International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Improving Soil Characteristics Using Machine Learning in Different Environment

T. Poovizhi[*1], S. Christy[2]

Submitted: 04/02/2024 **Revised**: 11/03/2024 **Accepted**: 17/03/2024

Abstract: Given that soil composition influences nutrient cycling, biodiversity support, and water regulation, it is critical to the ecosystem's health. It reduces greenhouse gas emissions by acting as a carbon sink and storing organic carbon. Maintaining arable land, preventing sedimentation in water bodies, and managing erosion are all significantly influenced by the texture and structure of the soil. Furthermore, because some soil constituents enhance the quality of the soil and water by absorbing and digesting pollutants, soil composition has an impact on pollution remediation. It is imperative to acknowledge the significance of soil content in order to sustain ecosystems and apply sustainable land management techniques. Research is mainly focused on evaluating precision, recall, true positive rate, and F-measure in order to forecast the Organic Matter Content in soil, such as that found in homes, farms, and forests. Machine learning methods like Naive Bayes, KNN, SVM, and Random Forest are employed in this study. The outcomes demonstrate that Random Forest outperforms other algorithms in the prediction of soil organic matter content..

Keywords: *Soil types, Organic matter content, Naïve bayes, KNN, SVM, Random Forest.*

1. Introduction

Forests are pivotal in combating climate change, as they absorb carbon dioxide from the atmosphere, storing it within their biomass. However, this stored carbon is released back into the atmosphere after the burning or felling of trees, which exacerbates the greenhouse effect and contributes to global warming. For biodiversity, carbon sequestration, and ecological health, forests are essential. For efficient forest management and conservation initiatives, it is essential to comprehend the soil qualities found in forests. Conventional soil examination techniques can be expensive, time-consuming, and difficult, labour-intensive, and costly. Machine learning methods have become increasingly effective in recent years at deciphering complicated datasets and drawing insightful conclusions. In order to better understand, manage, and conserve forest ecosystems, this research investigates the use of machine learning algorithms to analyze the features of forest soil. The significance of soil properties in forest ecosystems Plant development, species distribution, and ecosystem function are influenced by the texture, moisture content, nutrient levels, and composition of the soil. Assessing the productivity, resilience, and health of forests to environmental stresses including land-use change and climate change requires an understanding of soil properties.

[1] Department of CSE, School of Computing, Saveetha School of Engineering, Thandalam, Chennai-600 124,Tamil Nadu,India
ORCID ID : 0009-0001-1138-6100
[2]Department of CSE, School of Computing, Saveetha School of Engineering, Thandalam, Chennai-600 124,Tamil Nadu,India
ORCID ID : 0000-0002-2334-2025
** Corresponding Author Email: poovizhijaya@email.com*

Conventional techniques for analyzing soil include of data interpretation by hand, laboratory analysis, and field sampling. Large amounts of soil data can be quickly analyzed using machine learning techniques, which can also be used to predict soil parameters at different geographical scales and spot patterns and relationships. Labeled soil datasets can be used to train supervised learning algorithms, such as gradient boosting, random forests, and support vector machines, to predict soil attributes based on environmental variables.

Farmers can choose the best crops to plant and improve their irrigation and fertilization techniques by having a thorough understanding of the composition of the soil in sandy areas. Estimating the amount of soil in sand helps evaluate the quality of the soil and spot possible environmental problems like contamination, salinity, or soil erosion. For construction projects in sandy areas, understanding the nature of the soil is essential for ensuring stability, preventing foundation issues, and choosing the right building materials. Land use planning decisions, including zoning laws, urban development, and conservation initiatives, are influenced by predictions about the soil content of sandy regions. By providing insight into sedimentary deposition, soil formation processes, and landscape change, soil content prediction in sand advances geological research.

Large particles and air gaps in sandy soil provide for superior drainage. This keeps soil from becoming soggy and lets extra water evaporate fast, which is good for a lot of crops—especially those that are vulnerable to root rot. In the spring, sandy soil heats up faster than other soil types. This makes it easier to plant early and encourages quicker crop

germination and growth, which lengthens the growing season. Because of its loose nature, sandy soil is typically easier to deal with and grow. It is easier to till and aerate, which lowers labor and equipment expenses. Loam or clay soils are more prone to compaction than sandy soils. Better root penetration and growth are made possible by this, which promotes healthier plants that absorb nutrients more effectively. Sand soil's loose texture lessens the possibility of soil erosion brought on by water and wind. This keeps priceless topsoil from being lost and helps to preserve soil fertility. Because sandy soil drains efficiently, it can be less friendly to some soil-borne pathogens and diseases, which can lower the frequency of plant diseases. Although sandy soil may not hold onto moisture and nutrients as well as other soil types, it can be used to grow crops—like some fruits and vegetables and drought-tolerant plants—that do well in well-drained environments. Early planting is made possible by the sandy soil's rapid spring warming, which gives farmers an advantage over other crops by starting the growing season early.

These are only a few instances of the various applications of carbon dioxide in various industries. Carbon dioxide is an essential gas for many commercial, industrial, and scientific uses in addition to being a greenhouse gas. Photosynthesis is the process by which plants, algae, and certain bacteria transform CO2, water, and sunshine into oxygen and carbohydrates. Carbon dioxide is an essential part of this process. The production of oxygen, which is necessary for breathing for both humans and other creatures, depends on this process. To promote plant development, greenhouses and other controlled environments employ carbon dioxide. Increased crop yields and photosynthesis can result from higher CO2 levels, which will improve agricultural production and food security. Beverages like soda, beer, and sparkling water are carbonated with carbon dioxide to give them the pleasing fizz and bubbles that consumers love. Medical applications for carbon dioxide include laparoscopy (minimally invasive surgery), cryotherapy (freezing tissue for medical operations), and medicinal gas mixes used as respiratory stimulants. In fire extinguishers and fire suppression systems, carbon dioxide is utilized as a fire suppressant. It smothers the flames and inhibits combustion by displacing oxygen in the vicinity. Carbon dioxide from industrial sources is sequestered and kept underground or used for enhanced oil recovery (EOR) to lower emissions into the environment in an attempt to slow down climate change. Carbon dioxide traps heat and helps to maintain a stable climate that is conducive to human habitation, even while excess CO2 in the atmosphere also contributes to global warming and climate change.

Since nitrogen is a part of proteins, chlorophyll, and nucleic acids, it is a nutrient that is necessary for plant growth. Its advantages include increased crop yields, robust plant development, and overall ecosystem productivity when nitrogen levels are appropriate. Additionally, it increases the plant's ability to withstand stress and raises the caliber of harvested goods. As a key component of soil organic matter, carbon helps with nutrient cycling, soil structure, and water retention. Its advantages include supporting soil microbial activity, organic matter decomposition, and nutrient availability with appropriate carbon levels. Soils that are high in carbon typically hold water better, have better soil structure, and are more fertile. In proposed system we are categorizing the soil types such that Forest soil, Residential soil to predict which type of soil have more good chemical composition. In Table 1 shown the Characteristic of soil type.

Table 1. Characteristics of Each Soil Type

Soil Type	**Characteristics**
Forest Soil	- Dark color due to high organic matter content
	- Thick layer of decomposed leaves, twigs, and other plant material
	- Well-aerated with a spongy texture
	- Good water infiltration and retention
	- Slightly acidic pH due to decomposition of organic matter
Agricultural Soil	- Variable color, often darker than forest soil due to organic matter from crop residues and manure
	- Tilled to improve structure and root penetration
	- Can be compacted in some areas due to plowing or machinery use
	- May contain higher levels of nutrients due to fertilization practices
Residential Soil	- Varies widely in texture and composition depending on landscaping practices and land use history
	- May be enriched with organic matter from compost or mulch
	- Can contain contaminants such as heavy metals or pollutants from urban runoff depending on surrounding land use and history

The relative amounts of sand, clay, and silt particles in the soil influence its texture, which has an impact on nutrient availability, drainage, and water retention. Good soil structure, water infiltration, and root penetration are all encouraged by soil that has an ideal ratio of sand, clay, and silt particles. It promotes healthy microbial activity, nutrient exchange, and aeration, all of which enhance plant growth

and productivity. Another kind of inorganic nitrogen that is necessary for plant nutrition and that influences soil fertility and plant growth is ammonium. Sufficient starting ammonium levels offer a quick source of nitrogen for plant uptake, encouraging early development and growth. Additionally, it helps to increase soil fertility, ecosystem productivity, and nutrient availability. Reflects the amount of nitrogen that soil microorganism's store, showing soil fertility and microbial activity. Elevated amounts of nitrogen-containing microbial biomass suggest the presence of dynamic soil microbial communities involved in nitrogen fixation, organic matter breakdown, and nutrient cycling. Plant productivity, nutrient availability, and soil fertility are all enhanced by this. The process that transforms organic nitrogen into inorganic forms (like ammonium and nitrate) that plants may absorb is known as "net nitrogen mineralization." The release of accessible nitrogen from organic matter is indicated by positive net nitrogen mineralization rates, which promote plant production and growth. It helps to enhance nutrient cycling, soil fertility, and ecosystem health. The process by which nitrifying bacteria in the soil change ammonium to nitrate, affecting the availability of nitrogen for plant uptake, is known as net nitrification.

2. Materials and Methods

The previous study attempted to categorize soil texture in photos taken with an ultraviolet fluorescent camera (UFC). The result of semantic segmentation was then broken down into smaller tiles, and to improve the robustness of the data, texture-enhancing filters were applied to a subset of the tiles using Garbon filters. This highlighted the various soil patterns in each image. Following the input of these photos into a convolutional neural network (CNN) for texture classification, the accuracy of the results significantly improved.

The prior research made an effort to classify soil texture in images captured by an ultraviolet fluorescent camera (UFC). After semantic segmentation, the output was divided into smaller tiles. To increase the data's resilience, texture-enhancing Garbon filters were then applied to a portion of the tiles. This brought to light the different soil patterns in every picture. The accuracy of the results greatly increased once these images were fed into a convolutional neural network (CNN) for texture classification. (Forkuo et al. 2018) The caliber and representativeness of the training data have a significant impact on how well the naive Bayes classifier performs. Inaccurate categorization results may arise from the study's use of biased, small-scale, or inadequately diversified data. (Myers, Montgomery 2016) There may be a lack of depth or comprehensiveness in the paper's discussion of the biotic factors affecting potato tuber yield and quality. It can leave out crucial details or not go into enough detail on the mechanics underlying the impacts that are seen. (Lookman, Alexander 2018), There may be a lack of depth or comprehensiveness in the paper's discussion of the biotic factors affecting potato tuber yield and quality. It can leave out crucial details or not go into enough detail on the mechanics underlying the impacts that are seen. (Bisgaard and Kulahci 2011) this paper may lack original contributions and may not provide novel insights or practical applications for practitioners. (D. K. Muriithi 2018) the quality and quantity of data used in the optimization process have a significant impact on its correctness and dependability. Inadequate or subpar data may jeopardize the validity of the study's conclusions and the efficiency of the optimization procedure. The study may concentrate on particular circumstances or elements affecting potato tuber yield, which could restrict the applicability of the conclusions to different geographical areas, climatic conditions, or farming methods.

2.1. Problem Identification

In the earlier work, the classification of soil texture in images taken under UFC was the main focus. Following semantic segmentation, the output is divided into smaller tiles. Texture-enhancing filters, such as the Garbon filter, are then applied to a subset of the tiles, highlighting different soil patterns in each image and boosting data robustness. After that, the images are sent into a CNN texture classification system, which greatly increases the accuracy.

2.2. System Architecture

Fig. 1 shows the information preparation process as converting raw data into a comprehensible or useable format. Pre-processing data typically consists of three stages. They are data transformation, information reduction, and information purification. Normalization of the dataset is made possible by data cleansing, or the act of removing erroneous information. The mapping of the data's homogeneity is known as transformation. Data reduction is the process of transforming data from a disorganized to a simplified state. The data are arranged in this case in an unsupervised manner. Using machine learning methods like SVM and Random Forest, we applied classification approaches to forecast the Organic Matter Content. Accurately classified instances, incorrectly classified instances, recall, precision, false positive rate, and true positive rate were among the assessment criteria. Our investigation, which was made possible by the use of the machine learning tool WEKA, sought to determine the ideal

precision.

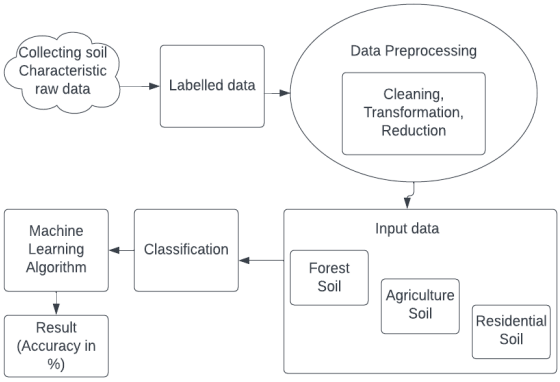

Fig 1. System Architecture

2.3. Data Collection

The Kaggle website will compile information about several categories of soil characteristics. The emphasis on agricultural, residential, and forest soil types and their corresponding advantages are shown in Figure 1. It will offer comprehensive data on the composition of the soil, including nitrogen, carbon, ammonium, nitrate, nitrite, and microbial biomass nitrogen. We have specifically gathered data on the properties of forest soil.

The year built, the age of the grass, bulk density, and past and present land uses are all considered independent variables. The dependent variables include the following: respiration, initial NO3 (+NO2), initial ammonium content, microbial biomass nitrogen, net nitrogen mineralization, and net nitrification; nitrogen percentage, carbon percentage, carbon to nitrogen ratio, nitrogen content in grams per square meter, carbon content in grams per square meter, percentage of sand, clay, and silt; and microbial biomass carbon.

The Kaggle website will be used to collect data on the various types of soil properties. The information in Fig. 1 will focus on the many types of soil that are present in agricultural, residential, and forest environments, as well as the benefits associated with each. It will include a list of the elements that make up the soil, such as carbon, microbial biomass, nitrogen, ammonium, nitrate and nitrite, and carbon. We collected information about the characteristics of the forest soil. The independent determinants are the bulk density, year of building, age of the grass, and past and present land uses. Respiration, starting NO3 (+NO2), starting ammonium content, percentages of sand, clay, and silt, nitrogen percentage, carbon percentage, and carbon to nitrogen ratio, as well as nitrogen and carbon contents in grams per square meter, are the dependent variables.

Classification is a data mining approach that helps with more accurate analysis and prediction by classifying a data set. In data mining, classification algorithms are frequently used to separate data into discrete groups. Many sectors employ classification techniques to identify the kind and category to which a given tuple belongs. Identification of microorganisms is aided by classification. Facilitates the formation of alliances between several types of organisms. Aids in comprehending the phylogeny and evolutionary history of species.

2.4. Classification

The process of classifying data into distinct classes or categories according to its attributes or features is known as machine learning classification. Machine learning classification is mostly used to automatically categorize fresh, unseen data into predetermined classes or categories. Machine learning classification algorithms use labeled training data to identify patterns, which they then use to new, unseen data to predict or classify it. These algorithms might be as basic as decision trees and logistic regression or as sophisticated as support vector machines, random forests, and neural networks. The size of the dataset, the type of data, and the required degree of accuracy or interpretability all influence the algorithm selection.

One example of a categorization challenge is the ability to identify spam in email service providers. In this classification, there are only two options: "spam" or "not spam." It is binary as a result. A classifier establishes the relationship between a given set of input variables and the class using training data. In this scenario, training data must include both known spam and non-spam emails. After the classifier is correctly trained, it can recognize an unfamiliar email. Classification is one type of supervised learning in which the objectives are also given input data. Classification can help with target marketing, credit approval, medical diagnosis, and other duties.

3. Machine Learning Algorithm

3.1. Naïve Bayes

A popular probabilistic classifier in machine learning for a variety of applications is Naive Bayes. It is frequently utilized in text mining tasks like document categorization, sentiment analysis, and spam filtering. Naive Bayes is a recommendation algorithm that can be used in recommendation systems to provide users with information or items based on past behavior or preferences.

The initial stage in applying the Naive Bayes algorithm is to gather labelled training data, which is made up of soil sample labels that correspond to different classes, including residential, agricultural, or forest. Subsequently, the pre-processed data is separated into training and testing sets, pertinent features are extracted, and categorical characteristics are encoded if needed. The class priors and class-conditional probabilities for every soil component given each class are then computed using the training data

to train the Naive Bayes classifier. The class with the highest posterior probability is chosen as the projected class label during prediction. The posterior probability of each class is computed for each new soil sample using the Naive Bayes method. Finally, the classifier is evaluated by comparing the predicted class labels with the true class labels from the testing set to measure the classification accuracy.

PSEUDOCODE:

1. Collect training data consisting of soil samples with associated class labels (e.g., forest, agricultural, residential).

2. Preprocess the data:

- Extract relevant features (soil components) from the soil samples.

- Encode categorical features if necessary.

- Split the data into training and testing sets.

3. Train the Naive Bayes classifier:

- Calculate class priors P(C) for each class (e.g., forest, agricultural, residential).

- For each class C:

- Calculate the class-conditional probabilities P(X|C) for each soil component X given class C using the training data.

- Use appropriate probability estimation methods such as Gaussian Naive Bayes for continuous features or Multinomial Naive Bayes for discrete features.

4. Predict the class label for new soil samples:

- For each new soil sample:

- Calculate the posterior probability P(C|X) for each class C using the Naive Bayes formula:

P(C|X) = P(C) * P(X1|C) * P(X2|C) * ... * P(Xn|C)

where X1, X2, ..., Xn are the soil components in the sample.

- Select the class with the highest posterior probability as the predicted class label.

5. Evaluate the classifier:

- Compare the predicted class labels with the true class labels from the testing set to measure the classification accuracy.

- Optionally, calculate other performance metrics such as precision, recall, and F1-score.

3.2. K Nearest Neighbor (KNN)

An efficient machine learning approach for both classification and regression problems is the k-Nearest Neighbors (k-NN) algorithm. The technique just commits the training dataset to memory throughout the training phase. The algorithm determines the distance between a new data point and every other point in the training dataset in order to predict the class of that new point.

The primary stage in putting the k-Nearest Neighbors (KNN) method into practice is gathering labeled training data, which consists of soil samples labeled with matching classes as residential, agricultural, or forest. After that, the data goes through preprocessing, during which pertinent characteristics are taken out of the soil samples and, if necessary, categorical features are encoded. Next, the dataset is split up into testing and training sets. The training data and related class labels are then stored, and the KNN classifier is trained. Using a selected distance metric, like the Manhattan or Euclidean distance, the algorithm determines the distance between each fresh soil sample and all training samples during the prediction phase. The k nearest neighbors are selected based on the calculated distances, and the majority class label among the k neighbors is assigned as the predicted class label for the new sample. Finally, the classifier's performance is evaluated by comparing the predicted class labels with the true class labels from the testing set to measure classification accuracy.

PSEDOCODE:

1. Collect training data consisting of soil samples with associated class labels (e.g., forest, agricultural, residential).

2. Preprocess the data:

- Extract relevant features (soil components) from the soil samples.

- Encode categorical features if necessary.

- Split the data into training and testing sets.

3. Define the distance metric:

- Choose a distance metric (e.g., Euclidean distance, Manhattan distance) to measure the similarity between soil samples.

4. Train the KNN classifier:

- Store the training data with associated class labels.

- No explicit training step is required for KNN, as it is a lazy learning algorithm.

5. Predict the class label for new soil samples:

- For each new soil sample:

- Calculate the distance between the sample and all training samples using the chosen distance metric.

- Select the k nearest neighbors based on the calculated distances.

- Determine the majority class label among the k nearest neighbors.

- Assign the majority class label as the predicted class label for the new sample.

6. Evaluate the classifier:

- Compare the predicted class labels with the true class labels from the testing set to measure the classification accuracy.

- Optionally, calculate other performance metrics such as precision, recall, and F1-score.

3.3. Support Vector Machine (SVM)

Support Vector Machines (SVMs) represent a potent category of supervised learning algorithms that find use in regression and classification problems.

To implement the Support Vector Machine (SVM) algorithm, the first step involves collecting labeled training data consisting of soil samples with associated class labels, such as forest, agricultural, or residential. The data is then preprocessed by extracting relevant features and encoding categorical features if necessary. Subsequently, the SVM model is trained by selecting a kernel function, defining the model parameters such as the regularization parameter (C), and training the model on the training data using optimization algorithms like gradient descent or quadratic programming. Once trained, the SVM model can predict the class label for new soil samples by mapping the input features to the same feature space used during training and using the trained model to predict the class label based on the decision function output. Finally, the classifier's performance is evaluated by comparing the predicted class labels with the true class labels from the testing set to measure classification accuracy.

PSEDOCODE:

1. Collect training data consisting of soil samples with associated class labels (e.g., forest, agricultural, residential).

2. Preprocess the data:

- Extract relevant features (soil components) from the soil samples.

- Encode categorical features if necessary.

- Split the data into training and testing sets.

3. Train the Random Forest classifier:

- Choose the number of trees (n_estimators) and other hyperparameters such as max_depth, min_samples_split, etc.

- For each tree in the forest:

- Randomly select a subset of features for each tree (feature bagging).

- Train the decision tree on a bootstrapped sample of the training data (bootstrap aggregating or bagging).

4. Predict the class label for new soil samples:

- For each new soil sample:

- Pass the sample through each tree in the forest and obtain a class prediction from each tree.

- Aggregate the predictions from all trees (e.g., by majority voting) to obtain the final predicted class label.

5. Evaluate the classifier:

- Compare the predicted class labels with the true class labels from the testing set to measure the classification accuracy.

- Optionally, calculate other performance metrics such as precision, recall, and F1-score.

3.4. Random Forest

Random Forest is a powerful ensemble learning method in machine learning that may be used for both classification and regression issues. It creates a huge number of decision trees during training, from which it extracts each tree's mean prediction (for regression) or mode (for classification).

To implement the Random Forest algorithm, start by collecting labeled training data containing soil samples and their corresponding class labels (e.g., forest, agricultural, residential). Preprocess the data by extracting relevant features and encoding categorical variables if necessary. Next, establish a Random Forest model by defining various hyperparameters like max_depth and min_samples_split, as well as the number of decision trees (n_estimators). Choose a subset of features at random for each decision tree in the forest to use as training data. Using a bootstrapped sample of the training data, teach each decision tree. Via combining the forecasts from each decision tree in the forest, determine the class label for fresh soil samples (e.g., via majority voting). Lastly, measure the classification accuracy of the classifier by comparing the true class labels from the testing set with the predicted class labels.

PSEDOCODE:

1. Collect training data consisting of soil samples with associated class labels (e.g., forest, agricultural, residential).

2. Preprocess the data:

- Extract relevant features (soil components) from the soil samples.

- Encode categorical features if necessary.

- Split the data into training and testing sets.

3. Train the Random Forest classifier:

- Choose the number of trees (n_estimators) and other hyperparameters such as max_depth, min_samples_split, etc.

- For each tree in the forest:

- Randomly select a subset of features for each tree (feature bagging).

- Train the decision tree on a bootstrapped sample of the training data (bootstrap aggregating or bagging).

4. Predict the class label for new soil samples:

- For each new soil sample:

- Pass the sample through each tree in the forest and obtain a class prediction from each tree.

- Aggregate the predictions from all trees (e.g., by majority voting) to obtain the final predicted class label.

5. Evaluate the classifier:

- Compare the predicted class labels with the true class labels from the testing set to measure the classification accuracy.

- Optionally, calculate other performance metrics such as precision, recall, and F1-score.

4. Experimental Result

After completing the data preprocessing outlined in Fig. 2, Weka, a machine learning program, was utilized to evaluate categorization methods for classifying soil types such as forest soil, residential soil, and agricultural soil based on components of chemical content. Table 2 presents the evaluated strategies, including metrics such as recall, precision, true positive (TP) rate, and false positive (FP) rate, considering both correctly and erroneously classified cases. The categorization techniques' output is depicted in Fig. 2.

The complexity of the dataset and the nature of the problem at hand determine which of the three algorithms—Naive Bayes, KNN, and SVM—to use. Each of these algorithms has its own set of advantages and disadvantages, and the choice of one over the other usually comes down to the particulars of the task. After a thorough analysis, Random Forest was shown to be the most accurate model, exhibiting an exceptional capacity to manage subtleties in the data. Additional proof of Random forest's efficacy in obtaining the required accuracy came from an SPSS box plot study.

The comparison of SVM, Random Forest, KNN, and Naive Bayes for classifying soil types depends on the particular problem being solved as well as the complexity of the dataset. Because each algorithm has advantages and disadvantages of its own, the choice between the two is frequently based on the particulars of the task. After a careful analysis, it was found that Random Forest outperformed the other options in terms of accuracy because of how well it handled the particular idiosyncrasies in the data. The results of the SPSS box plot analysis further affirmed the effectiveness of the Naive Bayes algorithm in achieving the desired level of accuracy.

Table 2. Measurement of Classification Techniques

Algorithm	TP rate	FP rate	Precision	Recall
Naïve bayes	0.869	0.915	0.875	0.816
KNN	0.813	0.841	0.916	0.875
SVM	0.916	0.951	0.878	0.915
Random Forest	1	0.989	0.978	1

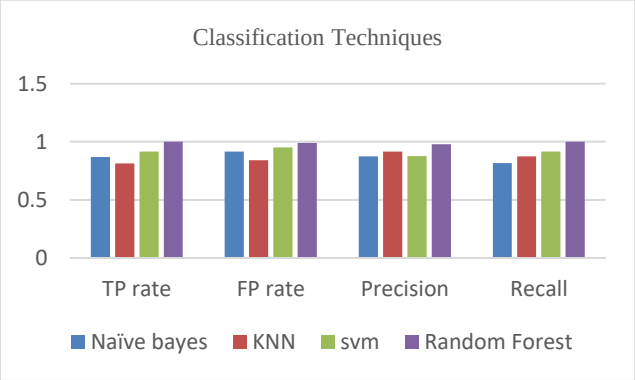

Fig 2. Classification Techniques

Furthermore, we are considering Mean absolute error, root mean square error, relatable mean square error, relatable root absolute error. When comparing the random forest algorithm to decision tables and linear regression, root mean square error yielded the best accuracy in the random forest. The measurement parameters can be found in Table 2. The output of the classified input is displayed in Fig. 2.

Table 3. Measurement of Parameters

	Forest soil	Residential Soil	Agriculture soil
Mean	83.75	83.25	87
Minimum	75	79	82
Maximum	89	90	92

Based solely on the provided statistics of mean, minimum, and maximum values for Forest Soil, Residential Soil, and Agricultural Soil, it's not appropriate to determine which soil type having more chemical content shown in Table 3. The choice of soil type depends on various factors such as the intended use, agricultural requirements, environmental conditions, and specific goals of the land management or agricultural practices. For example, agricultural soil may be

preferable if the objective is to develop crops because of its higher mean and maximum values, which may indicate superior nutrient levels or fertility for crop growth. However, because of its distinctive ecosystem functions and biodiversity, forest soil may be valued more for ecological or conservation reasons. Similarly, because of its balanced qualities, residential soil could be appropriate for urban gardening or landscaping.

Table 4. Descriptive Statistics

	Forest soil	Residential Soil	Agriculture soil	Total
Naïve bayes	75	79	82	78.67
KNN	84	85	89	86
SVM	87	79	85	83.67
Random Forest	89	90	92	90.33
Total	83.75	83.25	87	84.67

We notice variations between the performance of the Naïve Bayes, KNN, SVM, and Random Forest algorithms based on the data that are presented, which include mean, minimum, and maximum values for specific performance parameter shown in Table 4.

Based on the provided performance metrics for Naïve Bayes, KNN, SVM, and Random Forest algorithms, it is evident that Random Forest achieved the highest mean performance (90.33) compared to the other algorithms. Furthermore, Random Forest demonstrated the highest minimum and maximum values in comparison to the other algorithms, demonstrating its superior performance on a variety of assessment parameters. Consequently, it can be inferred from these numbers that Random Forest outperforms all other algorithms that have been assessed in this particular scenario. In Fig. 3 shown the graph of Descriptive Statistic.

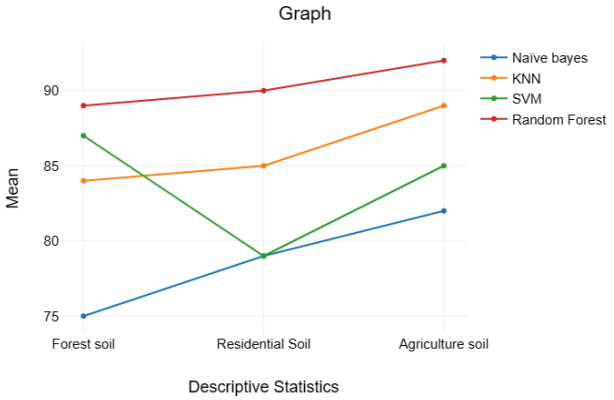

Fig 3. Graph and Descriptive Statistics

4. Conclusion

The comparison between Naïve Bayes, Support Vector Machine, K-Nearest Neighbor, and Random Forest relies on the particulars of the data and the type of problem you are attempting to solve. The decision between the four algorithms is frequently influenced by the specifics of the work at hand, as each has advantages and disadvantages. Finally Random Forest got the best accuracy.

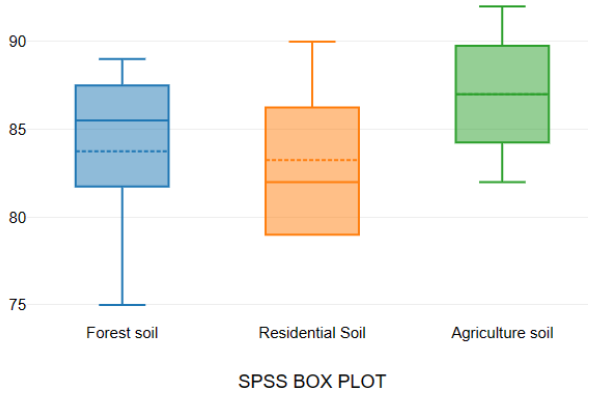

Fig 4. SPSS Box Plot

According to the experimental results displayed in Figure 4, agricultural soil has a higher chemical composition. This includes respiration, initial NO3 (+NO2), initial ammonium content, microbial biomass nitrogen, net nitrogen mineralization, and net nitrification; additionally, it has a higher percentage of carbon, nitrogen, and carbon to nitrogen ratios, as well as higher amounts of carbon, sand, clay, and silt.

In order to determine which kind of soil is ideal, we are using a soil dataset in this study to forecast the soil's good chemical composition. To do this, we have divided the different types of soil into three categories: residential, agricultural, and forest soil. We are using machine learning algorithms, such as Random Forest, SVM, KNN, and Naïve Bayes, to improve. Contrasting these algorithms to determine which is superior. The algorithm with the highest quality among the order systems was chosen to be the best algorithm. Given that Random Forest has the maximum

order-smoothness in this classification, it is regarded as the best classification computation.

5. Future work

In analyzing the loss of trees, it might be necessary in future study endeavors to take other aspects like climate change into account

Author contributions

T.Poovizhi initiated the research topic and provided guidance throughout the project. S. Christy actively participated in the design and implementation of the modelling system. All authors read and approved the final version of the article.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Bisgaard, Søren, and Murat Kulahci. 2011. *Time Series Analysis and Forecasting by Example*. John Wiley & Sons.

[2] Box, George E. P., and Norman R. Draper. 2007. *Response Surfaces, Mixtures, and Ridge Analyses*. John Wiley & Sons.

[3] Forkuo, Gloria S., Amanda N. Nieman, Revathi Kodali, Nicolas M. Zahn, Guanguan Li, M. S. Rashid Roni, Michael Rajesh Stephen, et al. 2018. "A Novel Orally Available Asthma Drug Candidate That Reduces Smooth Muscle Constriction and Inflammation by Targeting GABA Receptors in the Lung." *Molecular Pharmaceutics* 15 (5): 1766–77.

[4] Goos, Peter, and Bradley Jones. 2011. *Optimal Design of Experiments: A Case Study Approach*. John Wiley & Sons.

[5] Li, Dan-Ping, Si-Jie Cheng, Peng-Fei Cheng, Jian-Qiang Wang, and Hong-Yu Zhang. n.d. *A Novel Financial Risk Assessment Model for Companies Based on Heterogeneous Information and Aggregated Historical Data*. Infinite Study.

[6] Liu, Jiangang, Zhenjiang Zhou, and Bo Li. 2024. *Remote Sensing for Field-Based Crop Phenotyping*. Frontiers Media SA.

[7] Lookman, Turab, Francis J. Alexander, and Krishna Rajan. 2015. *Information Science for Materials Discovery and Design*. Springer.

[8] Mitran, Tarik, Ram Swaroop Meena, and Abhishek Chakraborty. 2020. *Geospatial Technologies for Crops and Soils*. Springer Nature.

[9] Myers, Raymond H., Douglas C. Montgomery, and Christine M. Anderson-Cook. 2016. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. John Wiley & Sons.

[10] Reichert, P., Dietrich Borchardt, Mogens Henze, Wolfgang Rauch, P. Shanahan, Laszlo Somlyody, and Peter A. Vanrolleghem. 2001. *River Water Quality Model*. IWA Publishing.

[11] "Sci-Hub." n.d. Accessed February 18, 2024. https://sci-hub.se/10.1109/CESYS.2018.8723956.

[12] Thenkabail, Prasad. 2018. *Remote Sensing Handbook - Three Volume Set*. CRC Press.

[13] "Website." n.d.

[14] Yada, Rickey Y. 2015. *Improving and Tailoring Enzymes for Food Quality and Functionality*. Elsevier.

[15] R. Jahan, ''Applying naive Bayes classification technique for classification of improved agricultural land soils,'' Int. J. Res. Appl. Sci. Eng. Technol., vol. 6, no. 5, pp. 189–193, May 2018.

[16] B. B. Sawicka and B. Krochmal-Marczak, ''Biotic components influencing the yield and quality of potato tubers,'' Herbalism, vol. 1, no. 3, pp. 125–136, 2017.

[17] B. Sawicka, A. H. Noaema, and A. Gáowacka, ''The predicting the size of the potato acreage as a raw material for bioethanol production,'' in Alternative Energy Sources, B. Zdunek, M. Olszáwka, Eds. Lublin, Poland: Wydawnictwo Naukowe TYGIEL, 2016, pp. 158–172.

[18] B. Sawicka, A. H. Noaema, T. S. Hameed, and B. Krochmal-Marczak, ''Biotic and abiotic factors influencing on the environment and growth of plants,'' (in Polish), in Proc. Bioróżnorodność Środowiska Znaczenie, Problemy, Wyzwania. Materiały Konferencyjne, Puławy, May 2017. [Online]. Available: https://bookcrossing.pl/ksiazka/321192

[19] R. H. Myers, D. C. Montgomery, G. G. Vining, C. M. Borror, and S. M. Kowalski, ''Response surface methodology: A retrospective and literature survey,'' J. Qual. Technol., vol. 36, no. 1, pp. 53–77, Jan. 2004.

[20] D. K. Muriithi, ''Application of response surface methodology for optimization of potato tuber yield,'' Amer. J. Theor. Appl. Statist., vol. 4, no. 4, pp. 300–304, 2015, doi: 10.11648/j.ajtas.20150404.20.

[21] M. Marenych, O. Verevska, A. Kalinichenko, and M. Dacko, ''Assessment of the impact of weather conditions on the yield of winter wheat in Ukraine in terms of regional,'' Assoc. Agricult. Agribusiness Econ. Ann. Sci., vol. 16, no. 2, pp. 183–188, 2014.

[22] J. R. Olędzki, ''The report on the state of remotesensing in Poland in 2011–2014,'' (in Polish),

Remote Sens. Environ., vol. 53, no. 2, pp. 113–174, 2015.

[23] K. Grabowska, A. Dymerska, K. Poáarska, and J. Grabowski, ‘‘Predicting of blue lupine yields based on the selected climate change scenarios,’’ Acta Agroph., vol. 23, no. 3, pp. 363–380, 2016.

[24] D. Li, Y. Miao, S. K. Gupta, C. J. Rosen, F. Yuan, C. Wang, L. Wang, and Y. Huang, ‘‘Improving potato yield prediction by combining cultivar information and UAV remote sensing data using machine learning,’’ Remote Sens., vol. 13, no. 16, p. 3322, Aug. 2021, doi: 10.3390/rs13163322.

[25] N. Chanamarn, K. Tamee, and P. Sittidech, ‘‘Stacking technique for academic achievement prediction,’’ in Proc. Int. Workshop Smart Info-Media Syst., 2016, pp. 14–17.

[26] W. Paja, K. Pancerz, and P. Grochowalski, ‘‘Generational feature elimination and some other ranking feature selection methods,’’ in Advances in Feature Selection for Data and Pattern Recognition, vol. 138. Cham, Switzerland: Springer, 2018, pp. 97–112.

[27] D. C. Duro, S. E. Franklin, and M. G. Dubé, ‘‘A comparison of pixelbased and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery,’’ Remote Sens. Environ., vol. 118, pp. 259–272, Mar. 2012.

[28] S. K. Honawad, S. S. Chinchali, K. Pawar, and P. Deshpande, ‘‘Soil classification and suitable crop prediction,’’ in Proc. Nat. Conf. Comput. Biol., Commun., Data Anal. 2017, pp. 25–29.

[29] J. You, X. Li, M. Low, D. Lobell, and S. Ermon, ‘‘Deep Gaussian process for crop yield prediction based on remote sensing data,’’ in Proc. AAAI Conf. Artif. Intell., 2017, vol. 31, no. 1, pp. 4559–4565.