SOIL FERTILITY CLASSIFICATION USING MACHINE LEARNING-BASED APPROACH

Thesis

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

by

Sujatha M

Register No. 197032IT002

DEPARTMENT OF INFORMATION TECHNOLOGY

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA
SURATHKAL, MANGALORE - 575025

June, 2024

Declaration

I hereby *declare* that the Research Thesis entitled "Soil Fertility Classification using Machine Learning-based Approach" which is being submitted to the National Institute of Technology Karnataka, Surathkal, in partial fulfillment of the requirements for the award of the Degree of Doctor of Philosophy in Information Technology is a *bonafide report of the research work carried out by me*. The material contained in this thesis has not been submitted to any University or Institution for the award of any degree.

Sujatha 3/6/2024

NITK Surathkal
Date: 3-6-2024

Sujatha M
Register No: 197032IT002
Department of Information Technology

Certificate

This is to *certify* that the Research Thesis entitled "Soil Fertility Classification using Machine Learning-based Approach" submitted by Sujatha M (Register Number: 197032IT002) as the record of the research work carried out by her, is *accepted as the Research Thesis submission* in partial fulfillment of the requirements for the award of the degree of Doctor of Philosophy.

03.06.2024

Dr. Jaidhar C D
Research Guide
Associate Professor
Department of Information Technology
NITK Surathkal - 575025

3/6/24

Chairman - DRPC
(Signature with Date and Seal)

CHAIRMAN - DRPC
Department of Information Technology
NITK Surathkal, Srinivasnagar P.O.
Mangaluru 575 025, INDIA

Acknowledgements

I express my sincere gratitude to my research supervisor for being constant motivation and providing valuable suggestions throughout my research journey. The support of my supervisor greatly helped in the completion of my research. I thank Prof. Sam Johnson P., Professor, MACS Department, and Dr. Nagamma Patil, Associate Professor, IT Department, NITK, for their technical suggestions.

I thank Dr. Mallikarjuna Lingappa, Scientist (Soil Science), Krishi Vigyan Kendra Dakshina Kannada, for his invaluable technical suggestions during my research work. I thank Dr. R P Sharma, Senior Scientist, ICAR-NBSS&LUP, Nagpur, for providing insights on the Soil-health dataset. I thank Dr. S M Hiremath, Scientist (Horticulture), Krishi Vigyan Kendra, Haveri, and Dr. Praveen T. Goroji, Senior Technical Officer, ICAR-KVK, Dharwad, for the invaluable timely help.

I thank all fellow research scholars and faculties of the Information Technology Department. I am indebted to my parents and friends for their support throughout my research.

(Sujatha M)

Abstract

Agriculture is the main source of economy and survival in many countries. To ensure sustainable agricultural development, it is crucial to promptly acquire soil fertility and apply accurate fertilizers. However, traditional laboratory methods for analyzing soil samples make it challenging to estimate soil fertility. Therefore, this research aims to develop a reliable Machine Learning (ML)-based classifier that can classify soil fertility as LOW, or MEDIUM, or HIGH. Additionally, prescribes fertilizers based on the classification results.

Soil fertility classification approach based on laboratory chemical parameters such as Electrical Conductivity (*EC*), Organic Carbon (*OC*), potential of hydrogen (*pH*), boron (*B*), copper (*Cu*), iron (*Fe*), manganese (*Mn*), phosphorus (*P*), potassium (*K*), sulphur (*S*), and zinc (*Zn*) have been proposed using ML approaches. The classifiers used in this study included Random Forest (RF), bagging, Boosted Regression Tree (BRT), J48 Decision Tree (J48), Logistic Regression (LR), Naive Bayes (NB), and Support Vector Machine (SVM). The experiments were conducted with a split dataset (75% of data for training and 25% for testing) and 10-fold cross-validation. The tree-based classifier RF, outperformed the other classifiers by producing an accuracy of 99.99% with 10-fold cross-validation test and a split dataset.

To avoid the need for laboratory analysis and obtain soil parameters specific to the site, this research relied on Sentinel-2 spectral data to determine *EC, pH, OC*, and *N*. The generated dataset was labeled using various clustering methods such as canopy, density-based, expectation-maximization, farthest-first, fuzzy C-means, and k-means and then compared with manual labeling. Among these, the canopy clustering approach achieved the highest accuracy of 75.99% on labeling dataset. Therefore, the proposed method for labeling the dataset uses the canopy-centered fuzzy C-means clustering. It was found that the proposed canopy-centered fuzzy-C-means clustering method achieved the highest accuracy of 78.42% in labeling the dataset. Furthermore, the performance of several ML-based classifiers, such as NB, SVM, J48, and RF were compared using datasets labeled with different clustering approaches. The RF classifier achieved the highest classification accuracy of 99.69% using the proposed approach and on 10-fold cross-validation.

To determine the best fertilizer for a given soil, a new fertilizer prescription approach was proposed. It uses an ensemble filter-based feature selection to classify soil fertility and prescribe the appropriate fertilizer. It was tested on two datasets from regions with varying climate conditions. Various tree-based classifiers, such as classification and regression tree, extra tree, reduced error pruning tree, RF, NB, and SVM, were compared using the first dataset with relevant soil parameters. The results showed that the RF classifier with relevant soil parameters was the most accurate, achieving a 99.96%

accuracy with dataset-1 and a 99.90% accuracy with dataset-2.

A soil fertility classifier and fertilizer prescription approach was proposed by utilizing 2D Convolutional Neural Networks (CNNs). The experiments were conducted on a split dataset with varying kernel sizes of 3×3 to 7×7 and input grid sizes from 11×11 to 13×13. The classifier showed an impressive accuracy of 97.24% and kappa statistics of 0.0938 with an input grid size of 11×11 and a kernel size of 3×3. To further improve the accuracy, the training data was oversampled using the Synthetic Minority Oversampling Technique (SMOTE). The proposed approach using oversampling achieved an accuracy of 97.52% and kappa statistics of 0.1397, with an input grid size of 12×12 and a kernel size of 3×3.

A 1D-CNN-based soil fertility classification approach was developed to simplify the 2D CNN-based classifier used for soil fertility classification. To improve the performance of the model, the dataset was normalized using Min-Max normalization, and training data was oversampled using SMOTE. The proposed approach was compared with the soil fertility classifiers based on Extreme Learning Machine (ELM) and Multi-Layer Perceptron (MLP). The proposed approach, with normalization and SMOTE, achieved an accuracy of 97.90% and kappa statistics of 0.2358.

A new method to classify soil fertility and prescribe fertilizers using symbolic deterministic finite automata, to overcome the limitations of traditional ML-based classifiers, which require large, unbiased datasets and are prone to errors. The proposed method was compared using ML-based classifiers using data from Sentinel-2 satellite imagery and laboratory-measured soil health data of Belgaum district. The data consisted of two sets: one with four soil parameters (Soil-health-1 dataset) and the other with twelve soil parameters (Soil-health-2 dataset). The results showed that the new approach was able to classify soil fertility with 100% accuracy using the Sentinel-2 and Soil-health-1 datasets, and with 98.37% accuracy using the Soil-health-2 dataset.

Satellite revisits to a specific site location are infrequent, hence, soil sensors are used to collect real-time values of *EC, pH, N, P*, and *K* in this study. The collected real-time data is tested using trained and saved ML-based classifiers, such as Classification and Regression Tree (CART), J48, RF, Reduced Error Pruning (REP), NB and SVM which were trained using the Soil-health dataset of Belgaum district. For the real-time test data RF and REP classifiers achieved highest test accuracy of 100%.

Keywords: Classification; Convolutional Neural Networks; Feature Selection; Fertilizer Prescription; Machine Learning; Precision Agriculture; Soil Fertility.

Contents

List of Figures

List of Tables

List of Abbreviations

ANN	Artificial Neural Network
BNN	Bayesian Neural Network
BPNN	Backpropagation Neural Network
BRT	Boosted Regression Tree
CEC	Cation Exchange Capacity
CART	Classification and Regression Tree
CNN	Convolutional Neural Network
CR	Cubic Regression
CRSI	Canopy Red-edge Spectral Index
CV	Cross-validation
DT	Decision Tree
DPSO	Dynamic Fitness inertia weighted Particle Swarm Optimization
EC	Electrical Conductivity
EEVI	Enhanced Exaggerated Vegetation Index
ELM	Extreme Learning Machine
ENDVI	Enhanced Normalized Difference Vegetation Index
ER	Elastic-net Regression
ET	Extra Trees
EVI	Enhanced Vegetation Index
EVI2	Enhanced Vegetation Index-2
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GainR	Gain Ratio
GA	Genetic Algorithms
GDVI	Green Difference Vegetation Index
GIS	Geographic Information System
GPR	Gaussian Process Regression
iPLSR	interval Partial Least Square Regression
INFOG	Information Gain
KNN	K Nearest Neighbors
LASSO	Least Absolute Shrinkage and Selection Operator
Lidar	Light detection and ranging
LinR	Linear Regression
LMI	Legates-McCabe's Index

LR	Logistic Regression
MAE	Mean Absolute Error
ML	Machine Learning
MLP	Multilayer Perceptrons
MLR	Multiple Linear Regression
MODIS	Moderate-resolution Imaging Spectroradiometer
MSI	Moisture Stress Index
MV	Multivariate
NB	Naive Bayes
NS	Nash-Sutcliffe's coefficient
NDVI	Normalized Difference Vegetation Index
NIR	Near-Infrared
NDMI	Normalized Difference Moisture Index
OC	Organic Carbon
OM	Organic Matter
PCA	Principal Component Analysis
PCR	Principal Component Regression
PLSR	Partial Least Square Regression
PRC	Precision-Recall Curve
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
ReliefF	Relief Feature
ReLU	Rectified Linear Unit
REP	Reduced Error Pruning
RF	Random Forest
RMSE	Root Mean Square Error
ROC	Receiver Operating characteristic Curve
RR	Ridge Regression
RRSE	Root Relative Square Error
RSI	Relative Stength Index
RTU	Remote Terminal Unit
SAVI	Soil Adjusted Vegetation Index
SDFA	Symbolic Deterministic Finite Automata
SFA	Symbolic Finite Automata
SMOTE	Synthetic Minority Oversampling Technique
SI	Salinity Index

SR	Stepwise Regression
SRatio	Simple Ratio
SVM	Support Vector Machine
SWIR	Shortwave Infrared
TN	True Negative
TP	True Positive
TPR	True Positive Rate
TWI	Topographic Wetness Index
VFI	Village-wise Fertility Index
VNIR	Visible and Near-Infrared
WI	Willmott's Index

List of Nomenclatures

Al	aluminium
B	boron
Ca	calcium
$CaCO_3$	calcium carbonate
Cl	chlorine
Cu	copper
Fe	iron
K	potassium
K_2O	potassium oxide
Mg	magnesium
Mn	manganese
Mo	molybdenum
N	nitrogen
N_2O	nitrous oxide
Na	sodium
Ni	nickel
P	phosphorous
P_2O_5	phosphorous pentoxide
pH	potential of hydrogen
S	sulphur
SO_4	sulfate
Zn	zinc

CHAPTER 1

INTRODUCTION

Agriculture is the primary source of income in many countries. India's development is largely dependent on the agricultural sector. The variations in soil nutrients negatively impact crop yield. The crops require a balanced proportion of macronutrients such as calcium (Ca), magnesium (Mg), nitrogen (N), phosphorus (P), potassium (K), sulfur (S), and micronutrients such as boron (B), chlorine (Cl), copper (Cu), iron (Fe), manganese (Mn), molybdenum (Mo), nickel (Ni) and zinc (Zn) (Lambot et al., 2017). In India, the traditional practice is to estimate soil fertility through laboratory analysis, and fertilizers are applied randomly. The laboratory analysis is time-consuming and also generates chemical residues. The variation in soil fertilization can either deteriorate or accelerate soil nutrients (Osman, 2012), and the excess fertilization and mineral fertilizers cause environmental pollution (Liu et al., 2023). Agricultural productivity can be improved by maintaining soil fertility balance. Sustainable crop production requires accurate soil fertility estimation (Fao, 2020a).

A systematic review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to identify previous research on soil fertility classification. Machine learning (ML)-based approaches were widely used by researchers to predict variations in fertility levels of soil parameters. Many authors predicted variations in soil chemical parameters and used remotely sensed environmental indices derived from Landsat spectral bands to accurately predict the organic carbon (OC) stock, OC, and salinity. Few authors derived soil parameters such as OC, salinity, P, potential of hydrogen (pH), and Fe using satellite spectral bands and predicted their variations based on climate. Additionally, ML-based regressors were used to predict soil fertility based on laboratory values of phosphorous pentoxide (P_2O_5) and potassium oxide (K_2O), and based on organic matter (OM), pH, K, N, P, S, B, Ca, Mg, and Zn. The use of Artificial Neural Networks (ANNs) was observed in predicting salinity distribution while Back Propagation Neural Networks (BPNN) predict changes in the levels of K, N, and P.

Very few researchers have used ML-based classifiers to classify soil fertility. Sirsat et al. (2017) classified soil fertility as HIGH or MEDIUM or LOW using Decision Tree (DT), K Nearest Neighbors (KNN), Random Forest (RF), Naive Bayes (NB), Reduced Error Pruning (REP), Support Vector Machine (SVM), Extreme Learning Machine (ELM), Multi-layer Perceptron (MLP), AdaBoost and Bagging based on pH, Electrical Conductivity (EC), OC, nitrous oxide (N_2O), potassium oxide (K_2O), sul-

fate (SO_4), *Zn, Fe, Mn*, and phosphorous pentoxide (P_2O_5). In Fernandes et al. (2019), fertility of *OM* was classified based on *Ca*, *pH*, *potential acidity*, *K*, *P*, and *Mg*. Gulhane et al. (2023) classified soil fertility using MLP and *pH, P*, and *Fe* as low, normal, and high. Khanal et al. (2018) employed MLP, RF, and SVM to classify soil fertility and predicted variations in Cation Exchange Capacity (CEC), *Mg*, and *OM* using Linear Regression (LinR), Cubic Regression (CR), and MLP.

Some researchers have recommended the use of fertilizers to improve the fertility of the soil. Chougule et al. (2019) prescribed specific fertilizers to boost the levels of *N, P* and *K* based on the predicted values of these nutrients. Ransom et al. (2019) suggested the use of nitrogen fertilizer for corn to improve *N*. Coulibali et al. (2020) recommended fertilizers based on variations in the fertility levels of these nutrients. Abera et al. (2022) suggested fertilizers to increase *K*, *P*, *N*, and *S*. Sirsat et al. (2017) recommended the use of fertilizers based on K_2O, N_2O, P_2O_5 for bajra, cotton, and soybean. Precision agriculture can optimize the use of fertilizers and water to enhance crop sustainability and productivity (Méndez-Vázquez et al., 2019). Some researchers have also derived Village-wise Fertility Indices (VFIs) or remotely sensed indices to classify soil fertility more accurately.

1.1 Influence of Soil Properties

The soil properties are categorized into three main categories: biological, chemical, and physical properties (Osman, 2012), as shown in Figure 1.1.

Biological Properties: Soil fertility is greatly influenced by the various organisms that live in the soil. These organisms secrete enzymes that break down organic matter into humus, providing essential nutrients to plants for their growth and development. Moreover, the presence of these organisms also affects several other soil properties. Several types of soil organisms play a role in determining soil fertility, such as Arthrobacter, Azospirillum, Bacillus, Enterobacter, Pseudomonas, and Serratia.

Chemical Properties: Soil is a mixture of organic and inorganic substances, soluble and insoluble, solid, liquid, and gas particles. The *pH* level of soil determines its acidity or alkalinity. The level of acidity or alkalinity is measured by its *pH*. The pH level of the soil is crucial for growing crops. A neutral soil with a *pH* level 6.5-8.5 is fertile. If the *pH* level is below 6.5, then the soil is acidic; if it is above 8.5, then the soil is alkaline. Both alkaline and acidic soils are less fertile, which can lead to poor crop yields. The soil is categorized as saline or non-saline based on the value of *EC*. Soil consists of *OC* or inorganic carbon. The *OC* is formed due to the decomposition of plants, while the inorganic carbon is from the reaction between soil minerals and carbon dioxide. The

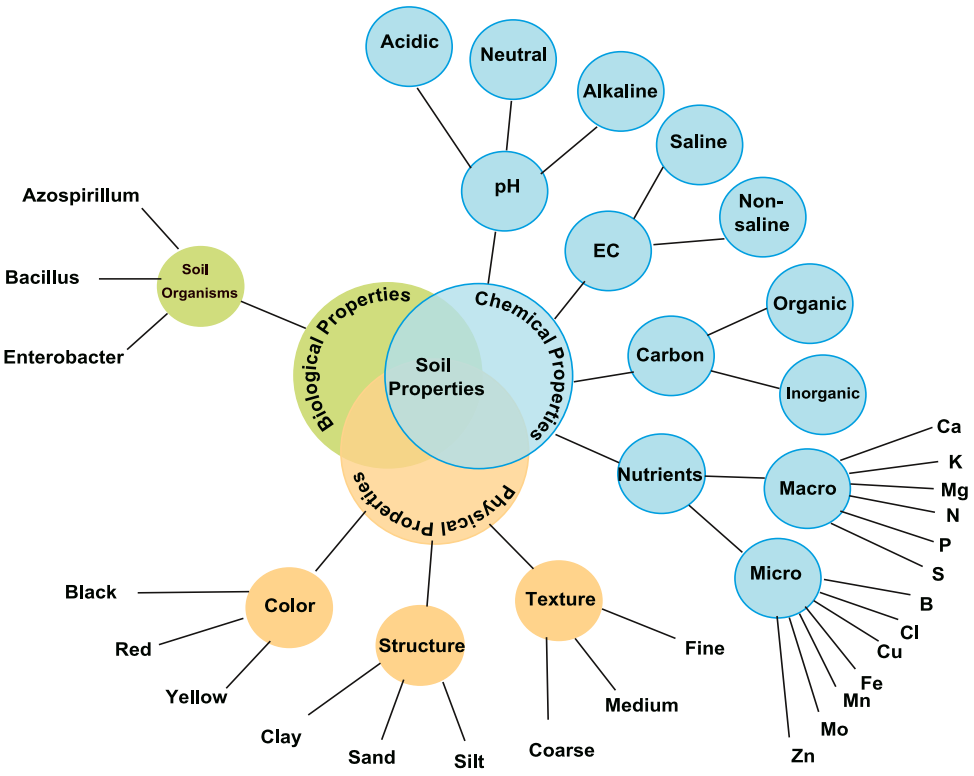


Figure 1.1: Soil property categorization

growth of crops depends on sixteen mineral elements, including carbon, hydrogen, oxygen, aluminum (*Al*), *B, Ca, Cl, Cu, Fe, K, Mg, Mn, Mo, N, P, S*, and *Zn*. The essential macronutrients, such as *Ca, K, Mg, N, P*, and *S*, are required in large quantities between 1 to 150 g/kg. The micronutrients, such as *B, Cl, Cu, Fe, Mn, Mo*, and *Zn*, are required in lesser quantities between 0.1 to 100 mg/kg. It is important to identify nutrient deficiencies early, as this can have a major impact on plant growth and yield.

Physical Properties: Soil characteristics include color, density, structure, texture, consistency, and temperature. The soil color can change due to the presence of *OM*, water, and redox conditions, while other properties depend on the type and arrangement of soil particles. Soil particles can be primary or secondary. Primary particles, such as sand, silt, and clay, are categorized based on their effective diameter, and their combination creates secondary particles. Sand has a low capacity to hold water and nutrients; silt soil has a medium capacity; and clay has a high capacity. Soil texture is determined by sand, silt, and clay composition and can be classified as coarse, medium, or fine. There are 12 subcategories: coarse (sand, loamy sand, sandy loam), medium (loam, silt loam, silt), and fine (sandy clay loam, clay loam, silty clay loam, sandy clay, silty clay, clay). Fine-textured soils typically contain nutrient holding capacity than coarse-textured soils. Coarsely textured sand has a low nutrient-holding capacity and therefore

requires frequent application of small amounts of fertilizer (Osman, 2012).

1.2 Significance of Nutrients in Soil and Prevention of Deficits

The growth cycle of crops is heavily influenced by soil chemical properties, including *EC, pH, OC*, macronutrients, and micronutrients (NRCS-USDA, 2020; Kalkhoran et al., 2019; Osman, 2012), as depicted in Figure 1.2.

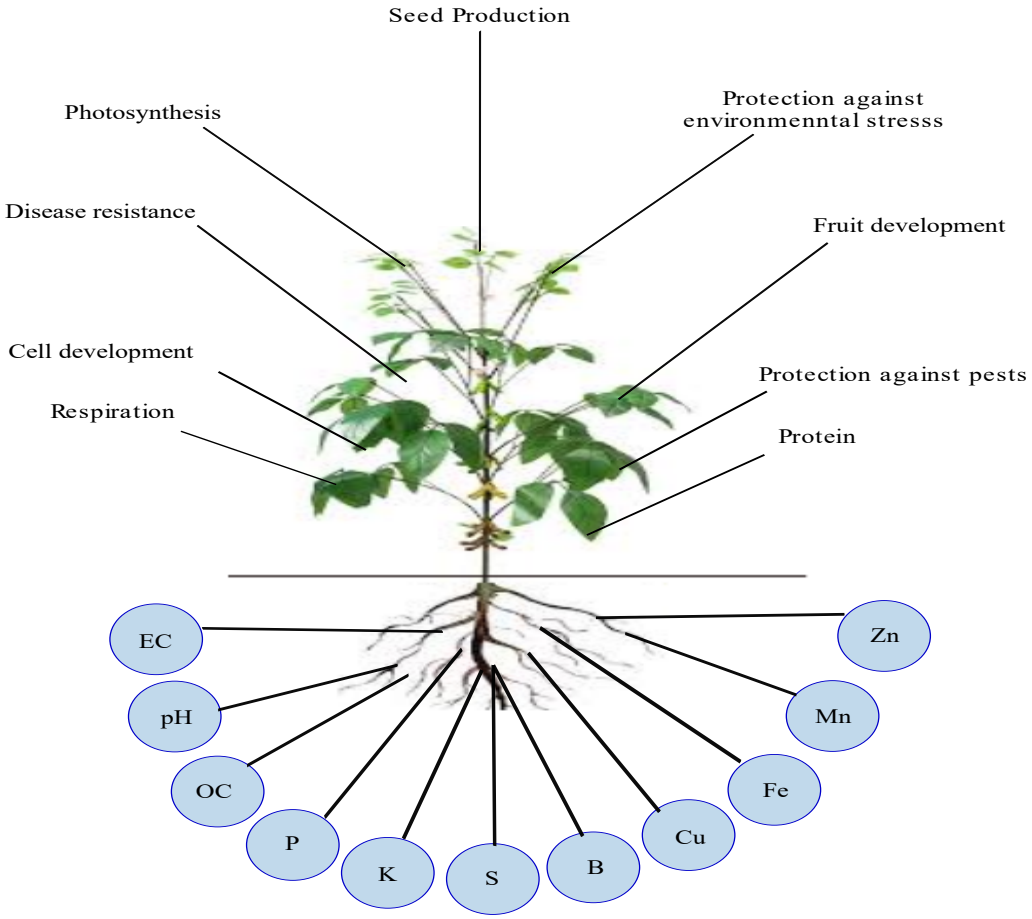


Figure 1.2: Significance of soil chemical parameters

Proper soil nutrient management is crucial for healthy plant growth and development. The soil *EC, pH*, and *OC* positively or negatively correlated to soil nutrients. The significance of soil nutrients can be outlined as follows:

Calcium (*Ca*) is crucial for plant growth and development, as it forms an essential component of cell walls. Insufficient levels of *Ca* can lead to leaf curling and stunted root development. An overabundance of *Ca* can reduce the uptake of other important nutrients such as *B, Cu, K, P, Mg, Fe*, and *Zn*.

Potassium (*K*) is essential for protein synthesis, enzyme and carbohydrate activation, crop growth, and fruit development. Insufficient *K* levels can result in stunted growth, reduced fruit color and sugar content. Conversely, excessive *K* levels can lead to re-

duced *Mg* and *Ca* levels.

Magnesium (*Mg*) increases the amount of vitamin C in fruits. A deficiency of *Mg* results in leaf curling or paleness. Excessive *Mg* causes a decrease in *K*.

Nitrogen (*N*) is crucial for photosynthesis and gives leaves their green color. A lack of *N* causes yellowing of leaves, reduced branching and leaves, and early fruiting and flowering. Excess *N* causes late flowering or fruiting (Osman, 2012).

Phosphorus (*P*) controls respiration and photosynthesis. Decreased *P* reduces root or shoot growth, photosynthesis, and leaf surface area. Increased *P* reduces *Fe, K*, and *Zn*.

Sulphur (*S*) is essential for chlorophyll production, protein synthesis, and activation of enzymes. It also plays a role in pest control. A deficiency in *S* can lead to reduced seed development and growth delays, while an excess can decrease *Mo* and *N*.

Boron (*B*) is a crucial element for the development of fruits and seeds, as well as for cell division. Insufficient *B* reduces yield and quality, Excessive amount of *B* acts as an inhibitor for the growth of roots and shoots as well as chlorophyll production, leading to yellowed leaf tips (Reid, 2007; Blevins & Lukaszewski, 1998).

Chlorine (*Cl*) plays a crucial role in photosynthesis and disease prevention (Chen et al., 2010). Insufficient *Cl* results in highly branched roots, while increased *Cl* reduces leaf size and plant growth.

Copper (*Cu*) is essential for cell walls and photosynthetic electron transport (Marschner, 2011). When there is not enough *Cu*, the leaves get twisted, and the yield is reduced. On the other hand, when there is excess *Cu*, the leaves become discolored (Osman, 2012).

Iron (*Fe*) is an essential element for plants as it plays a crucial role in the production of chlorophyll and the transfer of electrons. A deficiency can cause chlorosis and reduced leaf size, while excess can damage deoxyribose nucleic acid and lipids in plants.

Manganese (*Mn*), an essential nutrient for plants, plays a crucial role in the production of chlorophyll, assimilation of carbon dioxide, and nitrate. It also plays a crucial role in the process of germination and growth. A lack of *Mn* can cause yellowing between leaf veins, stunted roots, and stems. Meanwhile, excessive *Mn* results in leaves with red or black spots.

Molybdenum (*Mo*) is essential for plant nitrogen metabolism, and its deficiency leads to stunted growth and pale green leaves, while an excess of *Mo* leads to leaf browning or yellowing.

Zinc (*Zn*) is an essential nutrient for plant growth and seed production. A deficiency in *Zn* leads to stunted growth and smaller leaves, while excess *Zn* can affect other soil nutrients.

Farmers can maintain soil nutrient balance by using appropriate fertilizers for the crop type and location. Chemical and biofertilizers are both excellent sources of nutrients (Fao, 2020b). To prevent *Ca* deficiency, lime can be added to acidic soils, while gypsum or soluble calcium sources are suitable for non-acidic soils. Potassium levels can be increased by using potassium chloride or potassium sulfate or by incorporating crop residues and manures. Using potassium chloride also increases chloride. Adding dolomite limestone can improve *Mg* levels in the soil. Nitrogen levels can be increased by adding organic matter, rotating crops, and using nitrogen fertilizers such as neem-coated urea, ammonium nitrate, etc. Phosphorus fertilizers such as single superphosphate and diammonium phosphate increase soil *P* levels. Fertilizers such as single superphosphate, gypsum sulfur, and ammonium sulfate can help to increase the levels of *S*. *B* deficiency can be corrected by applying boric acid or borax. Copper levels in the soil can be increased by using copper-based fertilizers. Ferrous and ferric sulfate can increase *Fe* levels. *Mn* levels can be increased by applying a manganese sulfate solution, and sodium ammonium molybdate can improve *Mo* levels. *Zn* can improved by applying zinc sulfate (Kant & Kafkafi, 2020).

1.3 Impact of Climate on Soil Fertility

Climate change has the potential to negatively impact the fertility of soil. Figure 1.3 (Dutta & Rakshit, 2016) shows the effect of climate on soil quality and, thus, agricultural production. A significant increase in precipitation leads to higher leaching, nutrient loss, and acidification. Elevated soil surface temperature accelerates soil organic matter mineralization, reducing carbon and water retention and inhibiting plant growth (Fao, 2021). Increased evaporation and transpiration rates can lead to erosion by water and wind, resulting in reduced soil moisture levels. This can impact semi-arid or clay soils (Dutta & Rakshit, 2016). A reduced rainfall can result in alkalization or salinization of soil. Evaporation or drought increases alkalinity (raises pH) due to the higher availability of *Ca*, Sodium (*Na*), and *Mg* (Rengel, 2011). Soil salinization occurs due to high soil temperatures that increase evaporation rates, leading to the accumulation of salts (such as sodium chloride, magnesium sulfate, and chloride) in the top soil layers. This accumulation and soil dispersion or aggregation caused by rain can reduce soil porosity, limiting water infiltration and hydraulic conductivity. The high temperatures or increased carbon dioxide cause acidification of soil. Climate variability or change impacts other soil nutrients (Day AD, 1993). The nutrients *N, K*, and *S* are less affected by soil *pH* than other nutrients. In acidic soil, the solubility of *Fe, Mn, Al, Zn*, and *Cu* increases, while the solubility of *Mg, Ca*, and *Mo* decreases. Alkaline soil, on the

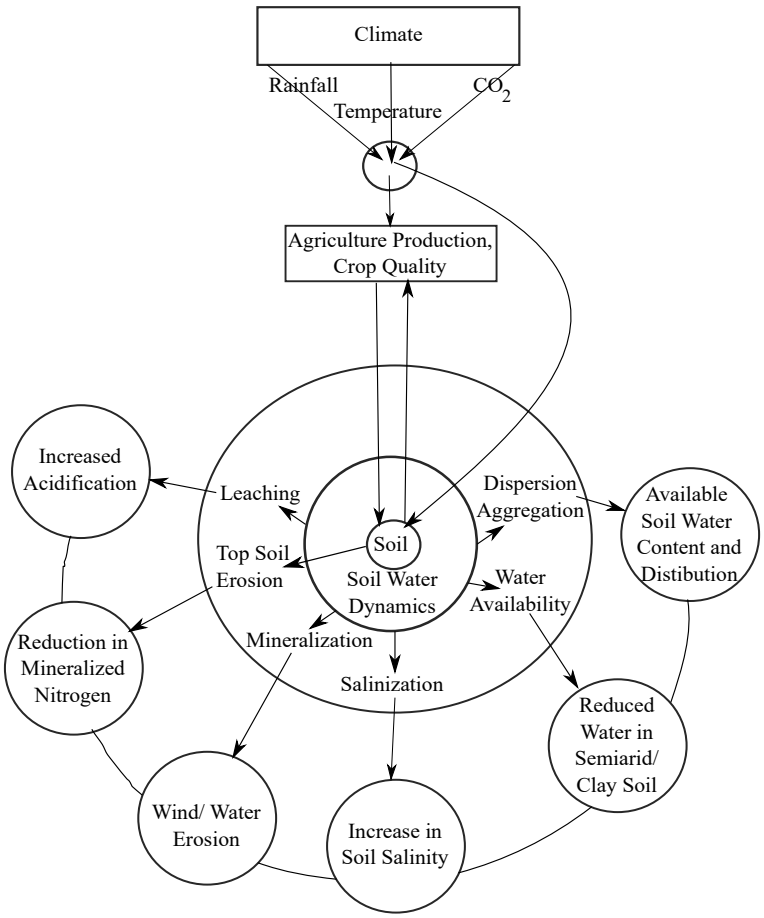


Figure 1.3: Impact of climate on soil

other hand, increases the solubility of *Ca, Mg, and Mo* but decreases the solubility of *Cu, Fe, Mn, Al*, and *Zn*. Although *Al* is highly available in the soil, it is not essential for plant growth. However, in soil with a *pH* less than 5, the solubility of *Al* with *Fe* and *Mn* increases, making the soil toxic to most plants. Soil acidification decreases essential minerals such as *Ca* and *Mg* in soils. The availability of *P* is also reduced in both acidic and alkaline soils. Highly acidic and alkaline soils can cause a deficiency in *B*. In highly acidic soils, macronutrients such as *P, Mg, K, N, Ca, S, Mo*, and *B* are significantly less available, whereas micronutrients such as *Cu, Fe, Mn*, and *Zn* significantly increase. Soil salinity degrades soil fertility and increases the *EC* value of soil. Salts may accumulate at the soil surface due to rainfall, increasing salinity. Soils with a higher *EC* value are less fertile, leading to reduced crop health (NRCS-USDA, 2020; Phonphan et al., 2014). An increase in soil *EC* results in an increase in hydrogen ions, leading to soil acidity and a subsequent decrease in *pH*. On the other hand, a temperature rise has an inverse effect on *OC*, which in turn affects nutrient availability and the soil's ability to retain moisture (Gan et al., 2013). Higher levels of *OC* imply increased carbon content in the soil, which can lead to a decrease in *pH* and an increase in *EC*.

1.4 Discussion on ML-based Approaches

The generic steps of ML-based soil fertility classification are as shown in Figure 1.4.

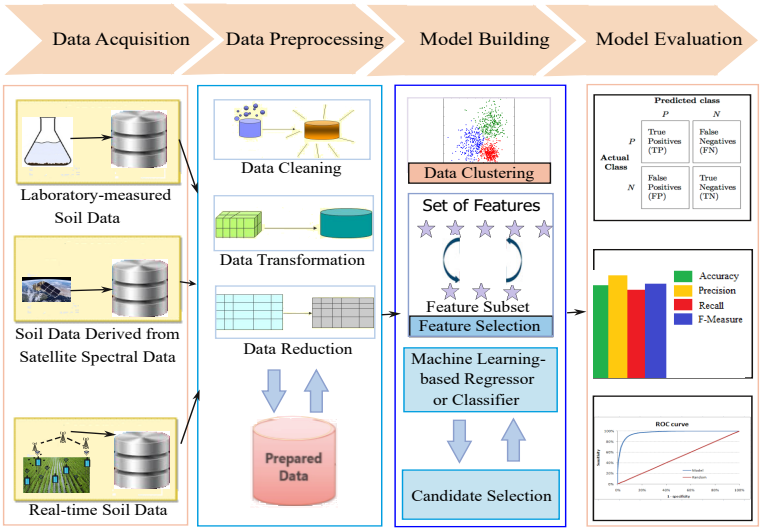


Figure 1.4: ML-based soil fertility classification

Data acquisition involves gathering data from various sources such as laboratory measurements, satellite images, or real-time soil sensing for use by an ML-based classifier.
Data Preprocessing Data preprocessing refers to converting raw data into a more usable and appropriate format that can be analyzed easily. This step is crucial to ensure the data is clean, consistent, and ready for further analysis. It involves several techniques, such as cleaning, normalization, and reduction, to prepare the data for ML models or other analytical tools. Data cleaning is the process of dealing with data that has irrelevant or missing values. The purpose of data cleaning is to ensure the accuracy and completeness of the data before it is used for analysis. When multiple values are missing, they can be excluded from the dataset, or the mean/median of the attribute can replace the missing values. The missing data can be filled manually or through regression. Data transformation prepares data for extraction. This involves data preprocessing, which includes normalization and attribute construction. Normalization can scale data values from -1 to 1 or 0 to 1. Data reduction can enhance storage efficiency and decrease analysis costs for large datasets. It typically involves attribute selection, dimensionality, and numerosity reduction. Attribute selection is used to select relevant attributes. Using numerosity reduction, storing a data model instead of the entire dataset is possible. This results in more efficient use of storage space. Dimensionality reduction is a technique that reduces the number of attributes by using encoding mechanisms. It helps to simplify complex data by transforming it into a lower-dimensional space without losing important information.

Model Building: The model building involves labeling the dataset manually or using clustering. The feature selection can be used to reduce the cost of the model (Pes, 2020). To ensure the accuracy and reliability of the model, the dataset is typically separated into training and test data. The classifier is trained on the training data and tested on the test data to evaluate performance (Müller & Guido, 2016). To obtain the most effective ML model, the model-building can be repeated several times through iteration.
Model Evaluation: The classifier performance can be evaluated by using widely used performance metrics such as confusion matrix, accuracy, precision, recall, F1-Score, and Root Mean Square Error (RMSE). Figure 1.5 shows various ML-based approaches.

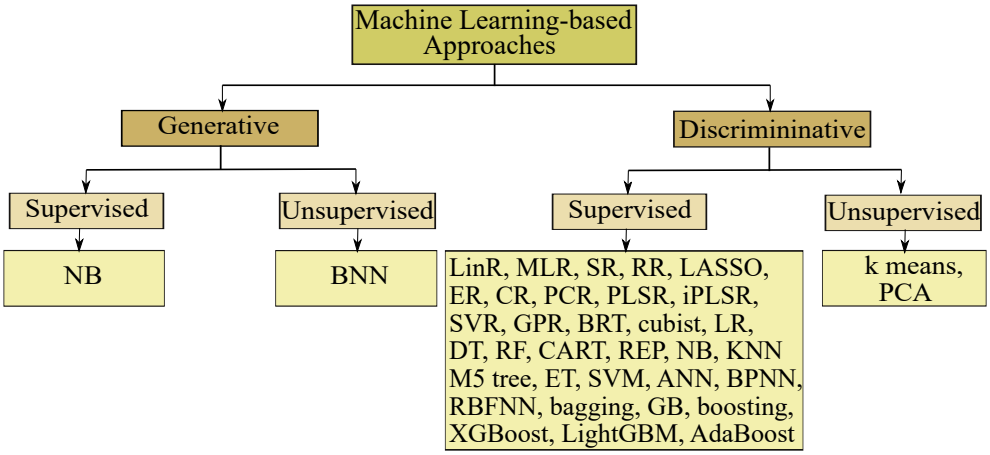


Figure 1.5: Various ML-based approaches

ML-based approaches can be broadly classified into generative and discriminative. Discriminative models use conditional probability, whereas generative models use joint distribution. Generative models can generate synthetic data using probabilistic distribution, while discriminative models cannot. Two categories of ML algorithms: supervised learning and unsupervised learning. Supervised learning involves using labeled data to train a model, while unsupervised learning involves finding patterns in unlabeled data. Regression techniques such as LinR, Multiple Linear Regression (MLR), Elastic-net Regression (ER), Ridge Regression (RR), Stepwise Regression (SR), Least Absolute Shrinkage and Selection Operator (LASSO), CR, Principal Component Regression (PCR), Partial Least Square Regression (PLSR), interval PLSR (iPLSR), Gaussian Process Regression (GPR), cubist, and classification approaches such as LR, DT, Boosted Regression Tree (BRT), RF, Classification and Regression Tree (CART), REP, M5 tree, Extra Tree (ET), BRT, NB, KNN, SVM are commonly used in supervised learning. Unsupervised learning is a type of ML that identifies hidden patterns in data without any prior labeling or supervision. Unsupervised learning includes Principal Component Analysis (PCA), and cluster analysis. Most generative models are created using unsu-

pervised learning techniques, while discriminative models are developed using supervised learning. In contrast, Principal Component Analysis and k-means algorithms are discriminative-unsupervised learning, while the NB classifier is generative-supervised learning (Jebara, 2004).

The classification approaches used in this research work are:

LR characterizes the connection between one or more independent variables and a dependent binary variable, primarily employed for classification tasks rather than regression. Its purpose is to model the probability of a binary outcome by transforming any real value into a range between 0 and 1 using the Sigmoid function. This transformed value signifies the probability of a data point belonging to a particular class. The learning algorithm iteratively adjusts the weights to precisely classify instances within the training set (Belyadi & Haghighat, 2021).

J48 represents data in a tree structure where the leaf nodes symbolize the target attribute, and decision nodes define tests to be conducted on specific observational characteristics (Castelli et al., 2018). As the decision tree classifier progresses from observations (depicted as branches) to conclusions about the target value (represented in the leaves), the decision node is evaluated until a leaf node is reached. The confidence factor determines the degree of pruning; higher values result in less pruning, and more folds are required to prevent overfitting. Randomly choosing a seed value can be advantageous to mitigate the risk of discarding valuable data during error pruning. Constructing the decision tree involves defining node attributes and conditions for splitting at each node. The algorithm selects the optimal option at each step to minimize information gain, employing equations based on entropy concepts to reduce entropy (Cohen, 2021).

BRT is an ensemble method that combines DTs and boosting techniques to enhance the accuracy of ML models. By fitting multiple DTs, BRTs mitigate the risk of outliers through repeated application. Each new tree is generated using a random subset of the entire dataset, which is then returned to the dataset for future tree selection. Boosting is a technique employed in BRTs to assign weights to input data in subsequent trees, thereby improving accuracy. Following fitting the initial tree, the model retains errors and utilizes them to construct subsequent trees. This involves assigning more weight to incorrectly classified instances, ultimately enhancing classification accuracy (De'Ath, 2007). The final classification is determined by the weighted majority of classifications made across the sequence of trees.

RF combines multiple Decision Trees (DTs) to decrease variance and maintain accuracy (Pham et al., 2020). During tree construction, the RF classifier identifies the

best split among selected features for each intermediate node. Enhancements in performance can be achieved by generating less connected and diverse trees (Han et al., 2020). In J48 classification, each tree node can be a decision node or a leaf node, with decision nodes specifying tests on individual features and leaf nodes specifying the target. RF utilizes bootstrap aggregating to create multiple random samples and aggregates them to enhance classification accuracy. From an initial dataset D1 with m rows and n columns, a new dataset D2 is generated by randomly selecting m instances with replacement. Approximately one-third of the rows from D1 are excluded to form out-of-bag samples, providing an unbiased error estimate. The model is trained on D2, and the out-of-bag samples are used for evaluation. The classification tree randomly selects P less than n columns, typically $\sqrt{n}$. RF generates multiple trees, and the final prediction is determined by averaging or voting. These trees are trained simultaneously through bootstrapping and aggregation, ensuring the uniqueness of each random forest and minimizing overall variance. The final decision of the RF classifier is derived by aggregating individual tree decisions. RF outperforms other classifiers without the risk of overfitting, doesn't require feature scaling, and is more reliable when training samples are selected (Misra et al., 2020).

CART is a data modeling technique that utilizes a tree-like structure, where each internal node represents a decision, and each leaf node represents a class label (for classification) or a continuous value (for regression) (Breiman et al., 2017). The tree's construction involves binary splitting at each internal node based on a chosen feature and a threshold value. The objective is to partition the data to minimize impurity (for classification) or error (for regression) in the resulting subsets. CART employs impurity measures such as Gini impurity and cross-entropy for classification tasks, evaluating the disorder or impurity of a set of class labels. For regression tasks, mean squared error serves as the impurity measure. The tree structure is created through a recursive partitioning process, starting with the entire dataset and iteratively splitting it into subsets based on selected features and thresholds until a stopping criterion is met. Stopping criteria include maximum tree depth, minimum samples required for a split, or a minimum impurity reduction threshold to prevent overfitting. Once these criteria are satisfied, the tree-building process halts, generating a leaf node. After constructing the full tree, CART may undergo pruning to enhance generalization by removing subtrees that do not significantly improve predictive accuracy. To make predictions, a new data point traverses the tree from the root to a leaf node.

REP constructs a decision or regression tree by employing information gain/variance reduction and prunes it using reduced-error pruning (Witten & Frank, 2002). The REP

tree features nodes representing decisions based on features and leaves representing class labels or distributions. The unique strategy of reduced error-pruning in REPTree involves eliminating branches that do not enhance performance on validation data after building an initial DT with training data. This approach helps prevent overfitting, making the tree more adaptable to new data. REPTree employs impurity measures such as Gini impurity or entropy to identify the optimal feature for splitting.

ET method constructs an ensemble of unpruned decision trees using a top-down approach. In contrast to other ensemble-based classifiers that utilize bootstrap replicas of the data sample, ET grows trees using the entire learning sample (Geurts et al., 2006). The algorithm forms an ensemble model by iteratively splitting the original learning sample. The key parameters governing the algorithm are k, denoting the number of attributes randomly selected at each node, and nmin, specifying the minimum sample size for node splitting.

NB classifier is a straightforward yet potent predictive modeling tool. It learns parameters by analyzing each feature individually and computing simple per-class statistics. The model incorporates both the probability and conditional probability for each class from the training data. Using the probability model, Bayes' theorem is employed to predict new data. When dealing with real-valued data, evaluating these probabilities is facilitated with the assistance of the Gaussian distribution (Shobha & Rangaswamy, 2018).

KNN selects the K nearest data points from the training set based on the distance between the new data point and all data points in the set. In classification, it tallies the class labels of these K neighbors and determines the most frequently occurring class label, assigning it as the predicted class label for the new data point. In regression, the algorithm calculates the average (mean) of the target values of these K neighbors, designating the mean as the predicted target value for the new data point (Rasjid & Setiawan, 2017).

SVM is a machine learning algorithm that employs decision planes to classify data based on decision boundaries. Each sequence's residual attributes are encoded to create a specific feature vector. The input vectors are transformed into a high-dimensional space during SVM model training. The algorithm then seeks to locate separating hyperplanes that maximize the margin between datasets for each available class in an N-dimensional space (Castelli et al., 2018). SVM utilizes radial basis or polynomial function kernels to transform input data into a higher-dimensional space. In regression, it focuses on fitting a hyperplane that captures a margin of error specified around the data points. The training set comprises input-output pairs, with each input corre-

sponding to a target value. The objective is to find the hyperplane that minimizes the deviation of predicted values from actual values within the designated error margin. SVM can use a kernel function to handle non-linear relationships between inputs and targets. It is versatile, handling both linear and non-linear data, and has been shown to outperform ANN and DT without requiring a large training dataset.

Bagging is an ensemble learning technique that helps reduce both dataset variance and bias. The approach consists of creating several random subsets of training data. This is achieved using bootstrap sampling, where data points are randomly selected from the original training dataset with replacement. After generating these bootstrap samples, each decision tree model is trained independently. Each model captures varying patterns and relationships in the data. During the prediction phase, each independently trained model makes predictions based on the test data. The predictions from each model are averaged for regression tasks, while a majority or weighted vote is used to combine the predictions for classification tasks. The use of Bagging effectively improves the model's performance and robustness by leveraging diverse models trained on different subsets of the data.

CNN is a deep neural network used for classification. It consists of one or more convolutional layers. The neurons in the convolutional layer perform discrete convolutions on the output of the preceding layer. The input layer takes a matrix of input data values. The dimensions of this matrix depend on the dataset. The convolutional layers apply filters (kernels) to the input to create feature maps. The pooling layers (e.g., max pooling) can be used to reduce the spatial dimensions of the input. This helps in reducing computational complexity and focusing on the most important features. The flattening layer can be used to flatten the output of the preceding layer into a one-dimensional vector. The fully connected layers or dense layers in which each neuron is connected to every neuron in the previous layer and neuron weights are adjusted during training to learn the patterns in the data. The output layer produces the final classification. The number of neurons in the output layer corresponds to the number of classes in the classification task. Common activation functions include softmax for multi-class classification. The activation function can be applied at the convolutional, dense, and output layers to introduce non-linearity. The activation function typically includes the Rectified Linear Unit (ReLU), Sigmoid, and Softmax. The ReLU activation function is represented in Eq. (1.1) (Tharsanee et al., 2021). If x is positive, it directly outputs the input, x; otherwise, it outputs zero, and the neurons will be disconnected from the network.

$$ReLU(x) = \begin{cases} x, & if \quad x \quad is \quad positive \\ 0, & otherwise \end{cases} \tag{1.1}$$

The Sigmoid activation function is represented by $\sigma(x)$ and is defined in Eq. (1.2) (Chakraborty et al., 2018). It maps an input to a value between 0 and 1.

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{1.2}$$

The Softmax activation function converts raw scores into probabilities, ensuring the sum of probabilities is 1. It is denoted as in Eq. (1.3) ((Gao et al., 2021)), where x is the vector of raw outputs from the neural network. The i^{th} entry in the output vector indicates the predicted probability of input belonging to class i.

$$Softmax(x)_i = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}} \tag{1.3}$$

ELMs builds an ensemble of unpruned decision trees using a top-down approach. It splits nodes by selecting cut points completely at random. Unlike other ensemble-based classifiers, rather than using a bootstrap replica of the data sample, the trees are grown using the full learning sample (Geurts et al., 2006). ELMs are feedforward neural networks with a single layer of hidden nodes or numerous layers, which converge faster than traditional methods and produce promising results. The ELM does not require hidden nodes/neurons to be adjusted, contrary to popular perception in neural network generalization theory, linear theory, and control theory. Unlike ANN, which distributes hidden nodes regularly, ELM assigns hidden nodes randomly, creates biases and input weights for hidden layers, and uses least-squares methods to estimate output weights. By randomly allocating weights, the ELM avoids iterative training that creates local minima (Kouadio et al., 2018).

MLP is a feedforward neural network that utilizes interconnected neurons organized into multiple layers. This architecture facilitates data flow from the input layer through one or more hidden layers, culminating in the output layer. The input layer comprises neurons representing the features of the analyzed data. Hidden layers, one or more in number, can intervene between the input and output layers. The quantity of hidden layers and neurons within each layer can be adjusted as hyperparameters. The output layer generates the final predictions, with each output node representing the probability of belonging to classes in classification tasks. In regression tasks, an output node typically produces a continuous numerical prediction. Within an MLP, each neuron conducts a weighted sum of its inputs, applies an activation function, and generates an output. The activation function introduces non-linearity, enabling the network to capture intricate relationships in the data. Available activation functions

include Sigmoid, Softmax, and ReLU. Connections between neurons are associated with weights, determining the strength of connections. Additionally, each neuron has an associated bias term contributing to the weighted sum shift. Data is fed into the network from input to output layers during the forward propagation phase. Neurons calculate their weighted sum, apply the activation function, and produce outputs. Backpropagation is employed to update weights and biases to train an MLP. This involves computing gradients of a loss function based on the network's parameters and adjusting parameters using optimization algorithms like gradient descent.

1.5 Motivation for Soil Fertility Classification

Maintaining soil fertility is crucial for sustainable development, encompassing environmental, economic, and social aspects. The decline in soil fertility and improper fertilizer management can lead to food crises, affecting the global population. Fertile soil is essential to support plant growth, yield and to produce healthy food. Each soil nutrient plays a specific role during the crop growth cycle. Inadequate or excessive nutrition can lead to crop diseases, reduced crop yield and quality, and the growth of bacteria or pests, ultimately limiting crop yield. In addition, climate conditions can also impact the soil parameters, reducing soil fertility and causing environmental pollution. An adequate amount of fertilizers must be applied appropriately to improve crop productivity and limit fertilization costs. While evaluating soil nutrient levels in the laboratory can be costly and time-consuming and generate chemical residues, non-chemical methods such as spectroscopy or hyperspectral airborne sensors require high initial equipment, maintenance, and labor costs. Therefore, there is a need for a cost-effective and robust approach to classify soil fertility.

1.6 Challenges in Classifying Soil Fertility

The soil fertility classification to increase agriculture productivity involves various challenges.

- Measuring soil nutrient levels in the laboratory can be time-consuming and expensive. It also has the potential to create chemical residues, which can harm the environment. While non-chemical techniques like hyperspectral airborne and multispectral sensors can measure soil nutrients, they require a significant initial investment in equipment, maintenance, and labor. Therefore, collecting data using real-time proximal sensors is necessary to address these challenges.

- The traditional practice of assessing soil fertility involves manual collection of soil samples, followed by chemical analysis and manual data entry. This process often results in missing values and irrelevant data, making data preparation insignificant.

- Soil fertility varies based on soil properties, and using a global taxonomy for data labeling is crucial.

- Soil testing is usually done before crops are cultivated using traditional methods. However, this approach may lead to imbalanced data, which can cause overfitting or underfitting of the learning model. The main challenge is to select an appropriate oversampling technique to increase the samples of the minority classes. Choosing and fine-tuning the hyperparameters is also necessary to ensure accurate classifications.

- Determining the right fertilizer prescription requires considering the region, crop type, cultivation season, and climate conditions. Additionally, organic or chemical fertilizers may be necessary to address nutrient deficiencies in the soil. Choosing the appropriate type and quantity of fertilizer can be a difficult task.

- Farmers need to receive early warning messages about soil fertility to ensure they can take the necessary steps to produce high-quality crops in large quantities. Additionally, there is a requirement for a low-cost and accurate model for estimating soil fertility based on real-time data.

- Soil nutrients play a crucial role in the growth cycle of crops. Each nutrient has specific functions, and both a deficiency and an excess of these nutrients can lead to crop disease and a reduction in yield and quality. The application of fertilizer depends on the crop type and the specific location. Therefore, applying an appropriate amount of fertilizer at the right time for each crop and geolocation is important.

1.7 Research Contributions

This research aims to analyze farming from the viewpoint of reducing the costs of soil fertilization with high crop productivity and quality. In this context, the prime objective is to classify soil fertility as LOW, MEDIUM, or HIGH based on the chemical parameters of the soil using ML-based and deep learning-based techniques.

The significant contributions of this research work are:

- It proposed ML-based soil fertility classifiers to classify soil fertility based on soil chemical parameters. The ML-based classifiers such as NB, LR, SVM, J48, Bagging, BRT, and RF were used to classify the soil as LOW, MEDIUM and HIGH fertile soil. The classifiers were developed using laboratory-measured soil chemical parameters such as *EC, pH, OC, P, K, S, Zn, B, Fe, Cu*, and *Mn*.

- The site-specific soil parameters are essential for precise soil fertility classification. In this research values for soil chemical parameters such as *EC, pH, OC*, and *N* are derived using Sentinel-2 data. The generated dataset is labeled using various clustering approaches. To increase the clustering accuracy, this study suggests using Canopy Center-based Fuzzy-C-Means clustering and compared it with manual labeling and other clustering techniques such as Canopy, Density-based, Expectation-Maximization, Farthest-first, k-Means, and Fuzzy-C-Means clustering. ML-based classifiers such as NB, SVM, J48, and RF were applied on the generated dataset, to classify the soil fertility as LOW, or, MEDIUM, or HIGH.

- This research work focuses on developing a robust machine learning-based classification approach by employing prominent features recommended by the ensemble filter-based feature selection. To overcome the inconsistency in generating different feature scores, an ensemble filter-based feature selection is devised using three different filter-based feature selection approaches: Information Gain (InfoG), Gain Ratio (GainR), and Relief Feature (ReliefF). Two different datasets of different climate zones were used to evaluate the robustness of the proposed approach. The proposed method selects relevant features among 11 soil parameters such as *EC, pH, OC, B, Cu, Fe, Mn, P, K, S*, and *Zn* present in the datasets. Furthermore, ML-based classifiers such as CART, ET, J48, RF, REP, NB, and SVM were employed to classify the soil fertility. The proposed method includes fertilizer prescriptions based on classification results.

- The research introduces a 2D CNN-based soil Fertility classifier and fertilizer prescription. The soil fertility is classified as HIGH, MEDIUM, or LOW fertile based on the chemical measurements of soil parameters, including *pH, EC, OC, P, K, S, Zn, B, Fe, Cu*, and *Mn*. The experiments were conducted by varying kernel size from 3×3 to 7×7 and input grid size from 11×11 to 13×13. Further, to improve the performance of the classifier the dataset was oversampled using Synthetic Minority Oversampling (SMOTE) technique, and the experiments were conducted using the oversampled dataset.

- To overcome the limitations of 2D CNN-based classifier, the research proposes a 1D CNN-based soil fertility classifier. To classify soil fertility, the classifier employs laboratory-measured soil data that encompasses *EC, pH, OC, K, P, S, B, Cu, Fe, Mn*, and *Zn*. The proposed approach employs MinMax normalization and the SMOTE to improve the classifier performance. The results of soil classification are used to recommend fertilizers. The performance of the proposed approach was compared with ELM and MLP classifiers.

- The Symbolic Deterministic Finite Automata (SDFA)-based soil fertility classifier was proposed to classify soil fertility as LOW, or MEDIUM, or HIGH based on *pH, EC, OC*, and *N*. The proposed approach was assessed using Sentinel-2 remotely sensed data and laboratory-measured soil-health data. The results of soil fertility classification were used to recommend fertilizers.

- The real-time site-specific soil chemical parameter values are acquired using soil proximal sensors such as NPK and EC-pH. The robustness of proposed ML-based classifiers was tested using the collected real-time soil data.

1.8 Thesis Outline

This thesis work is organized into several chapters. Chapter 2 explores the method used in the literature review and previous research on predicting or classifying soil fertility. Problem definition and objectives of the research work are outlined based on observed research gaps. In Chapter 3, ML-based classifiers for soil fertility classification are discussed. Chapter 4 explores the automated fertilizer prescription modules developed based on the classification results. Chapter 5 discuses on soil fertility classification using Real-Time soil data. Chapter 6 delves into the conclusions and future works.

CHAPTER 2

LITERATURE SURVEY

To identify nondestructive works dealing with soil fertility estimation, a review was conducted by following the steps described in PRISMA (Moher et al., 2009; Page et al., 2021). Using the following research questions, the study attempted to identify relevant articles.

1. What are the different ways of assessing soil fertility?
2. Which are predominantly used soil properties in predicting or classifying soil fertility?
3. What are the different ways of collecting soil data?
4. Which are different regions used for the study?
5. How was the performance of models evaluated?

This systematic review followed the methodology shown in Figure 2.1 (adapted from Page et al. (2021)).
Identification: The relevant articles in the Scopus database were searched using the following queries:

1. TITLE-ABS-KEY(soil AND (fertility OR ”spatial prediction”) AND (”machine learning” OR ”deep learning” OR ”artificial neural network” OR ”remote sensing”)) AND soil AND (chemical OR properties OR nutrient) AND (”machine learning” OR ”deep learning” OR ”neural networks” OR ”statistical analysis”)
2. TITLE-ABS-KEY((”satellite image” OR ”spectral features” OR ”soil properties” OR nutrient) AND (agriculture OR farming OR field OR agricultural OR agriculturally) AND (”remote sensing” OR sensing OR sensor) AND (estimation OR prediction OR classification)) AND soil AND (chemical OR properties OR nutrient) AND (”machine learning” OR ”deep learning” OR ”neural networks” OR ”statistical analysis”)
3. TITLE-ABS-KEY(soil AND fertilizer AND (recommendation OR regulation OR prescription OR management) AND (”machine learning” OR ”deep learning” OR

sensing OR algorithm OR (precision AND model)) OR ”manure recommendation”) AND soil AND (chemical OR properties OR nutrient) AND (”machine learning” OR ”deep learning” OR ”neural networks” OR ”statistical analysis”)

The queries filter articles based on approaches and soil properties used by searching for a combination of terms such as soil and (machine learning/ deep learning/ neural networks/ statistical analysis) and (chemical/ properties/ nutrient) in the article content. Additionally, each query searches for specific keywords, abstracts, or titles. In query 1, articles that discuss soil fertility or parameters derived from spatial prediction. Query 2 filters articles that use remote sensing or sensors to derive data and Query 3 searches for articles on fertilizer prescription. The execution of these three queries on September 8th, 2023, resulted in 492, 638, and 297 articles, respectively, resulting in 1427 articles from 2017 to 2023. Combining three queries removed the 105 duplicate articles.

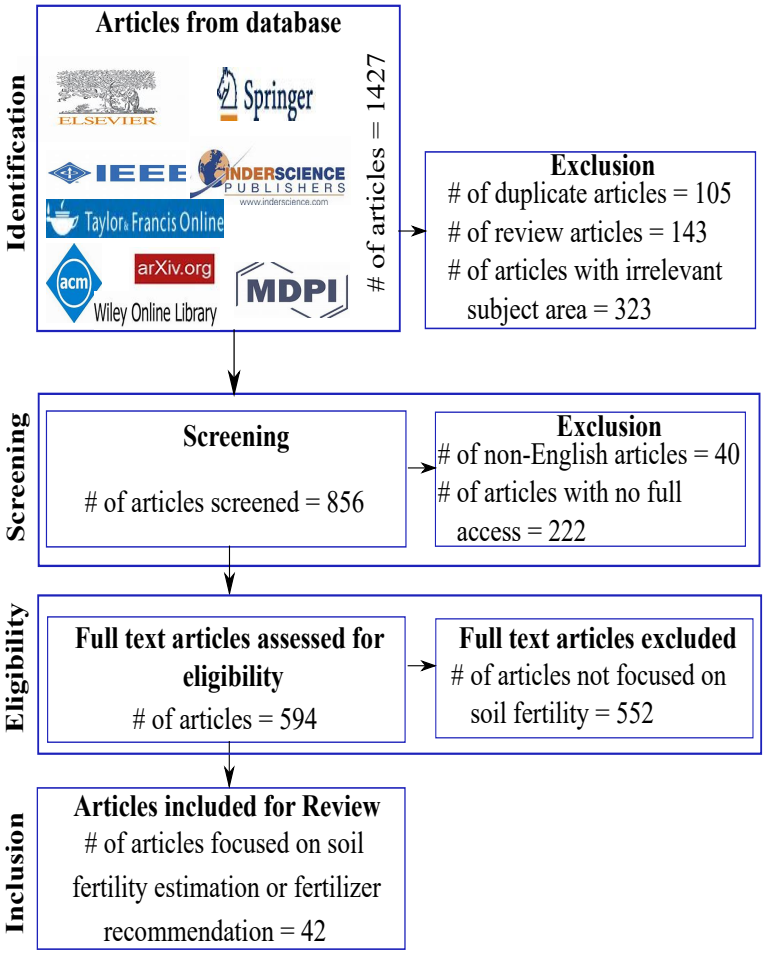


Figure 2.1: Methodology used in review process

Furthermore, the 143 review articles were removed by searching for the document type ”Review” and terms in the title such as case study, literature, review, overview, study, or

survey. The remaining 1179 articles were further analyzed for irrelevant subject areas and 323 articles belonging to irrelevant subject areas were removed.

Screening: A screening of the 856 articles resulted in 40 non-English articles and 816 English articles. In addition, 594 full-text English articles were downloaded through "Scopus Document Download Manager" (229 articles), followed by downloads from publishers' sites (365 articles).

Eligibility: The 594 articles were filtered to remove those not addressing soil fertility estimation. There were 552 articles focused on irrelevant topics, such as soil moisture landslides, crop yield, soil type, texture, and erosion.

Inclusion: A detailed study using 42 relevant articles was carried out to analyze parameters, study area, and methodology used. Figure 2.2 presents the number of articles reviewed.

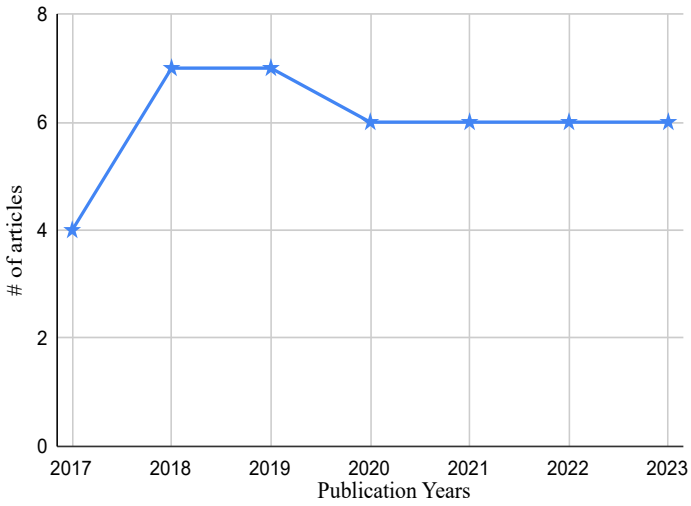


Figure 2.2: Number of articles reviewed

2.1 Overview on Prior Works on Estimation of Soil Fertility

The assessment of articles on soil fertility relies on the methodology employed for estimating soil fertility and the soil properties considered. Numerous studies applied ML methods for soil fertility estimation. Table 2.1 presents an overview of soil fertility prediction or classification research. Researchers utilized regression techniques such as LinR, MLR, ER, RR, SR, LASSO, CR, PCR, PLSR, iPLSR, GPR, BRT, and Cubist, along with classification approaches like LR and hybrid classification-regression approaches such as DT, RF, CART, REP, M5 tree, ET, NB, KNN, and SVM to predict variations in soil parameters or classify soil fertility.

Most prior works employed regression techniques to predict variations in soil parameters such as Cu, EC, K, N, OC, OC stock, OM, P, salinity, etc. Büchele et al.

(2019) predicted variations in *Fe* and *K*, while Tavares et al. (2023) predicted variations in *Ca* and *K*. Inoue et al. (2020) predicted soil fertility based on variations in *OC* and *N*. Furthermore, Chougule et al. (2019) and Coulibali et al. (2020) predicted variations in *N*, *P*, and *K*, while Abera et al. (2022) and Hu et al. (2023) predicted variations in *N, P, K, S* and *N, P, K pH, CEC, OM*, respectively. Hengl et al. (2017) predicted variations in *OC* and total *N*, total *P*, and extractable *P, K, Ca, Mg, S, Na, Fe, Mn, Zn, Cu, Al*, and *B*, whereas Kouadio et al. (2018) predicted soil fertility based on *pH, B, Ca, K, Mg, N, P, S, Zn*, and *OM*. Keshavarzi et al. (2023) predicted variations in *Fe, Mn, Zn*, and *Cu*. To enhance accuracy, Sirsat et al. (2018) utilized VFIs of *Fe, Mn,* P_2O_5*, OC*, and *Zn*.

Classification methods were also employed in soil fertility studies. For instance, Sirsat et al. (2017) utilized various methods such as DT, KNN, RF, NB, REP, SVM, MLP, RBF, bagging, and AdaBoost to classify soil fertility into low, medium, or high categories based on laboratory measurements of several factors including *EC, pH,* N_2O*,* K_2O*,* SO_4*, Zn, Fe, OC, Mn, and* P_2O_5. Bagherzadeh & Gholizadeh (2017) used MLP to classify soil fertility as high, moderate, or marginal based on laboratory measurements of *EC*, Calcium Carbonate ($CaCO_3$), *OC*, and *pH*. Additionally, Khanal et al. (2018) classified soil fertility using remotely sensed soil images, and Fernandes et al. (2019) employed the MLP classifier to classify the fertility of *OM* as good or excellent. Moreover, Delavar et al. (2020) classified soil fertility based on salinity into different classes, and Gulhane et al. (2023) used the MLP classifier to classify soil fertility into LOW, NORMAL, or HIGH categories based on laboratory measurements of *P, Fe*, and *pH*. Few researchers suggest fertilizers to increase soil fertility. Chougule et al. (2019) recommended fertilizers to increase *N, P* and *K*. Ransom et al. (2019) recommended nitrogen fertilizer. Coulibali et al. (2020) suggested the *N, P, K* fertilizer dosage. Abera et al. (2022) recommended fertilizers to improve *N*, *P*, *K*, and *S*. Sirsat et al. (2017) recommended fertilizers based on N_2O*,* K_2O*,* P_2O_5.

Table 2.2 summarizes the soil parameters employed in previous research. Researchers commonly used laboratory-measured soil parameters such as *EC, pH, OC*, soil minerals: *Al, B, Ca, Cl, Cu, Fe, K, Mg, Mn, N, P* and extractable *P, S, Zn, OC* stock, salinity, *OM, CEC*, sodium (*Na*), N_2O*,* K_2O*,* SO_4, and P_2O_5, potential acidity, carbon (*C*), inorganic *C*, carbonate, bicarbonate, $CaCO_3$, *Na* absorption ratio. Some researchers used multispectral images to predict variations in *pH, K, Mg, OM*, and *CEC* (Khanal et al., 2018).

Few studies derived soil chemical parameters using satellite spectral bands. For instance, Morgan et al. (2018), Gorji et al. (2020), and Delavar et al. (2020) derived

salinity index (SI) using Sentinel-2 and Landsat-ETM+ spectral bands. Li et al. (2021) derived *OC* using Sentinel spectral bands, while Peng et al. (2021) derived *N*, *P, K* using hyperspectral data. Gulhane et al. (2023) used Sentinel-2 and Landsat-8 spectral band information to derive *P, Fe*, and *pH*. Gholizadeh et al. (2018) compared soil texture and mapping *OC* using Sentinel-2 satellite imagery data with those obtained from airborne hyperspectral sensors and a spectroradiometer. Zia et al. (2019) proposed a real-time low-dimensional prediction model to predict the total loss of *N* by measuring *N* in fertilizer and manure using proximal sensors. In addition to soil chemical parameters, researchers also considered the physical properties of the soil, including soil texture, soil type, and soil density, among others.

Table 2.3 presents the study areas and auxiliary predictors used in the evaluated works. Various auxiliary predictors such as the Enhanced Vegetation Index (EVI), Canopy Red Edge Spectral Index (CRSI), Normalized Difference Vegetation Index (NDVI), Enhanced NDVI (ENDVI), Enhanced Exaggerated Vegetation Index (EEVI), Soil-Adjusted Vegetation Index (SAVI), Green Difference Vegetation Index (GDVI), Simple Ratio (SRatio), Enhanced Vegetation Index-2 (EVI2), Topographic Wetness Index (TWI), Near-Infrared (NIR), and Visible-NIR (VNIR) were utilized to assess geographic regions and their impact on crop growth or yield.

2.2 Detailed Discussion on Review Articles

The researchers focused on the prediction or classification of soil fertility.
Bagherzadeh & Gholizadeh (2017) employed an MLP to predict land suitability for alfalfa cultivation and classify its fertility. The predictors used in the model included laboratory measurements of important soil parameters such as *EC*, $CaCO_3$, *OC*, *pH*, and soil texture. Additionally, climate parameters such as mean annual precipitation and temperature and topological parameters, including depth, gravel content, slope, gypsum presence, drainage conditions, and flooding, were considered. Based on these predictors, the authors calculated a land suitability index that classified land fertility as highly suitable, moderately suitable, marginally suitable, marginally not suitable, or permanently unsuitable. The model's performance was evaluated using correlation coefficient values (R^2).

Table 2.1: Classification of soil fertility: an overview

Author(s)	Regression technique used (Yes/No)	Classification technique used (Yes/ No)	Learning model	Purpose	Fertilizers recommended (Yes/No)
Bagherzadeh & Gholizadeh (2017)	Yes	Yes	MLP	To predict land suitability to cultivate based on soil properties. The fertility of the land is classified as High, Moderate, Marginal.	No
Hengl et al. (2017)	Yes	No	XGBoost	Spatial prediction of soil macro and micro-nutrient	No
Schillaci et al. (2017)	Yes	No	BRT	To estimate the variation in *OC*	No
Sirsat et al. (2017)	No	Yes	AdaBoost, bagging, DT, KNN, RF, NB, REP, SVM, MLP, and RBFNN	To classify soil fertility as LOW, MEDIUM, or HIGH using soil chemical parameters. Fertilizer recommendation based on N_2O, K_2O, P_2O_5	Yes
Deng et al. (2018)	Yes	No	BRT, LinR, GPR and RF	To predict *OC* stock	No
Khanal et al. (2018)	Yes	Yes	LinR, CR, GB, MLP, and SVM	To predict soil parameters and crop yield. Remotely sensed soil images were classified into three classes: light, medium, and dark.	No
Kouadio et al. (2018)	Yes	No	RF and MLR	To analyze soil fertility and forecast coffee yield	No
Gholizadeh et al. (2018)	Yes	No	MV	To predict variations is OC and texture	No
Morgan et al. (2018)	Yes	No	MLR	To monitor salinity	No
Sirsat et al. (2018)	Yes	No	bagging, boosting, LASSO, RR, Bayesian Neural Networks (BNNs), ET, SVM, RF, MLP, DNN	Prediction of VFIs of *Fe, Mn,* P_2O_5*, OC*, and *Zn*	No
Wang et al. (2018)	Yes	No	BRT and RF	Estimate variation in *OC* stock	No
Büchele et al. (2019)	Yes	No	PCA with PLSR	Prediction of Fe and K	No
Chougule et al. (2019)	Yes	No	RF and k-means	Estimation of variations in *K, N, P*. Fertilizers recommendation based on K, *N*, and P.	Yes

Table 2.1: Classification of soil fertility: an overview Contd.

Author(s)	Regression technique used (Yes/No)	Classification technique used (Yes/ No)	Learning model	Purpose	Fertilizers recommended (Yes/No)
Fernandes et al. (2019)	Yes	Yes	MLP	Prediction of *OM* based on other soil parameters and classification of fertility of *OM* as Good or Excellent	No
Ghorbani et al. (2019)	Yes	No	MLP-FFA, MLP and GPR	Estimation of *EC*	No
Ransom et al. (2019)	Yes	No	ER, PLS, LASSO, RR, PCR, SR,DT, and RF	Estimation of change in fertility of N	Yes
Zhang et al. (2019)	Yes	No	PLS, MLR, and SVM.	Estimation of variation in *N*	No
Zia et al. (2019)	Yes	No	M5 tree, MLR, MLP, and REPTree	To predict the total loss of *N*	No
Coulibali et al. (2020)	Yes	No	KNN, RF, ANN	To predict NPK fertilizer dosage for crop	Yes
Delavar et al. (2020)	Yes	Yes	hybrid model using ANN and GA	Prediction of soil salinity and classification into classes from S0 to S5.	No
Gorji et al. (2020)	Yes	No	LinR and MLR	Prediction of variation in salinity	No
Inoue et al. (2020)	Yes	No	PLSR, and interval PLSR	assessment of soil fertility	No
Mahmoudzadeh et al. (2020)	Yes	No	Cubist, KNN, SVM, and XGBoost, RF	Prediction of variation in *OC*	No
Zhang et al. (2020)	Yes	No	MLR, and ANN	Estimation of *K*	No
Hossen et al. (2021)	Yes	No	SVM, and MLP	Estimation of *N*	No
Li et al. (2021)	Yes	No	RF and Cubist	Prediction of variation in *OC*	No
Parsaie et al. (2021)	Yes	No	RF, DT, and Cubist	Prediction of *N*	No
Peng et al. (2021)	Yes	No	BPNN, LASSO, gradient boosting, DT, MLR, RR, and SVM	Prediction of soil nutrients *N*, *P*, k	No
Shang et al. (2021)	Yes	No	GPR, boosting, bagging, and SVM	Estimation of *Cu*	No
Shi et al. (2021)	Yes	No	LASSO with PLSR, ANN, SVM, RF	Estimation of *OM*	No

Table 2.1: Classification of soil fertility: an overview Contd.

Author(s)	Regression technique used (Yes/No)	Classification technique used (Yes/ No)	Learning model	Purpose	Fertilizers recommended (Yes/No)
Abera et al. (2022)	Yes	No	RF	Prediction of variation in *N*, *P, K*, and *S*	Yes
Aksoy et al. (2022)	Yes	No	CART, RF, and SVM	To predict variation in salinity	No
Al Mas-moudi et al. (2022)	Yes	No	MLR, SVM, and RF with PCA, and hierarchical clustering	To predict fertility level of *OM, K_2O* and *P_2O_5*	No
Chang et al. (2022)	Yes	No	Dynamic fitness inertia weighted particle swarm optimization (DPSO) with BPNN	To monitor *OM*	No
Xu et al. (2022)	Yes	No	PLSR, RF, and SVM	To estimate spatial variation in *N*	No
Yang et al. (2022)	Yes	No	Particle Swarm Optimization, Ant Colony Optimization, and Simulated Annealing with PLSR, BPNN	To estimate OM	No
Chen et al. (2023)	Yes	No	SVM, DT, RF, XGBoost, and LightGBM	To estimate changes in fertility of *P*	No
Dos Santos et al. (2023)	Yes	No	LASSO, and PLS	To estimate *OC*	No
Gulhane et al. (2023)	Yes	Yes	MLP	Classification of fertility level of soil parameters into Low, Normal, or High. prediction of fertility indices of soil parameters.	No
Hu et al. (2023)	Yes	No	RF	To predict variations in OM, *N*, *P, K, pH*, and *CEC*	No
Keshavarzi et al. (2023)	Yes	No	RF, and SVM	To predict the content of *Fe, Mn*, *Zn* and *Cu*	No
Tavares et al. (2023)	Yes	No	LinR, MLR, PLSR, and RF	To predict *Ca* and *K*	No

Table 2.2: Parameters used in previous research works

Author(s)	Laboratory-measured soil chemical parameters	Satellite-derived soil chemical parameter indices	Real-time soil chemical parameters	Soil physical/ biological parameters
Bagherzadeh & Gholizadeh (2017)	*EC, $CaCO_3$, OC, pH*	-	-	texture
Hengl et al. (2017)	*OC*, *N*, total *P*, and extractable *P, K, Ca, Mg, S, Fe, Mn, Zn, Cu, Al, B, Na*	-	-	-
Schillaci et al. (2017)	OC	-	-	texture
Sirsat et al. (2017)	*EC, pH, N_2O, K_2O, SO_4, Zn, Fe, OC, Mn, P_2O_5*	-	-	soil type
Deng et al. (2018)	*OC* stock, *pH*, salinity, total *N*, *P, K, CEC*	-	-	moisture, texture, soil type, coarse fragment, and water table depth, evaporation, ground surface temperature
Khanal et al. (2018)	*pH, K, Mg, CEC, OM*	-	-	-
Kouadio et al. (2018)	*pH*, *B, Ca, K, Mg*, *N*, *P, S, Zn, OM*	-	-	-
Gholizadeh et al. (2018)	OC	OC	-	Texture: clay, silt, sand
Morgan et al. (2018)	salinity	SI	-	-
Sirsat et al. (2018)	*EC, OC, N_2O, P_2O_5, K_2O, SO_4, Cu, Fe, Mn, Zn, B*	-	-	-
Wang et al. (2018)	*OC*	-	-	-
Büchele et al. (2019)	*Fe, K*	-	-	texture: clay, loam, silt, and sand
Chougule et al. (2019)	*N, K, P*	-	-	soil type
Fernandes et al. (2019)	*OM*, potential acidity, *pH, Ca, Mg*	-	-	-
Ghorbani et al. (2019)	EC	-	-	-
Ransom et al. (2019)	*OC*, *N*, *CEC, C*, inorganic *C, OM, pH*	-	-	bulk density, texture: clay, sand, silt,
Zhang et al. (2019)	N	-	-	soil type
Zia et al. (2019)	-	-	N in fertilizer and manure	soil moisture
Coulibali et al. (2020)	pH, extractable *P, K, Al, Mg, Ca*	-	-	texture: clay, silt, and sand

Table 2.2: Parameters used in previous research works Contd.

Author(s)	Laboratory-measured soil chemical parameters	Satellite-derived soil chemical parameter indices	Real-time soil chemical parameters	Soil physical/ biological parameters
Delavar et al. (2020)	EC, pH, Ca, $CaCO_3$, K, soluble Na, Na absorption ratio	salinity indices: SI1, SI2, SI3	-	-
Gorji et al. (2020)	*Mg, Cl*, carbonate, bicarbonate	salinity indices: SI1, SI2, SI3	-	-
Inoue et al. (2020)	*OC, N*	-	-	-
Mahmoudzadeh et al. (2020)	*OC*	-	-	-
Zhang et al. (2020)	K	-	-	-
Hossen et al. (2021)	*N*	-	-	-
Li et al. (2021)	*OC*	*OC*	-	-
Parsaie et al. (2021)	*$CaCO_3$, OC, N*	-	-	texture: sand, silt, clay
Peng et al. (2021)	*N, P, K*	*N, P, K*	-	-
Shang et al. (2021)	*Cu*	-	-	-
Shi et al. (2021)	*OM*	-	-	-
Abera et al. (2022)	*N, P, K, S, OC, pH* and *CEC*	-	-	texture: clay, silt
Aksoy et al. (2022)	EC	-	-	-
Al Masmoudi et al. (2022)	*OM, K_2O and P_2O_5*	-	-	-
Chang et al. (2022)	*OM*	-	-	-
Xu et al. (2022)	*N*	-	-	-
Yang et al. (2022)	*OC*	-	-	-
Chen et al. (2023)	*P*	-	-	-
Dos Santos et al. (2023)	*OC*	-	-	-
Gulhane et al. (2023)	*P, Fe, pH*	*P, Fe*, and *pH*	-	-
Hu et al. (2023)	*OM, N*, *P*, K	-	-	-
Keshavarzi et al. (2023)	*Fe, Cu, Mn*, *Zn, pH*	-	-	clay, and sand
Tavares et al. (2023)	*Ca, K*	-	-	sandy-clay, sand-clay loam, and sandy-loam

Table 2.3: Region of study and auxiliary predictors used

Author(s)	Study region	Climate data used	Topographic data used	Crop type, yield data, or other satellite-derived indices used
Bagherzadeh & Gholizadeh (2017)	Joveyn plain, Khorasan-e-Razavi Province, northeast Iran	mean annual precipitation and mean annual temperature	depth, gravel, slope	Gypsum, drainage, flooding
Hengl et al. (2017)	Sub-Saharan Africa	-	-	landform, lithologic and land cover maps, EVI
Schillaci et al. (2017)	Sicily, Italy	rainfall, temperature	soil erosion or decomposition, land use	Landsat-5 indices: NDVI, EVI, SAVI, MSAVI, NDMI, NBR, NBR2, NDSI
Sirsat et al. (2017)	Marathwad, Maharashtra (State), India	-	-	crops: bajra, cotton, and soybean
Deng et al. (2018)	Zhejiang Province, East China	temperature, precipitation, air pressure, and sunshine	land cover, topography, geographical location, and farming practices	NDVI
Khanal et al. (2018)	Madison County, Ohio, USA	-	aspect, elevation, roughness, slope, flow direction	corn
Kouadio et al. (2018)	Central Highlands, Southern Vietnam	rainfall, temperature, and solar radiation	-	coffee
Gholizadeh et al. (2018)	Czech Republic	-	-	-
Morgan et al. (2018)	northwest Cairo	-	-	NDVI, MSI, SI
Sirsat et al. (2018)	Marathwad, Maharashtra (State), India	-	-	-
Wang et al. (2018)	eastern Australia	rainfall and temperature	slope, elevation,	NDVI, EVI, SAVI, MSAVI, NDMI, NBR, NBR2, latitude, longitude, fractional cover data, radiometric: potassium, uranium, thorium
Büchele et al. (2019)	12 different study sites all across Germany	rainfall, and temperature	aspect, TWI, partial contributing area, multi-resolution valley bottom flatness, multi-resolution ridge top flatness, plan curvature, profile curvature	NDVI, EVI, SAVI, modified-SAVI, NDMI, NBR, NBR2, fractional cover data, radiometric
Chougule et al. (2019)	Maharashtra (State), India	-	-	-

Table 2.3: Region of study and auxiliary predictors used Contd.

Author(s)	Study region	Climate data used	Topographic data used	Crop type, yield data, or other satellite-derived indices used
Fernandes et al. (2019)	São Paulo (State), Brazil	-	-	-
Ghorbani et al. (2019)	Soofiyan City, Tabriz regions	-	-	latitude and longitude
Ransom et al. (2019)	US Midwest	precipitation, rainfall, growing degree days, corn heat units, and Shannon diversity index of precipitation	-	-
Zhang et al. (2019)	North China	-	-	-
Zia et al. (2019)	University of Cork, Dripsey catchment, south of Ireland	precipitation		crop cover, total N applied, days since last application of *N*, day of the year
Coulibali et al. (2020)	Quebec, Canada	cumulative precipitation, Shannon Diversity Index for rainfall distribution, mean temperature, and number of growing degree days	-	Potato crops
Delavar et al. (2020)	South of Urmia Lake, Iran	-	-	NDVI
Gorji et al. (2020)	Western part of Urmia Lake, Iran	-	-	NDVI
Inoue et al. (2020)	paddy field in Tomioka-machi, Fukushima region in Japan	-	-	Paddy, RSI, NDSI
Mahmoudzadeh et al. (2020)	Kurdistan Province, Western Iran	rainfall, temperature	valley depth and terrain surface texture	NDVI, and SAVI
Zhang et al. (2020)	Chaobai River alluvial plain, Beijing, Tianjin City, and Hebei Province, China	-	-	Landsat TM and OLI bands

Table 2.3: Region of study and auxiliary predictors used Contd.

Author(s)	Study region	Climate data used	Topographic data used	Crop type, yield data, or other satellite-derived indices used
Hossen et al. (2021)	Sturgis, South Dakota, USA	humidity and temperature	-	NDVI and multispectral characteristics (Red, NIR, and Green)
Li et al. (2021)	Ebinur Lake Basin, Eurasia	-	slope, elevation, and depth	-
Parsaie et al. (2021)	Qorveh-Dehgolan Plain, Kurdistan province, Iran	-	elevation, aspect, valley depth, channel network distance, normalized height, convergence index, direct insulation, total insulation, midslope position	NDVI, B2, B11, SAVI,
Peng et al. (2021)	Conghua district, Guangdong Province, China	-	-	-
Shang et al. (2021)	Dexing City, Jiangxi Province, China	-	-	VNIR and SWIR
Shi et al. (2021)	Guangxi State-owned Huangmian Forest Farm, Liuzhou City, China	-	-	Red, Fast data lookup and Fraction of Data Remaining, VNIR
Abera et al. (2022)	Wheat growing regions in Ethiopia, East Africa	precipitation, maximum temperature, minimum temperature, and solar radiation.	elevation and topographic index.	wheat
Aksoy et al. (2022)	Western part of the Urmia Lake, Iran	-	-	NDVI, SAVI, EVI, GDVI, CRSI, SRatio, EVI2, ENDVI, and EEVI
Al Masmoudi et al. (2022)	Doukkala, Central Morocco	-	-	-

Table 2.3: Region of study and auxiliary predictors used Contd.

Author(s)	Study region	Climate data used	Topographic data used	Crop type, yield data, or other satellite-derived indices used
Chang et al. (2022)	northeast of Jiamusi City, Heilongjiang Province, Northeast China	-	-	Spectral reflectance feature bands of different wavelengths
Xu et al. (2022)	Kaifeng, north China Plain	-	elevation, and TWI	SWIR, NIR, MSI, NDSI, NDVI
Yang et al. (2022)	Aksu in Xinjiang, northwestern China	-	-	longitude and latitude
Chen et al. (2023)	Wheat and paddy farmlands in six climate zones of China	temperature, humidity, precipitation, and atmospheric pressure.	-	wheat, and paddy
Dos Santos et al. (2023)	Pernambuco, Northeastern Brazil	-	-	Spectroscopy reflectance data: VNIR and MIR
Gulhane et al. (2023)	Baggi, Ibrahimpur, Mogra, and Wai villages in Amravati (district), Maharashtra (State), India.	-	-	NDVI
Hu et al. (2023)	farmland soil of Jiangxi Province, southern China	mean annual temperature and precipitation	aspect, slope, TWI, and multi-resolution valley bottom flatness	population density
Keshavarzi et al. (2023)	Neyshabur Plain, Iran	mean annual temperature, precipitation, precipitation seasonality, total solar radiation	elevation, slope, TWI, Flow Accumulation, and Stream power index	Sentinel 2A: MSAVI2, topsoil grain size index, Saturation index, Normalized clay index, NDVI, and GNDVI
Tavares et al. (2023)	Field at the southeast and central-west region of Brazil	-	-	VNIR

Hengl et al. (2017) used XGBoost to predict variations in soil parameters, including *OC*, *N*, *P*, and extractable elements such as *P, K, Ca, Mg, S, Fe, Mn, Zn, Cu, Al, B*, and *Na* in soil samples collected from Sub-Saharan Africa. To enhance prediction accuracy, the researchers incorporated auxiliary predictors such as landforms, lithologic information, land cover maps, and EVI. The performance of the XGBoost model was evaluated using the R^2.

Schillaci et al. (2017) predicted the variation in soil *OC*. The study utilized a combination of laboratory measurements of *OC*, land use, and soil texture. Additionally, various environmental data, including Landsat indices, climate information (rainfall, temperature), and topographical data collected through Global Positioning System, were incorporated as predictors. The concentration of *OC* served as the target variable, and a total of 25 predictors were considered, encompassing factors such as rainfall, temperature, soil texture, topographic indices related to soil erosion or decomposition, land use, and Landsat-5 indices. Soil samples were collected from regions in Sicily, Italy, and Geographic Information System was utilized to collect environmental parameters. The performance of the approach was evaluated using standard deviation and pseudo-R^2. The results indicated that the remote sensing approach, incorporating factors such as rainfall, soil texture, temperature, land use, and topographic indices, was less uncertain compared to an approach without remote sensing. The study concluded that rainfall, soil texture, temperature, land use, and topographic indices were identified as key determinants of the concentration of soil *OC*. Moreover, remote sensing and topographic indices were effective in accurately estimating the concentration of *OC*.

Sirsat et al. (2017) compared twenty classifiers, including AdaBoost, Bagging, DT, KNN, RF, NB, and SVM. The aim of the study was to classify the VFI of four soil nutrients: *Fe, Mn, OC, and* P_2O_5. The classification was based on chemical measurements of EC, K_2O, N_2O, pH, SO_4, Zn, and soil type. The VFI for each soil nutrient was calculated by considering the number of patterns classified as LOW, MEDIUM, and HIGH in each village. The objective was to determine the fertility level of soil nutrients to provide recommendations for suitable crops and fertilizers. Fertilizer recommendations were tailored for three prominent crops in Marathawada, Maharashtra, namely, bajra, cotton, and soybean, based on measurements of K_2O, N_2O, and P_2O_5. The performance of each classifier was evaluated using Cohen kappa statistics, confusion matrix, and the Friedman rank test. The RF classifier demonstrated the highest accuracy in classifying soil type and OC-VFI, indicating its effectiveness in predicting soil fertility levels.

Deng et al. (2018) employed various regression models to estimate the amount of *OC* stock in the topsoil of cultivated regions in Zhejiang province, East China, throughout the year. The regression models included BRT, General Linear Model, kriging, and RF regressor. The study incorporated a wide range of predictor variables, encompassing *pH, salinity,* total *N, P, K*, CEC, soil moisture, soil texture, soil type, coarse fragments, water table depth, evaporation, ground surface temperature, air temperature, precipitation, air pressure, sunshine, geographical features land cover, topography, geographical location, and farming practices. The NDVI derived from Moderate-resolution Imag-

ing Spectroradiometer spectral bands was also considered a predictor variable. The RF model exhibited the best performance among the regression models, with the lowest RMSE and a high R^2. This indicates that RF was the most effective in estimating the *OC* stock in the topsoil based on the given set of predictor variables.

Khanal et al. (2018) utilized site-specific precision agriculture to generate high-quality and cost-effective yield and soil maps in a timely manner. They integrated ML techniques with remotely captured soil image maps to enhance the accuracy of predicting crop yield and fertility levels of *pH, K, Mg, OM*, and *CEC* in spatial dimensions compared to conventional methods. The data collection involved using multispectral aerial images captured by a Leica ADS80 airborne digital sensor and Lidar data. This information was used to derive topographic and yield data for corn crops. The field-based data included soil properties such as pH, K, Mg, CEC, and OM, collected from various fields in 2013. Several machine learning methods were employed, including LinR, CR, GB, MLP, and SVM with linear and radial kernels. Among these models, the MLP model exhibited superior performance, achieving the highest R^2 and the lowest RMSE. This indicates that the MLP model outperformed other machine learning models in accurately predicting crop yield and soil fertility levels based on the provided predictor variables.

Kouadio et al. (2018) aimed to assess the effectiveness of RF and MLR in analyzing soil fertility and predicting coffee yield, specifically focusing on Robusta coffee. The data collection spanned 2013 to 2014 and involved the analysis of various soil variables, including pH, B, Ca, K, Mg, N, P, S, Zn, and OM. The goal was to develop different models to select the most optimal soil parameters for predicting coffee yield. In addition to soil variables, the researchers considered climate variables such as rainfall, minimum and maximum temperature, and solar radiation to enhance the accuracy of yield predictions. Model performance was evaluated using several metrics, including RMSE, Mean Absolute Error (MAE), Nash-Sutcliffe's coefficient (NS), Willmott's Index (WI), and Legates-McCabe's Index (LMI). The results of the evaluation indicated that RF produced more accurate estimations compared to MLR. This suggests that the RF model outperformed MLR in predicting coffee yield based on the provided soil and climate variables.

Gholizadeh et al. (2018) compared the effectiveness of using different sources of data, including Sentinel-2 satellite imagery, a spectroradiometer, and airborne hyperspectral sensors, for mapping soil texture and *OC*. The data were collected from four agricultural sites situated in the Czech Republic. The researchers created a separate

model for each site using Multivariate (MV) regression techniques. To assess the performance of the models, they employed evaluation metrics such as the standard deviation, covariance, and correlation matrix. The experimental results indicated that the lab spectroscopy method yielded better prediction accuracy for *OC* data when compared to the models developed using airborne and Sentinel-2 data. However, the study emphasized the advantages of Sentinel-2 data, including frequent revisits and broader coverage, which can improve the model's performance. Despite the differences in prediction accuracy, the research highlighted the potential of utilizing satellite imagery, especially Sentinel-2 data, for soil mapping applications.

Morgan et al. (2018) employed MLR to monitor soil salinity. The study utilized soil data and Sentinel-2 satellite imagery collected from northwest Cairo. Three different approaches were employed for data preprocessing: reflectance data of six selected Sentinel-2 bands, PCA of the selected bands, spectral indices derived from Sentinel-2 bands. The researchers evaluated three spectral indices, including NDVI, Moisture Stress Index (MSI), and the derived SI, for predicting soil salinity. The results indicated that combining the reflectance data of the shortwave infrared band of Sentinel-2, NDVI spectral index, and PCA provided the best performance when designing the ANN. The accuracy of salinity estimation using MLR was compared with laboratory-measured estimation using R^2, and the results demonstrated promising performance.

Using a pedotransfer function, a mathematical relationship can be used to predict soil properties that are missing, expensive, or time-consuming to measure (Yao et al., 2015). Sirsat et al. (2018) utilized regression techniques along with pedotransfer to predict numeric values of VFI for various soil nutrients, including *Fe, Mn, OC,* P_2O_5, and *Zn*. The researchers employed regression techniques from 20 families, such as bagging, boosting, LASSO, ET, RF, RR, SVM, and among others. To evaluate the performance of the regression models for different nutrients, the researchers used various measures such as RMSE, R^2, WI test, and Friedman rank test. The results showed that ET performed better than other regression techniques.

Wang et al. (2018) employed BRT and RF to estimate the stock of *OC* in eastern Australia. They utilized field data and Geographic Information System-based environmental variables, including climate, topography, and Landsat-5 indices. Feature selection was performed using the SR method and Genetic Algorithms (GA) to identify predictors for training the BRT and RF models. The results indicated that the RF model, coupled with the GA for feature selection, outperformed the BRT model in predicting *OC* stocks. The RF model exhibited the lowest RMSE and the highest R^2, highlighting

its superior performance in estimating *OC* stocks in the specified region.

Chougule et al. (2019) developed an ontology-based system to recommend fertilizers and crops. The recommendations were made by considering soil type, as well as the levels of *N*, *P*, and *K*, along with historical data stored in an ontology. The system employed k-means clustering and RF regression to recommend crops and fertilizers. The data used for the recommendations were collected from various regions in Maharashtra, India. Crop recommendations were tailored to specific regions, while fertilizer recommendations were based on the levels of *N, P*, and *K*. This approach aimed to leverage ontology-based data storage and machine learning techniques to enhance the precision of crop and fertilizer recommendations.

Fernandes et al. (2019) aimed to assess the accuracy of predicting *OM* levels using different predictors, including potential acidity, pH, *Ca*, and *Mg*. The dataset for analysis comprised 8556 soil samples collected from São Paulo state in Brazil. To predict *OM* levels, the researchers employed a MLP with two hidden layers. The MLP had varying numbers of neurons in the hidden layers, ranging from 4 to 20. The performance of the MLP was assessed using several evaluation metrics, including R^2, RMSE, mean error, and confidence coefficient. The results indicated that the MLP exhibited superior performance compared to other models in both the calibration and validation phases. It achieved the lowest RMSE and the highest R^2 values, suggesting its effectiveness in predicting *OM* levels in the soil samples.

Ghorbani et al. (2019) aimed to predict *EC* using a hybrid model called MLP-FFA, which combines a MLP with the Firefly Algorithm. The performance of this hybrid model was compared with standalone MLP and kriging models. The study employed a grid sampling scheme, collecting 126 soil samples to measure the *EC* in the Soofiyan City, Tabriz regions. For evaluation, the researchers used different metrics, including RMSE, NS, WI, and LMI. The results indicated that the hybrid MLP-FFA model outperformed the other models in terms of accurately estimating *EC*, showing lower RMSE and higher WI, NS, and LMI values.

Ransom et al. (2019) focused on enhancing *N* prediction for corn by utilizing soil and weather data from 49 sites in the US Midwest. Their objective was to recommend an optimal nitrogen fertilizer for corn. To achieve this, various statistical-based approaches and machine learning-based regressors were compared to assess their performance in predicting optimal *N*. The methods evaluated included ER, PLSR, LASSO, RR, PCR, SR, DT, and RF. The results of the study indicated that the RF regressor outperformed other methods when soil and climate data were considered. RF exhibited

reduced RMSE and the highest R^2, suggesting its effectiveness in predicting optimal *N* for corn in the given context.

Zhang et al. (2019) aimed to determine the total *N* content in different types of soil using spectral wavebands obtained through a spectrometer. The soil samples were collected from various regions of North China. The researchers employed a method to select sensitive wavebands based on their relevance to the target variables, spectral information capability, and waveband redundancy. The selection process involved using the Mutual Information algorithm and a combination of Mutual Information and Ant Colony Optimization. Subsequently, the researchers developed PLSR based on full-spectral information, as well as MLR and SVM models based on the selected wavelengths. The comparison of these models was performed using metrics such as R^2 and RMSE. The study's results revealed that the MLR and SVM models outperformed the PLSR in predicting total *N* content in the soil.

Zia et al. (2019) recommended using real-time low dimensional prediction models to accurately predict total loss of *N*, instead of relying on models developed using large historical data sets with multiple parameters that are difficult to extract. They developed models using four different algorithms, namely, M5 tree, MLR, MLP, and REPTree. Villa-Vialaneix et al. (2012) utilized 11 parameters such as pH, precipitation, temperature, bulk density, carbon, clay, *N* in fertilizer and manure, *N* from precipitation and plant residue, and *N* from fixation to predict total *N* loss. Zia et al. (2019) used an abstraction of the parameters employed in Villa-Vialaneix et al. (2012), reducing the dataset parameters by 50%. They used parameters such as crop cover, precipitation, soil moisture, *N* in fertilizer and manure, total *N* applied, days since last *N* application, and day of the year. The real-time dataset they collected was from regions at the University of Cork in the Dripsey catchment in the south of Ireland. The M5 tree-based model outperformed the other algorithms, with high R^2, least RMSE, and least RRSE.

Büchele et al. (2019) employed PLSR in combination with PCA to predict the levels of *Fe* and *K* in soil samples. The soil samples were collected from twelve different regions in Germany. The authors considered laboratory measurements of *Fe* and *K* and soil texture variables, including clay, loam, silt, and sand. The predictive performance of the model was assessed using metrics such as R^2, average deviation, and MAE.

Coulibali et al. (2020) compared the performance of the hierarchical Mitscherlich model with ML regression models such as KNN, RF, ANN, and Gaussian processes. The goal was to predict the appropriate dosage of *N*, *P*, and *K* fertilizers for potato crops. The soil samples were collected from a field in Quebec, Canada. The laboratory

measurements included *pH*, extractable *K, P, Mg, Al*, and *Ca*. Soil texture parameters, such as clay, silt, and sand, were also considered. Additionally, various climate parameters, including cumulative precipitation, the Shannon Diversity Index for rainfall distribution, mean temperature, and the number of growing-degree days, were included in the analysis. The study found that the ML regression models, including KNN, RF, ANN, and Gaussian processes, outperformed the Mitscherlich model, achieving higher R^2 values.

Delavar et al. (2020) proposed a hybrid model that combined ANN and GA to predict soil salinity. The model utilized both supervised and unsupervised approaches. Various data sources were incorporated into the model, including a 45-year soil salinity map, remotely sensed data, SI, Salinity Ratio Index, and NDVI. The hybrid model demonstrated high accuracy, evidenced by a high R^2 value and the least RMSE. Additionally, the authors developed a time series salinity distribution map to predict the variation in salinity over time. This suggests that the combined use of ANN and GA provided an effective approach for predicting soil salinity and understanding its temporal dynamics.

Gorji et al. (2020) predicted salinity using LinR and MLR. The study focused on the western part of Urmia Lake in Iran, and data sources included laboratory-measured soil data and satellite data from Sentinel-2 and Landsat-8 OLI. The laboratory measurements encompassed variables such as *Mg*, *Cl*, carbonate, and bicarbonate. Additionally, satellite-derived salinity indices and the NDVI were incorporated into the models. The accuracy of the predictions was assessed using three performance indicators: RMSE, R^2, and MAE.

Inoue et al. (2020) compared PLSR and iterative iPLSR for evaluating soil fertility. The study focused on soil samples collected from Tomioka-machi in the Fukushima region of Japan. Laboratory measurements included the levels of *OC* and *N*. Additionally, Remote Sensing Indices and Normalized Difference Spectral Index were derived from a portable spectroradiometer. The performance of the two regression methods, PLSR and iPLSR, was assessed using metrics such as R^2 and RMSE. The results indicated that iPLSR outperformed PLSR in terms of accuracy for evaluating soil fertility.

Mahmoudzadeh et al. (2020) employed to predict variations in *OC* in the western region of Iran. The research utilized a range of auxiliary predictors including air temperature, rainfall, terrain surface texture, terrain vector roughness, and valley depth. Various machine learning algorithms, including Cubist, KNN, SVM, XGBoost, and RF, were compared in terms of their predictive performance. The results of the study

indicated that RF provided the most accurate spatial distribution of *OC*, exhibiting the lowest RMSE and higher R^2. Furthermore, the study estimated that the total *OC* stocks were highest in forest soils and lowest in bare land based on the spatial distribution predicted by the ML models.

Zhang et al. (2020) employed MLR and ANN to estimate variations in the parameter *K*. Soil samples were collected from the Chaobai River alluvial plain, located in the border area of Beijing, Tianjin City, and Hebei Province, China. Additionally, spectral data were gathered from multispectral images acquired from Landsat TM and OLI bands. The study focused on assessing the prediction accuracy of *K* over time using the MLR and ANN models.

Hossen et al. (2021) used SVM and MLP to predict the amount of *N* present in soils of Sturgis, South Dakota, USA. The researchers utilized laboratory estimations of *N* and incorporated environmental predictors such as temperature and humidity. Additionally, multispectral indices, including NDVI and multispectral bands (Red, NIR, and Green) were used as features. The performance of the SVM and MLP models was evaluated using the RMSE, and the results indicated that MLP outperformed SVM in terms of accuracy.

Li et al. (2021) employed RF and Cubist models were employed to predict the concentration of *OC*. The predictions were based on data collected from Sentinel-1A/2A/3A satellites over the Ebinur Lake Basin in Eurasia. The researchers investigated the variation in *OC* with respect to elevation, depth, and slope. The performance of both algorithms, RF and Cubist, was evaluated using metrics such as Lin's Concordance Correlation Coefficient, MAE, RMSE, and R^2. The study findings indicated that RF outperformed Cubist in terms of accuracy.

Parsaie et al. (2021) aimed to predict the amount of *N* using RF, DT, and Cubist. The research was conducted in Kurdistan province, Iran, on the Qorveh-Dehgolan plain lands. The researchers collected soil samples and obtained laboratory measurements of key parameters, including $CaCO_3$, *OC*, and *N*. Additionally, various soil characteristics such as texture (sand, silt, clay), elevation, aspect, valley depth, channel network distance, normalized height, convergence index, direct insulation, total insulation, mid-slope position, and Sentinel-2 spectral data (NDVI, B2, B11, SAVI) were considered as features for modeling. The results of the study indicated that RF exhibited the highest R^2, lowest RMSE, and standard deviation among the models, suggesting that RF outperformed DT and Cubist in predicting *N* levels.

Peng et al. (2021) employed hyperspectral remote sensing to monitor soil nutrients, specifically focusing on *K, N*, and *P*. The hyperspectral images used in the study were acquired from the Huan Jing-1A satellite in October 2017, with a spatial resolution of 100 meters and a total of 115 bands. To estimate soil nutrient levels, the researchers applied various regression models, including MLR, RR, SVM, and BPNN with a GA. These models were used to predict the concentrations of *K, N*, and *P* in the soil. The performance of these estimation approaches was evaluated using several metrics, including R^2, concordance correlation coefficient, ratio of performance to the interquartile range, RMSE. According to the study's findings, the BPNN with a GA demonstrated superior performance as compared to the other approaches in estimating soil nutrient levels.

Shang et al. (2021) assessed the potential of satellite hyperspectral data in estimating *Cu* levels in soil. To achieve this goal, the researchers employed GPR, BRT, bagging, and SVM, to predict the levels of *Cu*. The soil samples used for the study were collected from Dexing City in Jiangxi Province, China. Laboratory measurements were conducted to determine the actual *Cu* levels in the soil samples. Additionally, multispectral images were captured using the ZY1-02D satellite, which included bands in the VNIR and SWIR regions. The performance of the different techniques was evaluated using key metrics such as R^2 and RMSE. According to the study's findings, GPR exhibited the highest R^2 and the lowest RMSE, indicating that it outperformed the other techniques in accurately predicting *Cu* levels in the soil.

Shi et al. (2021) employed PLSR, ANN, SVM, and RF to predict soil *OM*. Additionally, Ranger and LASSO techniques were utilized for the selection of the most relevant hyperspectral bands. The soil samples used for the study were collected from Huangmian in Liuzhou City, China. The prediction of variations in *OM* was based on laboratory-measured *OM* and hyperspectral data, including Red, Fast data lookup, and Fraction of Data Remaining bands, as well as visible and near-infrared bands derived from images captured by analytical spectral devices. According to the study's findings, the ANN with FDR performed the most accurately in predicting soil organic matter (*OM*).

Abera et al. (2022) utilized a RF regressor to predict variations in nutrients, specifically *N* (nitrogen), *P*, *K*, and *S*. The dataset used for building the model consisted of yield response data collected over 31 years (1986-2017) from various wheat-growing environments in Ethiopia. To enhance the prediction model, the researchers incorporated diverse climate variables such as rainfall, temperature, and solar radiation. Additionally, topographic variables, including elevation and topographic index, were con-

sidered. Soil parameters such as *pH*, *OC*, soil texture, and *CEC* were also included in the dataset. The utilization of the RF regressor, combined with a comprehensive set of variables, aimed to capture the complex interactions influencing nutrient variations and provide accurate predictions for *N, P, K*, and *S* in wheat-growing environments in Ethiopia.

Aksoy et al. (2022) focused on predicting soil salinity using CART, RF, and SVM. Landsat-8 OLI and Sentinel-2A satellite bands were utilized to derive salinity and vegetation indices, with the primary goal of enhancing prediction accuracy. To assess and compare the performance of the models, the authors employed five salinity indices (SI1, SI2, SI3, SI4, and SI5) and nine vegetation indices (CRSI, EVI, EVI2, ENDVI, EEVI, GDVI, NDVI, SAVI, and SRatio) in their analysis. The evaluation metrics used to measure the performance of the models were RMSE and R^2. The results of the study indicated that the RF outperformed the other two algorithms, CART and SVM, in predicting soil salinity. The research contributes to the assessment of the suitability and effectiveness of different ML algorithms in predicting soil salinity based on satellite-derived data.

Al Masmoudi et al. (2022) focused on predicting the levels of *OM*, K_2O, and P_2O_5 in soil samples. To achieve this, they employed MLR, SVM, and RF. The soil samples were randomly collected from agricultural fields in Doukkala, located in central Morocco. The accuracy of the models was assessed using two commonly used metrics in regression tasks such as RMSE and the R^2. The study's findings indicated that all three models: MLR, SVM, and RF produced promising R^2 scores when predicting *OM* levels.

Chang et al. (2022) employed a combination of DPSO and BPNN to predict the levels of *OM* in soil samples. The soil samples were collected from the northeast of Jiamusi City, Heilongjiang Province, Northeast China. The methodology involved measuring the spectral reflectance feature bands at various wavelengths and obtaining laboratory measurements of *OM*. The combination of DPSO and BPNN was used as a predictive model for estimating *OM* levels in the soil. The results of the study demonstrated that the DPSO-BPNN method outperformed other approaches, achieving higher accuracy in predictions. This was evidenced by having the least RMSE and the highest R^2.

Xu et al. (2022) employed PLSR, RF, and SVM to estimate the spatial variability of nitrogen (*N*) in the soil. The soil samples were collected from Kaifeng, a city located in the North China Plain. The researchers utilized laboratory measurements of nitrogen (*N*) along with additional factors such as elevation, TWI, and satellite-derived data

from various sources including Sentinel-2, Landsat-8, and WorldView-2. The satellite-derived data encompassed bands such as SWIR, NIR, Multispectral Instrument, Normalized Difference Snow Index, NDVI, Normalized Difference Red-edge Index, Red-edge Chlorophyll Index, and MERIS Terrestrial Chlorophyll Index. The models based on Landsat-8 and Sentinel-2 data exhibited higher accuracy compared to those based on WorldView-2 data.

Yang et al. (2022) employed PLSR and BPNN to predict the *OM* content in soil samples collected from Aksu, located in northwestern China. The authors utilized various feature selection algorithms, including Ant Colony Optimization, Particle Swarm Optimization, and Simulated Annealing, to filter spectral features relevant to the prediction of *OM*. The results indicated that BPNN with Particle Swarm Optimization outperformed the other methods, achieving the highest R^2.

Chen et al. (2023) aimed to predict the variation in *P* levels in soil using both Linear Multiseriate and various ML regressors. The ML models included SVM, DT, RF, XGBoost, and LightGBM. Soil samples were collected from wheat and paddy farmlands in China at different depths. Two datasets were used for model learning and testing: Dataset 1, which included samples from all soil depths (0 to 170 cm), and Dataset 2, which only included samples from the topsoil layer (0 to 20 cm). The evaluation of model performance involved several metrics, including RMSE, Mean Deviation, MAE, and model effectiveness. The results indicated that ML methods, particularly DT, RF, XGBoost, and LightGBM, outperformed Linear Multiseriate in estimating total *P*. These ML techniques demonstrated lower RMSE, Mean Deviation, and MAE, indicating their effectiveness in predicting *P* levels compared to SVM.

Gulhane et al. (2023) employed Sentinel-2 and Landsat-8 data to derive indices for three soil parameters: *P*, *Fe*, and pH. The yellowness index obtained through Sentinel-2 was utilized for calculating *P*, while the Ferrous and Carbonate indices derived from Landsat-8 spectral bands were used for estimating *Fe* and *pH*, respectively. The researchers predicted the soil parameter values for four villages—Baggi, Ibrahimpur, Mogra, and Wai—in Amravati district, Maharashtra State, India. Soil fertility was classified into LOW, MEDIUM, or HIGH using MLP classifier. To assess the accuracy of predicting the satellite-derived parameters, the results were compared to laboratory-measured values, and the performance was evaluated using the R^2.

Dos Santos et al. (2023) compared the performance of LASSO and PLS in estimating *OC* using spectroscopy reflectance data from soil samples in Pernambuco, Northeastern Brazil. The study aimed to assess the ability of these two regression methods to

predict *OC* levels in soil. The findings indicated that LASSO outperformed PLS in the context of estimating *OC*. LASSO provided more accurate and reliable predictions of *OC* content compared to PLS.

Hu et al. (2023) utilized RF to predict variations in soil nutrients, including *OM*, *N*, *P*, *K*, pH, and *CEC*. The research was conducted in farmlands located in Jiangxi Province in southern China. The authors collected laboratory measurements of soil properties, including *OM, N, P, K, pH*, and *CEC*. Additionally, environmental variables such as mean annual precipitation, mean annual temperature, slope, aspect, TWI, topographic position index, multi-resolution valley bottom flatness, and population density were considered as predictors in the model. The study revealed significant spatial variability in soil properties, and the authors observed that climate factors had a dominant effect on soil nutrients and organic matter.

Keshavarzi et al. (2023) employed RF and SVM to predict variations in soil properties, specifically *Cu*, *Fe*, *Mn*, and *Zn*. The research focused on the Neyshabur Plain in Iran. The authors utilized a diverse set of predictors, including laboratory measurements of soil properties such as *Fe, Cu, Mn, and Zn, pH*, and texture (clay and sand). Additionally, they incorporated various environmental and topographic parameters, such as mean annual temperature, annual precipitation, total solar radiation, slope, profile curvature, TWI, Flow Accumulation, Stream power index, and satellite-derived indices from Sentinel-2: MSAVI2, topsoil grain size index, saturation index, normalized clay index, NDVI, and Landsat 8 OLI indices: topsoil grain size index, saturation index, normalized clay index, NDVI, GNDVI. The performance of the models was evaluated using normalized RMSE, and the results indicated that RF outperformed SVM in predicting the variations in *Cu, Fe, Mn*, and *Zn*.

Tavares et al. (2023) aimed to predict the levels of *Ca* and *K* in soil. The authors used regression models such as LinR, MLR, PLSR, and RF. The research focused on soil samples collected from fields in the southeast and central-west regions of Brazil. The laboratory measurements for *Ca* and *K*, and soil texture were classified as sandy-clay, sand-clay-loam, or sandy-loam. Additionally, the researchers analyzed spectral information in the VNIR range. The results of the study indicated that the RF model outperformed the other regression models, achieving the highest R^2 and the least RMSE in predicting the levels of *Ca* and *K*.

2.3 Gaps in Literature

The literature study directed the way toward identifying some research gaps. To the best of our knowledge, only a few research attempts have been made to classify soil fertility using ML-based approaches. Thus, there is a need for significant research work in this domain of research.

- Most of the research work has focused on using regression techniques to predict changes in soil parameters, with very few researchers exploring soil fertility classification. Soil fertility is dependent on various chemical parameters such as *EC, pH, OC*, macronutrients, and micronutrients. For optimal crop growth, soil nutrients must be present in balanced proportions (Marschner, 2011). By accurately classifying soil fertility, fertilizer recommendations can be tailored to meet specific soil needs. Previous research prescribed fertilizers to improve fertility levels of some soil nutrients, but precise recommendations can be made through soil fertility classification.

- The soil classification in previous studies involved N_2O, P_2O_5, K_2O, and SO_4. The estimation of N, P, K and S were performed using expensive methods such as alkaline permanganate, Olsen's, flame photometry, and atomic absorption spectrophotometry, respectively (Sirsat et al., 2017).

- The fertility of soil varies during the crop growth cycle. Previous research did not measure soil parameters during dynamic crop growth.

- The previous works mainly used historical data collected using destructive methods for experimental purposes. In most former works, real-time soil sensor data was not used.

- The fertility of soil varies based on climate The existing works have not measured the performance of the classifiers by selecting soil data from different climate zones.

- To reduce the cost of laboratory chemical analysis or the need for sensors, it's important to identify the most relevant soil parameters. Previous studies have

used feature selection methods to determine the environmental factors that affect soil fertility. However, studies did not use feature selection methods to select soil parameters.

- Most previous works have not used soil fertility index, which is helpful for the precise classification of soil fertility.

2.4 Problem Statement and Research Objectives

2.4.1 Problem Statement

"Design and development of Machine Learning-based soil fertility classification approach".

2.4.2 Research Objectives

The research objectives are defined as:

1) To propose a soil fertility classification using Machine Learning-based Classier(s). Conduct a set of experiments using a publicly available dataset(s) and measure its performance using standard performance metrics such as Accuracy, F1-Score, Precision, and Recall.

2) Measure the robustness of the proposed approach by using real-time data.

3) To propose an automated fertilizer prescription approach based on the outcome of the Machine Learning-based soil fertility classifiers.

CHAPTER 3

MACHINE LEARNING-BASED SOIL FERTILITY CLASSIFICATION

This research aimed to develop a non-destructive method for assessing soil fertility instead of traditional laboratory analysis. It employed ML-based classifiers that use soil chemical parameters to classify soil fertility as LOW, or MEDIUM, or HIGH. To achieve the precise classification of soil fertility, this work used soil chemical parameters, such as *pH, EC, OC*, macronutrients: *N, P, K, S*, and micronutrients: *B, Cu, Fe, Mn*, and *Zn*. This chapter discusses the proposed soil fertility classification approaches using ML-based classifiers. The performance of ML-based classifiers depends on the dataset used. The laboratory-measured historical soil data and satellite-derived data were used to train and test the proposed approaches.

3.1 Datasets Used

Initially, in the proposed method "Soil Fertility Classification using Machine Learning-based Classifiers" a ML-based classifiers were developed for Karnataka soil health dataset (ICRISAT and Government of Karnataka, 2016) containing the laboratory-measured soil data of farmlands of Karnataka (State), India. The dataset consists of 92832 instances with the attributes such as card_no, farmer_number, sampling authority unit, state, district, taluk, village, farmer_name, survey_number, soil_type, pH, EC, OC, P, K, S, Zn, B, Fe, Cu, and Mn. For experiments were conducted by using 11 soil chemical parameters such as *pH, EC, OC, P, K, S, Zn, B, Fe, Cu*, and *Mn*. A large number of values are missing values for the parameters such as Fe, Cu, and Mn. It is difficult to fill a large number of missing values manually. MLR was used to fill the missing values in the dataset. The dataset was labeled using Table 3.1, which resulted in 92794 instances of LOW fertile, 36 instances of MEDIUM and 2 of HIGH fertile soil.

It was observed that laboratory measured dataset was unbiased with more number of instances of LOW fertile and very less number of instances of HIGH fertile. In practice, the soil fertility will be analysed before the cultivation. Thus, the soil data including the chemical parameters such as *pH, EC, OC* and *N* was generated using Sentinel-2 spectral bands. The generated dataset was labeled based on values of *pH, EC, OC* and

N using Table 3.1, and fertility level of remaining soil nutrients were determined based on *pH* value using Table 3.2. The WEKA open source tool (WEKA, 2021) was used to eliminate redundant and missing data. The preprocessed Sentinel-2 dataset consists of 329 instances. After labeling, the Sentinel-2 dataset consists of 293 instances of LOW fertile, 25 of MED fertile, and 11 of HIGH fertile.

3.2 Dataset Labeling

Using Table 3.1, each instance of the dataset is labeled as LOW, or MEDIUM, or HIGH according to soil chemical parameters. The pH value influences nutrient availability. The value of *N* is proportional to *OC*; on the unavailability of *N*, the fertility level of *N* is determined based on the fertility level of *OC*. Furthermore, the *pH* values influence nutrient availability and are considered indicators of other soil parameters (Tharavathy, 2016). When the soil nutrient values are unavailable, their fertility levels can be determined based on *pH* value, using Table 3.2.

Table 3.1: Fertility level of soil parameters to label the soil data

Soil Parameters	**LOW**	**MEDIUM**	**HIGH**
EC (dS/cm)	>2.5	>1.6, <=2.5	<=1.6
OC (%)	<0.5	>=0.5, <0.75	>=0.75
pH	>8.5,<6.5	-	>= 6.5, <=8.5
K (kg/ha)	<141	>=141, <336	>=336
N (kg/ha)	<280	>=280, <560	>=560
P (kg/ha)	<10	>=10, <24.6	>=24.6
S (kg/ha)	<10	>=10, <20.3	>=20.3
B (ppm)	<0.6, >1.8	-	>=0.6, <=1.8
Cu (ppm)	<2	>=2, <4	>=4
Fe (ppm)	<2.5	>= 2.5, <4.5	>=4.5
Mn (ppm)	<2	>=2, <4	>=4
Zn (ppm)	<0.6, >0.8	-	>=0.6, <=0.8

Table 3.2: Fertility level of soil parameters based on *pH* value

Soil Parameters	**LOW**	**MEDIUM**	**HIGH**
K	pH<5.5	5.5<=pH<5.9	pH>=5.9
N	pH<5.1 or pH>=8.75	8<pH<8.5 or 5.1<pH<5.9	5.9<=pH<=8
P	pH<5.5 or 8.5<pH<9.0	5.5<=pH<5.9 or 7.5<pH<=8.5	5.9<=pH<=7.5 or pH>=9
S	pH<5.5	5.5<=pH<5.9	pH>=5.9
B	pH<5.1 or 8<pH<8.5	7.5<pH<=8 or 8.5<pH<=8.75	5.1<=pH<=7.5 or pH>=8.5
Fe	pH>8	7.5<pH<=8.0	pH<=7.5
Mn	pH<5	5<=pH<5.4 or 7.5<pH<=8	5.4<=pH<=7.5
Cu, Zn	pH<4.5 or pH>8	4.5<=pH<5 or 7.5<pH<=8	5<=pH<=7.5

3.3 Performance Evaluation Metrics

The performance of proposed ML-based soil fertility classification approaches were evaluated using different measures (Campesato, 2020). Some of which are:
Accuracy indicates percentage of correct predictions made by a machine learning classifier. It is defined as in Eq. (3.1).

$$Accuracy = (TP + TN)/N \tag{3.1}$$

where, *TP* indicates true positive, *TN* indicates true negative and *N* is the total number of instances.

Precision measures the frequency with which the model's classification is correct and is computed using Eq. (3.2).

$$Precision = TP/(TP + FP) \tag{3.2}$$

where, *TP* indicates true positive and *FP* indicates false positive.

Recall measures the frequency with which the classifier correctly identifies true positives and is calculated using Eq. (3.3).

$$Recall = TP/(TP + FN) \tag{3.3}$$

where, *FN* indicates false negative.

F1-Score (or F-measure) is a harmonic mean of the Precision and Recall and is calculated as given in Eq. (3.4).

$$F1 - Score = (2 * Precision * Recall)/(Precision + Recall) \tag{3.4}$$

Per-class Precision, Per-class Recall, and Per-class F1-Score are calculated using Eq. (3.5), Eq. (3.6), Eq. (3.7).

$$Per-class\ Precision = \frac{TP_i}{TP_i + FP_i} \tag{3.5}$$

where i indicates the class LOW, MEDIUM, or HIGH.

$$Per-class\ Recall = \frac{TP_i}{TP_i + FN_i} \tag{3.6}$$

$$Per-class\ F1-Score = \frac{2*Per-class}{Per-class} \frac{Precision_i*Per-class}{Precision_i+Per-class} \frac{Recall_i}{Recall_i} \tag{3.7}$$

FPR is the proportion of negative instances classified incorrectly and is computed using Eq. (3.8).

$$FPR = \frac{FP}{FP+TN} \tag{3.8}$$

Kappa statistic indicates how much better the classifier is performing based on the frequency of each class. The value can be less than or equal to 1. If the value is 0 or lesser than 0, it indicates that the classifier is not appropriate. It is calculated using Eq. (3.9).

$$Kappa\ statistics = \frac{pr_{\text{obs}} - pr_{\text{exp}}}{N - pr_{\text{exp}}} \tag{3.9}$$

where pr_{obs}, pr_{exp} are observed and expected predictions, respectively and N is total observations.

RMSE is the error difference between actual value and the predicted value (Bisonget al. , 2019) and is calculated using Eq. (3.10).

$$RMSE = \sqrt{(\sum_{i=1}^{N}(Predicted_{\text{i}} - Actual_{\text{i}})^2/N)} \tag{3.10}$$

where, $Predicted_i$ indicates predicted value in i$^{\text{th}}$ observation, $Actual_i$ is actual value in i$^{\text{th}}$ observation and N is the total number of instances.

RRSE is the differences between values predicted by a classifier and the values actually observed and is defined as in Eq. (3.11).

$$RRSE = \sqrt{(\sum_{i=1}^{N}(P_i - T_i)^2 / \sum_{i=1}^{N}(T_i - TMean)^2)} \tag{3.11}$$

where, P_i indicates the predicted value in ith observation, T_i indicates the target value in ith observation, TMean is mean of target values, and N is the number of instances.

ROC indicates a trade-off between True and False Positive rates at different classification threshold values. The Precision-Recall Curve (PRC) indicates a trade-off between Precision and Recall for different probability threshold values.

3.4 Proposed Soil Fertility Classification using Machine Learning-based Classifiers

A precise soil fertility classifier can be developed using several oxygen-free soil chemical parameters. Sirsat et al. (2017, 2018) used ML-based classifiers to classify soil fertility based on several soil parameters, including the oxidation forms of soil nutrients such as *K, N,* and *S*. The emission of N_2O relies on soil moisture, temperature, microbial activity, aeration, and organic matter content (Wilson et al., 2013). The P_2O_5 and K_2O are available fertilizer contents in the soil (Sander & Wiese, 1973). SO_4 availability is based on the distribution of *S* in the soil and soil microbial properties (Malik et al., 2020).

In this work, ML-based soil fertility classification approaches are proposed by using historical soil data of laboratory-measured chemical parameter values. The classifiers such as bagging, BRT, J48, LR, NB, RF, and SVM are used to classify soil fertility as LOW, MEDIUM, or HIGH. Karnataka (State) soil health dataset (ICRISAT and Government of Karnataka, 2016) containing the soil data collected from farmlands of Karnataka (State), India, has been used to implement the proposed approach. The dataset consists of 92832 instances with 22 attributes, namely, card_no, farmer_number, soil acquisition unit, state, district, taluk, village, farmer_name, survey_number, soil_type, authority, and 11 chemical parameters: *pH, EC, OC, K, P, S, B, Cu, Fe, Mn*, and *Zn*. The dataset was analyzed for irrelevant and missing data. For experimental purposes, 11

chemical parameters were selected, and LinR was used to fill the missing values. Based on the soil chemical parameter values, the dataset instances were labeled as LOW, or MEDIUM, or HIGH. The dataset was split into training and test data. The classifiers were trained using training data, and the performance of each classifier was measured using the test data.

3.4.1 Performance Evaluation of the Proposed Approach

The experiments were conducted using the WEKA (WEKA, 2021) open-source tool with 10-fold cross-validation and by dividing the dataset into 75% and 25% as training data and test data (split dataset), respectively. Bagging, BRT, J48, and RF employed a batch size of 100. BRT was configured with a bag size of 100 and number of folds set to 3. The process encompassed 10 iterations with a seed value of 1, a variance proportion of 0.001, a maximum of three folds, and a seed value of 1. For J48, a fold of three and a seed value of one were utilized, while RF utilized 100 bags and 100 iterations. LR, NB, and SVM classifiers used a batch size of 100. The SVM classifier employed a radial basis kernel function with a seed and cost value of 1, a gamma value of 0.1, epsilon of 0.001, a loss value of 0.1, and a degree of 3. The performance of the classifiers was evaluated using Accuracy, Precision, Recall, F1-Score, RMSE, RRSE, ROC, and PRC. Table 3.3 and Table 3.4 shows the performance of classifiers with 10-fold cross-validation and split dataset, respectively.

The bagging for 10-fold cross-validation achieved an Accuracy of 99.98% with RMSE 0.0076 and RRSE 12.747%, and for the split dataset, Accuracy was found to be 99.99% with RMSE 0.0074 and RRSE 11.575%. BRT for 10-fold cross-validation obtained an Accuracy of 99.98% with RMSE 0.0083 and RRSE 14.012%, and for the split dataset, the Accuracy was found to be 99.98% with RMSE 0.0101 and RRSE 15.671%. J48 classifier for 10-fold cross-validation achieved an Accuracy of 99.99% with RMSE 0.0079 and RRSE 13.220%, and using split dataset Accuracy was found to be 99.99% with RMSE 0.0063 and RRSE 9.947%. LR using 10-fold cross-validation achieved an Accuracy of 99.85% with RMSE 0.0026 and RRSE 43.747%, and for split dataset, an Accuracy of 99.85% with RMSE 0.0269 and RRSE 41.875%. NB classifier for 10-fold cross-validation achieved Accuracy is found to be 99.80% with RMSE 0.0032 and RRSE 53.821%, and for the split dataset Accuracy of 99.84%, RMSE 0.0318, and

Table 3.3: Performance of the proposed soil fertility classification using 10-fold cross-validation

Classifier	Accuracy (%)	RMSE	RRSE	Precision	Recall	F1-Score	ROC	PRC	Class
Bagging	99.98	0.0076	12.747%	1.000	1.000	1.000	1.000	1.000	LOW
				0.987	0.989	0.988	1.000	0.992	MEDIUM
				0.894	1.000	0.944	1.000	0.975	HIGH
BRT	99.98	0.0083	14.012%	1.000	1.000	1.000	1.000	1.000	LOW
				0.987	0.991	0.989	1.000	0.990	MEDIUM
				0.913	1.000	0.955	1.000	0.934	HIGH
J48	99.99	0.0079	13.220%	1.000	1.000	1.000	1.000	1.000	LOW
				0.987	0.993	0.990	0.999	0.993	MEDIUM
				0.993	1.000	0.966	1.000	0.915	HIGH
LR	99.85	0.0026	43.747%	0.999	1.000	0.999	1.000	1.000	LOW
				0.889	0.799	0.842	1.000	0.843	MEDIUM
				0.998	0.999	0.998	1.000	0.999	HIGH
NB	99.80	0.0032	53.821%	0.999	0.999	0.999	0.999	1.000	LOW
				0.777	0.845	0.810	0.999	0.756	MEDIUM
				1.000	1.000	1.000	1.000	1.000	HIGH
RF	99.99	0.0053	8.862%	1.000	1.000	1.000	1.000	1.000	LOW
				0.998	0.998	0..998	1.000	1.000	MEDIUM
				1.000	1.000	1.000	1.000	1.000	HIGH
SVM	99.99	0.0046	7.804%	1.000	1.000	1.000	0.999	1.000	LOW
				1.000	0.993	0.997	0.997	0.993	MEDIUM
				1.000	1.000	1.000	0.999	1.000	HIGH

Table 3.4: Performance of the proposed soil fertility classification using Split dataset

Classifier	Accuracy (%)	RMSE	RRSE	Precision	Recall	F1-Score	ROC	PRC	Class
Bagging	99.99	0.0074	11.575%	1.000	1.000	1.000	1.000	1.000	LOW
				0.992	0.992	0.9922	1.000	1.000	MEDIUM
				0.938	1.000	0.968	1.000	1.000	HIGH
BRT	99.98	0.0101	15.671%	1.000	1.000	1.000	1.000	1.000	LOW
				0.992	0.977	0.984	1.000	0.997	MEDIUM
				1.000	1.000	1.000	1.000	1.000	HIGH
J48	99.99	0.0063	9.947%	1.000	1.000	1.000	1.000	1.000	LOW
				0.994	0.994	0.994	1.000	0.999	MEDIUM
				1.000	1.000	1.000	1.000	1.000	HIGH
LR	99.85	0.0269	41.875%.	0.999	0.999	0.999	1.000	1.000	LOW
				0.900	0.837	0.867	1.000	0.868	MEDIUM
				1.000	1.000	1.000	1.000	1.000	HIGH
NB	99.84	0.0318	49.574%	0.999	0.999	0.999	0.999	1.000	LOW
				0.835	0.899	0.866	0.999	0.841	MEDIUM
				1.000	1.000	1.000	1.000	1.000	HIGH
RF	99.99	0.0059	9.145%	1.000	1.000	1.000	1.000	1.000	LOW
				0.999	0.999	0.999	1.000	1.000	MEDIUM
				1.000	1.000	1.000	1.000	1.000	HIGH
SVM	99.99	0.0054	8.356%	1.000	1.000	1.000	1.000	1.000	LOW
				1.000	0.992	0.996	0.996	0.992	MEDIUM
				0.938	1.000	0.968	1.000	0.938	HIGH

RRSE 49.574%. RF classifier with 10-fold cross-validation obtained an Accuracy of 99.99% with RMSE 0.0053 and RRSE 8.8622%, and for split dataset, an Accuracy was found to be 99.99% with RMSE 0.0059 and RRSE 9.145%. The SVM classifier for 10-fold cross-validation obtained an Accuracy of 99.99% with RMSE 0.0046 and RRSE 7.804%, and the Accuracy was found to be 99.99% for the split dataset with RMSE 0.0054 and RRSE 8.356%. The RF classifier with a split dataset outperformed other classifiers with high Accuracy, Precision, Recall, F1-Score, ROC, and PRC values. RMSE and RRSE for the RF classifier are less than Bagging, BRT, J48, LR, and NB. The Kappa statistic is used measure the robustness of the classifiers. With 10-fold cross validation test Bagging, BRT, J48, LR, NB, RF, SVM achieved Kappa statistics of 0.9889, 0.9899, 0.9909, 0.8553, 0.8243, 0.998, 0.997, respectively. With split data set Bagging, BRT, J48, LR, NB, RF, SVM achieved Kappa statistics of 0.993, 0.9861, 0.9947, 0.8811, 0.8765, 1, and 0.997, respectively.

3.4.2 Summary

The ML-based classifiers such as Bagging, BRT, J48, LR, NB, RF, and SVM effectively classified the soil fertility as LOW, MEDIUM, or HIGH. The laboratory measurements of eleven chemical parameters were used as input to the classifiers. The RF classifier outperformed other classifiers with an Accuracy of 99.99% using 10-fold cross-validation and using a split dataset. With 10-fold cross validation test RF achieved kappa statistic of 0.998 and with split dataset it achieved Kappa statistic of 1.

3.5 Proposed Soil Fertility Classification of Satellite-derived Data

An accurate soil fertility classification requires site-specific soil parameter values. Furthermore, laboratory measurements leave behind chemical residues and take extra time. The researchers have employed remotely sensed data to derive soil parameters. Gholizadeh et al. (2018) utilized Sentinel-2 to derive *OC*, Li et al. (2021) derived *OC* based on Sentinel-1A/2A/3A data. Delavar et al. (2020) and Gorji et al. (2020); Morgan et al. (2018) derived *SI* using Landsat-ETM+ and Sentinel-2, respectively. Peng et al. (2021) utilized Huan Jing-1A satellite images to derive *K, N*, and *P*. Gulhane et al. (2023) derived *Fe* and *pH* using Landsat-8 and *P* using Sentinel-2 spectral images. The multispectral Sentinel-2 satellite data offers an innovative method for collecting images with high spatial resolution, 290 km swath width, and high repetition rate. Thus in

this proposed method the experiments were conducted by generating the dataset using Sentinel-2 satellite spectral band information. The Sentinel-2 satellite revisits for every five days to a specific location. This is beneficial to collect more number of site-specific soil data.

Figure 3.1 presents the steps involved in soil fertility classification using Sentinel-2 data.

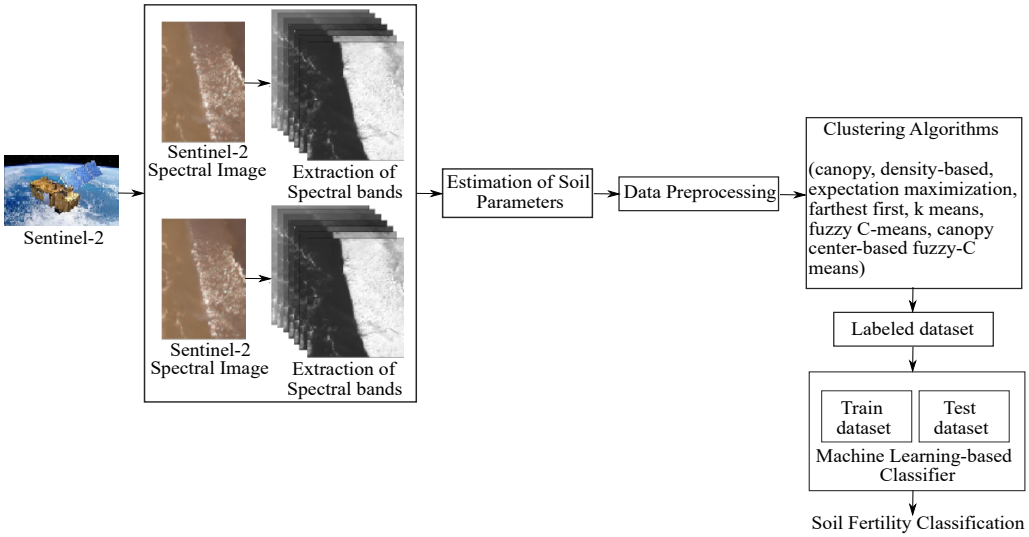
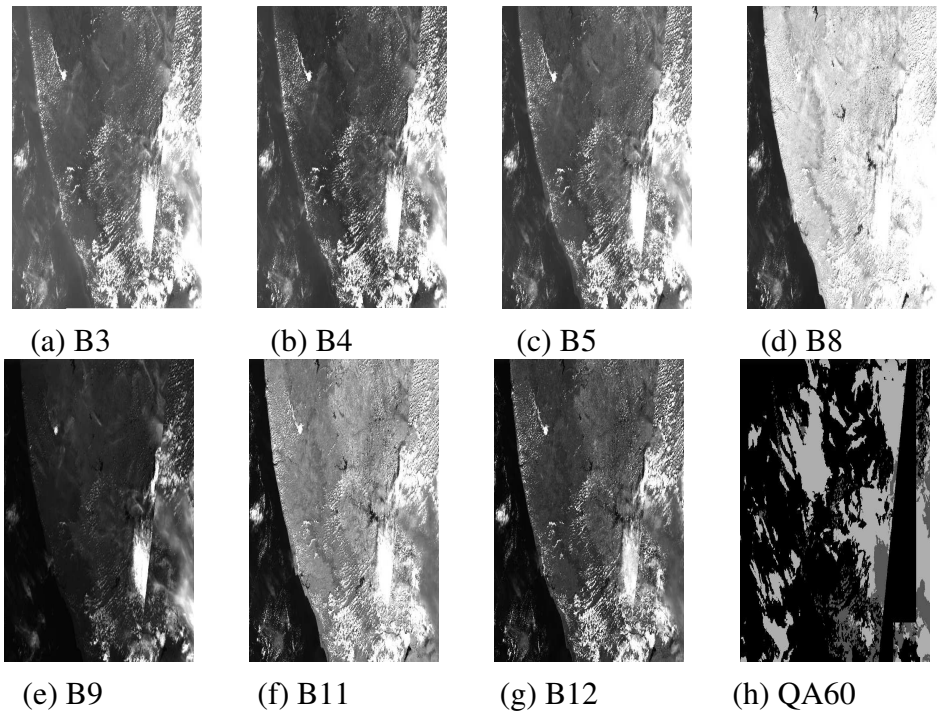


Figure 3.1: Steps to classify soil fertility using Sentinel-2 data

The study region employed in this work is located at Konaje, Mangalore town in Dakshina Kannada (District), Karnataka (State), India, with a latitude of 12.80593353 and a longitude of 74.906469. With the help of Google Earth Engine code and using the Python programming language Sentinel-2 (Sentinel-2, 2020), spectral band information of the chosen study area was extracted. The retrieved data involves spectral information from 14th November 2015 and 23rd October 2021. Seven spectral bands, B3, B4, B5, B8, B9, B11, and B12, were employed to get the soil parameters out of the 13 retrieved spectral bands. A combination of spectral bands was used to compute soil parameters, including *EC, pH, OC*, and *N*. The WEKA tool (WEKA, 2021) was used to preprocess data to remove redundant data. Various clustering methods were used to label the dataset, including canopy, density-based, expectation-maximization, farthest-first, fuzzy C-means, and proposed canopy center-based fuzzy C-means. Additionally, the dataset was labelled used by employing the state-of-the-art clustering methods, and using the proposed clustering method. Additionaly, the generated dataset was labelled based on level soil parameters *pH, EC, OC* and *N* using Table 3.1. The labeled dataset



(a) B3 (b) B4 (c) B5 (d) B8

(e) B9 (f) B11 (g) B12 (h) QA60

Figure 3.2: Visualization of Sentinel-2 spectral bands

was divided into training and test data. The NB, J48, RF, and SVM were used to classify the data. The clustering and classification results were compared with the results obtained using a manually labeled dataset.

3.5.1 Estimation of Soil Parameters using Sentinel-2 Spectral Bands

Sentinel-2, a multispectral satellite, offers 13 spectral bands: B1-Aerosol, B2-Blue, B3-Green, B4-Red, B5-Red Edge 1, B6-Red Edge 2, B7-Red Edge 3, B8-Near Infra-Red, B8A-Red Edge 4, B9-Water Vapor, B10-Cirrus, B11-Short Wave Infra-Red 1, B12-Short Wave Infra-Red 2. Additionally, it provides three cloud masks, QA10, QA20, and QA60, with pixel sizes of 10 meters, 20 meters, and 60 meters, respectively (Vaudour et al., 2019). As illustrated in Figure 3.2, spectral bands B3, B4, B5, B8, B9, B11, and B12 were utilized to estimate soil parameters along with cloud mask band QA60.

EC indicates the amount of salinity in soil (NRCS-USDA, 2020) and provides information on nutrient availability and loss, soil texture, and water availability (Al-Gaadi et al., 2021). Fertile soils have an *EC* value of 1.6 dS/cm or less, whereas low fertility soils have an *EC* value of greater than 2.5 dS/cm. Eq. (3.12) was used to get the *EC*

values for the spectral bands B3, B4, B8, and B12.

$$EC = 2.5 \times \frac{B12}{B8} - 1.6 \times \frac{B4}{B8} + 1.6 \times \frac{B3}{B8} \tag{3.12}$$

pH assesses the amount of hydrogen ions present in soil solutions (Tharavathy, 2016). A *pH* of 7 in the soil indicates a neutral soil and is considered to be very fertile. Acidic soil is caused by water or moisture in the soil, which lowers *pH* below 6.5. Low *pH* soil is considered to be less fertile. The spectral bands used for *pH* calculation were B4, B9 and B12 (Hengl et al., 2021) as in Eq. (3.13).

$$pH = 7 \times \frac{B4}{B12} - 6.5 \times \frac{B9}{B12} \tag{3.13}$$

To calculate *OC*, the spectral bands B4, B5, B11, and B12 (Hengl et al., 2021) were used as shown in Eq. (3.14). Soils with *OC* values above 0.75% are considered very fertile.

$$\begin{gathered} OC = 0.75 \times \frac{(B12-B4)}{B4} + 0.75 \times \frac{B5}{B11} + 0.75 \times \frac{B12}{B11} \\ -0.05 \end{gathered} \tag{3.14}$$

Soil *N* is proportional to the Red to Green Index, B4/B3 (Xu et al., 2018). The vegetation increases with *N* (Mashaba-Munghemezulu et al., 2021). As a result, NDVI is subtracted, and the normalized red edge index is added to the Red to Green Index. The NDVI can be calculated using the equation (B8-B4)/(B8+B4) and the Normalized Red-Edge Index by using (B8-B5)/(B8+B5). High fertility is defined as having a nitrogen content of 560 kg/ha or more, while low fertility is defined as having a nitrogen content of less than 280 kg/ha. Thus, as given in Eq. (3.15), *N* is determined using the spectral bands B3, B4, B5, and B8.

$$N = 560 \times \frac{B4}{B3} - \frac{280}{100} \times \frac{(B8-B4)}{(B8+B4)} + \frac{280}{100} \times \frac{(B8-B5)}{(B8+B5)} \tag{3.15}$$

3.5.2 Comparision of Sentinel-2 data with Laboratory-measured Soil data

The results were compared to Soil-health data (Soil-health, 2021), which includes laboratory-measured soil parameter values for the study area and three other regions of Dakshina Kannada (District), including Marpadi, Mangalore with Longitude: 75.720343, Latitude: 14.360019, Beluvai, Mangalore with Longitude: 74.900180, Latitude: 15.100230, and Attur, Mangalore with Longitude: 74.820230, Latitude: 15.100230 as shown in the Table 3.5.

Table 3.5: Comparision of derived values using Sentinel-2 with Soil-health data

Area Used	**Derived Values using Sentinel-2 Data**			**Soil health Data**			**Observed Variations (Difference)**		
	EC	**pH**	**OC**	**EC**	**pH**	**OC**	**EC**	**pH**	**OC**
Konaje	0.533	6.34	2.406	0.385	6.33	2.2	0.148	**0.01**	0.206
	0.975	6.44	1.522	0.962	6.4	1.16	0.013	0.04	0.362
	1.327	6.59	0.82	1.023	6.51	0.86	**0.304**	0.08	**0.04**
	0.282	6.57	1.128	0.264	6.71	1.41	0.018	0.14	0.282
	1.979	5.36	0.956	1.91	5.87	1.071	0.069	**0.51**	0.115
	1.633	5.64	1.082	1.659	5.76	1.552	0.026	0.12	**0.47**
	1.171	6.43	0.987	1.116	6.41	0.68	0.055	0.02	0.307
	0.282	6.57	1.128	0.275	6.22	1.54	**0.007**	0.35	0.412
Marpadi	1.357	5.63	1.209	1.258	5.69	1.852	**0.099**	0.06	0.643
	1.297	6.18	0.947	1.82	6.13	1.606	0.523	0.05	**0.659**
	1.161	5.74	1.039	1.329	5.77	1.21	0.168	0.03	**0.171**
	1.153	5.85	0.822	1.347	5.85	1.397	0.194	**0**	0.575
	1.129	5.99	0.905	1.331	5.9	1.312	0.202	0.09	0.407
	0.935	8.66	1.595	1.639	8.51	1.921	**0.704**	**0.15**	0.326
Beluvai	1.615	6.53	1.146	1.118	6.79	1.9	0.497	0.26	0.754
	1.615	2.17	1.175	2.202	2.51	1.37	0.587	**0.34**	0.195
	1.615	7.89	1.066	1.263	7.63	1.98	0.352	0.26	**0.914**
	1.615	5.72	2.679	1.024	5.81	2.469	**0.591**	0.09	0.21
	1.615	6.22	1.816	1.027	6	1.7	0.588	0.22	**0.116**
	1.615	5.99	1.825	1.367	5.93	1.496	**0.248**	**0.06**	0.329
Attur	1.846	5.58	1.03	1.24	5.47	1.924	0.606	0.11	0.894
	2.089	4.36	1.133	2.12	4.84	2.239	**0.031**	**0.48**	**1.106**
	1.035	5.22	1.598	1.997	5.12	2.241	**0.962**	**0.1**	0.643
	0.944	2.64	2.483	1.817	2.47	2.29	0.873	0.17	**0.193**

Eq. (3.16) is used to compute the observed variation.

$$Observed\ Variations = \left| Derived\ Values\ using\ Sentinel-2-Data - Soil\ health\ Data \right| \tag{3.16}$$

3.5.3 Dataset Labeling using Clustering Methods

Various clustering techniques such as canopy, density-based, expectation-maximization, farthest-first, fuzzy-C-means, and k means were used to group the data points in the Sentinel-2 dataset. For each clustering algorithm, the number of clusters was fixed to 3. Instances in the dataset were assigned a LOW, or MEDIUM, or HIGH label based on the way the data points were clustered. The dataset was also manually labeled using the estimated values of *pH, EC, OC, N*, and the fertility level of the remaining eight parameters such as *B, Fe, K, Mn, P, S, Cu*, and *Zn*, determined using the *pH* value. Using manual labeling, 293 instances in the dataset were labeled as LOW, 25 as MEDIUM, and 11 as HIGH soil fertility.

Canopy clustering was implemented by calculating the approximate distances between data point pairs and a distance threshold (T1, T2) with T1>T2. The algorithm starts with a set of data points, eliminates each at a time, and then repeats the process over the remaining points to produce a canopy that includes the removed points. If the farthest points are closer to the starting point than T1, they are included in the cluster. Additionally, the point was eliminated from the set if the distance was less than T2. The algorithm iterates until the initial set is empty, building a set of canopies, each with one or more points. By using T1=1.25 and T2=0.75, we obtained the highest clustering accuracy. The Canopy clustering resulted in 279 instances labeled LOW, 36 MEDIUM, and 14 HIGH. It was observed that 250 instances were clustered accurately, resulting in an accuracy of 75.99%.

Density-based clustering uses the local density of each data point to calculate an outlier score. If the local density of a particular data point is low compared to its neighbors, the data point is likely an outlier (Nozad et al., 2021). The data point with the highest local density is chosen as the cluster center (Gu et al., 2020). Density-based clustering identified 135 instances as LOW, 148 as MEDIUM, and 46 as HIGH. It was found that 164 instances were clustered accurately, with an accuracy of 50%.

Expectation-maximization clustering uses the probability that each data point is present in either cluster. This approach resulted in 126 instances classified as LOW, 161 as MEDIUM, and 42 as HIGH. This method of clustering correctly clustered 155 instances, with an accuracy of 47%.

Farthest-first clustering selects a random data point as the initial cluster center. During the cluster assignment phase, the data point farthest from the first center is chosen as the new center. This process is repeated until the 'k' number of centroids has been chosen. Each remaining data point is assigned to the cluster characterized by the centroid nearest to the data point, and the algorithm terminates. Farthest-first requires a single pass to cluster a set of data points (Devi et al., 2020). This clustering method resulted in 216 instances of LOW, 96 of MEDIUM, and 17 of HIGH soil fertility. This clustering method clustered 182 instances accurately, with an accuracy of 55.32%.

The k-means clustering algorithm selects k random data points from the cluster centroids, where k is equal to the number of clusters, and uses a distance metric (usually Euclidean) to assign all the points with the closest distances to the centroid. The algorithm iteratively computes the centroids of newly formed clusters and assigns the remaining data points to the cluster with the closest centroid until clusters are stable (Wang & Kumar, 2019). It is difficult to achieve ideal clusters since the k-means clustering is sensitive to outliers (Guo et al., 2021)). Using k-means clustering resulted in 137 instances of low soil fertility, 146 as having medium soil fertility, and 46 as having high soil fertility. This method of clustering has a 50.46% and correctly grouped 166 instances.

Fuzzy C-Means clustering is an unsupervised ML algorithm that assigns data points to clusters, with points belonging to the same cluster being as similar as possible, and each data point may belong to more than one cluster (Chen et al., 2022). The fuzz clustering can improve the classification speed, whereas it is sensitive to initial cluster centroids (Xue et al., 2016). The clustering is achieved by minimizing the objective function in Eq. (3.17).

$$\sum_{i=1}^{N}\sum_{j=1}^{k}\mu_{ij}^{m}\left|x_i - c_j\right|^2 \tag{3.17}$$

where N is the number of objects, and k is the number of clusters, μ_{ij} is the degree of membership of instance x_i in the j^{th} cluster, m is the fuzzy parameter, which indicates the degree of fuzzy overlap, x_i indicates i^{th} instance, and c_j represents the center of the j^{th} cluster. Initially, μ_{ij} is set randomly, then the c_j and updated μ_{ij} are calculated by

using Eq. (3.18) and Eq. (3.19), respectively.

$$c_j = \frac{\sum_{i=1}^{N} \mu_{ij}^{j} x_i}{\sum_{i=1}^{N} \mu_{ij}^{j}} \tag{3.18}$$

$$\mu_{ij}^{m} = \frac{1}{\sum_{k=1}^{N} \left(\frac{|x_i - c_j|}{|x_i - c_k|} \right)^{\frac{1}{m-1}}} \tag{3.19}$$

This clustering technique classified 155 instances as LOW, 90 as MEDIUM, and 84 as HIGH. The highest accuracy attained by this clustering, with the correct clustering of 145 cases, was 44%.

3.5.4 Proposed Canopy Center-based Fuzzy-C-Means Clustering

This study evaluated and compared the clustering algorithms' accuracy to manual labeling. The state-of-the-art fuzzy C-means clustering uses random cluster centers, which will limit the accuracy of clustering. Hence, the proposed approach selects the cluster centers using the best-performing clustering approach. From the experimental results, it was observed that Canopy clustering achieved better accuracy. Hence, a canopy center-based fuzzy C-means clustering is proposed to improve the clustering accuracy. Figure 3.3 and Algorithm 3.1 depicts the steps involved in the proposed approach. The fuzzy parameter (m) was employed to enhance the fuzzification process. By varying "m" from 1.1 to 1.9 with a 0.1 increment, the optimal values for "m" were discovered. The canopy approach was used to obtain the initial centroids employed in the membership function. The membership function and cluster centroids were updated iteratively until every data point had been allocated to a cluster. The clusters are also labeled as LOW, MEDIUM, or HIGH.

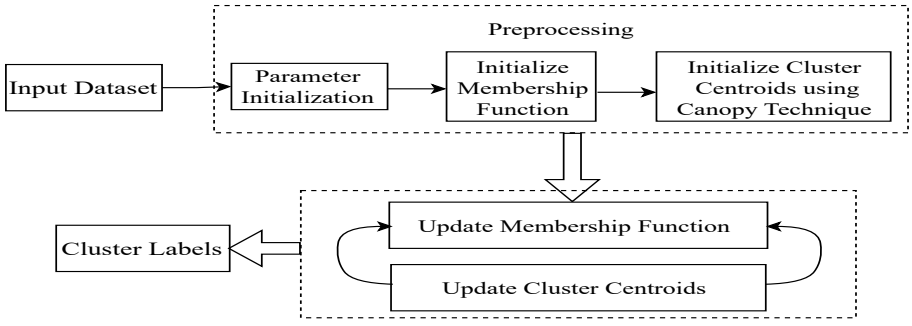


Figure 3.3: Proposed Canopy Center-based Fuzzy-C-Means clustering

Algorithm 3.1 Canopy Center-based Fuzzy-C-Means Clustering

Initialize X with the given dataset and assign the number of clusters, k
Assign fuzzy parameter, m with 1<m<2
Initialize membership matrix M with random values [0,1]

procedure CANOPYFUZZYCMEANSCLUSTERING(X,k, m,M)
    $cur\_iter \leftarrow 0$
    $cluster\_labels \leftarrow \{\}$
    $cluster\_centers \leftarrow$ CALCULATECANOPYCENTER(X)
    while cur_iter<MAX_ITER **do**
        $M \leftarrow$ UPDATEMEMBERSHIPMATRIX($M, cluster\_centers$)
        $clustercenters \leftarrow$ CALCULATECANOPYCENTER(M)
        for i=1 to N **do**
            $idx \leftarrow Indexof\_minimum\_M[i]$
            $cluster\_labels = cluster\_labels \cup idx$
        $cur\_iter \leftarrow cur\_iter + 1$
    return $cluster\_labels$
procedure CALCULATECANOPYCENTER(X)
    Initialize threshold T1, T2 such that T1>T2
    $canopies = \{\}$
    $dist[i,j] \leftarrow$ EUCLIDEANDISTANCE$(x_i, x_j), \forall (x_i, x_j) \in X$
    $canopy\_points = (x_i, x_j)$
    while $canopy\_points \neq \{\}$ **do**
        $point \leftarrow pop(canopy\_points)$
        $i \leftarrow Length(canopies)$
    if $dist[point] < T1$ **then**
        $canopies[i] \leftarrow point$
    if $dist[point] < T2$ **then**
        $X = X - point$
    return $canopies$
procedure UPDATEMEMBERSHIPMATRIX(M,cluster_centers)
    $p \leftarrow \frac{2}{m-1}$
    for i=1 to N **do**
        for j=1 to k **do**
            $distances = X[i] - cluster\_centers[j]$
        for j in k **do**
            for q=1 to k **do**
                $sum = sum + \frac{distances[j]^p}{distances[q]^p}$
                $M[i] \leftarrow \frac{1}{sum}$
    return M

Canopy centers were computed as initial centroids for the Fuzzy-C-Means clustering algorithm with a fixed number of clusters and fixed threshold values T1=1.25 and T2=0.75. Soil fertility can be LOW, MEDIUM, or HIGH. Hence, the number of clusters was fixed to 3 (k=3). We selected a value of fuzzy parameter m, such that 1.1<m <2, and the algorithm obtained better accuracy for m=1.7. The membership matrix, M (i.e., membership function), was initialized using random values and updated by calculating the Euclidean distance between a pair of data points and cluster centers using Eq. (3.20).

$$M_{ij} = \frac{1}{\sum_{k=1}^{N} \left(\frac{\left(x_i - c_j\right)^2}{\left(x_i - c_k\right)^2} \right)^{\frac{1}{m-1}}} \tag{3.20}$$

where M_{ij} indicates the degree to which an observation x_i belongs to cluster c_j. The value M_{ij} is inversely proportional to the distance from x to the cluster center.

The algorithm used a maximum of 100 iterations to obtain optimal clustering. In fuzzy C-means clustering, the fuzziness of the data point belonging to more than one cluster was avoided by selecting a maximum value from the membership function. But, soil fertility depends on the level of each soil parameter. The low fertility level of any soil chemical parameters makes the soil less fertile. Thus, the algorithm selects the minimum membership value to avoid overlap. The proposed clustering technique produced 254 instances of LOW, 60 instances of MEDIUM, and 15 instances of HIGH fertile soil. With an accuracy of 78.42%, this approach correctly clustered 258 instances. The comparison of the accuracy of all clustering techniques used in the study is depicted in Figure 3.4.

3.5.5 Performance of Classifiers using Dataset Labeled using Different Clustering Techniques

The classification accuracy using the dataset labeled using different methods is assessed using four ML-based classifiers, such as NB, SVM, J48, and RF. The classification results obtained after applying Canopy clustering are shown in Table 3.6. The RF classifier with a split dataset achieved the highest Accuracy of 98.78%, Precision of 0.989, Recall of 0.988, and F1-Score of 0.987. The classification results obtained after Density-based clustering are shown in Table 3.7. The RF classifier with a 10-fold

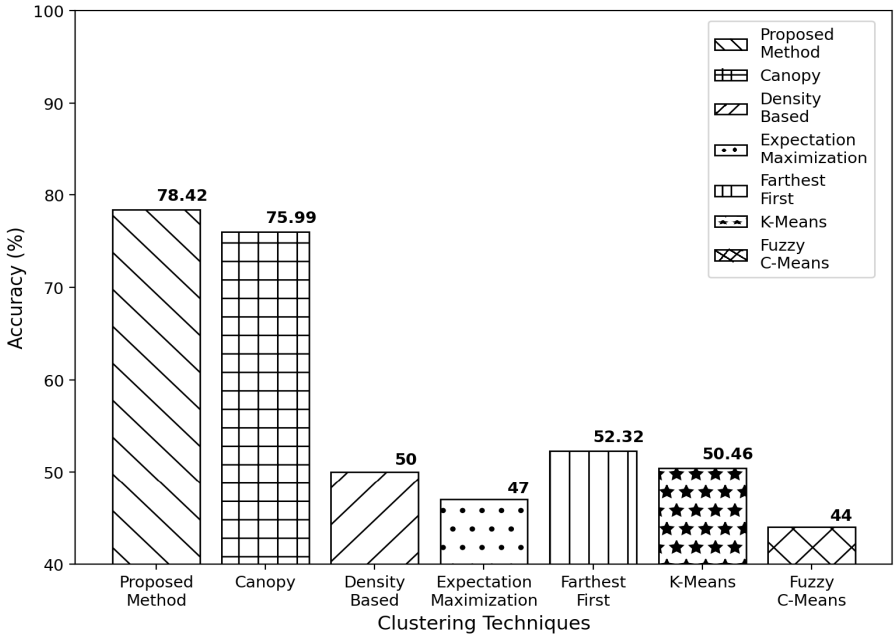


Figure 3.4: Analyzing accuracy of different clustering techniques

cross-validation test achieved the highest Accuracy of 96.96% and Precision, Recall, and F1-Score of 0.970. Table 3.8 displays the classification results achieved using Expectation-Maximization clustering. The RF classifier attained the best Accuracy of 97.57% and Precision, Recall, and F1-Score of 0.976 using a 10-fold cross-validation test. The results of classification obtained after the Farthest-first clustering are shown in Table 3.9. The RF classifier with a 10-fold cross-validation test of the dataset obtained the highest Accuracy of 97.87%, Precision, Recall, and F1-Score of 0.979. Table 3.10 displays the classification results achieved using k-means clustering. RF classifier with a 10-fold cross-validation test performed better with an Accuracy of 96.05% and with Precision, Recall, and F1-Score of 0.960. Table 3.11 depicts the classification results of fuzzy C-means clustering. The J48 classifier achieved the greatest Accuracy of 98.78%, Precision, Recall, and F1-Score of 0.988 using a split dataset as test data. Table 3.12 shows the classification results of the manually labeled dataset. The RF and J48 classifier with a 10-fold cross-validation test achieved the highest Accuracy of 98.48%, Precision, Recall, and F1-Score of 0.985. The proposed method achieved the highest clustering accuracy compared to other clustering methods. The results of applying classification are presented in Table 3.13. The RF classifier with 10-fold cross-validation of the dataset obtained the highest Accuracy of 99.69%, Precision, Recall, and F1-Score of 0.977.

Table 3.6: Data classification after Canopy clustering

Method Used	Classifier Name	Accuracy	Precision	Recall	F1-Score
10-fold Cross-validation	**NB**	95.44%	0.960	0.954	0.956
	SVM	90.27%	0.913	0.903	0.882
	J48	95.14%	0.955	0.951	0.953
	RF	97.87%	0.980	0.979	0.979
75% Split Training Data: 25% Split Test Data	**NB**	95.12%	0.953	0951	0.952
	SVM	91.46%	0.922	0.915	0.899
	J48	96.34%	0.963	0.963	0.963
	RF	98.78%	0.989	0.988	0.987

Table 3.7: Data classification after Density-based clustering

Method Used	Classifier Name	Accuracy	Precision	Recall	F1-Score
10-fold Cross-validation	**NB**	96.35%	0.964	0.964	0.964
	SVM	86.63%	0.881	0.866	0.860
	J48	97.87%	0.979	0.979	0.979
	RF	96.96%	0.970	0.970	0.970
75% Split Training Data: 25% Split Test Data	**NB**	96.34%	0.964	0.963	0.964
	SVM	85.37%	0.857	0.854	0.850
	J48	96.34%	0.964	0.963	0.963
	RF	96.34%	0.964	0.963	0.963

Table 3.8: Data classification after Expectation-Maximization clustering

Method Used	Classifier Name	Accuracy	Precision	Recall	F1-Score
10-fold Cross-validation	**NB**	95.44%	0.960	0.954	0.956
	SVM	82.98%	0.842	0.830	0.823
	J48	96.96%	0.970	0.970	0.970
	RF	97.57%	0.976	0.976	0.976
75% Split Training Data: 25% Split Test Data	**NB**	96.34%	0.965	0.963	0.964
	SVM	74.39%	0.784	0.744	0.738
	J48	95.12%	0.956	0.951	0.951
	RF	96.34%	0.966	0.963	0.963

Table 3.9: Data classification after Farthest-first clustering

Method Used	Classifier Name	Accuracy	Precision	Recall	F1-Score
10-fold Cross-validation	**NB**	89.67%	0.924	0.897	0.901
	SVM	84.50%	0.809	0.845	0.826
	J48	96.96%	0.970	0.970	0.970
	RF	97.87%	0.979	0.979	0.979
75% Split Training Data: 25% Split Test Data	**NB**	91.46%	0.954	0.915	0.926
	SVM	74.39%	0.750	0.744	0.718
	J48	91.46%	0.925	0.915	0.912
	RF	96.34%	0.964	0.963	0.963

Figure 3.5 shows the classification accuracy of different classifiers such as NB, SVM, J48, and RF using the dataset labeled using different techniques.

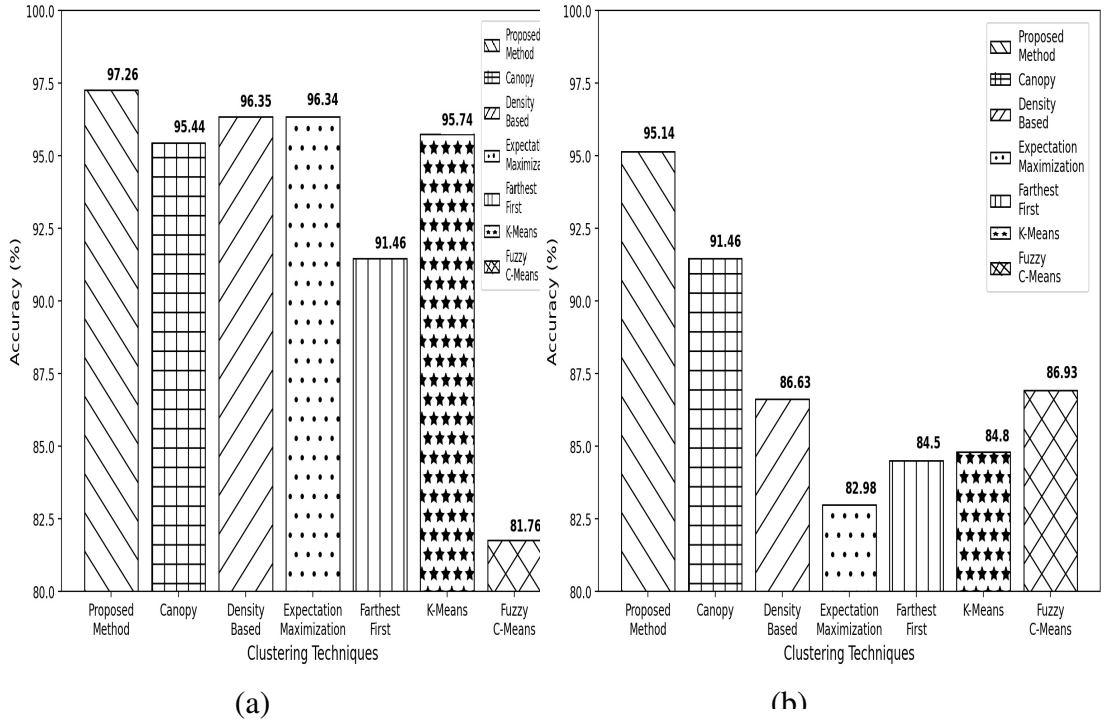


Figure 3.5: Comparision of clustering techniques based on classification accuracy achieved using a) NB, b) SVM, c) J48, d) RF

Table 3.10: Data classification after k-means clustering

Method Used	Classifier Name	Accuracy	Precision	Recall	F1-Score
10-fold Cross-validation	**NB**	95.74%	0.958	0.957	0.958
	SVM	84.80%	0.866	0.848	0.839
	J48	95.14%	0.951	0.951	0.951
	RF	96.05%	0.960	0.960	0.960
75% Split Training Data: 25% Split Test Data	**NB**	95.12%	0.952	0.951	0.951
	SVM	82.93%	0.839	0.829	0.825
	J48	95.12%	0.952	0.951	0.951
	RF	95.12%	0.952	0.951	0.951

Table 3.11: Data classification after fuzzy C-means clustering

Method Used	Classifier Name	Accuracy	Precision	Recall	F1-Score
10-fold Cross-validation	**NB**	81.76%	0.947	0.818	0.813
	SVM	86.93%	0.875	0.869	0.852
	J48	98.18%	0.982	0.982	0.982
	RF	97.57%	0.976	0.976	0.975
75% Split Training Data: 25% Split Test Data	**NB**	80.49%	0.834	0.805	0.800
	SVM	84.15%	0.869	0.841	0.808
	J48	98.78%	0.988	0.988	0.988
	RF	97.56%	0.976	0.976	0.976

Table 3.12: Classification using manually labeled dataset

Method Used	Classifier Name	Accuracy	Precision	Recall	F1-Score
10-fold Cross-validation	**NB**	93.92%	0.940	0.939	0.939
	SVM	89.06%	0.877	0.891	0.866
	J48	98.48%	0.985	0.985	0.985
	RF	98.48%	0.985	0.985	0.985
75% Split Training Data: 25% Split Test Data	**NB**	93.42%	0.983	0934	0.953
	SVM	90.24%	0.912	0.902	0.877
	J48	97.56%	0.984	0.976	0.976
	RF	97.56%	0.984	0.976	0.976

Table 3.13: Data classification after proposed Canopy Center-based Fuzzy-C-Means clustering

Method Used	Classifier Name	Accuracy	Precision	Recall	F1-Score
10-fold Cross-validation	**NB**	97.26%	0.973	0.973	0.973
	SVM	95.14%	0.956	0.951	0.951
	J48	99.39%	0.994	0.994	0.994
	RF	99.69%	0.997	0.997	0.997
75% Split Training Data: 25% Split Test Data	**NB**	96.34%	0.965	0.963	0.964
	SVM	92.68%	0.937	0.927	0.927
	J48	98.78%	0.988	0.988	0.988
	RF	98.78%	0.988	0.988	0.988

3.5.6 Summary

This study derived soil parameters such as *EC, pH, OC* and *N* from acquired Sentinel-2 spectral bands. The data points were clustered using a variety of cutting-edge cluster-

ing techniques. It was found that Canopy clustering achieved a clustering accuracy of 75.99%, and by utilizing an RF classifier with 10-fold cross-validation, Canopy clustering achieved a classification Accuracy of 98.78%. It was observed that using the proposed Canopy Center-based Fuzzy-C-Means clustering achieved the highest clustering accuracy of 78.42%, and by utilizing an RF classifier with a 10-fold cross-validation proposed approach obtained a classification Accuracy of 99.69%. The classification of soil fertility using satellite-derived near real-time data is more accurate than classification using laboratory-measured data.

CHAPTER 4

SOIL FERTILITY CLASSIFICATION WITH AUTOMATED FERTILIZER PRESCRIPTION

The automated fertilizer prescription system was proposed and developed to prescribe precise amounts of fertilizers for prevalent crops in specific districts of Karnataka (State), India such as paddy, black pepper, cucumber, black gram/ green gram, turmeric, and arecanut. The cereal crop paddy grows in coastal saline or alluvial soil. It is grown during the Kharif and Rabi seasons, sowing in June-July and October-November and harvesting in September-October and March-April, respectively (Reddy, 2018). Pulses such as black gram/ green gram are commonly grown during the Kharif or Zaid season in mixed red and black soil. Black pepper is a medicinal and spice crop cultivated in the Kharif season. The vegetable crop cucumber is typically grown in the Zaid. The arecanut is a horticulture crop that requires laterite or clay loam soil. The quantity of fertilizers for different crops based on fertility levels of *N, P,* and *K* is shown in Table 4.1, Table 4.2, and Table 4.3, respectively. The fertilizer prescription for any crop, on the deficiency of '*S*' and other micronutrients, is presented in Table 4.4. The study uses classification results to prescribe the fertilizers.

4.1 Proposed Ensemble Filter-based Feature Selection for Soil Fertility Classification with Fertilizer Recommendation

Soil fertility classifier needs to be robust to accurately classify soil fertility. The selected features are inconsistent across different feature selection techniques and depend on the datasets employed (Pes, 2020). This research aimed at developing a robust ML-based classifier based on relevant features recommended by the ensemble filter-based feature selection. An ensemble filter-based feature selection approach was created to address the issue of inconsistent feature scores. It utilized filter-based feature selection methods such as Information Gain (InfoG) (Beraha et al., 2019), Gain Ratio (GainR) (Pes, 2020), and Relief Feature (ReliefF) (Chandrashekar & Sahin, 2014).

Table 4.1: Recommended neem-coated urea quantity (kg/ha) based on 'N' fertility level

Fertility Level	Paddy		Black Pepper	Cucumber	Black Gram /Green Gram		Turmeric	Arecanut
	Kharif	Rabi	Kharif	-	Kharif	Zaid	-	-
LOW	289.13	360.87	289.13	173.91	36.96	71.74	434.78	434.78
MEDIUM	217.39	271.74	217.39	130.43	28.26	54.35	326.09	326.09
HIGH	145.65	180.43	145.65	86.96	17.39	36.96	217.39	217.39

Table 4.2: Recommended single superphosphate quantity (kg/ha) based on 'P' fertility level

Fertility level	Paddy		Black Pepper	Cucumber	Black Gram /Green Gram		Turmeric	Arecanut
	Kharif	Rabi	Kharif	-	Kharif	Zaid	-	-
LOW	418.75	518.75	331.25	418.75	206.25	418.75	1037.50	500
MEDIUM	312.5	393.75	250	312.50	156.25	312.50	781.25	375
HIGH	206.25	262.5	168.75	206.25	106.25	206.25	518.75	250

Table 4.3: Recommended potassium chloride quantity (kg/ha) based on 'K' fertility level

Fertility Level	Paddy		Black Pepper	Cucumber	Black Gram /Green Gram		Turmeric	Arecanut
	Kharif	Rabi	Kharif	-	Kharif	Zaid	-	-
LOW	111.67	138.33	310	176.67	55	55	555.00	465
MEDIUM	83.33	105	233.33	133.33	41.67	41.67	416.67	350
HIGH	55	70	155	88.33	28.33	28.33	278.33	233.33

Table 4.4: Quantity of fertilizers recommended (in kg/ha) on deficiency of soil nutrient 'S' and micronutrients

Soil Parameter	Fertilizer Name	Quantity
S	Sulphur / Gypsum	S: 20-40 / Gypsum: 140-280
B	Borax	5-10
Cu	Copper sulphate	5-10
Fe	Ferrous sulphate	25-50
Mn	Manganese sulphate	10 -25
Zn	Zinc sulphate	15-25

4.1.1 Dataset Used

In this proposed method soil health data (Soil-health, 2021) collected from farmlands in Dakshina Kannada and Gulbarga districts of Karnataka (State), India. Acidic soil is the dominant type of soil in Dakshina Kannada. Gulbarga, a non-coastal district, has either neutral or alkaline soil (Badrinath et al., 1995). The dataset consists of 19 attributes such as sample number, state name, district name, block name, village code, village name, latitude, longitude, and 11 soil chemical parameters. The classifiers are developed using *EC, pH, OC, P, K, S. B, Cu, Fe, Mn*, and *Zn*. The open-source WEKA tool is used to eliminate redundant data, missing parameter values, and parameters with a value of 0. The instances are labeled as LOW, or MEDIUM, or HIGH using the level of soil parameters as discussed in Section 3.2. After preprocessing, the data collected from Dakshina Kannada district (dataset-1) had 36796 instances of LOW fertile, 158 instances of MEDIUM, and 25 instances of HIGH fertile. The preprocessing of data collected from the Gulbarga district resulted in 40929 instances, of which 36979 were randomly selected as dataset-2. Dataset-2 contains 36775 instances of LOW fertile, 174 instances of MEDIUM, and 30 instances of HIGH fertile.

4.1.2 Proposed Approach

Figure 4.1 outlines the proposed soil fertility classification approach. In the proposed method The laboratory-measured soil chemical parameters are acquired from farmlands of two regions with different climate conditions. the features are ranked using three filter-based feature selection techniques: InfoG, GainR, and ReliefF. The feature ranked 1 is considered more relevant, whereas the higher rank is considered the least relevant. It was observed that the feature ranks depend on the datasets used, and the ranks obtained for a dataset using three different techniques were not constant. Table 4.5 and Table 4.6 shows the feature ranking using InfoG, GainR, and ReliefF approaches for dataset-1 and dataset-2, respectively. The least relevant feature in dataset-1, according to InfoG and GainR, was ‘S’, while ReliefF identified ‘B’ as the least relevant. In dataset-2, ‘P’ was identified as the least relevant feature by InfoG and GainR, while ‘S’ was the least relevant based on ReliefF. Figure 4.2 shows the proposed ensemble filter-based feature selection method designed to maintain stability in feature selection. The average of ranks obtained by InfoG, GainR, and ReliefF for each soil parameter ‘x’ is calculated

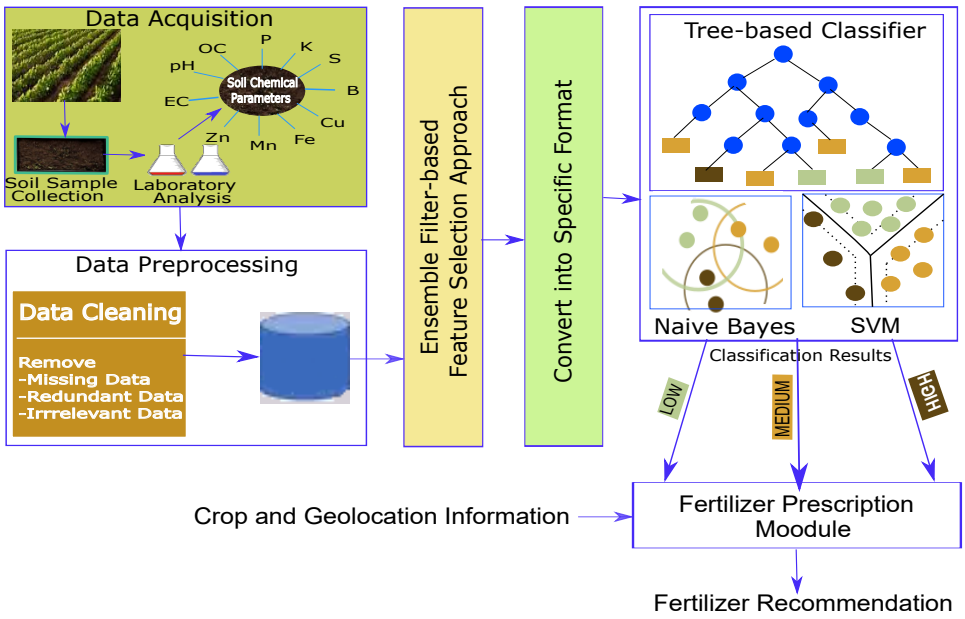


Figure 4.1: Steps involved in proposed ensemble filter-based soil fertility classification approach

using Eq. (4.1).

$$Average \quad Rank(x) = \frac{Rank_{InfoG}(x) + Rank_{GainR}(x) + Rank_{ReliefF}(x)}{3} \tag{4.1}$$

The features were arranged in ascending order of their rankings. The soil parameter with the highest average rank was chosen as the parameter to be eliminated. The subset of features by removing the feature with the highest average rank was given as input to the tree-based classifier. ML-based classifiers such as CART, ET, J48, RF, REPTree, NB, and SVM are employed to classify soil fertility. The robustness of the developed classifier is evaluated using two different datasets. The tree-based classifier classifies the input data instances as LOW, or MEDIUM, or HIGH fertility. Based on the classification results, the proposed model prescribes fertilizers for paddy, green gram/ black gram, black pepper, and cucumber, which are predominant in Dakshina Kannada and Gulbarga districts. The fertility level of N and S is determined based on pH value.

4.1.3 Experimental Results of Feature Selection

The open-source WEKA tool was employed for obtaining feature scores. The proposed method assigns rank 1 to the feature having the lowest average rank and the highest rank to the feature with the highest average rank. The generated feature rank is provided in Table 4.7. Both datasets ranked the feature 'S' as the highest, with a rank 11.

Table 4.5: Feature selection scores and ranks for dataset-1

Features (x)	$Score_{InfoG(x)}$	$Score_{GainR(x)}$	$Score_{ReliefF(x)}$	$Rank_{InfoG(x)}$	$Rank_{GainR(x)}$	$Rank_{ReliefF(x)}$	Average Rank
EC	0.00238	0.00193	0.018103	9	8	4	7
pH	0.01207	0.01483	0.055469	2	2	2	2
OC	0.0021	0.00163	0.004477	10	9	7	8.67
K	0.00241	0.00157	0.058884	8	10	1	6.33
P	0.00414	0.00209	0.000824	6	7	9	7.33
S	0.00174	0.00139	0.016596	11	11	5	9
B	0.0071	0.00499	0.000107	4	4	11	6.33
Cu	0.01497	0.02021	0.003144	1	1	8	3.33
Fe	0.00351	0.00311	0.023297	7	6	3	5.33
Mn	0.00641	0.00481	0.012952	5	5	6	5.33
Zn	0.01097	0.01138	0.000107	3	3	10	5.33

Table 4.6: Feature selection scores and ranks for dataset-2

Features (x)	$Score_{InfoG(x)}$	$Score_{GainR(x)}$	$Score_{ReliefF(x)}$	$Rank_{InfoG(x)}$	$Rank_{GainR(x)}$	$Rank_{ReliefF(x)}$	Average Rank
EC	0.02031	0.27375	0.012372	2	1	2	1.67
pH	0.01665	0.01017	0.032111	4	7	1	4
OC	0.01643	0.00885	0.005453	5	8	5	6
K	0.01397	0.01033	0.00374	8	6	7	7
P	0.00435	0.00297	0.001717	11	11	8	10
S	0.01042	0.00629	0.000475	10	10	11	10.33
B	0.01187	0.00751	0.001393	9	9	9	9
Cu	0.02379	0.04659	0.007797	1	2	3	2
Fe	0.01563	0.01639	0.005312	6	4	6	5.33
Mn	0.01535	0.0242	0.006917	7	3	4	4.67
Zn	0.01828	0.01518	0.001236	3	5	10	6

4.1.4 Performance Evaluation of Classifiers

The fertilizer prescription module uses classifiers implemented in Google Collab with Python and the python-weka-wrapper3 package (Python_weka_wrapper3, 2022). The experiments were conducted using 10-fold cross-validation and using split datasets.

The tree-based classifiers were created using a batch size of 100. Initially, all soil parameters from the datasets were included in the experiment. For dataset-1, CART produced a tree with size 25 and 13 leaf nodes using 5-fold pruning and a seed value of 1, and for dataset-2, CART produced a tree with size 39 and 20 leaf nodes. ET was

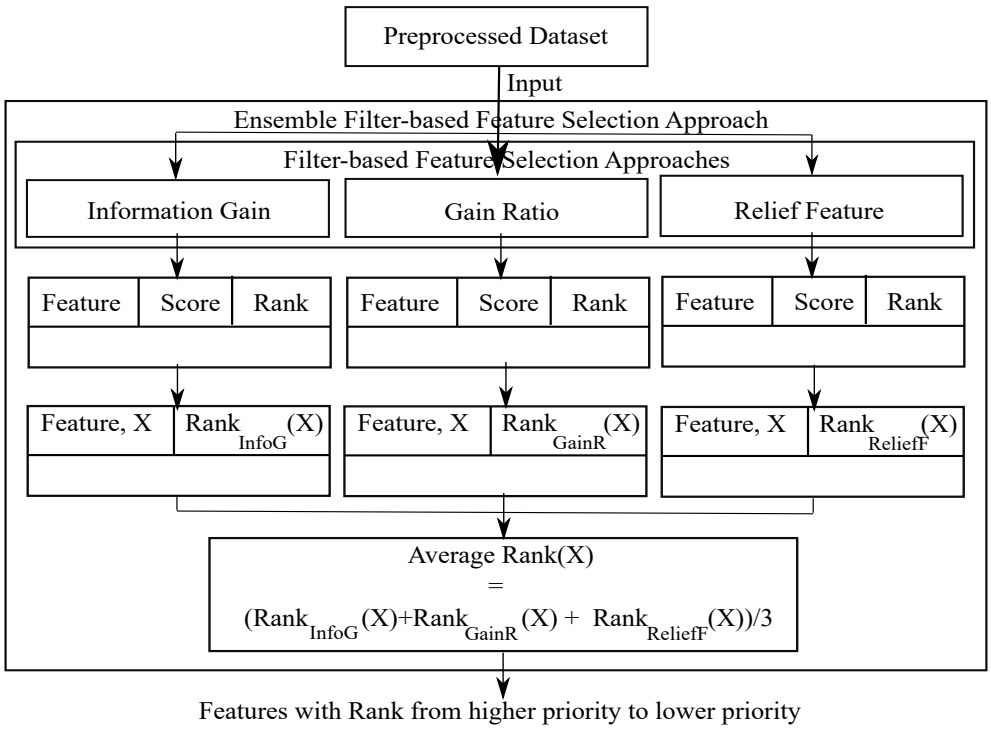


Figure 4.2: Proposed ensemble filter-based feature selection

created with a split of a minimum of two instances per node, using seed value of 1, and randomly selecting $\sqrt{(m-1)}$ attributes at each node. The tree generated for dataset-1 had 1301 nodes, and 561 nodes for dataset-2. The J48, using a fold of three and a seed value of one, generated a tree with 33 nodes and 17 leaves for dataset-1 and a tree of the size of 45 and 23 leaves for dataset-2. Datasets-1 and dataset-2 were classified using RF using 100 bags and 100 iterations. The REP used a variance proportion of 0.001, a maximum of three folds, and a seed value of 1. Using dataset-1 and dataset-2, REP generated trees of 27 and 39 sizes, respectively. The experiments were repeated without the feature 'S' in the datasets. With five folds of pruning and a seed value of 1, CART created a tree of size 25 with 13 leaf nodes for dataset-1 and a tree of size 43 with 22 leaves for dataset-2. ET created trees with 1033 and 655 nodes for datasets-1 and dataset-2, respectively. J48 produced a tree of size 33 with 17 leaves for dataset-1 and a tree of size 41 with 21 leaves for dataset-2. The RF classifier created the random forest using 100 bags and 100 iterations. For dataset-1 and dataset-2, REP produced trees with sizes of 27 and 45, respectively. NB and SVM classifiers used a batch size of 100. A radial bias kernel function is utilized to create an SVM classifier with a seed and cost value of 1, gamma value of 0.1, epsilon of 0.001, loss value of 0.1, and degree of 3.

Table 4.7: Ranking of features using the proposed approach

Dataset-1			Dataset-2		
Features	**Average Rank**	**Rank**	**Features**	**Average Rank**	**Rank**
pH	2	1	**EC**	1.67	1
Cu	3.33	2	**Cu**	2	2
Fe	5.33	3	**pH**	4	3
Mn	5.33	4	**Mn**	4.67	4
Zn	5.33	5	**Fe**	5.33	5
K	6.33	6	**OC**	6	6
B	6.33	7	**Zn**	6	7
EC	7	8	**K**	7	8
P	7.33	9	**B**	9	9
OC	8.67	10	**P**	10	10
S	**9**	**11**	S	**10.33**	**11**

Table 4.8 shows the performance of classifiers with 10-fold cross-validation on dataset-1 without removing feature 'S'. Without eliminating feature 'S,' for dataset-1, CART, J48, and RF classifiers obtained an Accuracy of 99.92% whereas REP, ET, NB, and SVM classifier achieved an Accuracy of 99.91%, 99.19%, 76.99%, and 99.51%, respectively. For dataset-2, J48, RF, and REP resulted in an Accuracy of 99.89% whereas CART, ET, NB, and SVM classifier was 99.84%, 99.64%, 99.14%, and 99.45%, respectively. The performance for dataset-1 and dataset-2 after removing feature 'S' are given in Table 4.10 and Table 4.11, respectively. With relevant features for dataset-1 CART, ET, J48, RF, REP, NB, and SVM classifier obtained an Accuracy 99.92%, 99.20%, 99.92%, 99.93%, and 99.92%, 76.45%, and 99.5%, respectively. For dataset-2, CART, ET, J48, RF, REP, NB, and SVM classifier achieved an Accuracy of 99.84%, 99.67%, 99.88%, 99.91%, 99.88%, 89.73%, and 99.45%, respectively. The Kappa statistic achieved by classifiers for dataset-1 and dataset-2 is depicted in Figure 4.3.

The performance of the classifiers with a split dataset for dataset-1 and dataset-2 without removing feature 'S' are as shown in Table 4.12 and Table 4.13, respectively. Without removing feature 'S,' for dataset-1 CART, ET, J48, RF, REP, NB, and SVM classifier obtained an Accuracy 99.96%, 99.29%, 99.96%, 99.94%, and 99.95%, 77.06%, and 99.56% respectively. For dataset-2, the CART, ET, J48, RF, REP, NB, and SVM classifier achieved an Accuracy of 99.81%, 99.62%, 99.87%, 99.88%, 99.83%, 99.84%, and 99.47%, respectively. The performance of classifiers after removing

Table 4.8: Performance of classifiers with 10-fold cross-validation for dataset-1 with all features

Classifier	Accuracy	Per-class Precision			Per-class Recall			Per-class F1-Score			Precision	Recall	F1-Score
		LOW	MED	HIGH	LOW	MED	HIGH	LOW	MED	HIGH			
CART	99.92%	1.000	0.907	0.862	1.000	0.924	1.000	1.000	0.915	0.926	0.999	0.999	0.999
ET	99.19%	0.996	0.172	0.000	0.996	0.171	0.000	0.996	0.171	0.000	0.992	0.992	0.992
J48	99.92%	1.000	0.900	0.889	1.000	0.911	0.960	1.000	0.906	0.923	0.999	0.999	0.999
RF	99.92%	1.000	0.928	0.885	1.000	0.899	0.920	1.000	0.913	0.902	0.999	0.999	0.999
REP	99.91%	1.000	0.897	0.857	1.000	0.937	0.960	1.000	0.916	0.906	0.999	0.999	0.999
NB	76.99%	1.000	0.017	0.065	0.770	0.911	0.200	0.870	0.033	0.098	0.995	0.770	0.866
SVM	99.51%	0.995	-	-	1.000	0.000	0.000	0.998	-	-	-	0.995	-

Table 4.9: Performance of classifiers with 10-fold cross-validation for dataset-2 using all features

Classifier	Accuracy	Per-class Precision			Per-class Recall			Per-class F1-Score			Precision	Recall	F1-Score
		LOW	MED	HIGH	LOW	MED	HIGH	LOW	MED	HIGH			
CART	99.84%	0.999	0.886	0.692	1.000	0.805	0.600	0.999	0.843	0.643	0.998	0.998	0.998
ET	99.64%	0.999	0.661	0.303	0.998	0.695	0.333	0.999	0.678	0.317	0.997	0.996	0.996
J48	99.89%	0.999	0.920	0.742	1.000	0.856	0.767	1.000	0.887	0.754	0.999	0.999	0.999
RF	99.89%	0.999	0.910	0.923	1.000	0.874	0.400	1.000	0.891	0.558	0.999	0.999	0.999
REP	99.89%	1.000	0.895	0.727	1.000	0.879	0.533	1.000	0.887	0.615	0.999	0.999	0.999
NB	88.14%	1.000	0.036	0.018	0.882	0.879	0.167	0.937	0.069	0.032	0.994	0.881	0.932
SVM	99.45%	0.994	-	-	1.000	0.000	0.000	0.997	-	-	-	0.994	-

Table 4.10: Performance of classifiers with 10-fold cross-validation for dataset-1 after removing feature 'S'

Classifier	Accuracy	Per-class Precision			Per-class Recall			Per-class F1-Score			Precision	Recall	F1-Score
		LOW	MED	HIGH	LOW	MED	HIGH	LOW	MED	HIGH			
CART	99.92%	1.000	0.907	0.862	1.000	0.924	1.000	0.915	1.000	0.926	0.999	0.999	0.999
ET	99.20%	0.996	0.195	0.040	0.996	0.196	0.040	0.996	0.196	0.040	0.992	0.992	0.992
J48	99.92%	1.000	0.912	0.862	1.000	0.918	1.000	1.000	0.915	0.926	0.999	0.999	0.999
RF	99.93%	1.000	0.918	0.875	1.000	0.924	0.840	1.000	0.921	0.857	0.999	0.999	0.999
REP	99.92%	1.000	0.897	0.857	1.000	0.937	0.960	1.000	0.916	0.906	0.999	0.999	0.999
NB	76.45%	1.000	0.017	0.068	0.764	0.918	0.200	0.866	0.032	0.101	0.995	0.765	0.862
SVM	99.51%	0.995	-	-	1.000	0.000	0.000	0.998	-	-	-	0.995	-

Table 4.11: Performance of classifiers with 10-fold cross-validation for dataset-2 after removing feature 'S'

Classifier	Accuracy	Per-class Precision			Per-class Recall			Per-class F1-Score			Precision	Recall	F1-Score
		LOW	MED	HIGH	LOW	MED	HIGH	LOW	MED	HIGH			
CART	99.84%	0.999	0.870	0.750	1.000	0.805	0.600	0.999	0.836	0.667	0.998	0.998	0.998
ET	99.67%	0.999	0.707	0.231	0.999	0.707	0.200	0.999	0.707	0.214	0.997	0.997	0.997
J48	99.88%	0.999	0.912	0.742	1.000	0.833	0.767	1.000	0.871	0.754	0.999	0.999	0.999
RF	99.91%	0.999	0.933	0.944	1.000	0.885	0.567	1.000	0.909	0.708	0.999	0.999	0.999
REP	99.88%	1.000	0.889	0.679	1.000	0.874	0.633	1.000	0.881	0.655	0.999	0.999	0.999
NB	89.73%	1.000	0.041	0.017	0.898	0.879	0.133	0.946	0.079	0.030	0.994	0.897	0.941
SVM	99.45%	0.994	-	-	1.000	0.000	0.000	0.997	-	-	-	0.994	-

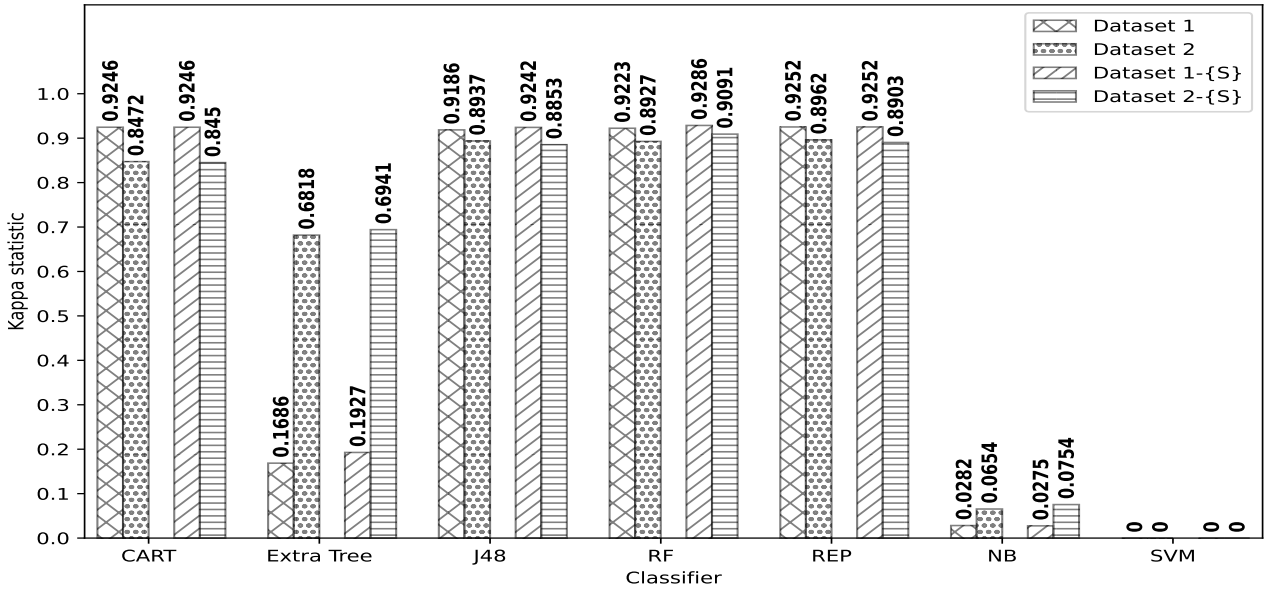


Figure 4.3: Kappa statistic achieved by ensemble filter-based approach with 10-fold cross-validation

Table 4.12: Performance of classifiers with Split dataset for dataset-1 with all features

Classifier	Accuracy	Per-class Precision			Per-class Recall			Per-class F1-Score			Precision	Recall	F1-Score
		LOW	MED	HIGH	LOW	MED	HIGH	LOW	MED	HIGH			
CART	99.96%	1.000	0.927	0.667	1.000	0.974	1.000	1.000	0.950	0.800	1.000	1.000	1.000
ET	99.29%	0.997	0.256	0.000	0.996	0.256	0.000	0.996	0.256	0.000	0.993	0.993	0.993
J48	99.96%	1.000	0.927	0.667	1.000	0.974	1.000	1.000	0.950	0.800	1.000	1.000	1.000
RF	99.94%	1.000	0.946	0.500	1.000	0.897	0.500	1.000	0.921	0.500	0.999	0.999	0.999
REP	99.95%	1.000	0.905	0.667	1.000	0.974	1.000	1.000	0.938	0.800	1.000	0.999	0.999
NB	77.06%	1.000	0.017	0.000	0.770	0.949	0.000	0.870	0.034	0.000	0.996	0.771	0.866
SVM	99.56%	0.996	-	-	1.000	0.000	0.000	0.998	-	-	-	0.996	-

Table 4.13: Performance of classifiers with Split dataset for dataset-2 with all features

Classifier	Accuracy	Per-class Precision			Per-class Recall			Per-class F1-Score			Precision	Recall	F1-Score
		LOW	MED	HIGH	LOW	MED	HIGH	LOW	MED	HIGH			
CART	99.81%	0.999	0.931	0.500	1.000	0.675	0.778	0.999	0.783	0.609	0.998	0.998	0.998
ET	99.62%	0.998	0.700	0.154	0.999	0.525	0.222	0.999	0.525	0.182	0.996	0.996	0.996
J48	99.87%	1.000	0.917	0.600	0.999	0.825	1.000	1.000	0.868	0.750	0.999	0.999	0.999
RF	99.88%	0.999	0.897	1.000	1.000	0.875	0.333	1.000	0.886	0.500	0.999	0.999	0.999
REP	99.83%	0.998	0.935	1.000	1.000	0.725	0.667	0.999	0.817	0.800	0.998	0.998	0.998
NB	99.84%	1.000	0.043	0.051	0.920	0.800	0.222	0.958	0.082	0.083	0.994	0.918	0.953
SVM	99.47%	0.995	-	-	1.000	0.000	0.000	0.997	-	-	-	0.995	-

Table 4.14: Performance of classifiers with Split dataset for dataset-1 after removing feature ‘S’

Classifier	Accuracy	Per-class Precision			Per-class Recall			Per-class F1-Score			Precision	Recall	F1-Score
		LOW	MED	HIGH	LOW	MED	HIGH	LOW	MED	HIGH			
CART	99.96%	1.000	0.927	0.667	1.000	0.974	1.000	1.000	0.950	0.800	1.000	1.000	1.000
ET	99.32%	0.997	0.282	0.000	0.996	0.282	0.000	0.997	0.282	0.000	0.994	0.993	0.993
J48	99.96%	1.000	0.927	0.667	1.000	0.974	1.000	1.000	0.950	0.800	1.000	1.000	1.000
RF	99.96%	1.000	0.927	0.667	1.000	0.974	1.000	1.000	0.950	0.800	1.000	1.000	1.000
REP	99.95%	1.000	0.905	0.667	1.000	0.974	1.000	1.000	0.938	0.800	1.000	0.999	0.999
NB	76.27%	1.000	0.017	0.000	0.762	0.949	0.000	0.865	0.033	0.000	0.996	0.763	0.861
SVM	99.56%	0.996	-	-	1.000	0.000	0.000	0.998	-	-	-	0.996	-

Table 4.15: Performance of classifiers with Split dataset for dataset-2 after removing feature 'S'

Classifier	Accuracy	Per-class Precision			Per-class Recall			Per-class F1-Score			Precision	Recall	F1-Score
		LOW	MED	HIGH	LOW	MED	HIGH	LOW	MED	HIGH			
CART	99.85%	0.999	0.889	1.000	1.000	0.800	0.778	0.999	0.842	0.875	0.998	0.998	0.998
ET	99.73%	0.999	0.743	0.375	0.999	0.650	0.333	0.999	0.693	0.353	0.997	0.997	0.997
J48	99.88%	0.999	0.971	0.667	1.000	0.825	0.889	1.000	0.892	0.762	0.999	0.999	0.999
RF	99.90%	0.999	0.923	1.000	1.000	0.900	0.444	1.000	0.911	0.615	0.999	0.999	0.999
REP	99.81%	0.999	0.872	1.000	1.000	0.850	0.667	1.000	0.861	0.800	0.999	0.999	0.999
NB	90.14%	1.000	0.035	0.054	0.902	0.800	0.222	0.949	0.068	0.087	0.994	0.901	0.944
SVM	99.47%	0.995	-	-	1.000	0.000	0.000	0.997	-	-	-	0.995	-

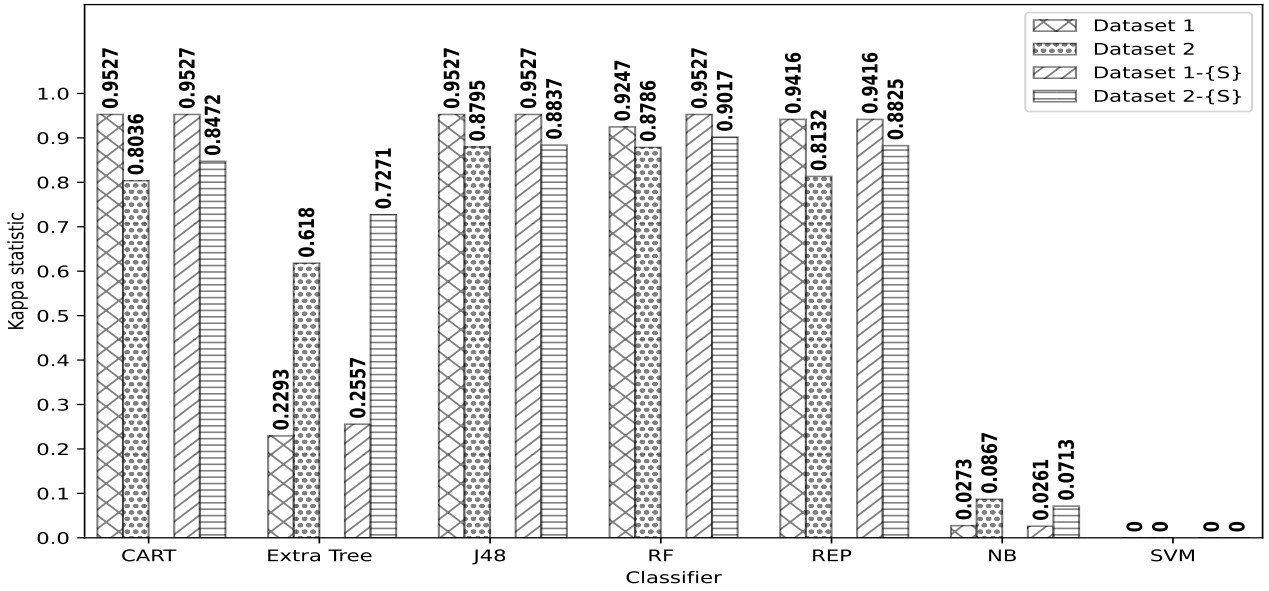


Figure 4.4: Kappa statistic achieved by ensemble filter-based approach with Split dataset

feature 'S' for dataset-1 and dataset-2 are shown in Table 4.14 and Table 4.15, respectively. After removing the feature 'S,' for dataset-1 CART, ET, J48, RF, REP, NB, and SVM classifier achieved an Accuracy 99.96%, 99.32%, 99.96%, 99.96%, 99.95%, 76.27%, and 99.56%, respectively. For dataset-2, CART, ET, J48, RF, REP, NB, and SVM classifier achieved an Accuracy of 99.85%, 99.73%, 99.88%, 99.90%, 99.81%, 90.14%, and 99.47%, respectively. The Kappa static achieved for dataset-1 and dataset-2 is shown in Figure 4.4.

4.1.5 Summary

The complexity of classifiers and lab chemical analysis costs can be reduced using a limited number of soil parameters. Three different feature selection approaches (InfoG, GainR, and ReliefF) were used to develop an ensemble filter-based feature selection-based classifier. The proposed approach eliminates the least relevant feature. The performance of the proposed soil fertility classifier was assessed using two datasets. The performance of tree-based classifiers such as CART, ET, J48, RF, and REP are compared with NB and SVM classifiers. The RF classifier achieved the highest Accuracy of 99.96% for dataset-1 and 99.90% for dataset-2 by excluding the soil parameter 'S' and using the remaining ten soil chemical parameters. Kappa statistics was improved by removing the least relevant feature, 'S,' using 10-fold cross-validation and split datasets. The RF classifier performed better than other classifiers when feature 'S' was eliminated from both datasets. Based on the classification results, a precise amount of fertilizers is recommended.

4.2 Proposed 2D CNN-based Soil Fertility Classification with Fertilizer Prescription

Despite widespread deep learning-based classifiers, few studies have employed MLP and ANN for soil fertility classification. CNN outperformed in various applications, such as sentiment analysis (Alatrash et al., 2022) and leaf disease (Zhang et al., 2019), and more. Thus, this research work uses CNN for soil fertility classification.

In the proposed method Dakshina Kannada district dataset (dataset-1) consisting of 11 soil chemical parameters was used (discussed in Section 4.1.1). Figure 4.5 shows the proposed CNN-based soil fertility classifier. The dataset consists of laboratory mea-

surements of eleven soil parameters, including *EC, pH, OC, K, P, S, B, Cu, Fe, Mn*, and *Zn*, collected from farmlands in Dakshina Kannada district, Karnataka (State), India. After preprocessing, the soil health dataset is partitioned into 75% training and 25% testing data. A SMOTE oversampling approach was implemented to mitigate class bias for resampling the training data. The instances in class HIGH and MEDIUM are resampled to 0.5% of the instances in class LOW. The resampled training data is used to train the proposed CNN-based classifier.

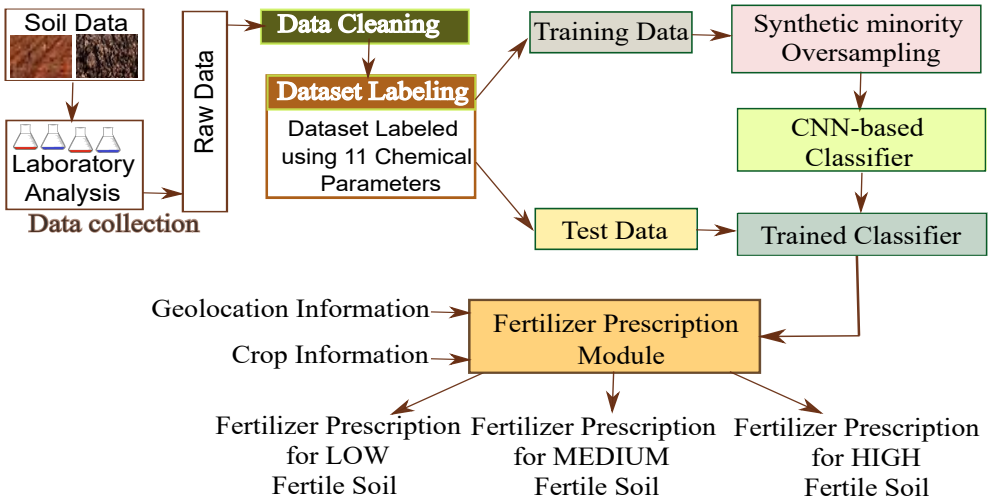


Figure 4.5: Steps involved in 2D-CNN-based soil fertility classification approach

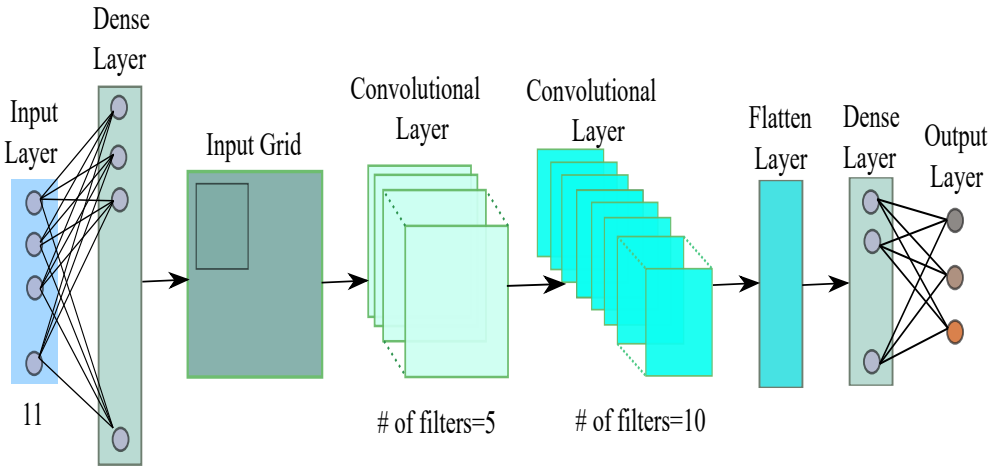


Figure 4.6: Proposed 2D-CNN-based soil fertility classifier

The proposed soil fertility classifier, shown in Figure 4.6 (adapted from (Ilango et al., 2022)), consists of an input layer, two dense layers, two convolutional layers, a flatten layer, and an output layer. The dense layer is connected to every neuron in

the input layer. The neurons in the dense layer perform matrix-vector multiplication of output received from the input layer and produce an output vector. The output vector is transformed into a matrix through reshaping. In this research, the number of neurons in the dense layer varied from 121 to 169 to produce an input grid of size 11×11, 12×12, and 13×13. The convolutional layer utilizes Eq (4.2) (Xie et al., 2023) to perform convolutions on the input matrix and extract feature maps.

$$y_{pq} = f(\sum_{m=1}^{z}\sum_{n=1}^{z} w_{m,n}x_{p+m,q+n} + b) \tag{4.2}$$

where y_{pq} is the output of node in the p^{th} row and q^{th} column of the output feature map; $x_{p+m,q+n}$ is the value in the p^{th} row and q^{th} column in the input matrix; $w_{m,n}$ is the weight value of the m^{th} row and n^{th} column in the convolution kernel; z is the size of the convolution kernel, and b is the bias term of the convolution kernel. The convolutional layers use the ReLU activation function.

The proposed approach involves two convolution layers, with 5 and 10 filters, respectively, a stride of [1,1], and padding. The flatten layer transforms the output of convolutional layer-2 into 1D data. The output of the flatten layer is fed into a dense layer that contains 20 neurons, which is further transmitted to an output layer comprising three neurons. The output layer utilizes the Sigmoid activation function, sparse categorical cross-entropy as the loss function, and Adam optimizer. It randomly chooses hidden nodes, calculates the output weights, and classifies the vector into LOW, or MEDIUM, or HIGH. The summary for the proposed approach with kernel size 3×3 and input grid size 11×11 is given in Table 4.16.

Table 4.16: Summary of the layers used in proposed approach

Layer	**Output Shape**	**Param #**
Input	[(None,11)]	0
Dense	(None, 121)	1452
Reshape	(None, 11, 11, 1)	0
Conv2D	(None, 11, 11, 5)	50
Conv2D	(None, 11, 11, 10)	460
Flatten	(None, 1210)	0
Dense	(None, 20)	24220
Dense	(None, 3)	63

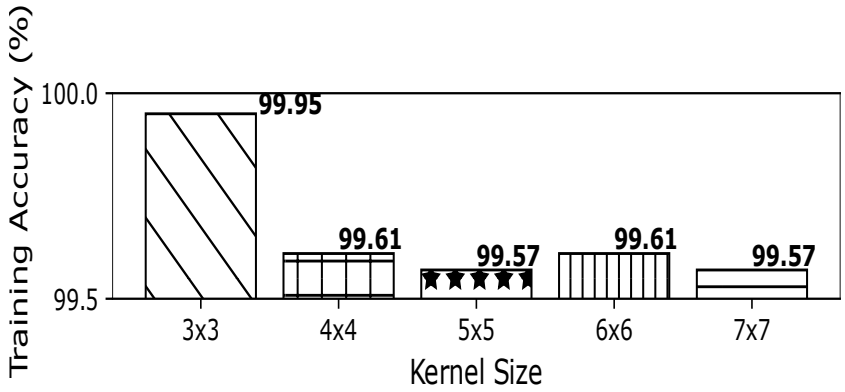


Figure 4.7: Training accuracy of 2D-CNN-based classifier without oversampling for 11×11 input grid

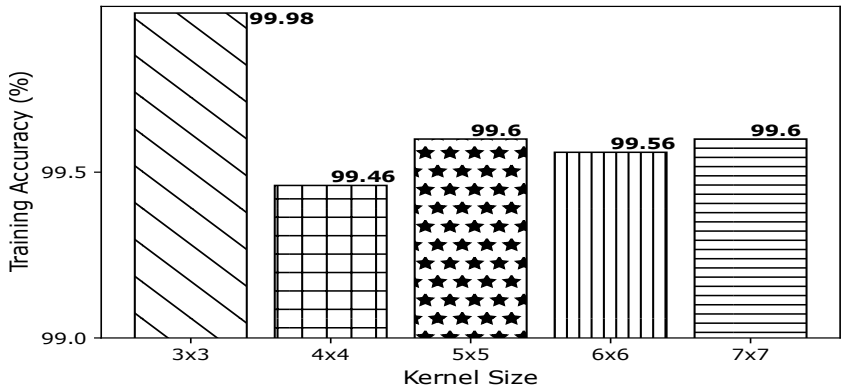


Figure 4.8: Training accuracy of 2D-CNN-based classifier without oversampling for 12×12 input grid

Based on the classification results, the proposed model prescribes fertilizers for crops grown in Dakshina Kannada, such as rice, black pepper, and cucumber.

4.2.1 Experimental Setup and Results

The experiments were carried out on the Google Collab platform using Python. The dataset is split into 75% for training and 25% for testing. The proposed soil fertility classifier was built using 500 epochs. The experiments involved varying the input grid from 11×11 to 13×13, using kernel sizes of 3×3, 4×4, 5×5, 6×6, and 7×7. Initially, the experiments were performed without the use of oversampling. The training accuracy obtained for different kernel sizes is depicted in Figure 4.7, Figure 4.8, and Figure 4.9 for an input grid size of 11×11, 12×12, and 13×13, respectively. The proposed approach obtained the highest training accuracy of 99.95% for kernel size 3×3 using an input grid size of 11×11. The performance of the proposed approach is compared for different kernel sizes using the dataset without oversampling. The results are

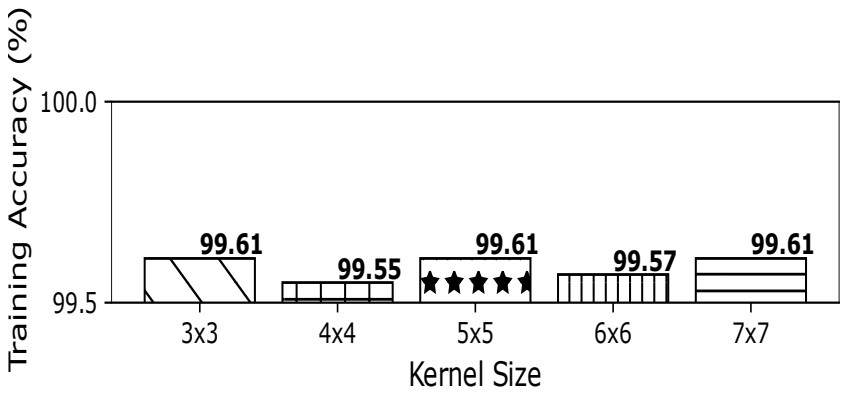


Figure 4.9: Training accuracy of 2D-CNN-based classifier without oversampling for 13×13 input grid

Table 4.17: Performance of proposed CNN-based Soil fertility classifier with 11×11 input grid

Kernel Size	Accuracy	Precision	Recall	F1-Score	Kappa Statistics
3x3	97.24%	0.987	0.974	0.980	0.0938
4x4	95.53%	0.986	0.960	0.973	0.0130
5x5	95.32%	0.986	0.962	0.974	0.0118
6x6	95.88%	0.986	0.957	0.971	0.0159
7x7	94.72%	0.986	0.916	0.950	0.0094

Table 4.18: Performance of proposed CNN-based Soil fertility classifier with 12×12 input grid

Kernel Size	Accuracy	Precision	Recall	F1-Score	Kappa Statistics
3x3	96.51%	0.987	0.965	0.976	0.0776
4x4	91.66%	0.987	0.917	0.950	0.015
5x5	96.70%	0.986	0.968	0.977	0.0162
6x6	95.81%	0.986	0.958	0.972	0.0098
7x7	95.18%	0.986	0.920	0.952	-0.0011

shown in Table 4.17, Table 4.18, and Table 4.19 for input grid sizes 11×11, 12×12, and 13×13, respectively. The proposed approach achieved the highest test Accuracy, Precision, Recall, F1-Score, and Kappa statistics of 97.24%, 0.987, 0.974, 0.980, 0.0938, respectively, for kernel size 3×3 using an input grid size of 11×11. It is observed that using the dataset without oversampling, most of the instances of the test dataset were misclassified as LOW fertile.

Table 4.19: Performance of proposed CNN-based soil fertility classifier with 13×13 input grid

Kernel Size	Accuracy	Precision	Recall	F1-Score	Kappa Statistics
3x3	95.85%	0.986	0.959	0.972	0.0196
4x4	96.30%	0.986	0.967	0.977	0.0261
5x5	94.74%	0.986	0.949	0.967	0.0095
6x6	95.50%	0.985	0.867	0.916	0.0004
7x7	96.48%	0.986	0.965	0.975	0.0138

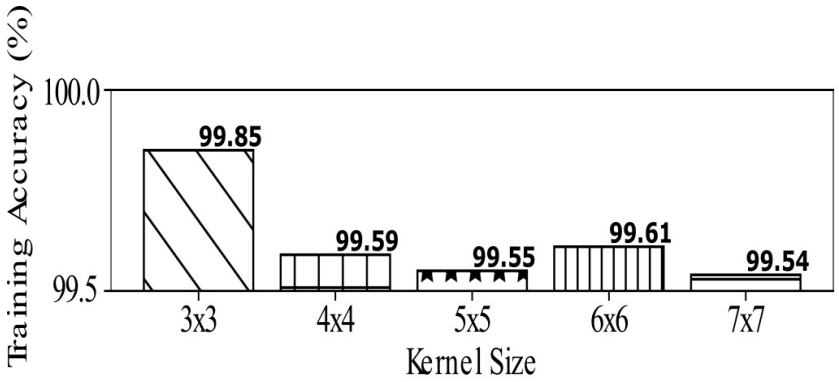


Figure 4.10: Training accuracy of 2D-CNN-based classifier for the oversampled dataset with 11×11 input grid

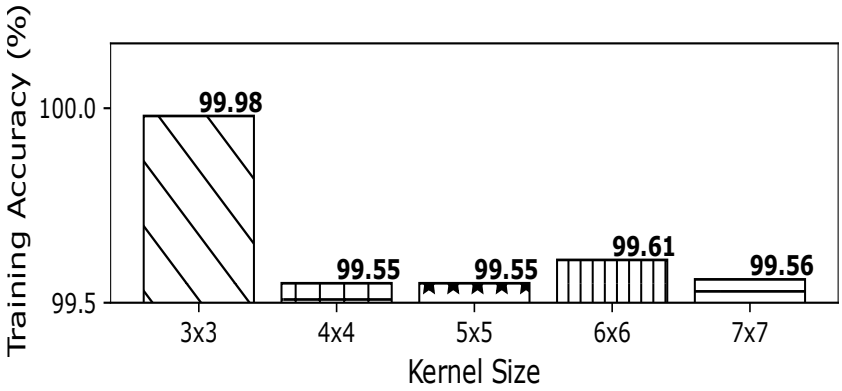


Figure 4.11: Training accuracy of 2D-CNN-based classifier for the oversampled dataset with 12×12 input grid

The dataset was oversampled using SMOTE to enhance the classification performance. The number of instances in MEDIUM and HIGH fertile classes was increased to 138 instances each. The training achieved for the oversampled dataset is depicted in Figure 4.10, Figure 4.11, and Figure 4.12 for an input grid size of 11×11, 12×12 and 13×13, respectively. When using a kernel size of 3×3 and an input grid size of 12×12, the highest test results were achieved, with Accuracy, Precision, Recall, F1-Score, and kappa statistics of 97.52%, 0.988, 0.978, 0.983, 0.1397, respectively.

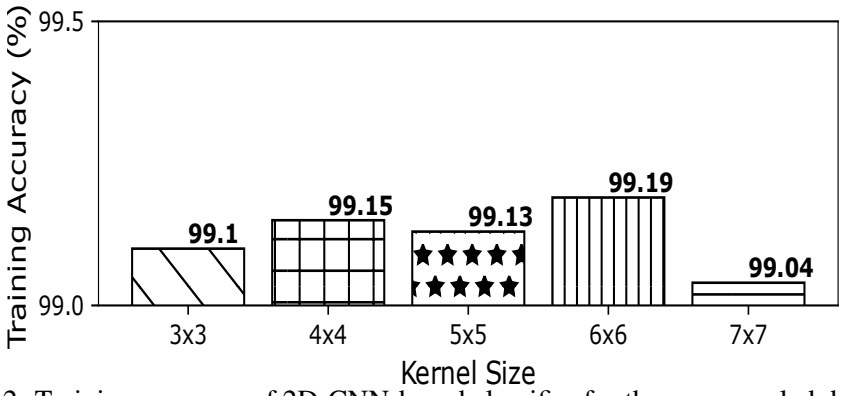


Figure 4.12: Training accuracy of 2D-CNN-based classifier for the oversampled dataset with 13×13 input grid

Table 4.20: Performance of proposed CNN-based soil fertility classifier using SMOTE oversampling with 11×11 input grid

Kernel Size	Accuracy	Precision	Recall	F1-Score	Kappa Statistics
3x3	97.03%	0.988	0.900	0.942	0.0361
4x4	95.91%	0.986	0.959	0.972	0.0047
5x5	94.22%	0.986	0.940	0.962	0.0131
6x6	96.48%	0.986	0.965	0.975	0.020
7x7	96.04%	0.985	0.817	0.893	-0.0045

Table 4.21: Performance of proposed CNN-based soil fertility classifier using SMOTE oversampling with 12×12 input Grid

Kernel Size	Accuracy	Precision	Recall	F1-Score	Kappa Statistics
3x3	97.52%	0.988	0.978	0.983	0.1397
4x4	93.96%	0.986	0.940	0.962	0.0195
5x5	95.73%	0.986	0.958	0.972	0.0190
6x6	97.32%	0.986	0.970	0.978	0.0359
7x7	95.67%	0.984	0.593	0.740	-0.0031

Table 4.22: Performance of proposed CNN-based soil fertility classifier using SMOTE oversampling with 13×13 input grid

Kernel Size	Accuracy	Precision	Recall	F1-Score	Kappa Statistics
3x3	94.99%	0.986	0.950	0.968	0.0167
4x4	94.59%	0.986	0.946	0.966	0.0124
5x5	94.84%	0.986	0.943	0.964	0.0077
6x6	94.42%	0.986	0.889	0.939	0.0061
7x7	91.99%	0.985	0.517	0.678	-0.0014

4.2.2 Summary

The proposed 2D-CNN-based classifier achieved the training and test Accuracy of 99.95% and 97.24%, respectively, for kernel size 3×3 and an input grid size 11×11. Furthermore, the performance of the classifier was improved using SMOTE oversampling. The proposed approach using an oversampled training dataset achieved the highest training and test Accuracy of 99.98% and 97.52%, respectively, for kernel size 3×3 and an input grid size 12×12. The classification results are used to recommend suitable fertilizers for specific crops.

4.3 Proposed 1D CNN-based Soil Fertility Classification and Fertilizer Prescription

CNNs require specialized hardware due to their complex computations and are impractical, with limited data (Jain et al., 2019; Kiranyaz et al., 2019). 1D-CNN utilizes cost-effective, simplified hardware for 1D convolutions (scalar additions and multiplications). The 1D-CNN is efficient, with fewer hidden layers and scarce labeled data (Kiranyaz et al., 2019). In this context, this research applied 1D-CNN to classify soil fertility. In the proposed method Dakshina Kannada district dataset (dataset-1) consisting of 11 soil chemical parameters was used (discussed in Section 4.1.1). Figure 4.13 presents the steps involved in 1D-CNN-based soil fertility classification. The research utilized soil-health data (Soil-health, 2021) collected from the Dakshina Kannada district of Karnataka (State), India. The experiment was performed using split dataset. Further, the dataset was converted into a compatible format using the MinMax scaling function (Jain et al., 2005; Lotfi & Pirnia, 2022)). Each feature 'x_i' value in this work is transformed by the MinMax scaler to a range of [0, 1] using the Eq. (4.3) (Somu et al., 2020).

$$x_i^{'} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{4.3}$$

The minimum and maximum values in the column are denoted as x_{min} and x_{max}, respectively. The dataset is split into 75% for training and 25% for testing without reshuffling. SMOTE oversampling is applied to the training data, and the trained classifier's effectiveness is assessed using test data. The fertilizer prescription module uses classification

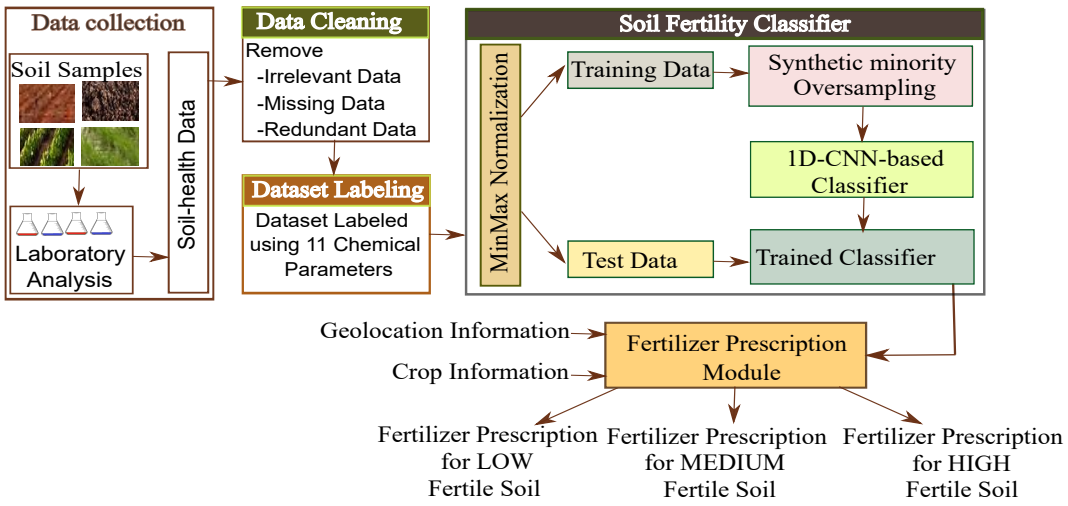


Figure 4.13: Steps in 1D-CNN-based soil fertility classification

results to determine the suitable fertilizers for specific crops.

The proposed soil fertility classification approach, depicted in Figure 4.14, consists of layers: input, convolutional, dense, 1D MaxPooling, flattening, and output. The

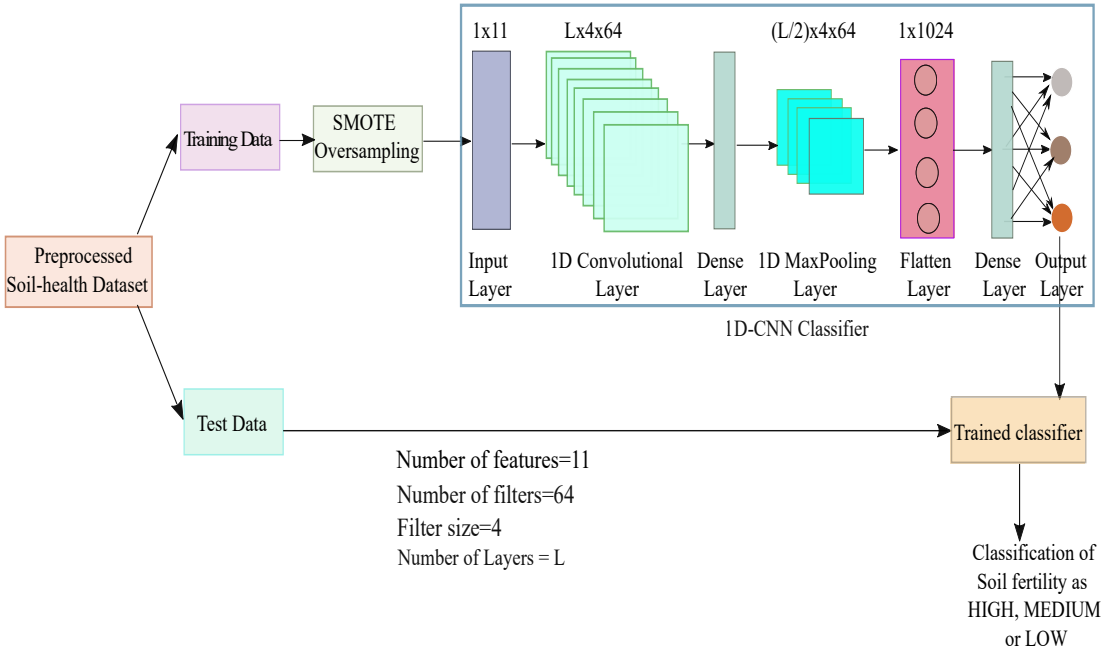


Figure 4.14: Proposed 1D-CNN-based soil fertility classifier

1D-CNN classifier receives the training data as input. The input layer utilizes 11 neurons, with randomly assigned weights for each connection. The method proposed uses a convolution layer with 64 filters, a filter size of 4, a stride of 1, and zero padding. The kernels slide over elements of a 1D input during convolution. Each 1D convolutional layer creates kernels of dimension $P \times Q$, where P is the temporal window (single

Table 4.23: Layers used in proposed 1D-CNN soil fertility classifier

Layer (type)	Output Shape	Param #
1D Convolutional	(None, 8, 64)	320
Dense	(None, 8, 64)	4160
1D MaxPooling	(None, 4, 64)	0
Flatten	(None, 256)	0
Dense	(None, 3)	771

spatial) covered by the filter and Q is the number of filters. The convolutional layer performs convolutions and produces the output 'y_i' of the ith node in the feature map using the Eq. (4.4).

$$y_i = f(\sum_{m=1}^{M}\sum_{q=1}^{Q} w_{mq}x_{i+m,i+q} + b) \tag{4.4}$$

where 'x' represents the input overlapping to the filter, 'w' represents the random weight of the convolutional filter connections, 'b' represents the bias, and 'f' represents the activation function.

After each convolution, the filter moves by one step. The convolutional layer utilizes the ReLU activation function. A dense layer with 64 neurons is connected to the 1D convolutional layer. Each neuron in the dense layer is connected to every neuron in the previous layer. Matrix-vector multiplication is performed by the dense layer neurons using the updated output from the preceding layer. Backpropagation is employed to calculate the gradient of the loss function with respect to the network weights. To reduce feature map size, the proposed approach employs L/2 max-pooling layers. By taking the maximum feature response within a neighborhood, the max-pooling layer reduces the output matrix to half of the input matrix. The flatten layer with four neurons transforms the previous layer's output into 1D data. The classification of the input is performed by the output layer using the sigmoid activation function. The output weights are determined by randomly selecting hidden nodes (Huang et al., 2006), and a vector of size is reduced into three classes. The 1D-CNN employs an Adam optimizer and a sparse categorical cross-entropy loss function. The layers used in the proposed 1D-CNN-based classifier are shown in Table 4.23.

The proposed approach recommends fertilizers for paddy, black gram, green gram, and arecanut based on the classification results.

4.3.1 Experimental Setup and Results

The proposed soil fertility classifier is implemented using Python programming on Google Collab. The soil fertility classification has been developed over 500 epochs. The proposed approach is compared with the ELM and MLP. The experiments were conducted: 1) using the raw dataset (i.e., without normalization and without oversampling), 2) with normalization and without oversampling, 3) without normalization and with oversampling, 4) with normalization and with oversampling.

The training and validation accuracy obtained using the raw dataset is as depicted in Figure 4.15. The performance comparision between the proposed approach and other

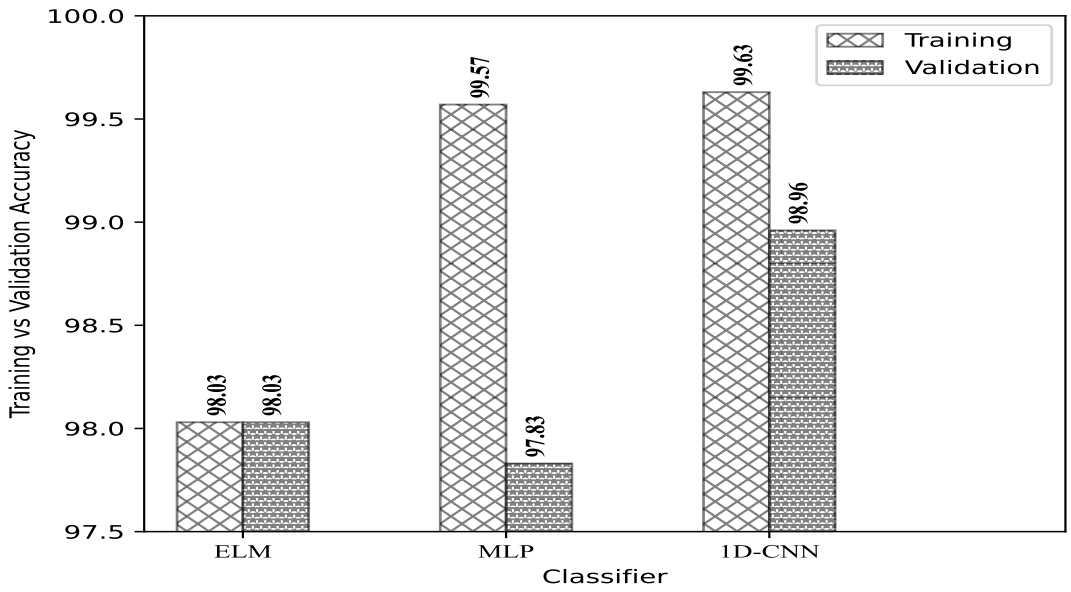


Figure 4.15: Comparision of training and validation accuracy using raw dataset

classifiers using a raw dataset is shown in Table 4.24. The proposed method obtained an Accuracy of 98.96%, with Recall, Precision, and F1-Score of 0.990. The Kappa statistics for the proposed approach and the MLP were extremely low. ELM obtained negative Kappa statistics, which indicates that classification is not appropriate. MinMax normalization was applied to the dataset to improve the classifier's performance. The training and validation accuracy of the proposed approach and the other two classifiers with normalization and without oversampling is depicted in Figure 4.16.

Table 4.25 compares the performances of the classifiers with normalization and without oversampling. The proposed approach outperformed ELM and MLP classi-

Table 4.24: Performance of the proposed approach, ELM and MLP classifiers using raw dataset

Classifier	Accuracy	Recall	Precision	F1-Score	Kappa Statistics
ELM	98.03%	0.980	0.985	0.983	-0.0070
MLP	97.84%	0.978	0.986	0.982	0.0014
Proposed approach	98.96%	0.990	0.990	0.990	0.0319

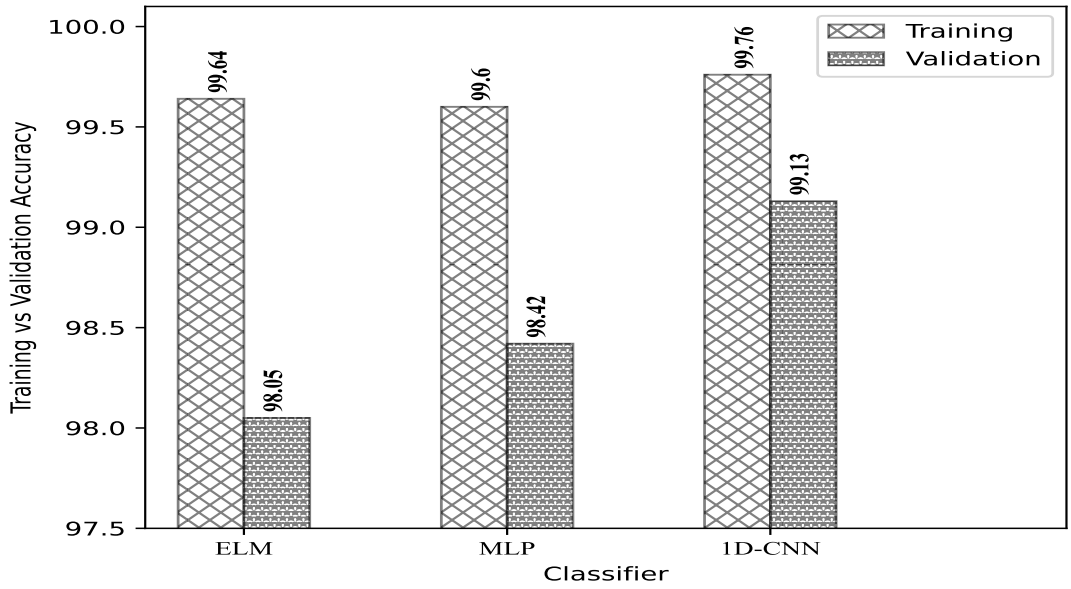


Figure 4.16: Comparision of training and validation accuracy with normalization and without oversampling

Table 4.25: Performance of the proposed approach, ELM and MLP classifiers with normalization and without oversampling

Classifier	Accuracy	Recall	Precision	F1-Score	Kappa Statistics
ELM	98.05%	0.981	0.987	0.984	0.0598
MLP	98.42%	0.984	0.986	0.985	0.0449
Proposed approach	99.13%	0.991	0.988	0.990	0.1270

fiers with the highest Accuracy of 99.13%, Recall of 0.991, Precision of 0.998, and F1-Score of 0.990. SMOTE oversampling is used to balance the dataset and reduce bias. Figure 4.17 shows the training and validation accuracy of the proposed approach and the other two classifiers without normalization and with oversampling.

Table 4.26 displays the performance comparision of classifiers without normalization and with oversampling. The proposed approach outperformed ELM and MLP classifiers with the highest Accuracy of 96.69%, Recall of 0.967, Precision of 0.992, and F1-Score of 0.978. The proposed approach obtained the highest Kappa statistics of 0.2358.

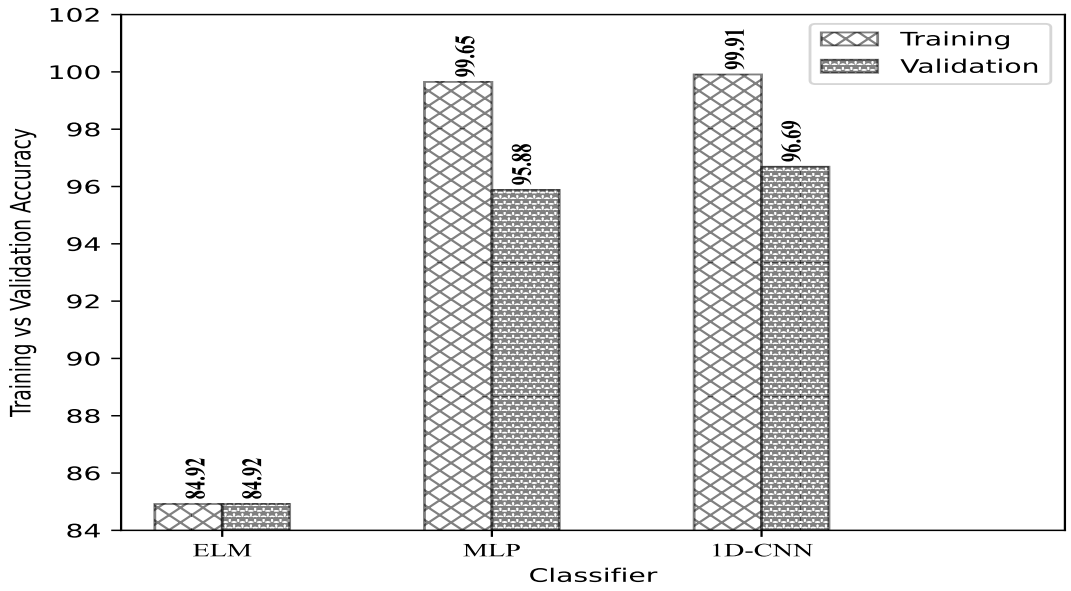


Figure 4.17: Comparision of training and validation accuracy without normalization and with oversampling

Table 4.26: Performance of proposed approach, ELM and MLP classifiers without normalization and with oversampling

Classifier	Accuracy	Recall	Precision	F1-Score	Kappa Statistics
ELM	84.92%	0.849	0.990	0.912	0.0402
MLP	95.88%	0.959	0.989	0.973	0.1261
Proposed approach	96.69%	0.967	0.992	0.978	0.2358

Furthermore, an experiment is performed with normalization and oversampling to enhance the efficiency of the classifiers. Figure 4.18 illustrates the training and validation accuracy of different classifiers using normalization and oversampling. The proposed 1D-CNN-based soil fertility classifier achieved the highest training accuracy of 99.91% and validation accuracy of 97.75%.

Table 4.27 shows the performance of the classifiers, with normalization and oversampling. The proposed approach outperformed ELM and MLP classifiers with an Accuracy of 97.90%, Recall of 0.979, Precision of 0.992, and F1-Score of 0.984. The Kappa statistics achieved for the classifiers are better than compared to the Kappa statistics obtained without using data normalization or oversampling. Further, the proposed 1D-CNN-based classification approach achieved better Kappa statistics of 0.2358.

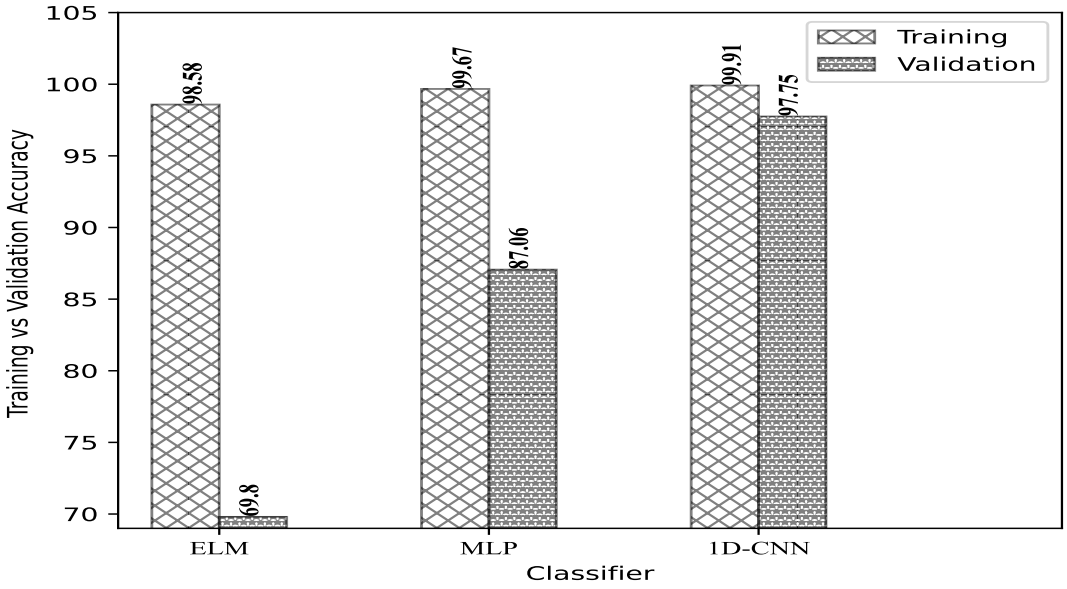


Figure 4.18: Comparison of training and validation accuracy with normalization and oversampling

Table 4.27: Performance of the proposed approach with normalization and with oversampling

Classifier	Accuracy	Recall	Precision	F1-Score	Kappa Statistics
ELM	69.80%	0.698	0.992	0.816	0.0264
MLP	87.06%	0.871	0.991	0.925	0.0668
Proposed approach	97.9%	0.979	0.992	0.984	0.2358

4.3.2 Summary

In this research, a 1D-CNN-based soil fertility classification is proposed. Experiments were performed to demonstrate the effectiveness of proposed approach with MinMax normalization and SMOTE oversampling. The proposed approach achieved the highest training and validation accuracies of 99.91% and 97.75% using the soil-health dataset. The proposed approach performed better than ELM and MLP with a classification Accuracy of 97.90%, Recall of 0.979, Precision of 0.992, F1-Score of 0.984, and kappa statistics of 0.2358. Additionally, the classification results are used to prescribe the fertilizers.

4.4 Proposed Finite Automata-based Soil Fertility Classification with Fertilizer Prescription

Numerous scientists have employed machine learning techniques to evaluate soil fertility. However, these approaches face significant challenges, including the need for

unbiased and massive datasets, more time and resource, and are susceptible to errors. Most ML-based classifiers exhibited low accuracy when applied to imbalanced datasets (Neyestani et al., 2021). Finite automata have found application in various contexts, including the design of lexical analyzer and test editor (Sunitha, 2013), malware detection (Ramesh & Menen, 2020), etc. The primary limitation of finite automata is their inability to handle infinite sets of alphabets. Symbolic Finite Automata (SFA) addresses this limitation by enabling the utilization of linear arithmetic predicates and functions over an alphabet. Consequently, SFAs extend the capabilities of finite automata to operate effectively over infinite alphabets (Van Noord & Gerdemann, 2001; Dalla Preda et al., 2015). The values of soil parameters belong to an infinite set of alphabets. In this study, SFA was employed to classify soil fertility and suggest fertilizers based on the deficiency of soil nutrients.

4.4.1 Datasets Used

In the proposed method Sentinel-2 dataset generated for Konaje village of Dakshina Kannada (District), Karnataka (State), India as discussed in Section 3.1. Additionally, the study incorporated laboratory-measured soil-health data (Soil-health, 2021) collected from farmlands of villages in Belgaum (District), Karnataka (State), India. The collected soil-health data consists of 5015 instances with attributes such as sample number, state name, district name, block name, village code, village name, longitude, latitude, and 12 soil chemical parameters. Initially, the Soil-health-1 dataset was generated by selecting attributes *EC, pH, OC*, and *N*, from soil-health data. The Soil-health-2 dataset consists of all 12 attributes of soil-health data, namely *pH, EC, OC, N, P, K*, *S*, *B*, *Cu*, *Fe, Mn*, and *Zn*. The WEKA open source tool (WEKA, 2021) was used to eliminate redundant and missing data. The preprocessed Soil-health-1 dataset consists of 4066 instances. After labeling, the dataset consists of 4,000 instances of LOW, 54 of MEDIUM, and 12 HIGH fertile soil. After preprocessing, the Soil-health-2 dataset consists of 4529 instances. The Soil-health-2 dataset was labeled using the levels of 12 soil parameters: *pH, EC, OC, N, K, P, S, B, Cu, Fe, Mn*, and *Zn*. Estimating the *pH, EC, OC*, and *N* levels was identical to the procedure for labeling the Sentinel-2 dataset. After labeling, the dataset consists of 4517 instances of LOW fertile and 12 of MED fertile.

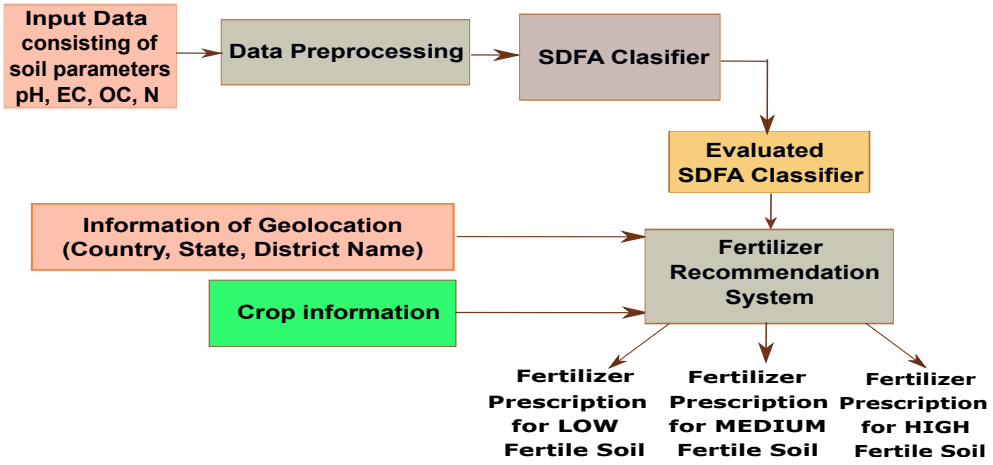


Figure 4.19: Proposed SDFA classification and fertilizer prescription approach

4.4.2 Proposed Fertilizer Recommendation System using SDFA

The sequence of steps in the proposed fertilizer recommendation system is depicted in Figure 4.19.

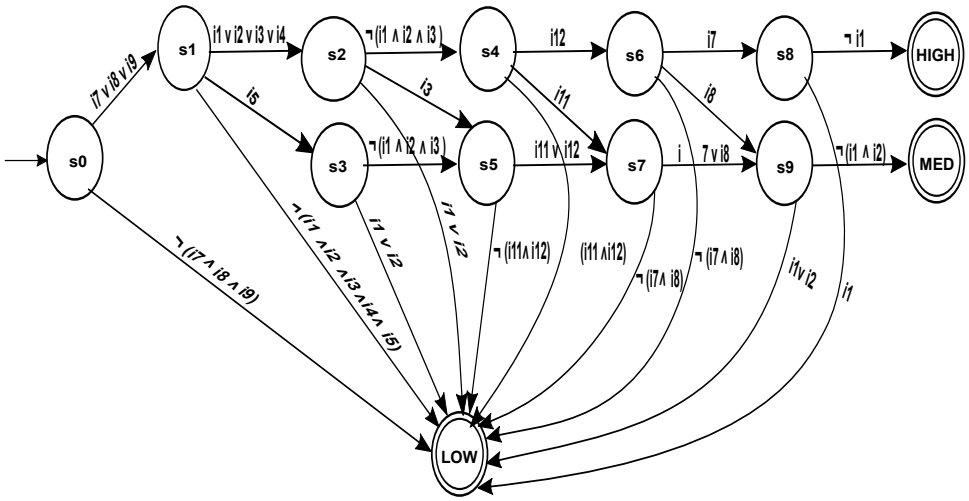


Figure 4.20: State transition diagram of proposed SDFA soil fertility classifier

SDFA is SFA that consists of states and transitions, with the transition edges labeled with predicates specified in terms of boolean algebra (Dalla Preda et al., 2015). Consider an infinite alphabet A = (D_A, ψ_A, $\wedge$,$\vee$,$\neg$), where D_A is set of infinite rational numbers, ψ_A is set of predicates which are closed under boolean operations $\wedge$, $\vee$, and $\neg$.

Let φ, and y are two predicates belonging to a set of predicates defined on alphabet A (i.e., ψ_A), then $\forall\varphi$, $y \in \psi_A$ $[\varphi \wedge y]=[\varphi \cap y]$, $[\varphi \vee y]=[\varphi \cup y]$, $[\neg\varphi]=[D_A - \varphi]$.

Table 4.28: Transition table of proposed approach

States	INPUTS											
	i1	**i2**	**i3**	**i4**	**i5**	**i6**	**i7**	**i8**	**i9**	**i10**	**i11**	**i12**
s0	LOW	LOW	LOW	LOW	LOW	LOW	s1	s1	s1	LOW	LOW	LOW
s1	s2	s2	s2	s2	s3	LOW	LOW	LOW	LOW	LOW	LOW	LOW
s2	LOW	LOW	s5	s4	s4	s4	s4	s4	s4	s4	s4	s4
s3	LOW	LOW	s5	s5	s5	s5	s5	s5	s5	s5	s5	s5
s4	LOW	LOW	LOW	LOW	LOW	LOW	LOW	LOW	LOW	LOW	s7	s6
s5	LOW	LOW	LOW	LOW	LOW	LOW	LOW	LOW	LOW	LOW	s7	s7
s6	LOW	LOW	LOW	LOW	LOW	LOW	s8	s9	LOW	LOW	LOW	LOW
s7	LOW	LOW	LOW	LOW	LOW	LOW	s9	s9	LOW	LOW	LOW	LOW
s8	LOW	HIGH	HIGH	HIGH	HIGH	HIGH	HIGH	HIGH	HIGH	HIGH	HIGH	HIGH
s9	LOW	LOW	MED	MED	MED	MED	MED	MED	MED	MED	MED	MED

SDFA model was designed using *pH, EC, OC*, and *N*. The transition diagram of the SDFA classifier is shown in Figure 4.20, and the transition table is shown in Table 4.28. The transitions are made in accordance with the levels of soil parameters. Moving to a final state within a sequence signifies the predicted soil fertility. The proposed SDFA classification approach uses D_A={i1, i2, i3, i4, i5, i6, i7, i8, i9, i10, i11, i12}, where ik indicates decimal input values such that i1=[0-0.2), i2=[0.2-0.5), i3=[0.5-0.75) i4=[0.75-1.6], i5=(1.6-2.5], i6=(2.5-6.5), i7=[6.5-7.5], i8=(7.5-8], i9=(8-8.5], i10=(8.5-280), i11=[280-560), i12=[560-9999].

SDFA is defined using five tuples as M=(A, Q, s0, F, δ),

where A is the alphabet such that, A = (D_A, φ_A,f, $\wedge$,$\vee$, $\neg$),

D_A is a set of infinite decimal values indicating the possible soil parameters.

φ_A is set of predicates. the predicates used in the proposed model are: φ_A={(i7$\vee$i8$\vee$i9),(i1$\vee$i2$\vee$i3$\vee$i4), i5,

$\neg$(i1$\wedge$i2$\wedge$i3$\wedge$i4$\wedge$i5), $\neg$(i1$\wedge$i2$\wedge$i3), (i1$\vee$i2), i3, i11, i12,

(i11$\vee$i12), $\neg$(i11$\wedge$i12), i7, i8, $\neg$(i7$\wedge$i8), i2, $\neg$i3, $\neg$i2}.

Q is a finite set of states, Q={s0, s1, s2, s3, s4, s5, s6, s7, s8, s9, LOW, MED, HIGH, REJECT}, where MED indicates MIDIUM.

s0 is an initial state.

F is a set of final states, F={LOW, MED, HIGH, REJECT}

δ: Qx $\varphi_A \rightarrow$Q is a finite set of transitions to a unique state.

The SDFA-based approach was designed by reading the pH value at the initial start

state, s0. With valid input, the transition has the potential to reach any final acceptance state: HIGH, MED, or LOW. Conversely, for invalid input values, the transition leads to the REJECT state. Table 4.29 describes the transitions to different states.

The proposed method is applied to classify Sentinel-2, Soil-health-1, and Soil-health-2 datasets. Using the classification results, the model suggests fertilizers for commonly cultivated crops in Dakshina Kannada and Belgaum, including paddy, black gram, green gram, and turmeric. The deficiency in S and micronutrients are determined based on pH value to recommend the fertilizers.

4.4.3 Experimental Results

The proposed SDFA-based soil fertility classifier was implemented using Python programming in the Google Collab platform.

4.4.3.1 Performance of Machine Learning-based Classifiers

The experiments involved ML-based classifiers such as NB, RF, J48, SVM, and KNN (k varying from 2 to 7). The performance of ML-based classifiers utilizing 10-fold cross-validation for the Sentinel-2 dataset, Soil-health-1 dataset, and Soil-health-2 dataset is presented in Table 4.30, Table 4.31, and Table 4.32, respectively. Figure 4.21, Figure 4.22, and Figure 4.23 display the TPR versus FPR for various datasets. A reliable classifier strives for a higher TPR and a lower FPR (Elfadel et al., 2019). In the case of the Sentinel-2 dataset, both the RF and J48 classifiers achieved an Accuracy of 97.87%, with Precision, Recall, and F1-Score reaching 0.979. Using Soil-health-1 dataset, J48 outperformed, with the highest Accuracy of 99.85%, with Precision, Recall, and F1-Score of 0.999. Similarly, using Soil-health-2 dataset, J48 attained the highest Accuracy of 99.85%, with Precision, Recall, and F1-Score values of 0.998.

4.4.3.2 Performance Evaluation of SDFA-based classifier

The confusion matrix for the proposed model employing the Sentinel-2 dataset is presented in Table 4.33, and the corresponding TP, FN, FP, and TN are shown in Table 4.34. The proposed classifier performed better with a TPR of 1.000 and an FPR of 0. The confusion matrix obtained for the Soil-heath-1 dataset is shown in Table 4.35, and the TP, FN, FP, and TN obtained are given in Table 4.36. SDFA performed better with a TPR of 1.000 and an FPR of 0. The LOW fertility class of Sentinel-2 obtained TP=293,

Table 4.29: Description of different transitions involved in the proposed approach

States	Scenario
s0–>s1	Read pH value is highly fertile.
s0–>LOW	Read pH value is low fertile level, indicating soil is LOW fertile.
s0–>s1–>s2	pH level of the soil is high, and the read EC value is at the high fertile level.
s0–>s1–>s3	pH level of the soil is high and read EC value is at medium fertile level.
s0–>s1–>LOW	pH level of the soil is high and read EC value is at low fertile level, indicates the soil is LOW fertile.
s0–>s1–>s2–>s4	pH, and EC level are high and read OC value is at high fertile level.
s0–>s1–>s2–>s5	pH, and EC level are high and read OC value is at medium fertile level.
s0–>s1–>s2–>LOW	pH, and EC level of the soil are high and read OC value is at low fertile level, indicates the soil is LOW fertile.
s0–>s1–>s3–>s5	pH level of the soil is high, EC value is at the medium fertile level, and read OC value level is medium.
s0–>s1–>s3–>LOW	pH level of the soil is high, EC value is at the medium fertile level, and read OC value level is low, indicates the soil is LOW fertile.
s0–>s1–>s2–>s4–>s6	pH, EC, OC level of the soil are high and read N value is at high fertile level.
s0–>s1–>s2–>s4–>s7	pH, EC, OC levels of the soil are high and read N value is at medium fertile level.
s0–>s1–>s2–>s4–>LOW	pH, EC, OC levels of the soil are high, and the read N value is at low fertile level, indicates the soil is LOW fertile.
s0–>s1–>s2–>s5–>s7	pH level of the soil is high, EC and OC value levels are at medium, and read N value is at the medium fertile level.
s0–>s1–>s2–>s5–>LOW	pH level of the soil is high, EC, and OC value level are at medium, and read N value is at low fertile level, indicates the soil is LOW fertile.
s0–>s1–>s2–>s4–>s6–>s8	pH, EC, OC, and N levels of the soil are high, and other soil parameter's level based on pH is high.
s0–>s1–>s2–>s4–>s6–>s9	pH, EC, OC, and N level of the soil are high and other soil parameter's level based on pH is medium.
s0–>s1–>s2–>s4–>s6–>LOW	pH, EC, OC, and N level of the soil are high, and any other soil parameter's level based on pH is LOW.
s0–>s1–>s2–>s4–>s7–>s9	pH, EC, OC level of the soil are high, N value is at medium fertile level, and other soil parameter's level based on pH is medium

Table 4.29: Description of different transitions involved in the proposed approach

States	Scenario
s0–>s1–>s2–>s5–>s7–>s9	pH level of the soil is high, EC, OC, and N value level are at medium, and other soil parameter's level based on pH is medium.
s0–>s1–>s2–>s5–>s7–>LOW	pH level of the soil is high, EC, OC, and N value level are at medium, and any other soil parameter's level based on pH is LOW.
s0–>s1–>s2–>s4–>s6–>s8–>HIGH	pH, EC, OC, N, and other soil parameter's level based on pH is high, and soil is not highly non-saline, indicates the soil is HIGH fertile.
s0–>s1–>s2–>s4–>s6–>s8–>LOW	pH, EC, OC, N, and other soil parameter's level based on pH is high, and soil is highly non-saline, indicates the soil is LOW fertile.
s0–>s1–>s2–>s4–>s6–>s9–>MED	pH, EC, OC, and N levels, of the soil are high, and other soil parameter's level based on pH is medium, and soil is slightly non-saline, indicates the soil is MEDIUM fertile.
s0–>s1–>s2–>s4–>s6–>s9–>LOW	pH, EC, OC, and N levels of the soil are high, and other soil parameter's level based on pH is medium, and soil is moderately or highly non-saline, indicates the soil is LOW fertile.
s0–>s1–>s2–>s4–>s7–>s9–>MED	pH, EC, OC levels of the soil are high, N value is at a medium fertile level, and other soil parameter's level based on pH is medium, and soil is slightly the non-saline indicates the soil is MEDIUM fertile.
s0–>s1–>s2–>s4–>s7–>s9–>LOW	pH, EC, OC level of the soil is high, N value is at a medium fertile level, and other soil parameter's level based on pH is medium, and soil is moderately or highly non-saline, indicates the soil is LOW fertile.
s0–>s1–>s2–>s5–>s7–>s9–>MED	pH level of the soil is high, EC, OC, and N levels are at medium, and other soil parameter's level based on pH is medium, the soil is slightly non-saline, the indicates the soil is MEDIUM fertile.
s0–>s1–>s2–>s5–>s7–>s9–>LOW	pH level of the soil is high, EC, OC, and N levels are at medium, and other soil parameter's level based on pH is medium, the soil is moderate or highly non-saline indicates the soil is LOW fertile.

FN=0, FP=0, and TN=36. The MED fertility obtained TP=25, FN=0, TN=304, and FP=0. The HIGH fertility achieved TP=11, FN=0, TN=318, and FP=0. The LOW fertility class of the Soil-health-1 dataset obtained TP=4000, FN=0, FP=0, TN=66. The results for the MED fertility class indicate TP=54, FN=0, TN=4012, and FP=0, while the HIGH fertility class achieved TP=12, FN=0, TN=4054, and FP=0. For the Soil-health-2 dataset, the LOW fertility class recorded TP=4451, FN=8, TN=4, and FP=66,

Table 4.30: Performance of ML-based classifiers using 10-fold cross-validation for Sentinel-2 dataset

Classifier	Accuracy	Per-class Precision			Per-class Recall			Per-class F1-Score			Precision	Recall	F1-Score
		LOW	MED	HIGH	LOW	MED	HIGH	LOW	MED	HIGH			
NB	93.62%	0.983	0.636	0.429	0.969	0.840	0.273	0.976	0.724	0.333	0.938	0.936	0.935
RF	97.87%	1.000	0.875	0.667	1.000	0.840	0.727	1.000	0.857	0.696	0.979	0.979	0.979
J48	97.87%	0.997	0.875	0.750	0.997	0.840	0.818	0.997	0.857	0.783	0.979	0.979	0.979
SVM	89.06%	0.898	0.600	0.000	0.990	0.120	0.000	0.942	0.200	0.000	0.845	0.891	0.854
KNN (k=2)	93.31%	0.960	0.652	0.333	0.993	0.600	0.091	0.977	0.625	0.143	0.916	0.933	0.922
KNN (k=3)	91.79%	0.963	0.583	0.333	0.973	0.560	0.273	0.968	0.571	0.300	0.913	0.918	0.915
KNN (k=4)	92.10%	0.957	0.652	0.167	0.980	0.600	0.091	0.968	0.625	0.118	0.907	0.921	0.913
KNN (k=5)	94.22%	0.970	0.714	0.500	0.983	0.800	0.182	0.976	0.755	0.267	0.935	0.942	0.936
KNN (k=6)	92.10%	0.953	0.667	0.000	0.980	0.640	0.000	0.966	0.653	0.000	0.900	0.921	0.910
KNN (k=7)	92.40%	0.960	0.607	0.000	0.980	0.680	0.000	0.970	0.642	0.000	0.901	0.924	0.912

Table 4.31: Performance of ML-based classifiers using 10-fold cross-validation for Soil-health-1 dataset

Classifier	Accuracy	Per-class Precision			Per-class Recall			Per-class F1-Score			Precision	Recall	F1-Score
		LOW	MED	HIGH	LOW	MED	HIGH	LOW	MED	HIGH			
NB	98.52%	0.985	1.000	1.000	1.000	0.037	0.333	0.993	0.071	0.500	0.985	0.985	0.979
RF	99.80%	0.999	0.945	1.000	0.999	0.963	0.750	0.999	0.954	0.857	0.998	0.998	0.998
J48	99.85%	0.999	0.930	1.000	0.999	0.981	0.917	0.999	0.955	0.957	0.999	0.999	0.999
SVM	99.43%	0.984	1.000	-	1.000	0.037	-	0.992	0.071	-	-	-	-
KNN (k=2)	98.94%	0.991	0.839	0.750	0.999	0.481	0.250	0.995	0.612	0.375	0.988	0.989	0.988
KNN (k=3)	98.72%	0.992	0.620	0.750	0.995	0.574	0.250	0.994	0.596	0.375	0.986	0.987	0.986
KNN (k=4)	98.70%	0.990	0.656	1.000	0.997	0.389	0.250	0.993	0.488	0.400	0.985	0.987	0.985
KNN (k=5)	98.84%	0.992	0.682	1.000	0.997	0.556	0.250	0.994	0.612	0.400	0.988	0.988	0.987
KNN (k=6)	98.89%	0.991	0.788	1.000	0.998	0.481	0.167	0.994	0.598	0.286	0.988	0.989	0.987
KNN (k=7)	98.99%	0.992	0.811	1.000	0.998	0.556	0.167	0.995	0.659	0.286	0.989	0.990	0.988

Table 4.32: Performance of ML-based classifiers using 10-fold cross-validation for Soil-health-2 dataset

Classifier	Accuracy	Per-class Precision		Per-class Recall		Per-class F1-Score		Precision	Recall	F1-Score
		LOW	MED	LOW	MED	LOW	MED			
NB	62.02%	0.998	0.003	0.621	0.500	0.765	0.007	0.995	0.620	0.763
RF	99.78%	0.998	0.667	1.000	0.333	0.999	0.444	0.997	0.998	0.997
J48	99.85%	0.999	0.857	1.000	0.500	0.999	0.632	0.998	0.998	0.998
SVM	99.74%	0.997	-	1.000	-	0.999	-	-	-	
KNN (k=2)	99.69%	0.997	0.000	1.000	0.000	0.998	0.000	0.995	0.997	0.996
KNN (k=3)	99.69%	0.997	0.000	1.000	0.000	0.998	0.000	0.995	0.997	0.996
KNN (k=4)	99.74%	0.997	-	1.000	-	1.000	-	-	-	-
KNN (k=5)	99.74%	0.997	-	1.000	-	0.999	-	-	-	-
KNN (k=6)	99.74%	0.997	-	1.000	-	0.999	-	-	-	-
KNN (k=7)	99.74%	0.997	-	1.000	-	0.999	-	-	-	-

Table 4.33: Confusion matrix obtained for Sentinel-2 dataset

		Reached State (Predicted Class)		
		LOW	**MED**	**HIGH**
Actual Class	**LOW**	293	0	0
	MED	0	25	0
	HIGH	0	0	11

Table 4.34: TP, FN, TN, and FP obtained for Sentinel-2 dataset

Class	TP	FN	TN	FP
LOW	293	0	36	0
MED	25	0	304	0
HIGH	11	0	318	0

Table 4.35: Confusion matrix obtained for Soil-health-1 dataset

		Reached State (Predicted Class)		
		LOW	**MED**	**HIGH**
Actual Class	**LOW**	4000	0	0
	MED	0	54	0
	HIGH	0	0	12

Table 4.36: TP, FN, TN, and FP obtained for Soil-health-1 dataset

Class	TP	FN	TN	FP
LOW	4000	0	66	0
MED	54	0	4012	0
HIGH	12	0	4054	0

and the MED fertility class obtained TP=4, FN=52, TN=4455, and FP=8. In the HIGH fertility class, TP=0, FN=14, TN=4515, and FP=0. Table 4.39 depicts the performance of proposed approach. When applied to the Sentinel-2 and Soil-health-1 datasets, the proposed model achieved a perfect Accuracy of 100%, with Precision, Recall, and F1-Score values of 1.000. When evaluating the Soil-health-2 dataset, the proposed approach achieved an Accuracy of 98.37%, with Precision, Recall, and F1-Score values of 0.984. Table 4.37 displays the confusion matrix for Soil-health-2, providing the TP, FN, FP, and TN values presented in Table 4.38. The SDFA performed with a TPR of 0.984 and an FPR of 0.008.

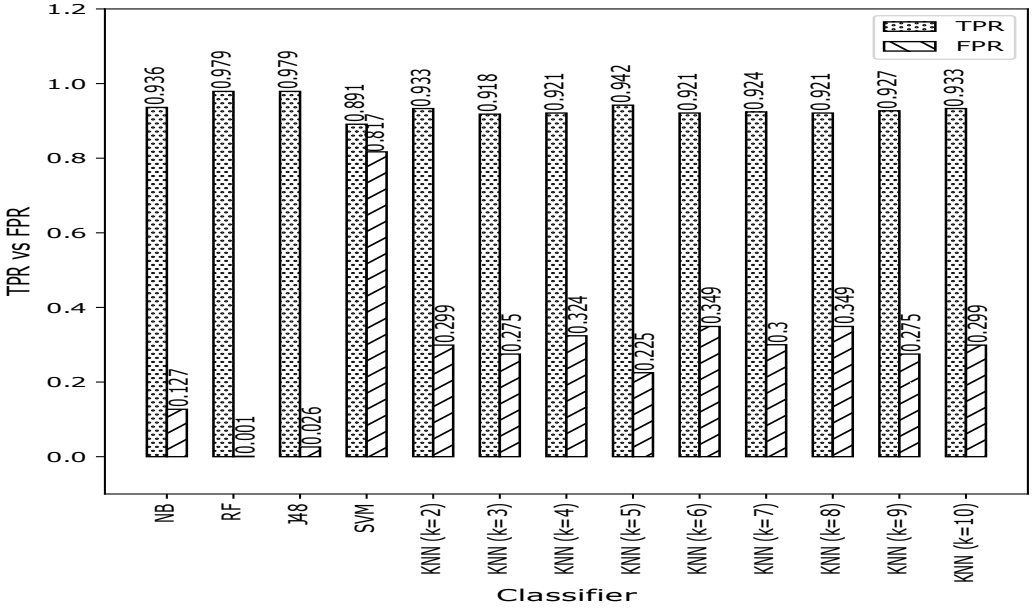


Figure 4.21: TPR versus FPR using Sentinel-2 dataset

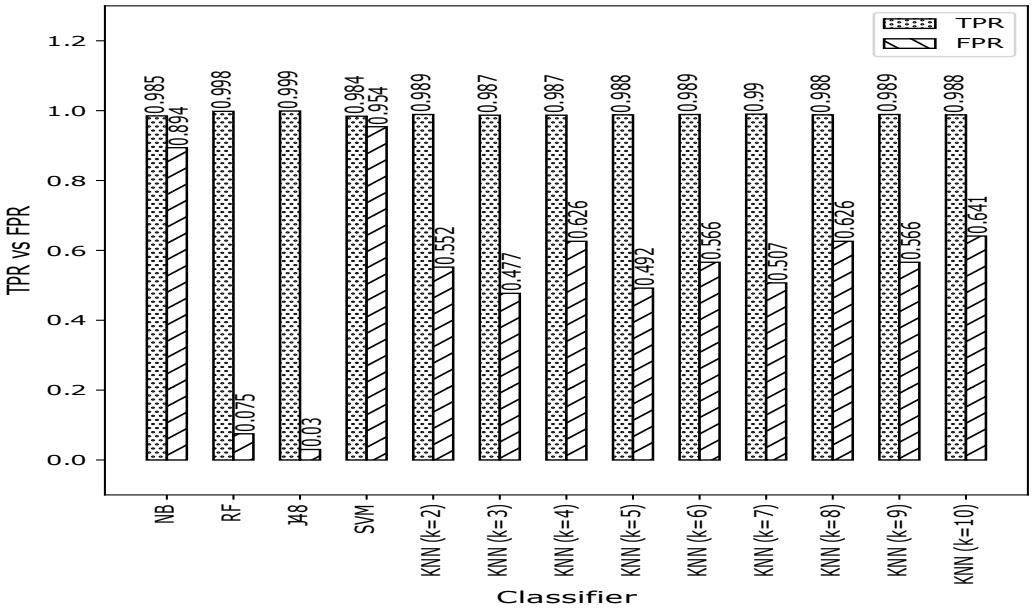


Figure 4.22: TPR versus FPR using Soil-health-1 dataset

Table 4.37: Confusion matrix obtained for Soil-health-2 dataset

		Reached State (Predicted Class)		
		LOW	**MED**	**HIGH**
Actual Class	**LOW**	4451	52	14
	MED	8	4	0
	HIGH	0	0	0

4.4.4 Summary

Accurate soil fertilization is essential for enhancing crop yields. In this research work, an SDFA-based soil fertility classification approach is proposed to dynamically classify

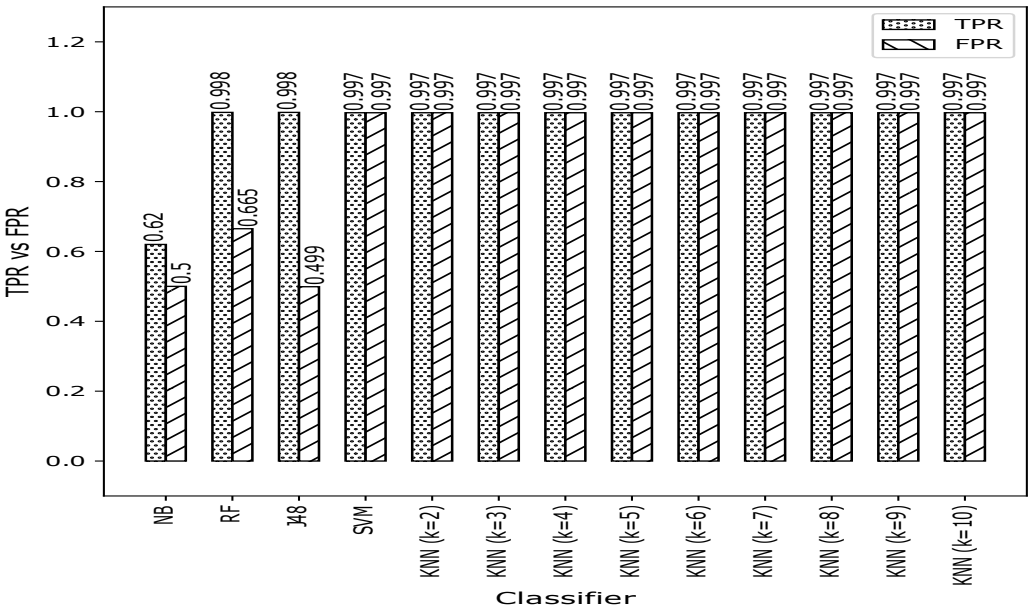


Figure 4.23: TPR versus FPR using Soil-health-2 dataset

Table 4.38: TP, FN, TN, and FP obtained for Soil-health-2 dataset

Class	TP	FN	TN	FP
LOW	4451	8	4	66
MED	4	52	4455	8
HIGH	0	14	4515	0

Table 4.39: Performance of proposed SDFA

Dataset Used	Accuracy	Precision	Recall	F1-Score
Sentinel-2	100%	1.000	1.000	1.000
Soil-health-1	100%	1.000	1.000	1.000
Soil-health-2	98.37%	0.984	0.984	0.984

soil data as LOW, or MEDIUM or HIGH fertile. The effectiveness of the approach is assessed using the Sentinel-2 dataset, which incorporates four soil parameters. It is observed that an Accuracy of 100% was achieved using Sentinel-2 dataset. The performance of the proposed method is further evaluated using laboratory-measured soil-health data, resulting in a 100% Accuracy for the Soil-health-1 dataset and 98.37% for the Soil-health-2 dataset. The performance of the proposed method is compared with ML-based classifiers. It is observed that the proposed method outperformed ML-based classifiers. The classification results are used to prescribe fertilizers for paddy, black gram, green gram, and turmeric.

CHAPTER 5

SOIL FERTILITY CLASSIFICATION USING REAL-TIME DATA

The revisit of the satellite to a specific location to gather data is infrequent. Revisit of Sentinel-2 is for every five days, whereas Landsat is for ten days. Frequent satellite revisits are necessary to increase the data samples and classification accuracy. The proximal soil sensors allow for the dynamic collection of soil data to determine the variation in soil fertility based on crop development and environmental factors. Thus, this research uses soil sensors, namely the Soil NPK sensor and EC-pH sensor, to collect real-time soil data.

5.1 Datasets Used

In this proposed method laboratory-measured Soil-health data (Soil-health, 2021) was collected from the farmlands of Belgaum (District), Karnataka (State), India and used to train the ML classifiers. The collected data had 5015 instances with 20 attributes such as sample number, state name, district name, block name, village code, village name, latitude, longitude, and 12 soil chemical parameters: *EC, pH, OC, N, P, K, S. B, Cu, Fe, Mn*, and *Zn*. *pH* indicates the soil's hydrogen ion concentration. *EC*, *OC*, macronutrients (N, P, K, and S) and micronutrients (B, Cu, Fe, Mn, and Zn) are measured in deciSeimen/cm (dS/cm), percentage (%), kilograms per hectare (kg/ha), and parts per million (ppm), respectively.

During data preprocessing, five soil parameters such as *EC, pH, OC, N, P, K* were selected from the dataset. The duplicates were removed using WEKA open source tool (WEKA, 2021). After removing duplicates, dataset consists of 4443 instances. The datasets were labelled by determining the fertility level of *EC, pH, OC, N, P, K* based on their value and fertility level of *S. B, Cu, Fe, Mn*, and *Zn* based on the value of *pH*. After labelling the dataset, it consists of 3269 instances of LOW fertile, 1017 of MEDIUM and 157 of HIGH.

5.2 Proposed Soil Fertility Classification using Real-time data

Figure 5.3 depicts the proposed soil fertility classification. The proposed approach uses

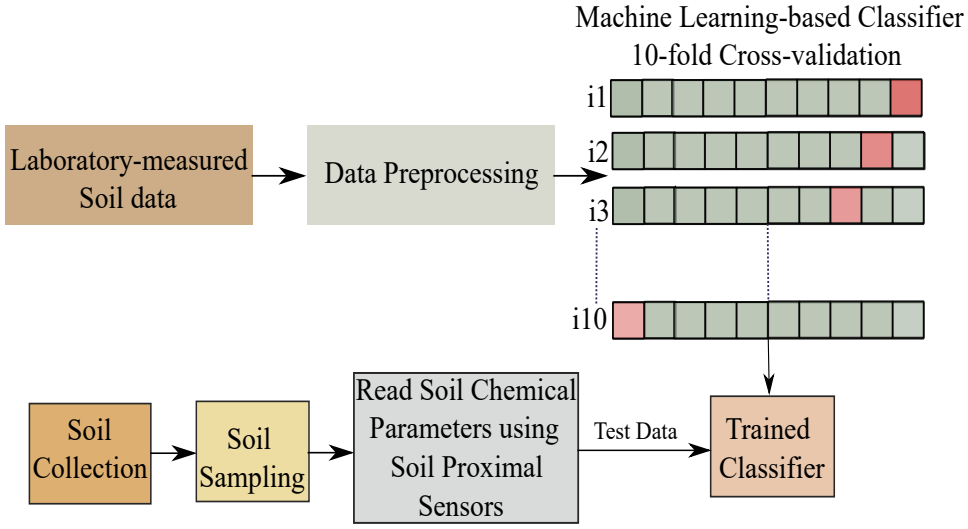


Figure 5.1: Proposed soil fertility classification using real-time test data

laboratory-measured soil data to train ML-based classifier. The ML-based classifier with 10-fold cross-validation test were employed by using the Soil-health data. The trained classifiers were utilized to test the real-time soil parameters acquired using soil proximal sensors.

5.3 Steps involved in Real-time Soil data Collection

Based on the expert's suggestion, the soil was collected after removing foreign materials such as roots and stones. The sub-surface soil samples were collected at a depth of 15cm from the bare land of a region in Dakshina Kannada (District) of Karnataka state, India, having a latitude of 13.009506 and a longitude of 74.788982. The collected 58 soils were placed in a transparent reusable plastics. The study uses soil sensors, namely the Soil EC-pH sensor and NPK sensor, to collect soil parameters. The EC-pH sensor was used to collect *EC* and *pH* values, and the NPK sensor was used to collect *N, P*, and *K* values. The circuit designed to collect the soil chemical parameters is depicted in Figure 5.2 and described in Table 5.1. Two serial communication ports are used, each with the baud rate=4800, parity=none, Timeout=10000 msec, slave id=1, and Modbus Remote Terminal Unit (RTU) protocol to read the data from the registers in sensors.

The *pH* and *EC* values were read from the register addresses 0x03 and 0x09 of

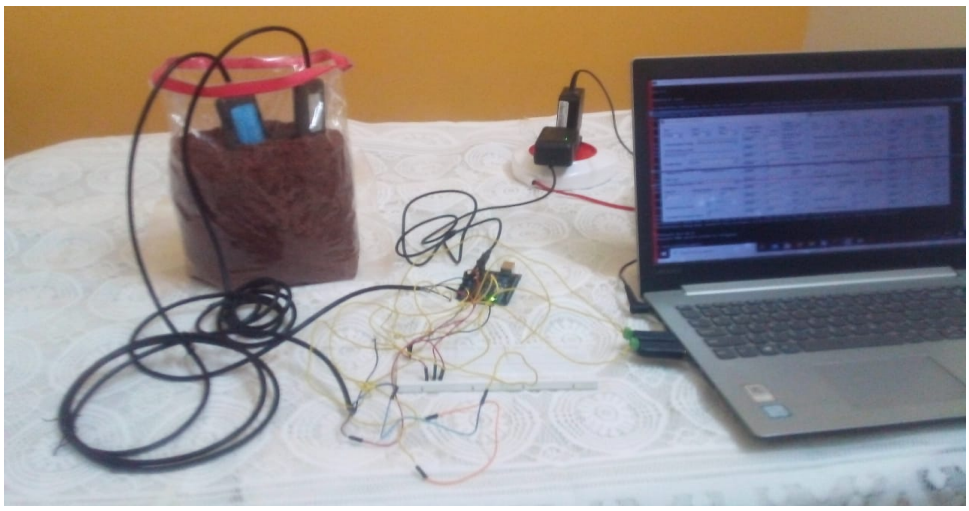


Figure 5.2: Circuit to collect soil chemical parameters

Table 5.1: Components used to collect the data

Sl. No.	**Components Used**	**Quantity**
1	NPK 3-in-1 fertility sensor (output type RS485)	1
2	EC-pH 2-in-1 fertility sensor (output type RS485)	1
3	USB CH340/341 to RS485 Converter	2
4	Arduino Uno	1
5	DC 12V 2A Power Supply Adapter	1

the EC-pH sensor, whereas *N, P*, and *K* values were read from the register addresses 0x09, 0x0D, and 0x0E, respectively. The data was collected using Python programming language with the MinimalModbus Python module. MinimalModbus is a module that uses Modbus protocol to talk to instruments from a computer (Jonas, 2023). The data read from sensors using Python programming are written to ThingSpeak as shown in Figure 5.3 using the Urllib package of Python programming. The steps involved in acquiring real-time soil parameters are:

1. Create ThingSpeak channel with 4 fields *pH, EC, N, P, K*.
2. Connect EC-pH sensor to port-1 using modbus-1
3. Connect NPK sensor to port-2 using modbus-2
4. Call minimalmodbus.instrument(port-1, slave_id=1)
5. Call minimalmodbus.instrument(port-2, slave_id=1)
6. Assign baud_rate, byte_size, parity, stop_bits, timeout for instrument1 and instrument2
7. Assign protocol for instrument, instrument2 as RTU

8. Clear instrument1_buffer, instrument2_buffer before transaction

9. For each soil sample do
   pH← instrument1.read_register(0x03)
   EC← instrument1.read_register(0x09)
   N← instrument2.read_register(0x09)
   P← instrument2.read_register(0x0B)
   K← instrument2.read_register(0x0D)
   Write pH, EC, N, P, K to ThingSpeak fields

The accuracy of data values is tested using two simulators: CAS Modbus scanner (CHIPKIN, 2023) and Generic Modbus/Jbus tester (Schneider, electric, 2023) as shown in Figure 5.5 and Figure 5.7, respectively.

5.3.1 Unit Conversion of Soil parameters

This study used the trained model developed using laboratory-measured soil chemical parameters. Separate testing was performed using laboratory-measured data and real-time soil data using the trained classifier developed using laboratory-measred Soil-health data. In Soil-health dataset, the *EC* was measured in dS/cm, whereas *N, P*, and *K* used kg/ha. The EC-pH sensor measures *EC* in terms of mS/cm= 0.01dS/0.01m= 1dS/m. Hence, the *EC* value after unit conversion remains the same. The real-time values of *N, P,* and *K* extracted using NPK sensor are measured in terms of mg/kg. To mg/kg can be converted into kg/ha by calculating the mass of the soil layer in terms of kg/ha using the Eq (5.1).

$$Mass\ of\ soil\ layer = volume\ of\ soil\ layer \times soil\ density \tag{5.1}$$

where the volume of the soil layer can be calculated using Eq. (5.2), and soil density is measured using Eq. (5.3).

$$Volume\ of\ soil\ layer\ (in\ m^3) = area\ (in\ m^2) \times depth\ (in\ m) \tag{5.2}$$

$$Soil\ Density(in\ kg/m^3) = \frac{Weight\ of\ soil\ (in\ kg)}{Volume\ of\ soil\ (in\ m^3)} \tag{5.3}$$

where area is 1 ha= 10,000m^2, soil depth is 15 cm= 0.15m.

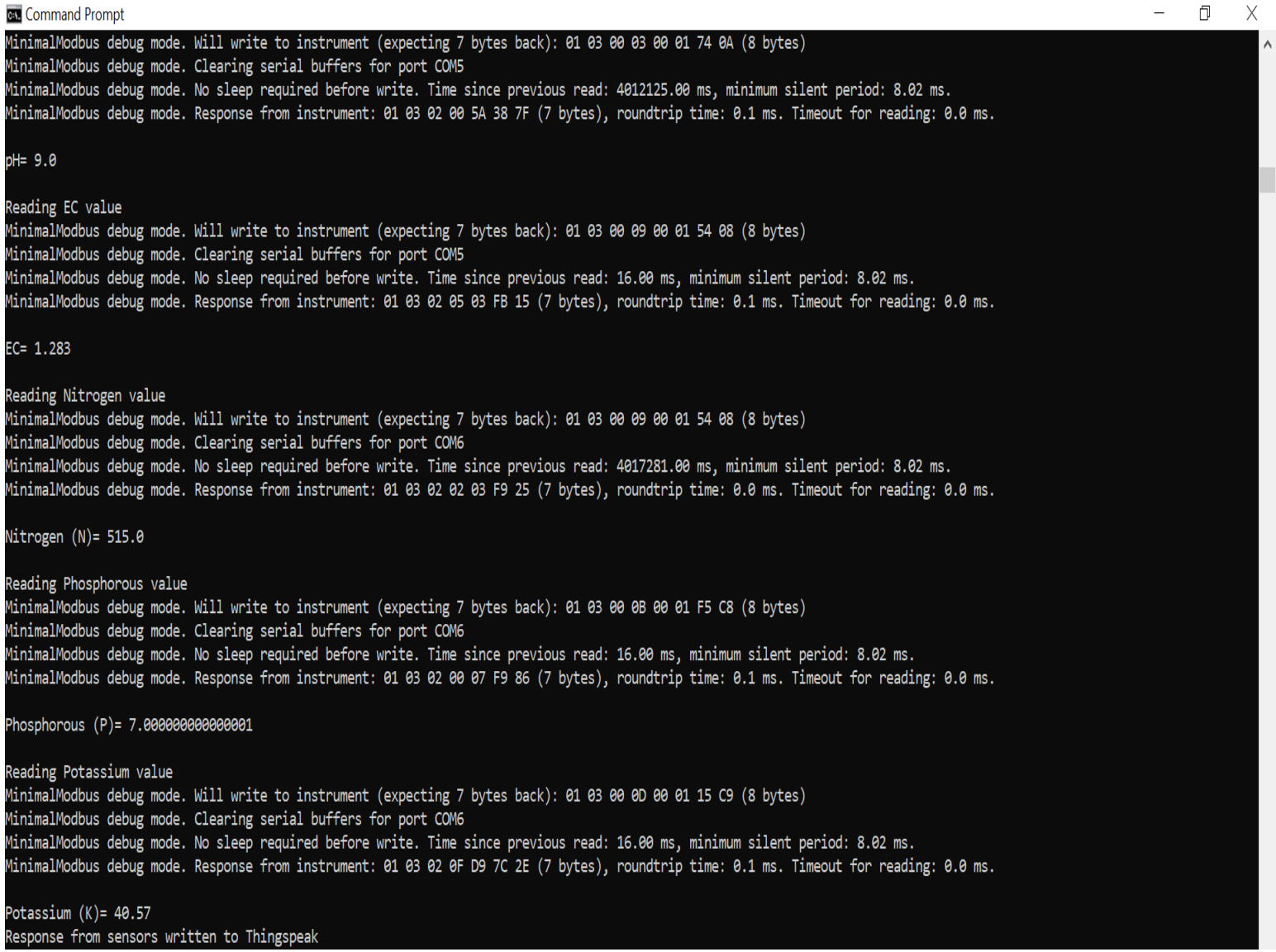


Figure 5.3: Sensor Readings using Python.

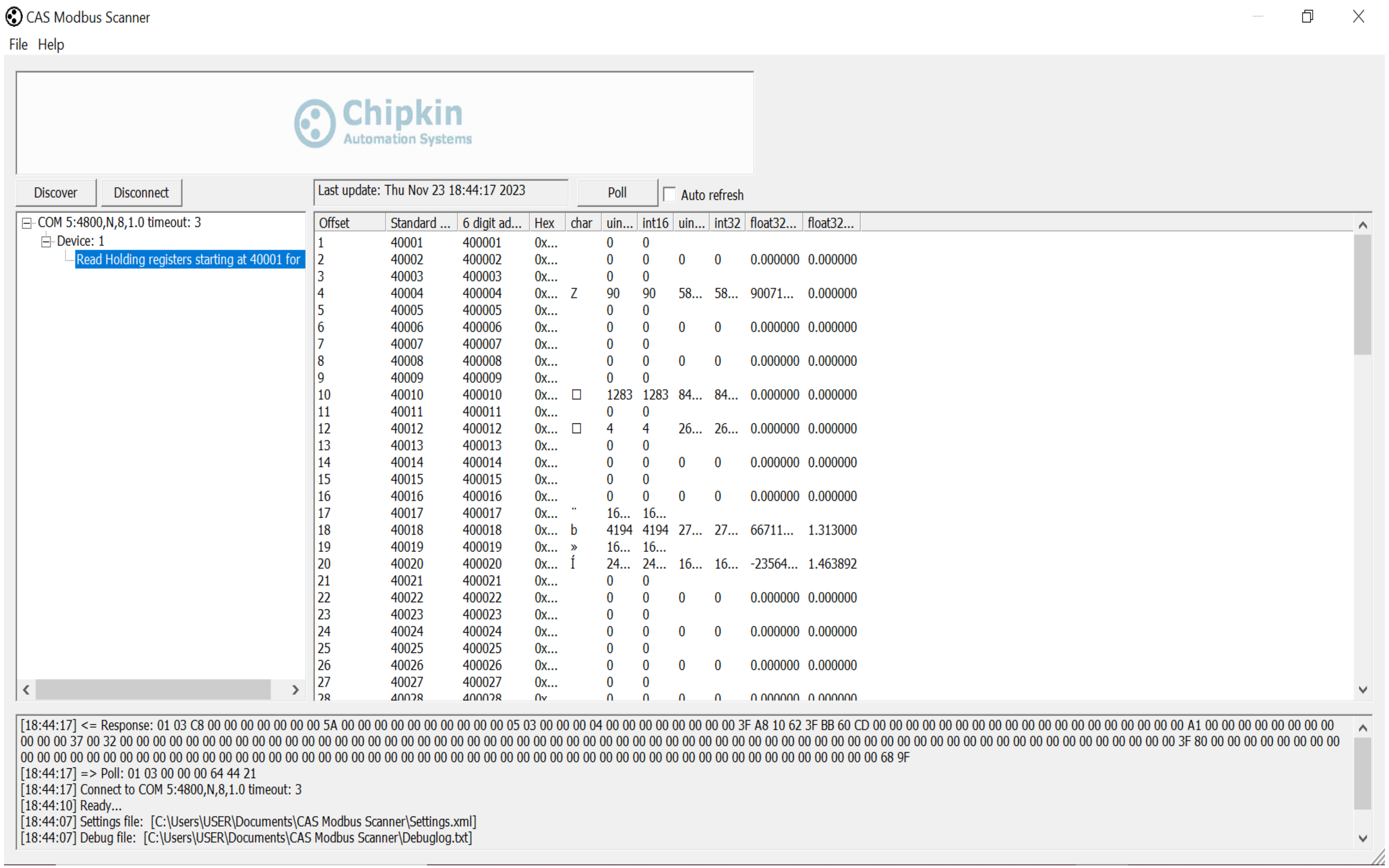


Figure 5.4: EC-pH Sensor readings using CAS Modbus scanner

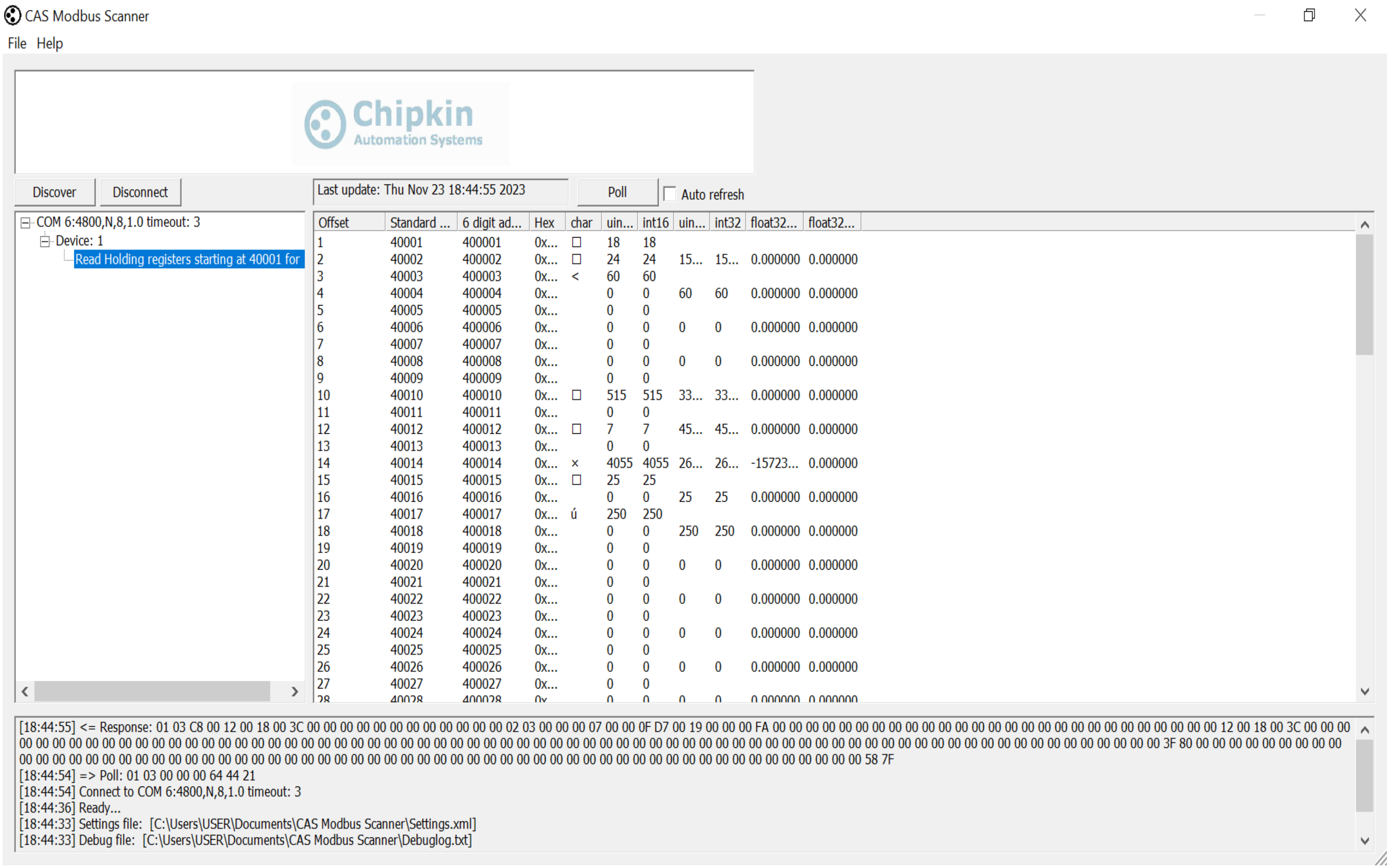


Figure 5.5: NPK Sensor readings using CAS Modbus scanner

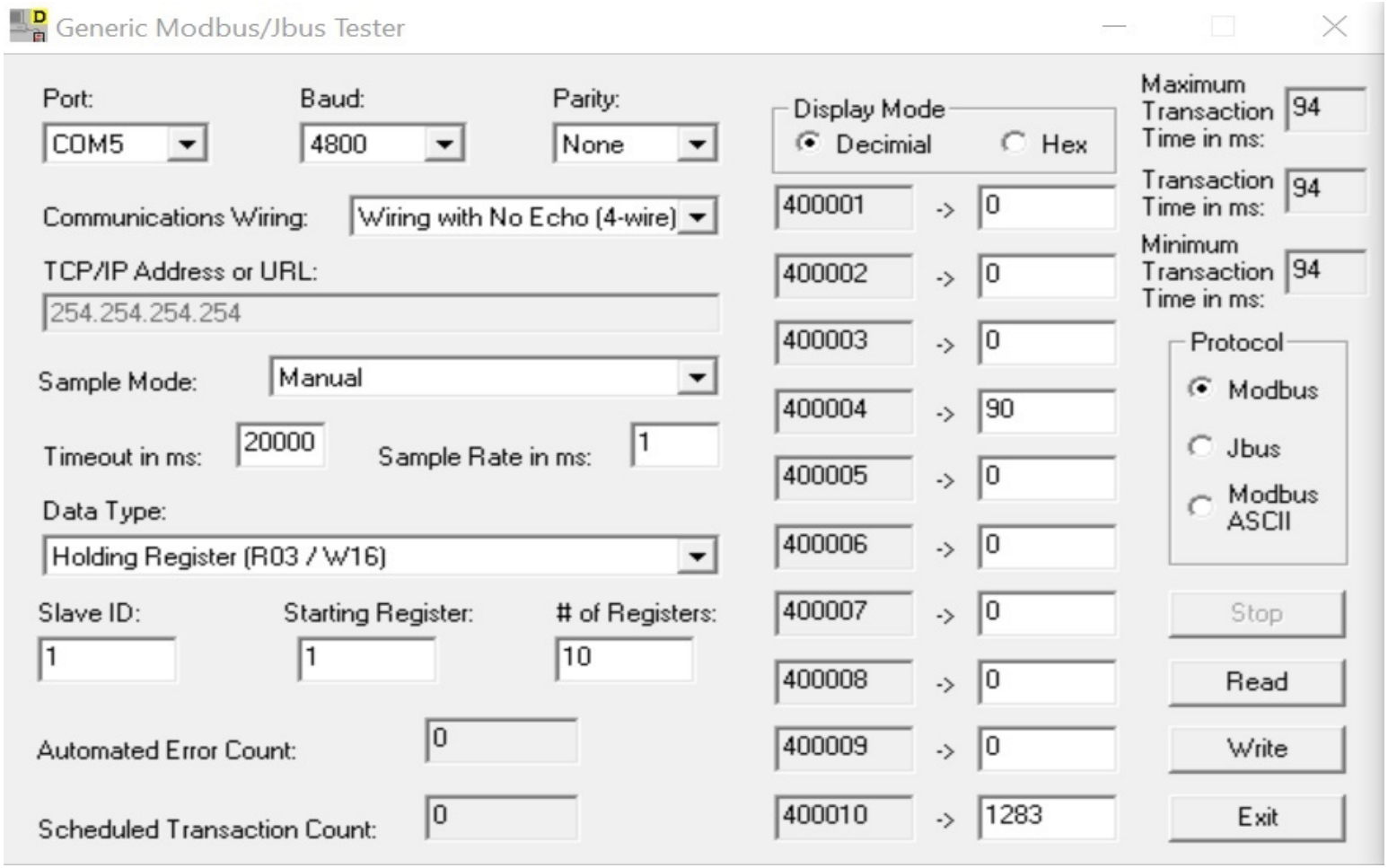


Figure 5.6: EC-pH Sensor readings using Generic Modbus/Jbus tester

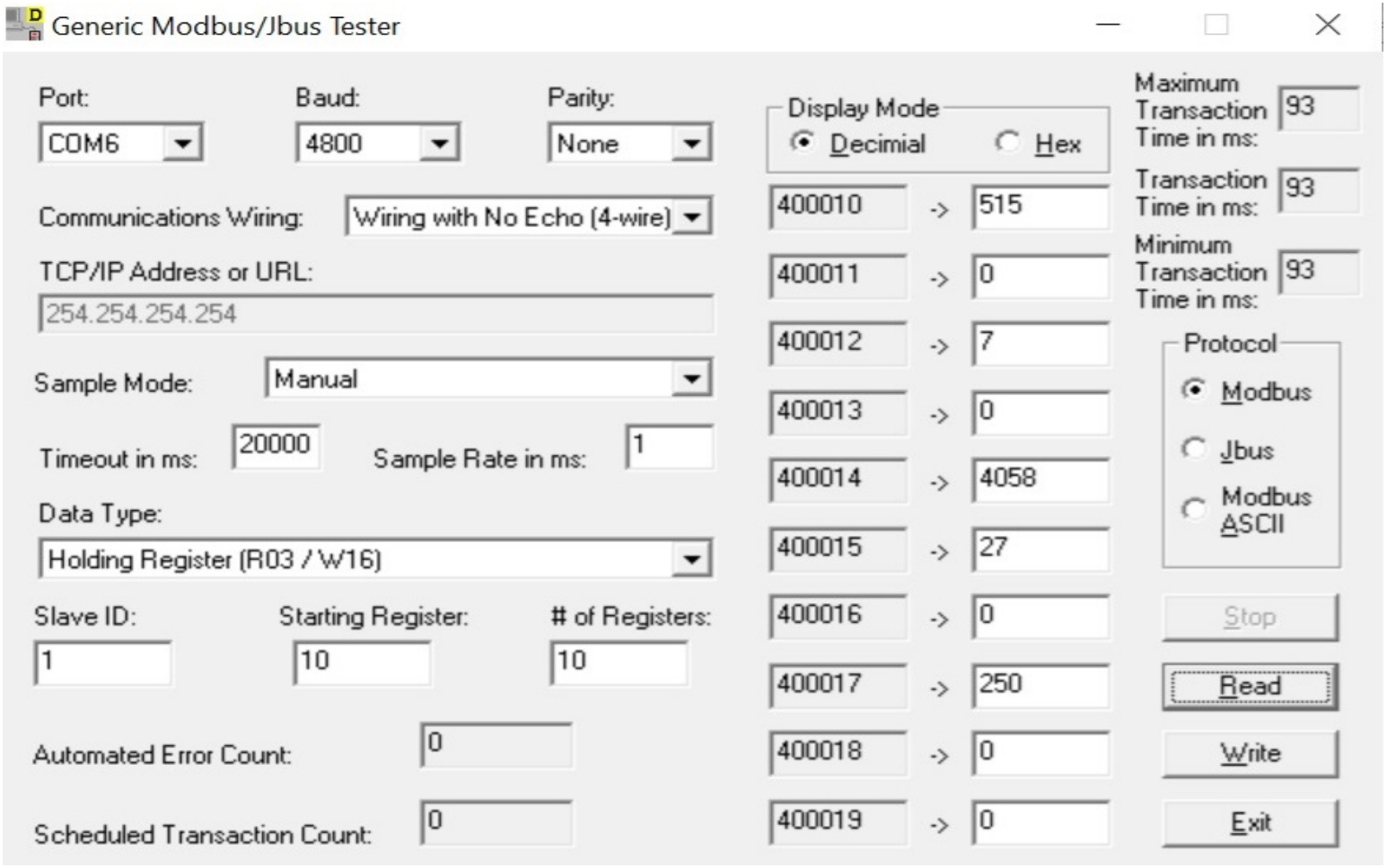


Figure 5.7: Sensor readings using Generic Modbus/Jbus tester

5.3.2 Data preprocessing of Real-time data

The redundant data were removed using the WEKA open-source tool, resulting in 50 instances. The collected data instances are labeled as LOW, MEDIUM, or HIGH based on the level of soil parameters. The fertility levels of *pH, EC, N, P*, and *K* are measured based on their values. Each soil parameter plays a significant role in soil fertility. The fertility level of remaining soil parameters such as *S*, *B*, *Cu*, *Fe*, *Mn*, and *Zn* is determined based on the value of *pH*. After labeling, 42 instances in the dataset were of LOW ferrtile soil, 1 instance of MEDIUM and 7 instances of HIGH fertile soil.

5.4 Experimental Results and Discussions

The experiments were conducted using WEKA (WEKA, 2021) open source tool. The ML-based classifiers such as CART, J48, RF, REP, NB, SVM were trained by using laboratory-measured Soil-health data with 10-fold cross-validation. The collected real-time data are tested using trained classifiers. The performance of the trained classifiers were measured using performance metrics such a Accuracy, Precision, Recall, F1-score. The dataset used in this study is imbalanced dataset. Thus, the performance of the classifiers were measured using per-class precision, per-class recall, per-class F1-score, kappa statistic and FPR.

A batch size of 100 was used to create the tree classifiers such as CART, J48, RF, and REP. The CART and J48 created tree of size 31 with number of leaves 16. RF was set with 100 iterations. For all nodes, REP utilized a minimum variance proportion of 0.001, a maximum number of 3 folds, and a seed value of 1 and produced tree of size 27. NB and SVM classifiers employed a batch size of 100. The SVM classifier was constructed by utilizing a radial basis kernel function, featuring a seed and cost value set at 1, a gamma value of 0.1, epsilon of 0.001, loss value of 0.1, and a degree of 3. With 10-fold cross-validation test on Soil-heath data CART, J48, RF, REP, NB and SVM missclassified 14, 9, 7, 19, 2748, and 1160 instances, respectively. Using real-time data as test data the trained CART, J48, RF, REP, NB and SVM classifier missclassified 7,1, 0, 0, 8, and 8 instances, respectively. The performance of the classifiers with 10-fold cross validation test using laboratory-measured data and using real-time soil data as test data are depicted in Table 5.2 and Table 5.3, respectively. The performance comparision of classifiers based on kappa statistics, and FPR achieved is presented in Table 5.4.

Table 5.2: Performance of classifiers using laboratory-measured Soil-health data

Classifier	Accuracy	Precision	Recall	F1-Score	Per-class Precision			Per-class Recall			Per-class F1-Score		
					LOW	MEDIUM	HIGH	LOW	MEDIUM	HIGH	LOW	MEDIUM	HIGH
CART	99.68%	0.997	0.997	0.997	1.000	0.991	0.975	0.997	0.996	1.000	0.998	0.994	0.987
J48	99.8%	0.998	0.998	0.998	1.000	0.996	0.975	0.998	0.996	1.000	0.999	0.996	0.987
RF	99.84%	0.998	0.998	0.998	1.000	0.997	0.981	0.999	0.997	1.000	0.999	0.997	0.991
REP	99.57%	0.996	0.996	0.996	0.999	0.995	0.934	0.997	0.991	0.994	0.998	0.993	0.963
NB	38.15%	0.749	0.381	0.366	0.909	0.260	0.589	0.216	0.912	0.401	0.349	0.405	0.477
SVM	73.89%	0.696	0.739	0.652	0.745	0.512	0.875	0.983	0.063	0.045	0.848	0.112	0.085

Table 5.3: Performance of classifiers using real-time test data

Classifier	Accuracy	Precision	Recall	F1-Score	Per-class Precision			Per-class Recall			Per-class F1-Score		
					LOW	MEDIUM	HIGH	LOW	MEDIUM	HIGH	LOW	MEDIUM	HIGH
CART	86%	-	0.860		1.000	0.125	-	1.000	1.000	0.000	1.000	0.222	-
J48	98%	0.990	0.980	0.983	1.000	0.500	1.000	1.000	1.000	0.857	1.000	0.667	0.923
RF	100%	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
REP	100%	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
NB	84%	-	-	-	0.84	-	-	1.000	-	-	0.913	-	-
SVM	84%	-	-	-	0.84	-	-	1.000	-	-	0.913	-	-

Table 5.4: Performance comparision using kappa statistic, TPR and FPR

	Laboratory-measured\ Soil-health data			**Real-time Soil data**		
Classifier	**Kappa Statistic**	**TPR**	**FPR**	**Kappa Statistic**	**TPR**	**FPR**
CART	0.9922	0.997	0.001	0.5192	0.860	0.003
J48	0.995	0.998	0.001	0.9277	0.980	0.000
RF	0.9961	0.998	0.001	1	1.000	0.000
REP	0.9895	0.996	0.002	1	1.000	0.000
NB	0.1	0.381	0.221	0	-	0.840
SVM	0.0667	0.739	0.692	0	-	0.840

Using Soil-health data as test data CART, J48, RF, REP, NB and SVM classifiers achieved an accuracy of 99.68%, 99.8%, 99.84%, 99.57%, 38.15%, 73.89%, respectively. Using real-time soil data as test data CART, J48, RF, REP, NB and SVM classifiers achieved an accuracy of 86%, 98%, 100%, 100%, 84%, 84%, respectively. The RF classifier achieved kappa statistic, TPR and FPR of 0.9961, 0.998 and 0.001 using Soil-health test data and it achieved kappa statistic, TPR and FPR of 1, 1.000, and 0.000, respectively. The soil parameter values are correlated to each other, resulting in poor performance of NB classifier. The dataset utilized to create the classifiers is biased dataset, resulting in reduced performance of SVM classifier.

5.5 Summary

Precise real-time soil fertility classification is essential for sustainable agriculture production. In this research work ML-based classifiers such as CART, J48, RF, REP, NB and SVM were developed using laboratory-measured soil data and 10-fold cross-validation test was performed. Furthermore, real-time soil parameters were collected and used as test data to evaluate the performance of trained classifiers. It was observed that the tree-based classifiers such as CART, J48, RF and REP performed better as compared to NB and SVM classifiers. The RF classifier outperformed other classifiers by producing an accuracy of 99.84% using Soil-health data. The RF and REP performed better than other classifiers using real-time soil test data.

CHAPTER 6

CONCLUSIONS AND FUTURE WORKS

A balanced soil fertility is vital for crop growth and yield. Precise classification of soil fertility is a significant requirement to enhance sustainable agricultural production. Using ML-based classifiers, soil fertility can be classified as LOW, or MEDIUM or HIGH fertile without chemical analysis. The ML-based soil fertility classification saves time and minimizes chemical analysis costs. The precise application of fertilizers reduces the cost of fertilization and reduces environmental pollution. The aim of this research work is to propose and implement a soil fertility classification approach with fertilizer recommendation module. Precise classification of soil fertility is a significant requirement to enhance sustainable agricultural production.

Initially, various ML-based classifiers such as NB, LR, J48, Bagging, BRT, RF and SVM were employed to classify soil fertility based on laboratory measurements of *pH, EC, OC, P, K, S, Zn, B, Fe, Cu,* and *Mn*. The experiments were conducted using 10-fold cross-validation test and using split dataset. The study examined the performance of aforesaid ML-based classifiers in classifying soil fertility levels as LOW, or MEDIUM, or HIGH. It was observed that tree-based RF classifier achieved the highest Accuracy of 99.99% using 10-fold cross-validation test as well as using split dataset.

The laboratory-measured soil data that was used in the previous work was highly imbalanced. To overcome this drawback, dataset was created using satellite spectral bands. This study effort employed remotely sensed Sentinel-2 spectral bands to predict soil parameters such as *EC, pH, OC* and *N* in order to overcome the limitations. The data points were clustered using a variety of cutting-edge clustering techniques. It was found that Canopy clustering achieved a clustering accuracy of 75.99%, and by utilizing an RF classifier with 10-fold cross-validation, Canopy clustering obtained a classification accuracy of 98.78%. It was observed that using the proposed Canopy Center-based Fuzzy-C-Means clustering achieved the highest clustering accuracy of 78.42%, and by utilizing an RF classifier with a 10-fold cross-validation proposed approach obtained a classification accuracy of 99.69%. Decisions on soil fertility are more precise using the proposed clustering technique.

Using a limited number of soil parameters reduces the training time of the classifiers and laboratory chemical analysis costs. An ensemble filter-based feature selection was proposed using three different feature selection approaches: InfoG, GainR, and ReliefF. The proposed approach removes the least relevant feature. The proposed soil fertility classification approach's performance is evaluated using two datasets. The performance of tree-based machine learning classifiers such as CART, Extra Tree, J48, RF, and REP are compared with NB and SVM classifiers. With the elimination of soil parameter 'S' from both the datasets, and using a subset of features consisting of ten soil parameters, *EC, pH, OC, K, P, B, Cu, Fe, Mn, and Zn* the RF classifier outperformed the other classifiers. The RF achieved the highest accuracy and kappa statistics of 99.96%, 0.9286 for dataset-1 and the highest accuracy and kappa statistics of 99.90%, 0.9091 for dataset-2, respectively. A significant improvement in kappa statistics were observed after removing the least relevant feature '*S*', with both 10-fold cross-validation and split dataset. Using both datasets, the RF classifier's performance increased compared to other classifiers after removing feature '*S*'. An adequate amount of fertilizers is recommended based on the obtained classification results.

Recent studies have proven that machine learning-based classifiers and deep learning-based classifiers such as ELM and MLP can successfully classify soil fertility based on chemical parameters. In this research work, CNN-based soil fertility classification is proposed to classify soil fertility. It uses 11 soil chemical parameters as input to classify soil fertility as HIGH, or MEDIUM, or LOW. The proposed approach obtained the training and test accuracy of 99.95% and 97.24%, respectively, for kernel size 3×3 and an input grid size 11×11. Using SMOTE oversampling, the proposed approach achieved the highest training and test accuracy of 99.98% and 97.52%, respectively, for kernel size 3×3 and an input grid size 12×12. The classification results are used to recommend suitable fertilizers for specific crops.

This research proposes a 1D-CNN-based classifier to classify soil fertility based on soil chemical parameters. The MinMax normalization and SMOTE oversampling techniques are also used to improve the classifier's performance. The proposed approach is compared with ELM and MLP. Experiments were conducted to show the effectiveness of a 1D-CNN-based classifier with MinMax normalization and SMOTE oversampling for classifying imbalanced datasets. Based on the soil-health dataset results, the pro-

posed approach attained the highest training and validation accuracies of 99.91% and 97.75%, respectively. The proposed approach performed better than ELM and MLP with a classification accuracy of 97.90%, highest recall of 0.979, precision of 0.992, F1-score of 0.984, and kappa of 0.2358. Furthermore, the fertilizers are prescribed using the classification results. The performance of the 1D CNN-based classifier is better than the 2D CNN-based classifier. However, the performance of 1D CNN was degraded compared to the tree-based classifier such as RF.

This work proposes an SDFA-based soil fertility classification approach to dynamically classify soil data as HIGH, or MEDIUM, or LOW fertile. The Sentinel-2 dataset containing four soil parameters is used to evaluate the efficacy of the proposed approach. It is observed that an accuracy of 100% was obtained using the Sentinel-2 dataset. The performance of the proposed approach is also measured using laboratory-measured soil-health data. The accuracy was found to be 100% using Soil-health-1 dataset and 98.37% using Soil-health-2 dataset. Furthermore, the proposed method outperformed machine learning-based classifiers. The fertilizers were recommended to site-specific crops based on classification results.

Precise real-time soil fertility classification is essential for sustainable agriculture production. In this research work ML-based classifiers were trained by using laboratory-measured soil data. The ML-based classifiers such as CART, J48, RF, REP, NB and SVM with 10-fold cross-validation test were employed to classify soil fertility. Furthermore, real-time soil collected and used as test data to evaluate the performance of trained classifiers. It was observed that the tree-based classifiers such as CART, J48, RF and REP performed better compared to NB and SVM classifiers. The RF classifier outperformed other classifiers obtaining an accuracy of 99.84% using Soil-health data. The RF and REP were performed better than other classifiers using real-time soil test data with accuracy of 100% with precision, recall and F1-Score of 1.000.

It is noted that classifiers based on decision trees outperformed other classifiers such as NB, SVM, KNN, 1D-CNN, and 2D-CNN. NB classifier relies on the assumption of feature independence, which does not hold true in datasets where soil parameters exhibit correlations. Additionally, the dataset's imbalance posed a challenge for SVMs, which strive to establish a decision boundary that optimizes class margin. Similarly,

KNN's performance was hindered by class imbalance, as the prevalence of the majority class could bias nearest neighbor computations, leading to skewed predictions and diminished accuracy for minority classes. Moreover, CNNs, including both 1D and 2D variants, necessitate extensive training data for effective learning. In the context of imbalanced datasets, CNNs may exhibit a bias toward the majority class. RF classifier performed better than other classifiers in most of the cases. RF classifier, as an ensemble learning method, combines the predictions of multiple decision trees trained on different subsets of the data. This ensemble approach helps mitigate overfitting and improves generalization performance compared to other classifiers. In very few scenarios J48 performed better than RF because of dataset size and class imbalance.

There are numerous opportunities to work on soil fertility classification. As future work, the spectral bands of different satellites can be used to derive more soil parameters. A fertilizer prescription can be developed to recommend different combinations of fertilizers or manures for crops grown at different soil types and climatic conditions. The real-time soil data can be collected during the growth period of crops, and fertilizers can be prescribed in real-time. A precise soil fertility classifier can be developed using physical and biological soil parameters in combination with chemical parameters. The early prediction of soil fertility based on climate conditions can help the farmers to reduce the cost of fertilization and avoids environmental pollution.

REFERENCES

Abera, W., Tamene, L., Tesfaye, K., Jiménez, D., Dorado, H., Erkossa, T., Kihara, J., Ahmed, J. S., Amede, T., & Ramirez-Villegas, J. (2022). A data-mining approach for developing site-specific fertilizer response functions across the wheat-growing environments in ethiopia. *Experimental Agriculture*, *58*. https://doi.org/10.1017/S0014479722000047.

Aksoy, S., Yildirim, A., Gorji, T., Hamzehpour, N., Tanik, A., & Sertel, E. (2022). Assessing the performance of machine learning algorithms for soil salinity mapping in google earth engine platform using sentinel-2a and landsat-8 oli data. *Advances in Space Research*, *69*(2), 1072–1086. https://doi.org/10.1016/j.asr.2021.10.024.

Al-Gaadi, K. A., Tola, E., Madugundu, R., & Fulleros, R. B. (2021). Sentinel-2 images for effective mapping of soil salinity in agricultural fields. *Currurent Science*, *121*, 384–390. https://doi.org/10.18520/cs/v121/i3/384-390.

Al Masmoudi, Y., Bouslihim, Y., Doumali, K., Hssaini, L., & Ibno Namr, K. (2022). Use of machine learning in moroccan soil fertility prediction as an alternative to laborious analyses. *Modeling Earth Systems and Environment*, *8*(3), 3707–3717. https://doi.org/10.1007/s40808-021-01329-8.

Alatrash, R., Priyadarshini, R., Ezaldeen, H., & Alhinnawi, A. (2022). Augmented language model with deep learning adaptation on sentiment analysis for e-learning recommendation. *Cognitive Systems Research*, *75*, 53–69. https://doi.org/10.1016/j.cogsys.2022.07.002.

Badrinath, M., Chidanandappa, H., Ali, H., & Chamegowda, T. (1995). Impact of lime on rice yield and available potassium in coastal acid soils of karnataka. *Agropedology*, *5*, 43–46. http://isslup.in/wp-content/uploads/2018/09/Impact-of-Lime-on-Rice-Yield-and-Available-Potassium-in-Coastal.pdf.

Bagherzadeh, A. & Gholizadeh, A. (2017). Parametric-based neural networks and topsis modeling in land suitability evaluation for alfalfa production using gis. *Modeling Earth Systems and Environment*, *3*, 1–11. https://doi.org/10.1007/s40808-016-0263-y.

Belyadi, H. & Haghighat, A. (2021). Machine learning guide for oil and gas using python: A step-by-step breakdown with data, algorithms, codes, and applications. United States: Gulf Professional Publishing.

Beraha, M., Metelli, A. M., Papini, M., Tirinzoni, A., & Restelli, M. (2019), 1–9. 1–9., Feature selection via mutual information: New theoretical insights. In *2019 international joint conference on neural networks (IJCNN)*, 1–9. IEEE. https://doi.org/10.1109/IJCNN.2019.8852410.

Bisong, E. et al. (2019). Building machine learning and deep learning models on google cloud platform. Springer. https://doi.org/10.1007/978-1-4842-4470-8.

Blevins, D. G. & Lukaszewski, K. M. (1998). Boron in plant structure and function. *Annual review of plant biology*, *49*(1), 481–500. https://doi.org/10.1146/annurev.arplant.49.1.481.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). Classification and regression trees. Routledge.

Büchele, D., Chao, M., Ostermann, M., Leenen, M., & Bald, I. (2019). Multivariate chemometrics as a key tool for prediction of k and fe in a diverse german agricultural soil-set using edxrf. *Scientific Reports*, *9*(1), 17588. https://doi.org/10.1038/s41598-019-53426-5.

Campesato, O. (2020). Python 3 for machine learning. Sterling: Stylus Publishing, LLC.

Castelli, M., Vanneschi, L., & Largo, Á. (2018). Supervised learning: classification. *por Ranganathan, S., M. Grisbskov, K. Nakai y C. Schönbach*, *1*, 342–349. https://doi.org/10.1016/B978-0-12-809633-8.20332-4.

Chakraborty, K., Bhattacharyya, S., Bag, R., & Hassanien, A. (2018). Sentiment analysis on a set of movie reviews using deep learning techniques. *Soc. Netw. Anal. Comput. Res. Methods Tech*, *7*, 127–147. https://doi.org/10.1016/B978-0-12-815458-8.00007-4.

Chandrashekar, G. & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024.

Chang, R., Chen, Z., Wang, D., & Guo, K. (2022). Hyperspectral remote sensing inversion and monitoring of organic matter in black soil based on dynamic fitness inertia weight particle swarm optimization neural network. *Remote Sensing*, *14*(17), 4316. https://doi.org/10.3390/rs14174316.

Chen, H., Das, S., Morgan, J. M., & Maharatna, K. (2022). Prediction and classification of ventricular arrhythmia based on phase-space reconstruction and fuzzy c-means clustering. *Computers in Biology and Medicine*, *142*, 105180. https://doi.org/10.1016/j.compbiomed.2021.105180.

Chen, W., He, Z. L., Yang, X. E., Mishra, S., & Stoffella, P. J. (2010). Chlorine nutrition of higher plants: progress and perspectives. *Journal of Plant Nutrition*, *33*(7), 943–952. https://doi.org/10.1080/01904160903242417.

Chen, Y., Jia, J., Wu, C., Ramirez-Granada, L., & Li, G. (2023). Estimation on total phosphorus of agriculture soil in china: a new sight with comparison of model learning methods. *Journal of Soils and Sediments*, *23*(2), 998–1007. https://doi.org/10.1007/s11368-022-03374-x.

CHIPKIN (2023), Cas modbus scanner. https://store.chipkin.com/products/tools/cas-modbus-scanner [Accessed on July 19, 2023].

Chougule, A., Jha, V. K., & Mukhopadhyay, D. (2019). Crop suitability and fertilizers recommendation using data mining techniques. In *Progress in Advanced Computing and Intelligent Engineering* 205–213. Springer. https://doi.org/10.1007/978-981-13-0224-4_19.

Cohen, S. (2021). The basics of machine learning: strategies and techniques. In *Artificial Intelligence and Deep Learning in Pathology* 13–40. United States: Elsevier. https://doi.org/10.1016/B978-0-323-67538-3.00002-6.

Coulibali, Z., Cambouris, A. N., & Parent, S.-É. (2020). Site-specific machine learning

predictive fertilization models for potato crops in eastern canada. *PloS one*, *15*(8), e0230888. https://doi.org/10.1371/journal.pone.0230888.

Dalla Preda, M., Giacobazzi, R., Lakhotia, A., & Mastroeni, I. (2015), 329–341. 329–341., Abstract symbolic automata: Mixed syntactic/semantic similarity analysis of executables. In *Proceedings of the 42nd Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, 329–341. https://doi.org/10.1145/2676726.2676986.

Day AD, L. K. (1993). Soil alkalinity. In *Plant Nutrients in Desert Environments. Adaptations of Desert Organisms* 35–37. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-77652-6_9.

De'Ath, G. (2007). Boosted trees for ecological modeling and prediction. *Ecology*, *88*(1), 243–251. https://doi.org/10.1890/0012-9658(2007)88[243:BTFEMA]2.0.CO;2.

Delavar, M. A., Naderi, A., Ghorbani, Y., Mehrpouyan, A., & Bakhshi, A. (2020). Soil salinity mapping by remote sensing south of urmia lake, iran. *Geoderma Regional*, *22*, e00317. https://doi.org/10.1016/j.geodrs.2020.e00317.

Deng, X., Chen, X., Ma, W., Ren, Z., Zhang, M., Grieneisen, M. L., Long, W., Ni, Z., Zhan, Y., & Lv, X. (2018). Baseline map of organic carbon stock in farmland topsoil in east china. *Agriculture, ecosystems & environment*, *254*, 213–223. https://doi.org/10.1016/j.agee.2017.11.022.

Devi, R. D. H., Bai, A., & Nagarajan, N. (2020). A novel hybrid approach for diagnosing diabetes mellitus using farthest first and support vector machine algorithms. *Obesity Medicine*, *17*, 100152. https://doi.org/10.1016/j.obmed.2019.100152.

Dos Santos, E. P., Moreira, M. C., Fernandes-Filho, E. I., Demattê, J. A. M., dos Santos, U. J., da Silva, D. D., Cruz, R. R. P., Moura-Bueno, J. M., Santos, I. C., & de Sá Barreto Sampaio, E. V. (2023). Improving the generalization error and transparency of regression models to estimate soil organic carbon using soil reflectance data. *Ecological Informatics*, *77*, 102240. https://doi.org/10.1016/j.ecoinf.2023.102240.

Dutta, D. & Rakshit, A. (2016). Potential effects of climate change on soil properties: A review. *Science International*, *4*(2). https://doi.org/10.17311/sciintl.2016.51.73.

Elfadel, I. M., Boning, D. S., & Li, X. (Eds.). (2019). Machine learning in VLSI computer-aided design. Basel, Switzerland: Springer International Publishing.

Fao (2020a), Agriculture, with its allied sectors. http://www.fao.org/india/fao-in-india/india-at-a-glance/en [Accessed on October 25, 2020].

Fao (2020b), Global soil partnership. http://www.fao.org/global-soil-partnership/areas-of-work/soil-fertility/en/ [Accessed on October 25, 2020].

Fao (2021), Sustainable soil and land management for csa. http://www.fao.org/climate-smart-agriculture-sourcebook/production-resources/module-b7-soil/b7-overview/en/ [Accessed on 15th October 2021].

Fernandes, M. M. H., Coelho, A. P., Fernandes, C., da Silva, M. F., & Marta, C. C. D. (2019). Estimation of som content by modeling with artificial neural networks. *Geoderma*, *350*, 46–51. https://doi.org/10.1016/j.geoderma.2019.04.044.

Gan, Y., Siddique, K. H., Turner, N. C., Li, X.-G., Niu, J.-Y., Yang, C., Liu, L., & Chai, Q. (2013). Ridge-furrow mulching systems—an innovative technique for boosting crop productivity in semiarid rain-fed environments. In *Advances in agronomy*, volume 118, 429–476. Elsevier.

Gao, F., Li, B., Chen, L., Shang, Z., Wei, X., & He, C. (2021). A softmax classifier for high-precision classification of ultrasonic similar signals. *Ultrasonics*, *112*, 106344. https://doi.org/10.1016/j.ultras.2020.106344.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, *63*(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1.

Gholizadeh, A., Zizala, D., Saberioon, M., & Boruvka, L. (2018). Soc and texture retrieving and mapping using proximal, airborne and sentinel-2 spectral imaging. *Remote*

Sensing of Environment, *218*, 89–103. https://doi.org/10.1016/j.rse.2018.09.015.

Ghorbani, M. A., Deo, R. C., Kashani, M. H., Shahabi, M., & Ghorbani, S. (2019). Artificial intelligence-based fast and efficient hybrid approach for spatial modelling of soil electrical conductivity. *Soil and Tillage Research*, *186*, 152–164. https://doi.org/10.1016/j.still.2018.09.012.

Gorji, T., Yildirim, A., Hamzehpour, N., Tanik, A., & Sertel, E. (2020). Soil salinity analysis of urmia lake basin using landsat-8 oli and sentinel-2a based spectral indices and electrical conductivity measurements. *Ecological Indicators*, *112*, 106173. https://doi.org/10.1016/j.ecolind.2020.106173.

Gu, X., Peng, J., Cheng, Y., Zhang, X., & Liu, K. (2020). Energy replenishment optimisation via density-based clustering. *International Journal of Computational Science and Engineering*, *21*(2), 271–280. https://doi.org/10.1504/IJCSE.2020.105735.

Gulhane, V. A., Rode, S. V., & Pande, C. B. (2023). Correlation analysis of soil nutrients and prediction model through iso cluster unsupervised classification with multispectral data. *Multimedia Tools and Applications*, *82*(2), 2165–2184. https://doi.org/10.1007/s11042-022-13276-2.

Guo, Y., Wu, Y., Zhang, X., Bo, A., & Li, X. (2021). The frck clustering algorithm for determining cluster number and removing outliers automatically. *International Journal of Computational Science and Engineering*, *24*(5), 485–494. https://doi.org/10.1504/IJCSE.2021.118097.

Han, S., Kim, H., & Lee, Y.-S. (2020). Double random forest. *Machine Learning*, *109*(8), 1569–1586. https://doi.org/10.1007/s10994-020-05889-1.

Hengl, T., Leenaars, J. G., Shepherd, K. D., Walsh, M. G., Heuvelink, G. B., Mamo, T., Tilahun, H., Berkhout, E., Cooper, M., Fegraus, E., & Wheeler, I. (2017). Soil nutrient maps of sub-saharan africa: assessment of soil nutrient content at 250 m spatial resolution using machine learning. *Nutrient Cycling in Agroecosystems*, *109*, 77–102. https://doi.org/10.1007/s10705-017-9870-x.

Hengl, T., Miller, M. A., Krizan, J., Shepherd, K. D., Sila, A., Kilibarda, M., Antonijevic, O., Glusica, L., Dobermann, A., Haefele, S. M., & McGrath, S. (2021). African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning. *Scientific Reports*, *11*(1), 6130. https://doi.org/10.1038/s41598-021-85639-y.

Hossen, M. A., Diwakar, P. K., & Ragi, S. (2021). Total nitrogen estimation in agricultural soils via aerial multispectral imaging and libs. *Scientific Reports*, *11*(1), 12693. https://doi.org/10.1038/s41598-021-90624-6.

Hu, B., Xie, M., Li, H., He, R., Zhou, Y., Jiang, Y., Ji, W., Peng, J., Xia, F., Liang, Z., et al. (2023). Climate and soil management factors control spatio-temporal variation of soil nutrients and soil organic matter in the farmland of jiangxi province in south china. *Journal of Soils and Sediments*, *23*(6), 2373–2395. https://doi.org/10.1007/s11368-023-03471-5.

Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K. (2006). Extreme learning machine: theory and applications. *Neurocomputing*, *70*(1-3), 489–501. https://doi.org/10.1016/j.neucom.2005.12.126.

ICRISAT and Government of Karnataka (2016), Karnataka soil health data. https://doi.org/10.21421/D2/QYCEGR.

Ilango, H. S., Ma, M., & Su, R. (2022). A feedforward–convolutional neural network to detect low-rate dos in iot. *Engineering Applications of Artificial Intelligence*, *114*, 105059. https://doi.org/10.1016/j.engappai.2022.105059.

Inoue, Y., Saito, T., Iwasaki, A., Nemoto, T., & Ono, T. (2020). Hyperspectral assessment of soil fertility in farm fields in fukushima decontaminatedf after the radioactive fallout. *Soil Science and Plant Nutrition*, *66*(6), 820–827. https://doi.org/10.1080/00380768.2020.1753237.

Jain, A., Nandakumar, K., & Ross, A. (2005). Score normalization in multimodal biometric systems. *Pattern recognition*, *38*(12), 2270–2285. https://doi.org/10.1016/j.patcog.2005.01.012.

Jain, R., Jain, N., Aggarwal, A., & Hemanth, D. J. (2019). Convolutional neural network based alzheimer's disease classification from magnetic resonance brain images. *Cognitive Systems Research*, *57*, 147–159. https://doi.org/10.1016/j.cogsys.2018.12.015.

Jebara, T. (2004). Machine learning: discriminative and generative, volume 755. New York: Springer Science & Business Media. https://doi.org/10.1007/978-1-4419-9011-2.

Jonas, B. (2023), Usage general on modbus protocol. https://minimalmodbus.readthedocs.io/en/stable/usage.html [Accessed on July 11, 2023].

Kalkhoran, S. S., Pannell, D. J., Thamo, T., White, B., & Polyakov, M. (2019). Soil acidity, lime application, nitrogen fertility, and greenhouse gas emissions: Optimizing their joint economic management. *Agricultural Systems*, *176*, 102684. https://doi.org/10.1016/j.agsy.2019.102684.

Kant, D. S. & Kafkafi, U. (2020), Impact of mineral deficiency stress. http://plantstress.com/mineral-deficiency/[Accessed on October 25, 2020].

Keshavarzi, A., Kaya, F., Basayigit, L., Gyasi-Agyei, Y., Rodrigo-Comino, J., & Caballero-Calvo, A. (2023). Spatial prediction of soil micronutrients using machine learning algorithms integrated with multiple digital covariates. *Nutrient Cycling in Agroecosystems*, *127*, 1–17. https://doi.org/10.1007/s10705-023-10303-y.

Khanal, S., Fulton, J., Klopfenstein, A., Douridas, N., & Shearer, S. (2018). Integration of high resolution remotely sensed data and machine learning techniques for spatial prediction of soil properties and corn yield. *Computers and electronics in agriculture*, *153*, 213–225. https://doi.org/10.1016/j.compag.2018.07.016.

Kiranyaz, S., Ince, T., Abdeljaber, O., Avci, O., & Gabbouj, M. (2019), 8360–8364. 8360–8364., 1-d convolutional neural networks for signal processing applications. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8360–8364. IEEE. https://doi.org/10.1109/ICASSP.2019.8682194.

Kouadio, L., Deo, R. C., Byrareddy, V., Adamowski, J. F., Mushtaq, S., & Phuong Nguyen, V. (2018). Artificial intelligence approach for the prediction of robusta coffee yield using soil fertility properties. *Computers and Electronics in Agriculture*, *155*, 324–338. https://doi.org/10.1016/j.compag.2018.10.014.

Lambot, C., Herrera, J. C., Bertrand, B., Sadeghian, S., Benavides, P., & Gaitan, A. (2017). Cultivating coffee quality—terroir and agro-ecosystem. In *The craft and science of coffee,* 17–49. Academic Press. https://doi.org/10.1016/B978-0-12-803520-7.00002-5.

Li, X., Ding, J., Liu, J., Ge, X., & Zhang, J. (2021). Digital mapping of soil organic carbon using sentinel series data: A case study of the ebinur lake watershed in xinjiang. *Remote Sensing*, *13*(4), 769. https://doi.org/10.3390/rs13040769.

Liu, Y., Li, C., Cai, G., Sauheitl, L., Xiao, M., Shibistova, O., Ge, T., & Guggenberger, G. (2023). Meta-analysis on the effects of types and levels of n, p, and k fertilization on organic carbon in cropland soils. *Geoderma*, *437*, 116580. https://doi.org/10.1016/j.geoderma.2023.116580.

Lotfi, A. & Pirnia, M. (2022). Constraint-guided deep neural network for solving optimal power flow. *Electric Power Systems Research*, *211*, 108353. https://doi.org/10.1016/j.epsr.2022.108353.

Mahmoudzadeh, H., Matinfar, H. R., Taghizadeh-Mehrjardi, R., & Kerry, R. (2020). Spatial prediction of soc using machine learning techniques in western iran. *Geoderma Regional*, *21*, e00260. https://doi.org/10.1016/j.geodrs.2020.e00260.

Malik, K. M., Khan, K. S., Akhtar, M. S., & Ahmed, Z. I. (2020). Sulfur distribution and availability in alkaline subtropical soils affected by organic amendments. *Journal of Soil Science and Plant Nutrition*, *20*, 2253–2266. https://doi.org/10.1007/s42729-020-00292-0.

Marschner, H. (2011). Marschners mineral nutrition of higher plants. London: Academic press.

Mashaba-Munghemezulu, Z., Chirima, G. J., & Munghemezulu, C. (2021). Modeling the spatial distribution of soil nitrogen content at smallholder maize farms using machine learning regression and sentinel-2 data. *Sustainability*, *13*(21), 11591. https://doi.org/10.3390/su132111591.

Méndez-Vázquez, L. J., Lira-Noriega, A., Lasa-Covarrubias, R., & Cerdeira-Estrada, S. (2019). Delineation of site-specific management zones for pest control purposes: Exploring precision agriculture and species distribution modeling approaches. *Computers and Electronics in Agriculture,*, *167*, 105101. https://doi.org/10.1016/j.compag.2019.105101.

Misra, S., Li, H., & He, J. (2020). Noninvasive fracture characterization based on the classification of sonic wave travel times. In *Machine Learning for Subsurface Characterization* 243–287. Amsterdam, The Netherlands: Gulf Professional Publishing.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group (2009). Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Annals of internal medicine*, *151*(4), 264–269. https://doi.org/10.7326/0003-4819-151-4-200908180-00135.

Morgan, R., El-Hady, M. A., & Rahim, I. (2018). Soil salinity mapping utilizing sentinel-2 and neural networks. *Indian Journal of Agricultural Research*, *52*(5), 524–529. https://doi.org/10.18805/IJARe.A-316.

Müller, A. C. & Guido, S. (2016). Introduction to machine learning with python: a guide for data scientists. O'Reilly Media, Inc.

Neyestani, M., Sarmadian, F., Jafari, A., Keshavarzi, A., & Sharififar, A. (2021). Digital mapping of soil classes using spatial extrapolation with imbalanced data. *Geoderma Regional*, *26*, e00422. https://doi.org/10.1016/j.geodrs.2021.e00422.

Nozad, S. A. N., Haeri, M. A., & Folino, G. (2021). Sdcor: Scalable density-based clustering for local outlier detection in massive-scale datasets. *Knowledge-based systems*, *228*, 107256. https://doi.org/10.1016/j.knosys.2021.107256.

NRCS-USDA (2020), Soil electrical conductivity. https://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/nrcs142p2_053280.pdf. Accessed on October 10, 2020.

Osman, K. T. (2012). Soils: principles, properties and management. New York: Springer Science & Business Media.

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., & McKenzie, J. E. (2021). Prisma 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ*, *372*. https://doi.org/10.1136/bmj.n160.

Parsaie, F., Farrokhian Firouzi, A., Mousavi, S. R., Rahmani, A., Sedri, M. H., & Homaee, M. (2021). Large-scale digital mapping of topsoil total nitrogen using machine learning models and associated uncertainty map. *Environmental Monitoring and Assessment*, *193*(162), 1–15. https://doi.org/10.1007/s10661-021-08947-w.

Peng, Y., Wang, L., Zhao, L., Liu, Z., Lin, C., Hu, Y., & Liu, L. (2021). Estimation of soil nutrient content using hyperspectral data. *Agriculture*, *11*(11), 1129. https://doi.org/10.3390/agriculture11111129.

Pes, B. (2020). Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains. *Neural Computing and Applications*, *32*(10), 5951–5973. https://doi.org/10.1007/s00521-019-04082-3.

Pham, A.-D., Ngo, N.-T., Truong, T. T. H., Huynh, N.-T., & Truong, N.-S. (2020). Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability. *Journal of Cleaner Production,*, *260*, 121082. https://doi.org/10.1016/j.jclepro.2020.121082.

Phonphan, W., Tripathi, N. K., Tipdecho, T., & Eiumnoh, A. (2014). Modelling electrical conductivity of soil from backscattering coefficient of microwave remotely

sensed data using artificial neural network. *Geocarto International*, *29*(8), 842–859. https://doi.org/10.1080/10106049.2013.868040.

Python_weka_wrapper3 (2022), Python with weka. https://pypi.org/project/python-weka-wrapper3/,[Accessed on October 10, 2022].

Ramesh, G. & Menen, A. (2020). Automated dynamic approach for detecting ransomware using finite-state machine. *Decision Support Systems*, *138*, 113400. https://doi.org/10.1016/j.dss.2020.113400.

Ransom, C. J., Kitchen, N. R., Camberato, J. J., Carter, P. R., Ferguson, R. B., Fernández, F. G., Franzen, D. W., Laboski, C. A., Myers, D. B., Nafziger, E. D., Sawyer, J. E., & Shanahan, J. F. (2019). Statistical and machine learning methods evaluated for incorporating soil and weather into corn nitrogen recommendations. *Computers and Electronics in Agriculture*, *164*, 104872. https://doi.org/10.1016/j.compag.2019.104872.

Rasjid, Z. E. & Setiawan, R. (2017). Performance comparison and optimization of text document classification using k-nn and naïve bayes classification techniques. *Procedia computer science*, *116*, 107–112. https://doi.org/10.1016/j.procs.2017.10.017.

Reddy, S. (2018), Cropping seasons of india: Kharif, rabi, and zaid(zayid). Accessed on September 10, 2022, https://learnnaturalfarming.com/cropping-seasons-of-india-kharif-rabi-and-zaid/.

Reid, R. (2007). Physiology and metabalism of boron in plants. In *Advances in Plant and Animal Boron Nutrition* 83–90. Springer. https://doi.org/10.1007/978-1-4020-5382-5_7.

Rengel, Z. (2011). Soil ph, soil health and climate change. In *Soil health and climate change,* 69–85. Springer. https://doi.org/10.1007/978-3-642-20256-8_4.

Sander, D. & Wiese, R. (1973). Ec73-197 fertilizer know how. *Historical Materials from University of Nebraska-Lincoln Extension.4186*. http://digitalcommons.unl.edu/extensionhist/4186.

Schillaci, C., Acutis, M., Lombardo, L., Lipani, A., Fantappie, M., Märker, M., & Saia, S. (2017). Spatio-temporal topsoc mapping of a semi-arid mediterranean region: The role of land use, soil texture, topographic indices and the influence of remote sensing data to modelling. *Science of the total environment*, *601*, 821–832. https://doi.org/10.1016/j.scitotenv.2017.05.239.

Schneider, electric (2023), Generic modbus/jbus tester. https://www.se.com/us/en/faqs/FA180037/ [Accessed on July 11, 2023].

Sentinel-2 (2020), Sentinel-2 msi: Multispectral instrument, level-1c. Accessed on October 23, 2021, https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2.

Shang, K., Xiao, C., Gan, F., Wei, H., & Wang, C. (2021). Estimation of soil copper content in mining area using zy1-02d satellite hyperspectral data. *Journal of Applied Remote Sensing*, *15*(4), 042607–042607. https://doi.org/10.1117/1.JRS.15.042607.

Shi, Y., Zhao, J., Song, X., Qin, Z., Wu, L., Wang, H., & Tang, J. (2021). Hyperspectral band selection and modeling of soil organic matter content in a forest using the ranger algorithm. *PloS one*, *16*(6), e0253385. https://doi.org/10.1371/journal.pone.0253385.

Shobha, G. & Rangaswamy, S. (2018). Chapter 8 - machine learning. In V. N. Gudivada & C. Rao (Eds.), *Computational Analysis and Understanding of Natural Languages: Principles, Methods and Applications*, volume 38 of *Handbook of Statistics* 197–228. Elsevier. https://doi.org/10.1016/bs.host.2018.07.004.

Sirsat, M., Cernadas, E., Fernández-Delgado, M., & Barro, S. (2018). Automatic prediction of village-wise soil fertility for several nutrients in india using a wide range of regression methods. *Computers and electronics in agriculture,*, *154*, 120–133. https://doi.org/10.1016/j.compag.2018.08.003.

Sirsat, M., Cernadas, E., Fernández-Delgado, M., & Khan, R. (2017). Classification of agricultural soil parameters in india. *Computers and electronics in agriculture*, *135*, 269–279. https://doi.org/10.1016/j.compag.2017.01.019.

Soil-health (2021), Soil health card india, nutrient status-sample wise (for geo coordinates updation). Accessed on July 1, 2021, https://soilhealth.dac.gov.in/PublicReports/nutrientstatussamplesurveywise.

Somu, N., MR, G. R., & Ramamritham, K. (2020). A hybrid model for building energy consumption forecasting using long short term memory networks. *Applied Energy*, *261*, 114131. https://doi.org/10.1016/j.apenergy.2019.114131.

Sunitha, K. V. N. (2013). Compiler construction. India: Pearson Education.

Tavares, T. R., de Almeida, E., Junior, C. R. P., Guerrero, A., Fiorio, P. R., & de Carvalho, H. W. P. (2023). Analysis of total soil nutrient content with x-ray fluorescence spectroscopy (xrf): Assessing different predictive modeling strategies and auxiliary variables. *AgriEngineering*, *5*(2), 680–697. https://doi.org/10.3390/agriengineering5020043.

Tharavathy, N. (2016). A study on soil characteristics in urban and rural areas of mangalore, karnataka. *International Journal of Research in Environmental Science*, *2*(2), 5–8. https://doi.org/10.20431/2454-9444.0202002.

Tharsanee, R., Soundariya, R., Kumar, A. S., Karthiga, M., & Sountharrajan, S. (2021). Deep convolutional neural network–based image classification for covid-19 diagnosis. In *Data Science for COVID-19* 117–145. Elsevier. https://doi.org/10.1016/B978-0-12-824536-1.00012-5.

Van Noord, G. & Gerdemann, D. (2001). Finite state transducers with predicates and identities. *Grammars*, *4*(3), 263–286. https://doi.org/10.1023/A:1012291501330.

Vaudour, E., Gomez, C., Fouad, Y., & Lagacherie, P. (2019). Sentinel-2 image capacities to predict common topsoil properties of temperate and mediterranean agroecosystems. *Remote Sensing of Environment*, *223*, 21–33. https://doi.org/10.1016/j.rse.2019.01.006.

Villa-Vialaneix, N., Follador, M., Ratto, M., & Leip, A. (2012). A comparison of eight metamodeling techniques for the simulation of n2o fluxes and n leaching from corn

crops. *Environmental Modelling & Software*, *34*, 51–66. https://doi.org/10.1016/j.envsoft.2011.05.003.

Wang, B., Waters, C., Orgill, S., Cowie, A., Clark, A., Li Liu, D., Simpson, M., McGowen, I., & Sides, T. (2018). Estimating soc stocks using different modelling techniques in the semi-arid rangelands of eastern australia. *Ecological indicators*, *88*, 425–438. https://doi.org/10.1016/j.ecolind.2018.01.049.

Wang, K. & Kumar, P. (2019). Characterizing relative degrees of clumping structure in vegetation canopy using waveform lidar. *Remote Sensing of Environment*, *232*, 111281. https://doi.org/10.1016/j.rse.2019.111281.

WEKA (2021), Weka: Machine learning software in java. https://waikato.github.io/weka-wiki/downloading_weka/,[Accessed on October 25, 2020].

Wilson, T. M., Warren, J., & Arnall, B. (2013). Nitrous oxide emissions from soil. Technical report, Oklahoma Cooperative Extension Service. https://extension.okstate.edu/fact-sheets/print-publications/pss/nitrous-oxide-emissions-from-soil-pss-2269.pdf.

Witten, I. H. & Frank, E. (2002). Data mining: practical machine learning tools and techniques with java implementations. *Acm Sigmod Record*, *31*(1), 76–77. https://doi.org/10.1145/507338.507355.

Xie, Y., Sun, W., Ren, M., Chen, S., Huang, Z., & Pan, X. (2023). Stacking ensemble learning models for daily runoff prediction using 1d and 2d cnns. *Expert Systems with Applications*, *217*, 119469. https://doi.org/10.1016/j.eswa.2022.119469.

Xu, Y., Li, B., Shen, X., Li, K., Cao, X., Cui, G., & Yao, Z. (2022). Digital soil mapping of soil total nitrogen based on landsat 8, sentinel 2, and worldview-2 images in smallholder farms in yellow river basin, china. *Environmental Monitoring and Assessment*, *194*(4), 282. https://doi.org/10.1007/s10661-022-09902-z.

Xu, Y., Smith, S. E., Grunwald, S., Abd-Elrahman, A., Wani, S. P., & Nair, V. D. (2018). Estimating soil total nitrogen in smallholder farm settings using remote sensing spectral

indices and regression kriging. *Catena*, *163*, 111–122. https://doi.org/10.1016/j.catena.2017.12.011.

Xue, Y., Zhao, B., & Ma, T. (2016). Performance analysis for clustering algorithms. *International Journal of Computing Science and Mathematics*, *7*(5), 485–493. https://doi.org/10.1504/IJCSM.2016.080089.

Yang, P., Hu, J., Hu, B., Luo, D., & Peng, J. (2022). Estimating soil organic matter content in desert areas using in situ hyperspectral data and feature variable selection algorithms in southern xinjiang, china. *Remote Sensing*, *14*(20), 5221. https://doi.org/10.3390/rs14205221.

Yao, R.-J., Yang, J.-S., Wu, D.-H., Li, F.-R., Gao, P., & Wang, X.-P. (2015). Evaluation of pedotransfer functions for estimating saturated hydraulic conductivity in coastal salt-affected mud farmland. *Journal of Soils and Sediments*, *15*(4), 902–916. https://doi.org/10.1007/s11368-014-1055-5.

Zhang, J., Ji, D., Du, D., Miao, J., Liu, H., & Bai, Y. (2020). Temporal paradox in soil potassium estimations using spaceborne multispectral imagery. *Catena*, *194*, 104771. https://doi.org/10.1016/j.catena.2020.104771.

Zhang, S., Huang, W., & Zhang, C. (2019). Three-channel convolutional neural networks for vegetable leaf disease recognition. *Cognitive Systems Research*, *53*, 31–41. https://doi.org/10.1016/j.cogsys.2018.04.006.

Zhang, Y., Li, M., Zheng, L., Qin, Q., & Lee, W. S. (2019). Spectral features extraction for estimation of soil total nitrogen content based on modified ant colony optimization algorithm. *Geoderma*, *333*, 23–34. https://doi.org/10.1016/j.geoderma.2018.07.004.

Zia, H., Harris, N. R., Merrett, G. V., & Rivers, M. (2019). A low-complexity machine learning nitrate loss predictive model–towards proactive farm management in a networked catchment. *IEEE Access,*, *7*, 26707–26720. https://doi.org/10.1109/ACCESS.2019.2901218.

PUBLICATIONS

Journal Papers

1. Sujatha M, Jaidhar C D, and Mallikarjuna Lingappa (2023). "1D Convolutional Neural Networks-based Soil Fertility Classification and Fertilizer Prescription." *Ecological Informatics*, Vol. 78, 102295, Pages 1-19. DOI: 10.1016/j.ecoinf.2023.102295.

2. Sujatha M and Jaidhar C D. (2024). "Canopy Centre-based Fuzzy-C-Means Clustering for Enhancement of Soil Fertility Prediction." *International Journal of Computational Science and Engineering (IJCSE)*, Vol. 27, No. 1, Pages 90 - 102. DOI: 10.1504/IJCSE.2022.10058486.

3. Sujatha M and Jaidhar C D. (2024). "Machine Learning-based Approaches to Enhance the Soil Fertility-A Review." *Expert Systems With Applications*, Vol. 240, 122557, Pages 1-23, DOI: 10.1016/j.eswa.2023.122557.

4. Sujatha M and Jaidhar C D. "Machine Learning-Based Soil Fertility Classification Based on Real-Time Soil Chemical Parameters.", *IEEE Sensors Journal*, IEEE Publisher. (Under Review).

Conference Papers

1. Sujatha M and Jaidhar C D. "Classification of Soil Fertility using Machine Learning-based Classifier." *Proc., 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC 2021)*, held at NIT, Jalandhar, India, Pages 138-143. DOI: 10.1109/ICSCCC51823.2021.9478169.

2. Sujatha M and Jaidhar C D. "CNN-based Soil Fertility Classification with Fertilizer Prescription." *Proc., 3rd International Conference on Secure Cyber Computing and Communication (ICSCCC 2023)*, held at NIT, Jalandhar, India, Pages 439-444. DOI: 10.1109/ICSCCC58608.2023.10176841.

3. Sujatha M and Jaidhar, C D. "Fertilizer Recommendation Using Ensemble Filter-Based Feature Selection Approach." *Proc. 1st International Conference on Agriculture-Centric Computation, (ICA 2023) held at IIT, Ropar, Chandigarh, India in Agriculture-Centric Computation, Communications in Computer and Information Science*, Vol. 1866, Pages 43-57, Springer Publisher. DOI:10.1007/978-3-031-43605-5_4.

4. Sujatha M and Jaidhar C D. "Symbolic Deterministic Finite Automata-based Automated Fertilizer Prescription." *Proc. 14th International Conference on Computing Communication and Networking Technologies (ICCCNT 2023)*, held at IIT, Delhi, India, Pages 1-7, DOI: 10.1109/ICCCNT56998.2023.10307774.

CURRICULUM VITAE

Ms. Sujatha M
Full-Time Ph.D. Research Scholar
Department of Information Technology
National Institute of Technology Karnataka
P.O. Srinivasanagar, Surathkal
Mangalore-575 025
Email: sujatham.197it002@nitk.edu.in

Permanent Address

Sujatha M
#1-350, D/o Krishnamoorthy
Near Vishnumoorthy temple
P.O. Moodushedde
Via Vamanjoor,
Mangalore-575 028
Dakshina Kannada
Karnataka
Email: smsujatha23@gmail.com
Mobile: +91-8088586029

Academic Records

1. M.Tech. in Computer Science and Engineering, N.M.A.M. Institute of Technology (NMAMIT), Nitte, Karkala, Udupi (2010).
2. B.E. in Computer Science and Engineering, St. Joseph Engineering College (SJEC), Mangalore, Dakshina Kannada (2006).

Research Interests

Machine Learning, Deep Learning, Internet of Things, Computer Networks, Network Security