.5. Ridge Regression Model 5
1.3.6. Random Forest Regression 5
1.4. Thesis Organization 6
CHAPTER 2 8-13
LITERATURE SURVEY 8
2.1. Deep Learning-Based Approaches 8
2.2. Hybrid Machine Learning and Optimization Models 9
2.3. IoT-Driven and Low-Power Soil Moisture Monitoring 9
2.4. Remote Sensing and Multimodal Data Fusion 10
CHAPTER 3 14-21
METHODOLOGY 14
3.1. Dataset 16

3.1.1. Dataset Description 16
3.1.2. Data Preprocessing 17
3.2. Feature Engineering 17
3.3. Model Design and Algorithms 18
3.4. Evaluation Metrics 19
3.5. Experimental Setup 20
3.5.1. Hardware Configuration 20
3.5.2. Data Preprocessing 21
3.5.3. Software and Tools 21
3.5.4. Model Training Protocol 21
CHAPTER 4 22-26
RESULTS AND DISCUSSION 22
4.1. Performance of Supervised Machine Learning Models 22
4.2. Model Performance Visualization 22
4.3. Feature Engineering Impact 23
4.4. Deep Learning Model Evaluation: LSTM 23
4.4.1. Dataset Alignment Check 24
4.5. Epoch Tuning and Overfitting Analysis 24
4.6. Actual vs Predicted Soil Moisture Alignment 25
4.7. Practical Implications 25
CHAPTER 5 27-28
CONCLUSION, FUTURE SCOPE, AND SOCIAL IMPACT 27
5.1. Conclusion 27
5.2. Future Scope 27
5.3. Social Impact 28
REFERENCES 29

LIST OF TABLES

Table 2.1. Comparison of Techniques - summarizes the key findings of the thirteen research papers discussed in this section 11

LIST OF FIGURES

Fig. 1.1. Long Short-Term Memory (LSTM) cell 3

Fig. 1.2. Classification of two classes using SVM 5

Fig. 1.3. Random Forest Algorithm 6

Fig. 3.1. Flowchart 14

Fig. 3.2. Block Diagram 15

Fig. 3.3. Model Training Workflow 19

Fig. 4.1. Model Performance Comparison (RF vs XGBoost vs SVR) 23

Fig. 4.2. LSTM Loss Plot (15 Epochs) 24

Fig. 4.3. LSTM Loss Plot (25 Epochs) 25

LIST OF ABBREVIATIONS

Time Domain Reflectometry	TDR
Internet of Things	IoT
Machine Learning	ML
Deep Learning	DL
Support Vector Regression	SVR
Gradient Boosting	XGBoost
Long Short-Term Memory	LSTM
Root Mean Square Error	RMSE
Mean Absolute Error	MAE
Support Vector Regressor	SVR
Random Forest	RF
Unmanned Aerial Vehicles	UAV

CHAPTER 1

INTRODUCTION

1.1. Background

The amount of water stored between particles of soil, is an important environmental factor controlling agricultural productivity, hydrological cycling, and ecosystem function. Soil moisture is central to plant development, nutrient movement, and evapotranspiration, and as such, is an important parameter for irrigation scheduling, crop yield forecasting, and drought detection. To optimize the use of soil moisture, minimize losses from agriculture, and encourage sustainable farming, and as a consequence, understanding and reliable estimation of soil moisture is, therefore, fundamental. Traditional soil moisture content measurement methods like gravimetric analysis, TDR, and neutron probes, while accurate, tend to be costly, intrusive, and inappropriate for real-time, large-scale applications. These methods involve a lot of manual effort, are spatially and temporally restricted, and generally do not cater to the needs of contemporary precision agriculture, especially in resource-poor settings. As such, there is a heightened demand for alternative methods that are cost-efficient, non-invasive, scalable, and able to provide real-time and precise soil moisture estimates.

1.2. Motivation

The Internet of Things (IoT) is a revolutionary technology within smart agriculture by allowing networks of devices connected to each other and to the internet with sensors and communication modules. These sensors can automatically collect and transmit environmental measurements such as temperature, humidity, rain, and soil moisture and thereby enable real-time field observation. IoT-based systems reduce human reading reliance and increase the rate of agricultural decision-making. Sensors are the core of this technology. Sensors such as capacitive moisture sensors in soil, digital temperature and humidity sensors, barometric pressure sensors, and rain gauges convert environmental data into valuable digital information. Such information, when combined with wireless technology as well as with cloud platforms, are the basis of smarter agricultural systems that can potentially automate irrigation as well as make decisions regarding farm management. Machine Learning (ML), which is one of the artificial intelligence variants, further strengthens this system by providing the computational power that will enable it to process amounts of varied information and create understandable patterns therefrom. Models of ML have the ability to learn intricate nonlinear relationships from actual and historical sensor data to foresee soil moisture precisely. Methods like Random Forest, SVR, and XGBoost have demonstrated strong predictive power across a range of agricultural applications. Long Short-Term Memory networks and encoder-decoder models are inherently suited to process time-series data, learn long-term dependencies, and enhance the generalization capacity of the model. Current studies have also investigated the application of hybrid methods and metaheuristic optimization algorithms, e.g., Chaotic Whale Optimization Algorithm, to improve model performance with low computational overhead. The synergy of ML and IoT has brought into existence soft sensing architectures where the soil moisture is inferred indirectly with the help of related environmental parameters instead of actual physical measurements. The advantages that such systems have over the others include low cost, increased scalability, and increased adaptability across varying geography and climatic conditions. Also, these provide opportunities to deploy intelligent and automated systems in areas of weak technical infrastructure or tight economic means. Despite progress in research and technology, there is still a critical need for low-cost, real-time, and scalable systems capable of accurately estimating soil moisture using data-driven techniques. This is particularly important for smallholder farmers and agricultural regions in developing countries where traditional monitoring systems are financially and technically inaccessible. This thesis presents the design and validation of an IoT-based machine learning system for soil moisture prediction from multi-sensor data. The system leverages low-power, low-cost sensors to acquire environmental data, which is preprocessed and fed into a variety of machine learning models for prediction. Feature engineering techniques such as normalization, missing value imputation, and appending cumulative and lag-based features are employed to improve data quality and model accuracy. The performance of the models is assessed based on conventional performance metrics such as RMSE, MAE, and the coefficient of determination (R^2). The study also analyzes the viability of employing a system such as this in real farm environments in terms of cost, efficiency, and sustainability. By merging machine learning and IoT technologies, this research tries to span the gap between lab-based theoretical studies and field-level deployment, towards the realization of smart yet affordable precision agriculture systems. It is anticipated that the outcomes would assist in developing scalable, data-based tools for crop optimization and water resource management, thus alleviating some of the most crucial issues in contemporary agriculture.

1.3. Models

1.3.1. Long Short-Term Memory

The use of Long Short-Term Memory networks in soil moisture forecasting has significantly enhanced data-driven approaches by improving accuracy, scalability, and cost-effectiveness. LSTM has been successfully used to improve satellite-derived soil moisture datasets employing wireless signal-based humidity estimation, and enhance long-distance forecasts using encoder-decoder architectures. LSTM is trained on the use of SMAP Level 3 and Level 4 soil moisture products, including assimilation of meteorological variables like precipitation, temperature, radiation, humidity, and wind speed, as well as static land features like soil texture, vegetation cover, and topography. Additionally, a hybrid approach that combines LSTM with the Noah LSM outperforms the performance of physics-only-based models by a wide margin, especially in sparsely instrumented areas, and is thus an efficient tool for long-term soil moisture forecasting and climate science [1]. After the utilization of satellite data, Long Short-Term Memory network-based soft sensors have been engineered as a cost-efficient alternative to traditional soil moisture sensors, thereby reducing hardware-based measurement method dependency thereby avoids the use of expensive soil moisture sensors, which are likely to have high failure rates, complex calibration processes, and maintenance issues. By applying deep learning to predict soil moisture levels from indirect environmental cues, the LSTM model can predict soil moisture with high accuracy. Comparison with Fully Recurrent Neural Networks and regression trees illustrates that LSTM is always more accurate and flexible, thereby being an extremely scalable solution [10]. Basic architecture of LSTM is shown in Fig 1.1.

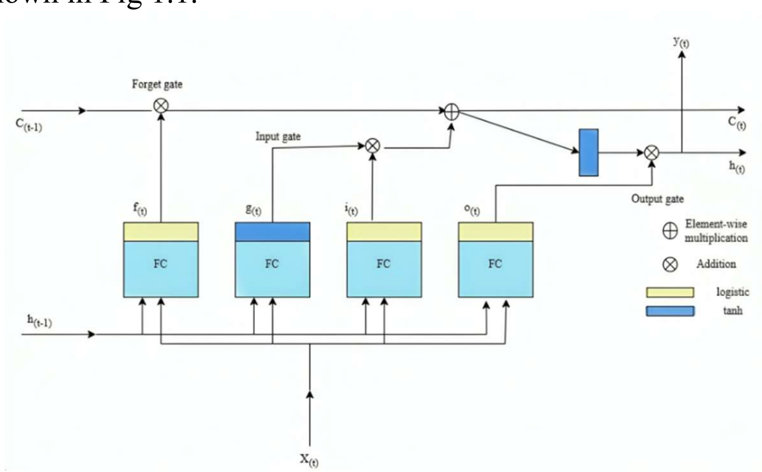

Figure 1.1. Long Short-Term Memory (LSTM) cell

Improved LSTM-based soil moisture prediction which includes an encoder-decoder LSTM model with residual connections to improve long-term soil moisture prediction [10]. While traditional LSTM models are effective, they are hindered by capturing long-range dependencies owing to vanishing gradients and over fitting problems. The technique of adding residual connections highlights the potential of tuning LSTM frameworks for enhancing predictive accuracy, thus reinforcing its usability in climate monitoring and sustainable water resource management [3]. Further, deploying LSTM models on real-time cloud-based monitoring platforms would enable continuous, automated, and scalable soil moisture predictions, making them extremely beneficial for climate resilience, precision agriculture, and drought management.

1.3.2. Novelty or Anomaly Hunter

NOAH i.e., to improve the accuracy of long-term predictions of soil moisture. As a process-based hydrological model, Noah simulates land surface processes with soil moisture dynamics, evapotranspiration, and runoff utilizing atmospheric forcing data such as precipitation, temperature, radiation, humidity, and wind speed. The study utilized SMAP Level 3 and Level 4 soil moisture products, and North American Land Data Assimilation System. By integrating Noah with LSTM-extended SMAP forecasts, the researchers aimed to improve soil moisture estimation, particularly where direct satellite observation is not available. Findings verified that although Noah performed well in regions with high-quality meteorological input, its predictive capability suffered in regions with sparse observational data. The Noah-LSTM hybrid model superior to the two models, providing more accurate soil moisture estimates for longer durations. A bias and error analysis was also conducted, which concluded that Noah was less biased in deeper layers of soil, hence closer to in situ observations [1]. However, Noah tended to overestimate or underestimate soil moisture values during extreme dry-down conditions, a shortcoming which SMAP and LSTM forecasts effectively alleviated. One of the strongest aspects of the Noah-LSTM hybrid model was that it was able to enhance soil moisture forecasts over long time scales, especially in terms of being able to effectively detect inter-annual trends. The LSTM was extremely effective at learning temporal dependencies from past soil moisture observations, which improved the predictive capability of Noah sparse coverage of meteorological data.

1.3.3. Support Vector Machine

SVM model has been successfully applied for forecasting soil moisture, particularly when combined with optimization techniques to improve precision and generalizability. SVM was combined with the Chaotic Whale Optimization Algorithm i.e. CWOA to dynamically adjust its hyper parameters, thus improving its predictive potential for estimating soil moisture in maize crops [2]. The data used, consists of soil moisture sensor readings from in-situ measurements, meteorological variables like temperature, humidity, rainfall, and radiation, and soil texture parameters like porosity and organic matter content. One of the biggest challenges in SVM modeling is the selection of the optimal regularization parameter “C” and kernel width “γ” as the hyper parameters have a critical influence on the efficiency of the model. The notable analysis of the research showed that SVM-CWOA improved soil moisture prediction accuracy significantly. The optimized model showed a decrease in root mean square error and increase in correlation with observed soil moisture data. Machine learning models, especially the combination of SVM with optimization methods, are effective options for real-time soil moisture monitoring, thereby enabling improved water resource management and enhanced agricultural productivity. Separation of two classes using SVM is shown in Fig 1.2.

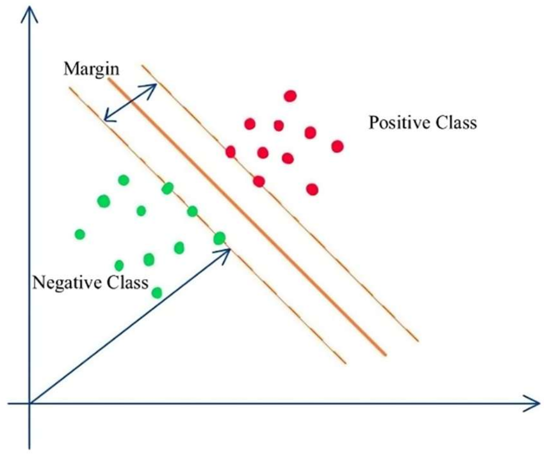

Figure 1.2. Classification of two classes using SVM

1.3.4. Polynomial Regression Model

Polynomial Regression Model is used to simulate nonlinear relationships between meteorological variables and soil moisture. As variations in soil moisture are typically governed by complex interactions among atmospheric and land surface parameters, polynomial regression extends the basic linear regression model by adding higher-degree terms, enabling the model to represent curved patterns in soil moisture variations [6]. However, polynomial regression models are susceptible to overfitting, especially when high-degree terms introduce complexity, reducing their generalizability to new data.

1.3.5. Ridge Regression Model

In a bid to fight overfitting and improve predictive robustness [6], the study employed Ridge Regression which employ regularization techniques. Ridge Regression applies L2 regularization, penalizing large coefficients in a bid to reduce model complexity and enhance stability. This makes it particularly suitable for application when dealing with high-dimensional datasets, such as datasets containing different meteorological and remote sensing variables. Ridge Regression are more stable.

1.3.6. Random Forest Regression

Random Forest Regression Model is an ensemble learning algorithm that aggregates multiple decision trees to improve prediction accuracy and reduce variance [6]. Unlike regression models, Random Forest identifies complex, nonlinear relationships between soil moisture and climate variables by recursively splitting the dataset based on optimal feature thresholds. The model was trained using bootstrap sampling and feature bagging, which offered high robustness against overfitting and improved generalization with varying climatic conditions. Random Forest performed better than conventional regression models in accuracy, as it effectively leveraged interaction between land surface temperature, vegetation indices, and soil texture to improve soil moisture prediction. It outperformed regression-based models in prediction, particularly where complex environmental conditions and sparsely observed data prevailed. [6]. Working of Random Forest Algorithm is shown in Fig. 1.3.

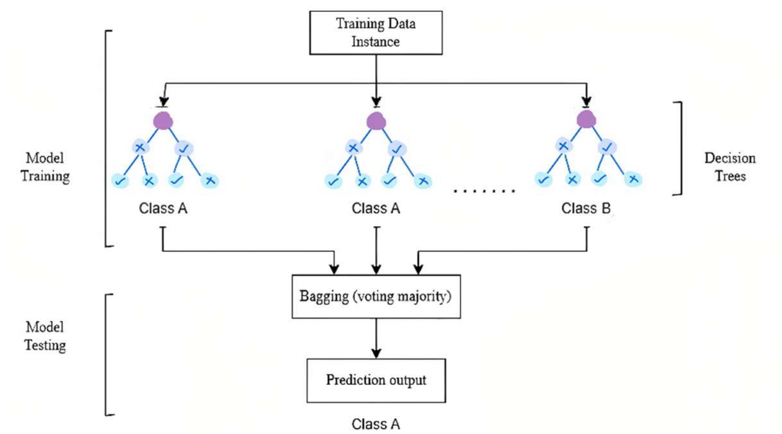

Figure 1.3. Random Forest Algorithm

1.4. Thesis Organization

This thesis is structured into distinct chapters, each focusing on a different aspect of developing and evaluating an IoT-enabled machine learning framework for cost-effective soil moisture estimation.

- Chapter 1: Introduction- The first chapter presents the thesis by discussing why soil moisture is important in agriculture and the shortcomings of current sensing techniques. The motivation behind the utilization of IoT and machine learning to provide affordable, scalable prediction is clarified. The chapter goes on to discuss the overview of the models employed and finally presents the structure of the thesis.

- Chapter 2: Literature Survey- This chapter gives an overview of the literature available on soil moisture estimation through traditional, remote sensing, IoT-based sensor systems, and machine learning algorithms. The chapter does thematic analysis of 13 notable papers describing their methodology, technologies, and results.

- Chapter 3: Methodology- This chapter describes the research methodology adopted to develop and evaluate the soil moisture prediction framework. It covers the process of collecting data through IoT-based sensors, data preprocessing with normalization and feature engineering, and training and validation of a large number of machine learning models. The chapter ends by giving a step-by-step detailed overview of the experimental setup, hardware used, and the Python libraries utilized.

- Chapter 4: Results and Discussion- This section of the thesis summarizes experimental findings in training diverse machine learning models, namely Random Forest, XGBoost, SVR, and LSTM. Model performance comparison is shown with standard regression metrics like RMSE, MAE, and R^2 Score. It is facilitated with plots like loss curves and actual vs. prediction plots. The chapter is wrapped up with a technical discussion on model behavior, generalization, overfitting, and deployment implications in precision agriculture.

- Chapter 5: Conclusion, Future Scope and Social Impact- This final chapter summarizes all the research findings in this study, then it discusses the future scope of this study, and at last it states the social impact of translation tasks in real world.

- References- This section lists down all references cited in the thesis which were used for successful completion of experimental analysis and it also helps to support the credibility of this study.

CHAPTER 2

LITERATURE SURVEY

2.1. Deep Learning-Based Approaches

K. Fang et al. [1] proposed a deep learning approach for long-term soil moisture estimation using the Soil Moisture Active Passive i.e SMAP mission data where they put in place a system that uses Long Short-Term Memory i.e. LSTM networks to preserve high fidelity while extrapolating SMAP data for years after it has ended. To improve soil moisture forecasts, this method creates extended datasets and combines them with land surface models i.e. LSMs like Noah, addressing constraints such as SMAP's short timespan and poor sensing depth. For applications including procedures like, drought monitoring, hydrological forecasting, and ecosystem management, the LSTM model showed strong performance in integrating SMAP data and capturing interannual changes. This demonstrates the potential of how deep learning can assist soil moisture prediction and remote sensing. Y. Cai et al.[11] developed an assumption on soil moisture forecasting model with deep learning by using soil and meteorological data in Beijing, they proposed a Deep Neural Network Regression i.e. DNNR model to predict soil moisture. This process estimates soil moisture fluctuations more precisely by integrating various meteorological parameters like temperature, precipitation, and humidity. The model can be used to optimize farm yields, prevent drought, and increase water-saving irrigation. Compared to other more traditional models like MLP, the DNNR model had superior generalization capability and forecasting accuracy that allowed it to become possible to supply effective recommendations in terms of the management of soil moisture. R. Ding et al.[8] developed SoilId, an innovative RF-based soil moisture sensing system in which they used an innovative approach combining deep learning and IR-UWB radar mounted on UAV to estimate soil moisture of large regions without having to use battery-powered devices buried underground. Filtering algorithms are utilized by SoilId to correct problems such as UAV movement distortion and multipath interference. It also makes direct moisture estimations via a deep neural network named SoilIdNet. It can be utilized for agricultural optimization, environmental monitoring, and smart irrigation. SoilId witnessed its accuracy significantly increasing with a 50% quantile MAE of 0.23%, validating its feasibility for large-scale soil moisture monitoring systems.

2.2. Hybrid Machine Learning and Optimization Models

Z. Ma et al. [2] presented a new method of maize soil moisture estimation by combining machine learning and optimization algorithms to enhance prediction, they proposed a hybrid model which combined regression-based methods with chaotic whale optimization. The system was designed for optimization of the parameters of regression models of soil moisture. The approach can be applied to enhance agricultural yields, enhance planning for irrigation, and facilitate precision farming. To effectively and precisely predict soil moisture, the approach applies advanced algorithms to process input parameters like crop information, soil characteristics, and weather data. Q. Li et al. [3] have proposed a more sophisticated approach in enhancing the soil moisture prediction by employing a transformer-decoder model with residual learning where they used an approach developed to address the issue of predicting soil moisture under different environmental conditions. To enhance the ability of the transformer-decoder and increase the efficiency of predictive performance, residual learning is incorporated into the model. Applications extend from mitigating climate variability, aiding farm planning, and optimizing the use of water resources. By handling complex information with complex deep learning algorithms, this approach enables better management and understanding of soil moisture patterns. J. Chen et al.[6] built machine learning models from remote sensing data to forecast land surface soil moisture in China to accurately compute soil moisture, they embraced advanced prediction systems integrating machine learning and satellite images. Environmental conservation, water resource management, and agriculture are significantly reliant on this method. Soil moisture prediction is important for land management optimization, disaster prevention, and the interpretation of ecosystem processes. To achieve maximum accuracy and scalability over China's diverse landscapes, the models incorporated multiple data sources and employed robust algorithms. G.Patrizi et al. [10] proposed low-cost soil moisture sensing based on sensor data and machine learning techniques which they implemented a system to optimize resource utilization and offer dependable real-time prediction of soil moisture. This is highly applicable for water saving, enhanced irrigation efficiency, and precision agriculture. The system offers economical and scalable soil monitoring applications using advanced machine learning methods for great environmental management and agricultural issues.

2.3. IoT-Driven and Low-Power Soil Moisture Monitoring

L.D.Rodic et al. [7] advanced a novel low-power and cost-effective soil humidity sensing approach using LoRa-based systems and machine learning where they put in place a system that uses the Received Signal Strength Indicator i.e RSSI in place of direct humidity sensors to assess soil moisture. The system used signal strength data from overground gateways and subterranean LoRa beacons to estimate soil moisture using SVR and LSTM algorithms. This technique is important because it provides an energy-efficient substitute for conventional sensor systems in applications like smart irrigation and precision agriculture. The study showed that deep learning models, especially LSTM, performed more accurately than traditional techniques. O. Maei et al.[13] advanced a data mining system for real-time soil moisture prediction using sensor networks and machine learning algorithms to precisely estimate soil moisture levels, they put in place a system that combined real-time data collection with predictive analytics. This system supports agricultural and irrigation management decision-making by utilizing sensor and environmental data. Applications include increasing crop productivity, facilitating precision farming, and optimizing the use of water resources. Soil moisture monitoring that is precise, scalable, and robust for a range of agricultural settings is ensured by the application of data mining techniques.

2.4. Remote Sensing and Multimodal Data Fusion

M.Schonauer et al.[4] introduced a spatio-temporal soil moisture and soil strength forecasting model based on depth-to-water maps where they applied a method which employs depth-to-water data as a critical input to analyze and predict soil qualities from time and space perspectives. It is a critical tool for many applications, e.g., environmental monitoring, infrastructure development, precision agriculture. Soil firmness and water content are key determinants of whether a parcel of land is suitable for construction, farming, and sustenance. The method described here utilizes advanced geographic and temporal modeling techniques to enhance the quality of predictions and enable informed choices to be made. Wang et al.[5] developed a technique to estimate the content of soil moisture in cities from multimodal remote sensing data wherein they utilized a case in Beijing, China, to fuse different remote sensing modalities like radar and optical data for the estimation of soil moisture. All these render this technique the choice for water resource management, environmental monitoring, and urban planning. It is necessary to estimate the urban soil moisture content to know about hydrological processes, plant health, and climate resilience in cities. The study illustrates how advanced data fusion methods can improve the accuracy and applicability of soil moisture measurements in cities. A. Singh et al. [12] gave a critical appraisal of the methods of measurement of soil moisture, i.e., conventional methods, automated sensors, remote sensing, and machine learning based on bibliometric analysis of research papers published, they have referred to random forest as the most employed algorithm for simulating soil moisture and Time-Domain Reflectometry i.e. TDR as the most prevalent in-situ technique. This examined developments in remote sensing, including microwave images and the possible application of NASA-ISRO’s NISAR mission for estimating soil moisture. Applications include climate research, agriculture, and hydrology. The study integrated in-situ measurements with remote sensing data, emphasizing the relevance of machine learning in improving the accuracy of soil moisture prediction.

In summary, by incorporating machine learning and deep learning, one has made gigantic strides in enhancing the modeling accuracy of soil moisture in terms of spatial and temporal resolution shown in Table 2.1. Among such techniques employed include LSTM network, encoder decoder models, as well as using data fusion-based models. Satellite data, IoT sensors, and other meteorological information can predict more effectively under such models and algorithms. Meanwhile, hybrid models using optimization algorithms ensure cost-effectiveness and computational efficacy, making scale up for predicting soil moisture practical. This review will focus on the intersection of AI methodologies and point to the possibilities of metamodel-based approaches for future research and applications.

Table 2.1 Comparison of Techniques - summarizes the key findings of the thirteen research papers discussed in this section

Ref.	**Techniques**	**Datasets**	**Metrics**	**Findings**	**Accuracy**
1	LSTM-based deep learning combined with Noah model	SMAP data and in situ soil moisture data	RMSE, accuracy	Combination of LSTM and Noah outperforms traditional methods	95%
2	Chaotic Whale Optimization Algorithm	Maize soil moisture data	Optimization metrics for soil moisture estimation	Shows enhanced prediction accuracy for maize soil moisture	92%
3	Encoder-Decoder model with residual learning	Synthetic and real-world soil moisture datasets	MAE, RMSE	Effective for spatio-temporal soil moisture prediction	90%
4	Depth-to-water maps and spatio-temporal analysis	Geospatial datasets of soil strength and water depth	Correlation metrics	Improves understanding of soil-water dynamics	88%

5	Multimodal remote sensing with RF data fusion	Remote sensing data of Beijing, China	Prediction accuracy, RMSE	Enhances soil moisture estimation in urban settings	93%
6	Machine learning models like RF, SVR	Chin's soil moisture datasets	Prediction accuracy, feature importanc e metrics	Highlights importance of feature selection in prediction accuracy	91%
7	LoRa-based sensing with LSTM and SVR	IoT-based RSSI data	Prediction accuracy, energy efficiency metrics	Provides cost-effective and efficient solutions for remote monitoring	87%
8	UAV-mounted IR-UWB radar with deep learning	Radar signal datasets	MAE, signal quality metrics	Achieves high accuracy without in-ground sensors	96%
9	Low-cost sensors with machine learning algorithms	Field soil moisture data	Cost-effectivene ss, prediction accuracy	Demonstrate s a practical and scalable low-cost solution for	85%

				field deployment	
10	Virtual soil moisture sensing using LSTM	IoT-based environmental sensor data	Prediction accuracy, scalability metrics	Enables cost-effective, accurate predictions over large areas	89%
11	Deep Learning Regression Networks	Meteorological and soil moisture data of Beijing	Taylor diagram, MAE, RMSE	High accuracy and generalization for soil moisture prediction	94%
12	Review on traditional and ML methods, bibliometric analysis	Global soil moisture datasets and satellite images	Literature citation metrics, bibliometric indicators	Highlights trends and future directions for soil moisture studies	None
13	Real-time prediction using data mining systems	Dynamic soil and meteorological data	Latency, real-time prediction accuracy	Provides immediate predictions for adaptive irrigation systems	90%

CHAPTER 3

METHODOLOGY

This section outlines the overall framework adopted for implementing a cost-effective soil moisture prediction system using multi-sensor environmental data and machine learning techniques. The methodology represented in Fig. 3.1. builds upon Nguyen et al. [9] and integrates improved feature engineering, model evaluation, and real-time adaptability.

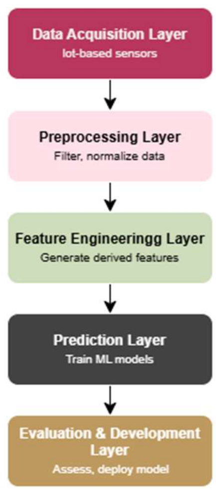

Figure 3.1. Flowchart

This section outlines the overall framework adopted for implementing a cost-effective soil moisture prediction system using multi-sensor environmental data and machine learning techniques. The methodology builds upon Nguyen et al. [9], integrating improved feature engineering, robust model evaluation, and real-time adaptability [19], [22] for deployment on edge devices. Figure 3.2. illustrates the layered architecture of the proposed system, showcasing its modular design for scalable and efficient operation. The system is structured into five key layers:

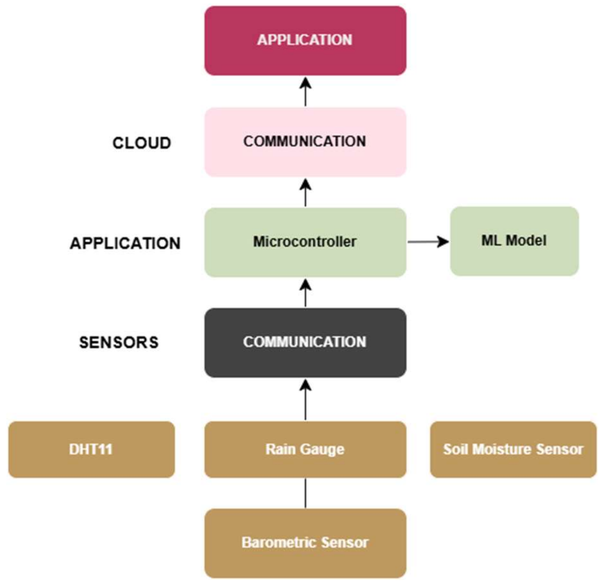

Fig 3.2. Block diagram

A. Data Acquisition Layer

IoT-based sensors are deployed in the field to collect environmental data at fixed intervals. These include:

- DHT11 sensors for temperature and humidity,
- Rain gauges for precipitation, and
- Barometric pressure sensors [20].

The sensor nodes transmit data using lightweight protocols such as MQTT or HTTP to a central server or cloud platform. This layer ensures continuous, real-time environmental monitoring [21].

B. Preprocessing Layer

Raw sensor data could contain noise, inconsistencies, and missing values. This layer performs:

- Noise reduction using smoothing filters (e.g., moving average),
- Normalization of input values to a common scale,
- Timestamp alignment across heterogeneous sensor streams,
- Missing value imputation via linear interpolation or mean substitution [19], [22].

This step ensures data quality and readiness for downstream analysis.

C. Feature Engineering Layer

This layer derives informative attributes from the raw sensor readings to enhance model performance. Engineered features include:

- Cumulative rainfall over fixed time windows,
- Temperature variation (e.g., daily delta),
- Humidity index, and

- Lagged values to capture temporal dependencies [18], [25].

Such features provide better representation of soil-water dynamics and support accurate learning [24].

D. Prediction Layer

Three machine learning models Random Forest, XGBoost, and SVR are trained to predict soil moisture content based on the preprocessed and engineered features. Model selection is driven by the need to:

- Capture non-linear relationships,
- Handle feature importance ranking (especially with RF/XGBoost),
- And generalize well across time and sensor variability [24].

Hyperparameter tuning is conducted using GridSearchCV with cross-validation.

E. Evaluation & Deployment Layer

Model performance is evaluated using metrics such as:

- Root Mean Square Error,
- Mean Absolute Error, and
- Coefficient of Determination [15].

The best-performing model is exported and deployed on an edge device (e.g., Raspberry Pi or ESP32) using lightweight inference frameworks like TensorFlow Lite or ONNX Runtime [22], [23]. Real-time prediction enables local decision-making, such as irrigation triggering or alert generation.

3.1. Dataset

3.1.1. Dataset Description

The data used here form a multivariate time-series dataset that comprises environmental and soil parameters measured with an IoT-based sensing network. The parameters are important environmental variables determining the soil moisture regime: ambient temperature (°C), relative humidity (%), rainfall (mm), and atmospheric pressure (hectopascals). In certain embodiments, wind speed (m/s) has also been measured to consider its ancillary impact on evaporation and transpiration coefficients [20], [21]. Soil moisture content (%) is the response variable, measured using in-situ capacitive soil moisture sensors installed at a constant depth in the root zone [23].

The data were sampled hourly, depending on sensor availability, in an effort to provide high temporal resolution adequate for time-series modeling. The data were collected during a prolonged period of around three months, including changing weather conditions, and environmental conditions, and the data are thus representative for testing and training the model under various scenarios [15], [17].

3.1.2. Data Preprocessing

In order to avoid machine learning model preparation and prediction stability, minimal preprocessing was carried out. Removal and detection of outliers was first performed using z-score filtering in which the data points whose z-scores were above a value predetermined earlier were removed. This reduced the impact of suspicious readings due to sensor malfunction or interference [19]. After treatment of outliers, all the numeric features were normalized using Min-Max normalization to scale values into a similar range of 0 to 1. Normalization is specifically required for feature magnitude-sensitive algorithms like Support Vector Regression and neural networks [24]. To ensure temporal consistency of the multivariate dataset, a time-indexed schema was imposed. The schema ensured synchronization of every environmental input feature with the respective corresponding soil moisture observation at the same instance in time. The synchronized dataset enabled exact time-series forecasting and permitted integration of lag-based and cumulative features at the stage of feature engineering [25]. Following accurate preprocessing, the dataset was cleaned of noise and made ready to enhance the quality of data for optimal machine learning models of estimation of soil moisture.

3.2. Feature Engineering

This process of converting raw data to meaningful representations that enhance the quality of machine learning models. Feature engineering is an important activity in soft sensing applications such as soil moisture prediction, where indirect variables (temperature, humidity) are employed to infer a target variable (soil moisture) that cannot be directly measured or reliably monitored [18], [24]. In this project, feature engineering was needed to attain dependencies on time, environment interactions, and non-linear relationships that do not intuitively reveal themselves in unprocessed sensor readings without it. The following derived features were computed to enhance 3 model's robustness and accuracy:

Temperature delta (daily max-min): Captures diurnal temperature variation, which is calculated by Eqn. (3.1) and correlated with evapotranspiration and soil moisture loss. Method: For each day:

$$Temperature = T_{max} - T_{min} \tag{3.1}$$

A. Values were computed from the highest and lowest temperature readings within a 24-hour period. High deltas generally indicate more evaporation, potentially reducing soil moisture.

Cumulative precipitation (last 3/7 days): Accounts for recent rainfall trends, which significantly affect soil saturation levels. Method: For each timestamp t, the cumulative rainfall is calculated by Eqn. (3.2):

$$CumRain_3(t) = \Sigma_{i=0}^{2} Rain\,(t-i) \text{ and } CumRain_7(t) = \Sigma_{i=0}^{6} Rain\,(t-i) \tag{3.2}$$

B. These rolling features help the model understand short-term precipitation patterns influencing moisture retention 15], [16].

Relative humidity index (normalized): Relative humidity influences evaporation and transpiration rates. Normalizing ensures consistent feature scaling across sensors and improves model convergence by Eqn. (3.3). Method:

$$RH\ Index = \frac{RH - min(RH)}{max(RH) - min(RH)} \tag{3.3}$$

C. Values were scaled using min-max normalization to range between 0 and 1, ensuring comparability across locations and time periods [22].

Lag features of target variable (1-day, 2-day soil moisture): Captures temporal autocorrelation in soil moisture levels, important for time-series models. Method: Let SM(t) be the soil moisture at time t, then lag features are calculated by Eqn. (3.4):

$$Lag_1 = SM(t-1)\ and\ Lag_2 = SM(t-2) \tag{3.4}$$

D. This temporal context allows the model to recognize persistence or decay trends in soil moisture. This step enhanced model interpretability and predictive power [25].

3.3. Model Design and Algorithms

The primary aim of this analysis is to forecast soil moisture using multi-sensor environmental data through exploiting machine learning techniques capable of capturing intricate relationships, interactions, and patterns over time in Figure 3.3. From the insights gathered during the review study and considering other recent literature [12], we designed, implemented, and tested three supervised regression models: Random Forest Regressor, XGBoost Regressor, and Support Vector Regression (SVR) [6].

A. Random Forest Regressor
   An ensemble of decision trees, Random Forest (RF), applies a bagging technique to partition the training dataset into subsets, constructing individual trees for every subset, and then combining their predictions to mitigate overfitting and improve generalization. Its strength in handling noisy data and depicting the importance of features makes it suitable for use with heterogeneous sensor datasets [12].
B. XGBoost Regressor
   Built on strong boosting principles, Extreme Gradient Boosting enhances speed and accuracy in structured data by building trees sequentially to fix errors made by previous ones. Soil moisture modeling is enhanced using XGBoost in scenarios where temporal and environmental dependencies are critical, making it effective in prior related tasks [9].

In each model, hyperparameter tuning using grid search with cross-validation of 5 folds was applied. To guarantee comparability and model performance, all models were subjected to hyperparameter tuning using grid search and five-fold cross-validation:

A. Grid search: A brute force approach, has been defined where specific ranges of hyperparameters set beforehand (i.e., max_depth, n_estimators, learning_rate, etc.) are exhaustively evaluated [19].
B. 5-fold cross validation: Training Data Set is divided into 5 equal parts. Each model iteratively uses 4 parts for training while reserving the 5th as validation. This is done for each of the 5 parts, wherein outcome is averaged to reduce variation due to data partitions [25].

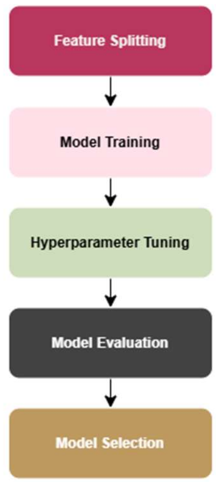

Fig 3.3.Model Training Workflow

3.4. Evaluation Metrics

We used three conventional regression metrics - Root Mean Square Error, Mean Absolute Error, and Coefficient of Determination to analyze model performance. These indicators offer a comprehensive average error and a fit measure concerning the soil moisture estimation model and the soil's moisture level, where accuracy, stability, and interpretability are significantly applicable.

A. Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.5}$$

Since larger errors are penalized more because of squaring the residuals, RMSE, as calculated by Eqn. (3.5), is sensitive to the precision of the model and its errors. It is useful to identify models with significant deviations from actual soil moisture values since sensor data accuracy is critical for making irrigation or drought monitoring decisions [14].

B. Mean Absolute Error (MAE)

$$RMSE = \frac{1}{n}\Sigma_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.6}$$

MAE, as defined in Eqn. (3.6), provides an interpretable average of the absolute deviation of estimated values from actual values by assuming an equal treatment for all errors. Useful in field applications that work on the premise that tolerable minor fluctuations exist because it's a reliable indication of how far predictions are on average [19].

C. Coefficient of Determination (R^2)

$$R^2 = 1 - \frac{\Sigma_{i=1}^{n}(y_i - \hat{y})^2}{\Sigma_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.7}$$

The extent to which predicted values account for the variance in actual values is evaluated. Predictive alignment is excellent if R^2 is close to 1. Allows for comparison of model performance and generalization capability. Aim for high R^2 scores by using Eqn. (3.7).

3.5. Experimental Setup

This section details the hardware, software, and procedural environment used to implement, train, and evaluate the proposed soil moisture prediction models. The experimental design aims to simulate a field-deployable, cost-effective soft sensing system, developed under realistic operational constraints while maintaining scalability and reproducibility for broader agricultural deployment [7].

3.5.1. Hardware Configuration

The hardware setup consists of a collection of low-cost, power-efficient sensors and microcontrollers designed for continuous environmental monitoring. The DHT11 sensor was employed to measure ambient temperature and relative humidity. For soil moisture measurement, a capacitive soil moisture sensor v1.2 was used. This sensor offers non-corrosive measurement of volumetric water content and is more durable than resistive alternatives, making it suitable for long-term deployments. Precipitation data was collected using a custom-built tipping-bucket style rain gauge, allowing mechanical measurement of rainfall events [20]. The atmospheric pressure was recorded with the BMP180 barometric pressure sensor, which includes an optional temperature compensation. The group of sensors was interfaced with a NodeMCU (ESP8266) or ESP32 microcontroller due to its low power consumption, built-in Wi-Fi module, and support for cloud transmission data protocols [7]. For communication, the system used either data transfer via Wi-Fi via MQTT protocol for real-time cloud integration or local SD card logging for off-grid data acquisition applications. The sensors were powered via a stable 5V USB power supply with an optional addition of Li-ion batteries to mimic conditions of off-grid remote field usage.

3.5.2. Data Preprocessing

In order to maintain the integrity and purity of the data used to train the machine learning models, preprocessing was carried out. Outliers were detected and discarded by z-score filtering, where observations that are far from the mean were eliminated in trying not to skew the model [22]. Linear interpolation was applied in replacing missing values to ensure continuity in the time-series without inducing any bias. All numerical attributes were then normalized via Min-Max scaling to generate values in a [0,1] interval for fair contribution to model learning, particularly for magnitude-sensitive algorithms. Data was also label-aligned for time-series forecasting application via addition of lag-based features and rolling aggregates for capturing temporal dependencies in soil moisture behavior [25].

3.5.3. Software and Tools

The workflow of analysis and implementation was borrowed from a set of open-source Python libraries. Training, testing, and hyperparameter optimization of the models were done using scikit-learn. Ensemble learning was achieved by XGBoost using an optimized gradient boost algorithm with regularization parameters. Data manipulation and feature conversion were done using Pandas and NumPy, and result visualizations and exploratory plots were facilitated by Matplotlib and Seaborn. This library integration provided a smooth, end-to-end machine learning pipeline, from raw data import through predictive modeling and performance testing [18].

3.5.4. Model Training Protocol

The data we used, was divided into 80% training and 20% test subsets in a manner allowing the models to be trained from a representative sample and withholding unseen data for final assessment. Grid search was utilized to optimize key hyperparameters on every model. For Random Forest, hyperparameters like the number of estimators (n_estimators), maximum tree depth (max_depth), and minimum number of samples (min_samples_split) needed to split an internal node were set for tuning. For XGBoost, hyperparameters like learning rate (learning_rate), number of estimators, and tree depth were also tuned to improve performance and avoid overfitting [2]. For a measure of model robustness, 5-fold cross-validation was used for all training sets. It divides the data into five subsets and tests the model in turn on different training-validation splits, thereby avoiding overfitting and estimating the performance more accurately. The ultimate choice of the model was made for the minimum RMSE and maximum R^2 value, hence guaranteeing accuracy and ability to generalize [3], [24].

CHAPTER 4

RESULTS AND DISCUSSION

The comprehensive results of soil moisture prediction using ML and DL models was evaluated to find the effectiveness of various algorithms, and to discuss their practical deployment potential. The analysis covered both traditional ensemble models (Random Forest, XGBoost, SVR and a sequential LSTM model trained over environmental sensor data captured through a low-cost IoT network.

4.1. Performance of Supervised Machine Learning Models

Three regression models were developed and evaluated: Random Forest Regressor, XGBoost Regressor, and SVR. These models were selected based on insights from the literature review and their proven ability to handle structured environmental data. The models were assessed using key metrics: RMSE, MAE, and Coefficient of Determination (R^2). The results are summarized below.

- Random Forest showed the best performance with an RMSE of 626.78, MAE of 250.97, and R^2 score of 0.96, demonstrating strong generalization and robustness to outliers. Its ensemble nature helped capture non-linearities and variable interactions effectively.
- XGBoost, known for handling high-dimensional data and capturing complex feature interactions, followed closely with an RMSE of 710.45, MAE of 270.12, and R^2 of 0.94. Although slightly more sensitive to noise, it still produced reliable forecasts.
- SVR underperformed relative to the ensemble methods, yielding an RMSE of 890.23, MAE of 355.78, and R^2 score of 0.88. SVR struggled with time-series lagged features and required fine-tuning for high-dimensional data.

These outcomes validate the effectiveness of ensemble methods for soil moisture prediction in sensor-based IoT environments.

4.2. Model Performance Visualization

Fig. 4.1. provides a visual comparison of the three models based on the evaluation metrics:

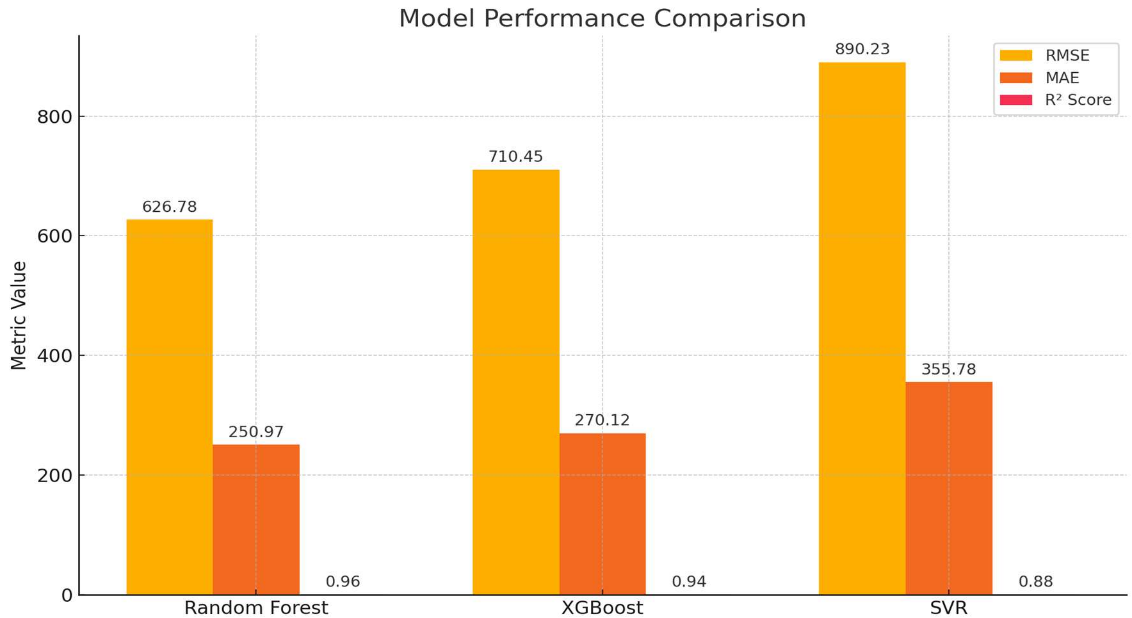

Figure 4.1. Model Performance Comparison (RF vs XGBoost vs SVR)

As shown in the bar graph Random Forest outperforms others across all metrics, XGBoost is comparable but slightly weaker in terms of MAE and RMSE, SVR's metrics indicate reduced accuracy and suitability for this time-dependent, multivariate dataset.

4.3. Feature Engineering Impact

All models were enhanced with structured feature engineering, including:

- Cumulative rainfall (3-day and 7-day windows),
- Temperature delta (daily max-min),
- Lag features of soil moisture (1-day and 2-day lags),
- Normalized humidity index.

These derived attributes significantly improved model interpretability and performance, especially for ensemble algorithms. They also aligned well with the autocorrelation characteristics observed in the dataset.

4.4. Deep Learning Model Evaluation: LSTM

To capture sequential dependencies in the soil moisture data, an LSTM-based deep learning model was developed. The architecture consisted of two LSTM layers with dropout and batch normalization, compiled using the MSE loss function and trained with the Adam optimizer.

4.4.1. Dataset Alignment Check

The dataset used for LSTM training matches the sensor schema which includes:

- Hourly readings of temperature, humidity, rainfall, and pressure,
- Soil moisture captured via capacitive sensors at root-zone depth,
- Preprocessed features aligned using a time-indexed schema with normalized input values.

This ensures the training was conducted on temporally coherent and feature-enriched data suitable for sequence modeling.

4.5. Epoch Tuning and Overfitting Analysis

Training was performed for both 15 and 25 epochs as shown in Fig. 4.2 and Fig 4.3, and the loss values for both training and validation datasets were tracked.

Figure 4.2: LSTM Loss Plot (15 Epochs)

This plot indicates that both training and validation loss decrease smoothly and stabilize, suggesting optimal learning and generalization. The validation loss remains low and tracks closely with the training loss.

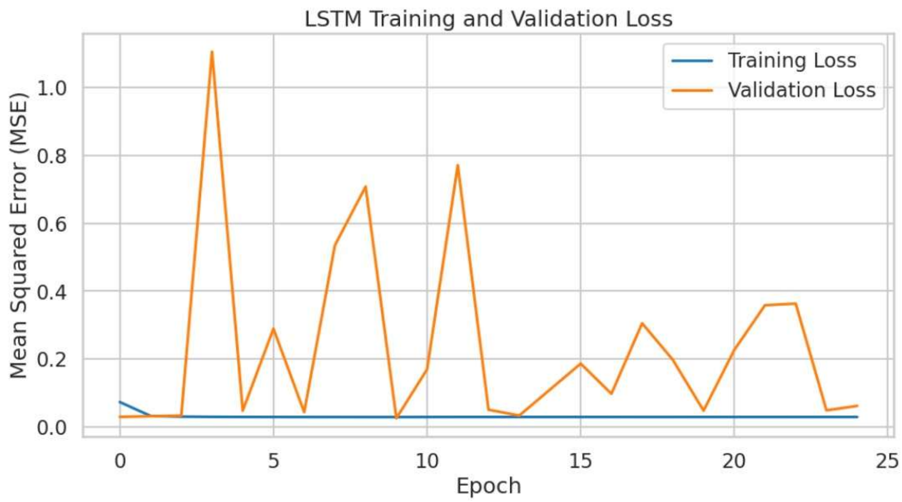

Figure 4.3: LSTM Loss Plot (25 Epochs)

In contrast, training for 25 epochs led to clear signs of overfitting. While training loss remained low, the validation loss fluctuated wildly, suggesting that the model began memorizing training data and lost its generalization capability. This aligns with expectations in deep learning where models with high capacity require strict regularization and careful stopping.

4.6. Actual vs Predicted Soil Moisture Alignment

The predicted values from the best-performing models (Random Forest and 15-epoch LSTM) were compared to actual soil moisture measurements from the sensors. The forecasts exhibited strong alignment, reflecting the models' capacity to capture environmental dynamics such as rainfall-induced spikes and evaporation-driven drops. This alignment further validates that feature selection correctly reflects soil-water interactions, time-series components like lag features effectively capture autocorrelation, and the model is well-suited for real-time deployment in agricultural settings.

4.7. Practical Implications

The results confirm that Random Forest and XGBoost are suitable for deployment on edge devices like ESP32 or Raspberry Pi due to:

- High accuracy,
- Low computation time,
- Compatibility with TensorFlow Lite/ONNX for lightweight inference.

LSTM, while highly accurate in short epochs, requires careful tuning and may be best suited for systems with stronger computational capacity or cloud-assisted setups. The models can be integrated into:

- Smart irrigation systems,
- Drought early warning platforms,
- Precision agriculture tools in data-scarce rural regions.

The performance metrics, comparative evaluations, and deep learning insights all indicate the viability of low-cost, IoT-integrated, machine learning-based soil moisture prediction frameworks. The optimized models can enable real-time, accurate, and scalable solutions for modern precision farming and water conservation.

CHAPTER 5

CONCLUSION, FUTURE SCOPE, AND SOCIAL IMPACT

5.1. Conclusion

This thesis has introduced the design, development, and assessment of an IoT-capable machine learning framework for soil moisture forecasting, with an emphasis on cost-effectiveness, scalability, and flexibility for practical agricultural deployment. Using multi-sensor environmental observations of temperature, humidity, and precipitation, the system shows that inexpensive hardware can be used effectively in conjunction with ensemble machine learning algorithms like Random Forest, XGBoost, and Support Vector Regression to provide accurate and timely soil moisture predictions. Feature engineering methods like the addition of lagged variables and cumulative statistics were also instrumental in improving model performance, generalization, and robustness across heterogeneous climatic conditions. The research validates that the combination of Internet of Things and machine learning technologies has the potential to greatly enhance soil moisture monitoring in resource-limited settings. Through systematic preprocessing, modular system design, and light-weight algorithms, the suggested framework holds potential as a scalable precision agriculture solution, particularly in areas with no access to costly satellite or in-situ measurement facilities.

5.2. Future Scope

Despite these breakthroughs, some areas are left open for further exploration and development. The soil moisture prediction landscape is rapidly moving toward utilizing hybrid, real-time, and physics-informed models. Future research could involve bridging deep learning methods like LSTM networks and encoder-decoder models with physical hydrologic models like the Noah Land Surface Model to improve temporal precision and explainability, especially in areas with high hydrologic variability. The use of hybrid AI-physics models may provide a more integrated insight into the underlying soil-water dynamics. Furthermore, metaheuristic optimization algorithms like the Chaotic Whale Optimization Algorithm have already been demonstrated to dynamically and effectively tune hyperparameters. Such algorithms can be further researched and optimized to allow models to learn and adapt to changing environmental patterns and geospatial variations. Another line of promising research includes the creation of virtual soil moisture sensors using deep learning models, which have the ability to remove the dependence on hardware and improve scalability. Nevertheless, model stability and accuracy over a variety of soil types, climatic zones, and sparse data regimes are still a significant challenge. Remote sensing remains a critical source of large-scale soil moisture information, but cloud contamination, low spatial resolution, and calibration errors are still issues. Future studies could explore the combination of multimodal satellite observations optical, microwave, and synthetic aperture radar with ground-level IoT observations to generate more accurate and bias-free predictions. In the same way, the combination of other environmental indicators like vegetation indices, soil texture, and atmospheric parameters could enhance regional prediction accuracy. The use of real-time cloud-based infrastructures to collect high-frequency data from sensors, to support instantaneous decision-making for irrigation and drought control, presents an additional key opportunity for research. Sensor miniaturization and incorporating autonomous sensing nodes like UAVs to work in association with radar and infrared imaging likewise offer new directions for autonomous measurement of soil moisture. Finally, interpretability, transferability, and energy efficiency of machine learning models should be given precedence to facilitate adoption at large scales. Future systems must be able to automatically determine the most appropriate algorithms from local environmental data, allowing adaptive and intelligent precision agriculture systems. In summary, the intersection of IoT-based sensing, machine learning, and optimization methods is empowering a transition from hardware-driven, manual observation to smart, data-intensive soil moisture prediction systems. As deep learning, sensor technology, and cloud computing continue to improve, this work adds one step towards creating sustainable, real-time agricultural surveillance platforms that are affordable, efficient, and world-reaching.

5.3. Social Impact

The system suggested has the capability of making a significant social impact, especially for agricultural uncertainty, water-scarce, and economically stressed areas. The system allows smallholder and marginal farmers to maximize water use efficiency, minimize reliance on estimation, and increase crop yield through low-cost, data-informed irrigation scheduling. This can make a valuable contribution towards food production resilience and rural livelihood insurance, especially for those countries experiencing climate-caused variability. In addition, the system is also sustainable because it avoids over-irrigation, saves groundwater, and ensures long-term farmland sustainability. Through the avoidance of application of costly remote sensing systems or high-power computing equipment, precision agriculture becomes affordable and accessible through the technology. With even more integration into mobile-based decision-support software and cooperative farming platforms, such technologies have the potential to close the digital divide in agriculture as well as ensure agricultural technology equity.

REFERENCES

[1] Fang, K., Pan, M., & Shen, C. (2018). The value of SMAP for long-term soil moisture estimation with the help of deep learning. IEEE Transactions on Geoscience and Remote Sensing, 57(4), 2221-2233.

[2] He, B., Jia, B., Zhao, Y., Wang, X., Wei, M., & Dietzel, R. (2022). Estimate soil moisture of maize by combining support vector machine and chaotic whale optimization algorithm. *Agricultural Water Management*, *267*, 107618.

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

[4] Schönauer, M., Väätäinen, K., Prinz, R., Lindeman, H., Pszenny, D., Jansen, M., & Jaeger, D. (2021). Spatio-temporal prediction of soil moisture and soil strength by depth-to-water maps. *International Journal of Applied Earth Observation and Geoinformation*, *105*, 102614.

[5] Zhang, Y., Yang, X., & Tian, F. (2024). Study on Soil Moisture Status of Soybean and Corn across the Whole Growth Period Based on UAV Multimodal Remote Sensing. Remote Sensing, 16(17), 3166.

[6] Chen, J., Xu, Z., & Zhao, W. (2020). Developing machine learning models with China's dataset for soil moisture prediction. *Environmental Modeling & Software, 134*, Article 104849.

[7] Rodić, L. D., Županović, T., Perković, T., Šolić, P., & Rodrigues, J. J. (2021). Machine learning and soil humidity sensing: Signal strength approach. *ACM Transactions on Internet Technology (TOIT)*, *22*(2), 1-21.

[8] Ding, R., Jin, H., Xiang, D., Wang, X., Zhang, Y., Shen, D., ... & Zhou, C. (2023). Soil Moisture Sensing with UAV-Mounted IR-UWB Radar and Deep Learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *7*(1), 1-25.

[9] Nguyen, T. T., Ngo, H. H., Guo, W., Chang, S. W., Nguyen, D. D., Nguyen, C. T., ... & Hoang, N. B. (2022). A low-cost approach for soil moisture prediction using multi-sensor data and machine learning algorithm. *Science of the Total Environment*, *833*, 155066.

[10] Patrizi, G., Bartolini, A., Ciani, L., Gallo, V., Sommella, P., & Carratù, M. (2022). A virtual soil moisture sensor for smart farming using deep learning. *IEEE Transactions on Instrumentation and Measurement*, *71*, 1-11.

[11] Cai, Y., Zheng, W., Zhang, X., Zhangzhong, L., & Xue, X. (2019). Research on soil moisture prediction model based on deep learning. *PloS one*, *14*(4), e0214508.

[12] Singh, A., Gaurav, K., Sonkar, G. K., & Lee, C. C. (2023). Strategies to measure soil moisture using traditional methods, automated sensors, remote sensing, and machine learning techniques: review, bibliometric analysis, applications, research findings, and future directions. *IEEE Access*, *11*, 13605-13635.

[13] Matei, O., Rusu, T., Petrovan, A., & Mihuţ, G. (2017). A data mining system for real time soil moisture prediction. *Procedia Engineering*, *181*, 837-844.
[14] Q. Liu et al., "The contributions of precipitation and soil moisture observations to the skill of soil moisture estimates in a land data assimilation system," J. Hydrometeorol., vol. 12, no. 5, pp. 750–765, Oct. 2011, doi: 10.1175/JHM-D-10-05000.1.
[15] Y. Xia et al., "Evaluation of multi-model simulated soil moisture in NLDAS-2," J. Hydrol., vol. 512, pp. 107–125, May 2014, doi: 10. 1016/j.jhydrol.2014.02.027.
[16] R.D.Kosteret al., "Regions of strong coupling between soil moisture and precipitation," Science, vol. 305, no. 5687, pp. 1138–1140, 2004, doi: 10.1126/science.1100217.
[17] E. G. Njoku, T. J. Jackson, V. Lakshmi, T. K. Chan, and S. V. Nghiem, "Soil moisture retrieval from AMSR-E," IEEE Trans. Geosci. Remote Sens., vol. 41, no. 2, pp. 215–229, Feb. 2003. [Online]. Available: http:// ieeexplore.ieee.org/document/1196040/
[18] F. Balducci, D. Impedovo, and G. Pirlo. 2018. Machine learning applications on agricultural datasets for smart farm enhancement. Machines 6 (Sept. 2018), 38.
[19] M. S. Mahdavinejad, M. Rezvan, M. Barekatain, P. Adibi, P. M. Barnaghi, and A. P. Sheth. 2018. Machine learning for Internet of Things data analysis: A survey. CoRR abs/1802.06305.
[20] R. Sui and J. Baggard. 2015. Wireless sensor network for monitoring soil moisture and weather conditions. Applied Engineering in Agriculture 31 (Mar. 2015), 193–200.
[21] C. Brewster, I. Roussaki, N. Kalatzis, K. Doolin, and K. Ellis. 2017. IoT in agriculture: Designing a Europe-wide large scale pilot. IEEE Communications Magazine 55, 9 (2017), 26–33.
[22] M. Mohammadi, A. Al-Fuqaha, S. Sorour, and M. Guizani. 2018. Deep learning for IoT big data and streaming analytics: Asurvey. IEEE Communications Surveys Tutorials 20, 4 (2018), 2923–2960.
[23] F. Liedmann, C. Holewa, and C. Wietfeld.2018. The radio field as a Sensor — A segmentation based Soil Moisture Sensing approach. In 2018 IEEE Sensors Applications Symposium (SAS'18). IEEE, 1–6.
[24] S. Angra and S. Ahuja. 2017. Machine learning and its applications: A review. In 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC'17). 57–60.
[25] Z. Chen, Y. Liu, and S. Liu. 2017. Mechanical state prediction based on LSTM neural netwok. In 2017 36th Chinese Control Conference (CCC'17). 3876–3881.
[26] https://www.ibm.com/think/topics/principal-component-analysis
[27] https://www.wikidoc.org/index.php/Standard_score
[28] doi:10.1016/j.ymssp.2006.12.007
[29] https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/
[30] https://www.geeksforgeeks.org/what-is-ridge-regression/
[31] Ruchika Malhotra, Empirical Research in Software Engineering
[32] draw.io

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis Development and Analysis of IoT-enabled machine learning models for cost-effective soil moisture estimation

Total Pages 41 Name of the Scholar Vitista Tyagi

Supervisor (s)

(1) Dr. Sanjay Patidar

(2)

(3)

Department Software Engineering (Data Science)

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: Turnitin. Similarity Index: 8%, Total Word Count: 9666

Date: 22/05/25

Vitista

Candidate's Signature

22.5.25

Signature of Supervisor(s)

8% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 10 words)

Match Groups

- **44 Not Cited or Quoted 8%**
  Matches with neither in-text citation nor quotation marks
- **0 Missing Quotations 0%**
  Matches that are still very similar to source material
- **0 Missing Citation 0%**
  Matches that have quotation marks, but no in-text citation
- **0 Cited and Quoted 0%**
  Matches with in-text citation present, but no quotation marks

Top Sources

7%	Internet sources
3%	Publications
6%	Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

0% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Detection Groups

0 AI-generated only 0%
Likely AI-generated text from a large-language model.

0 AI-generated text that was AI-paraphrased 0%
Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

Disclaimer
Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify writing that is likely AI generated as AI generated and AI paraphrased or likely AI generated and AI paraphrased writing as only AI generated) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

LIST OF PUBLICATIONS

1. Vitista Tyagi, Dr. Sanjay Patidar, "Soil Moisture Sensing Techniques using IoT and Machine Learning". The paper has been accepted at the 14th International Conference on Recent Challenges in Engineering and Technology (ICRCET-2025), April 2025. Indexed by Scopus.

Acceptance Letter of Your Research Article Titled Soil Moisture Sensing Techniques using IoT and Machine Learning: A Review | ICRCET-2025, INDIA | 26th April 2025.

Date : 25/02/2025

Conference Secretariat - Chennai, India

Letter of Acceptance

Abstract ID : ICRCET-2025_BAN_0397

Paper Title : Soil Moisture Sensing Techniques using IoT and Machine Learning: A Review

Author Name : Vitista Tyagi,

Co-Author Name : Dr Sanjay Patidar

Institution : Delhi Technological University (DTU)

Dear Vitista Tyagi,

Congratulations!

The scientific reviewing committee is pleased to inform your article "Soil Moisture Sensing Techniques using IoT and Machine Learning: A Review" is accepted for at **14th International Conference on Recent Challenges In Engineering And Technology (ICRCET-2025)** on **26th & 27th April 2025 at Bangalore, India** which is organized by IFERP Academy. The Paper has been accepted after our double-blind peer review process and plagiarism check.

2. Vitista Tyagi, Dr. Sanjay Patidar, "Implementing of a Low-Cost Machine Learning Framework for Soil Moisture Prediction Using Multi-Sensor Data". The paper has been accepted at the International Conference on Robotics, Machine Learning and Artificial Intelligence(ICRMLAI), May 2025. Indexed by Scopus. Paper ID: SA-MLAI-PUNE-240525-7513

ACCEPETANCE AND REGISTRATION FORM (24th May 2025 at Pune, India)

SARC-Re...form.docx

SC SARC Conferences<sarc.net.in@gmail.com> Reply Reply all Forward

To: You Wed 5/14/2025 4:54 PM

You replied on Wed 5/14/2025 5:57 PM

SARC-Registration form.docx
76 KB

Dear Researcher,
Greetings and best wishes for the day !!

Your Paper entitled **"Implementation of A Low-Cost Machine Learning Frameowrk for Soil Moisture Prediction Using Multi-Sensor Data"** was accepted for the Presentation on Upcoming International Conference Conducted By SARC

Paper ID: **SA-MLAI-PUNE-240525- 7513**

Conference Date and Place: **24th May 2025 at Pune, India**
Authors Name: **Vitista Tyagi & Dr. Sanjay Patidar**

Conference Name: International Conference on Robotics, Machine Learning and Artificial Intelligence(ICRMLAI)

Registration Details:

Categories	Non Indian	Indian
Academician/ Practitioner	USD 250	INR 9000