Ensemble classifier to support decisions on soil classification

Sanjay Motia¹, SRN Reddy²

¹USICT, Guru Gobind Singh Indraprastha University, Dwarka, Delhi, India
²Deptt. of CSE, Indira Gandhi Delhi Technical University for Women, Kashmere Gate, Delhi, India
¹motia.sanjay@gmail.com
²rammallik@yahoo.com

Abstract. Soil performs a significant role in the agricultural ecosystem by supplying essential nutrients and a conducive environment for plants' growth and crop yield. Inside the agribusiness space, the soil classification is a crucial work that gives good classification results for different soil types. The taxonomy provides an excellent rating for inherent soil elements. This work investigates the accuracy of three well-known classification models like K-Nearest Neighbor (k-NN), Naive Bayes (NB) and, Decision Tree (DT) using a publicly available agricultural soil dataset. Post investigation, an Ensemble Classifier (EC) is proposed by fusing the above mentioned three classifiers. The experimental results indicate that EC has the highest accuracy of 84% in comparison to the NB (72.90%), k-NN (73.56%), and DT (80.84%). So it performs better than the other classifiers. The results infer that EC would be useful for accurate classification of soil types in the agricultural domain.

Keywords. Ensemble Classifier; Machine Learning Algorithms; Soil Classification, Performance Analysis, Agriculture.

1. Introduction

Soil is a dynamic, non-renewable, and most essential natural resource for the agriculture production system. It provides essential elements in the form of nutrients, water, and air to plants that help them develop, strongly grow, survive, and flourish physically. Fertile soil always serves as a foundation for healthy and nutritious crops. It meets different productive functions without any degradation and adverse effects on the environment [1]. The key factors that directly affect soil fertility are its inherent physical, biochemical, biological, and mineralogical properties [2]. The evaluation and assessment of the soil properties, as mentioned above, is usually achieved through a chemical testing of manually collected soil samples. The process is complicated and time-consuming, which justifies the need to develop methods that determine or estimate some of the properties with knowledge of previously identified features [3]. Before identification, classification of soil in different homogenous classes is the prerequisite. Soil analysis without a proper rating is like doing field experimentation with green plants or laboratory experimentation with minimum soil elements [4]. Therefore, soil classification is an essential aspect of soil analysis.

The classification process splits the soil data into different classes according to some pre-defined conditions. It also manages the systematic categorization of soils based on distinct qualities and the criteria that define the choices and options. Besides, it also helps forecast the behaviour and capacity of the land for growing crops, reduction in soil [3], diminishing the damages to the environment, and enhancement of productivity. The soil classification enhances knowledge, understanding, and results in better communication [4]. The development of a classification model that classifies the soils by using soil properties as health indicators would improve fertilizer use and reuse of farmland for different crop types.

From the last decade, ML's use in solving multifaceted problems is in high demand. ML is in use for building classification models in a variety of applications. Because of the above, this paper discusses
the implementation of an ensemble classifier to classify soil types by using machine learning (ML) techniques. The research work covers:

a) An extensive review of the ML techniques used in soil classification purposes for different applications.

b) Comparative Analysis of the performance of three well-known classifiers KNN, DT, and NB methods) to understand how accurately each model classifies the soil.

c) Development of Ensemble Classification (EC) model with higher accuracy than the other three classifiers.

The investigation helps to get the domain knowledge of soil science and ML techniques to solve various soil research problems. The section on related work discusses the outline of the relevant literature. Post capturing enough domain knowledge, a publically available soil data set from the online database is accessed and prepared maintained by ICRISAT [5]. For study purposes, the dataset undergoes standard data processing and validation steps followed by a dataset division into 80:20 ratios of training and test dataset from the processed dataset. Three classifiers, i.e., KNN, DT, and NB method, have been implemented and their performance in terms of resultant classification accuracy is analysed and compared. The presented work also discusses developing an ensemble or hybrid classifier using the fusion of selected three ML algorithms and its performance compared with the rest of the three methods.

The paper is arranged into sections where Section 2 discusses the comprehensive review of the existing work supported by the motivation to conduct the research work (section 3). Section 4 explains the materials and methods adopted to perform the research work that includes a brief description of the data set, tools, and classifiers used in the presented work. This section also compares the performance of existing classifiers with each other. The paper's main objective is to design and develop a novel ensemble classifier, the fusion of existing well-known classifiers. Section 5 contains the complete details of the proposed algorithm, followed by a comparative analysis of the same with three well-known classifiers. Section 6 elaborates on the discussion on results obtained and the last part, i.e., section 7, discusses the concluding remarks with the future scope of research.

**Analogous Work**

Nowadays, ML algorithms are showing tremendous potential in solving numerous problems in soil research and analysis [6]. The wetting/drying process [7] of soil was modelled by using KNN and BP techniques with 91–94% accuracy. Through this kind of approach, farming the choices can be empowered distantly, without the time and cost of on-location appearances. The SVM was used to estimate the soil's physio-chemical properties and the soil type classification [3],[26] An MDT method proposed by Shastry et al. [8] is more accurate in comparison to C4.5 and CART algorithms. The performance of a suite of ten ML-based classifiers (CART, CART with bagging, RF, KNN, NSC, ANN, MLR, LMT, and SVM) was compared for classification in soil analysis [9]. In addition, the taxonomic units of soil were also estimated that might assist in futuristic DSM and GM [9]. Gambill et al. [10] estimated the classification of soil using soil properties in USCS via RF models that results into an overall PER of less than 5%.

An image processing and SVM based efficient classifier for soil was developed and tested by Chandan and Thakur [11] using soil surface images. About seven categories of soil, namely clay, peat, silty sand, variants of clay soil (humus, sandy, peat, sand clay), were explored for soil classification. The performance of four supervised ML classifiers k-NN, ANN, SVM, and RF have been compared for the classification of crop types and enhanced land cover for better agriculture production monitoring and farmland use patterns [12]. In this case study, SVM and RF are the most robust methods whereas SVM is the most (94.4%), and k-NN was the least (92%) accurate classifier, especially for crop types [12]. A model utilizes k-NN, BT, and GKSVM for land type classification and prediction of soil series [13]. In addition, the model also recommends suitable crop type for given land type. A collection of twenty very diverse classifiers of families bagging, boosting, DT, NN, NNet, RF, and SVM have been used for the solution of ten classification problems [14]. The RF performs better than others in six
problems out of ten. The other techniques, like AdaBoost, SVM, and GELM, follow RF i.e., they have less accuracy as compared to RF [14].

From the discussion given above, it is evident that soil classification is an area of soil research in which much has already been done, and a lot needs to be done. The classification of soil is useful in many applications, such as to suggest a suitable amount of fertilizers and suitable crop and identification of village wise fertility indices [14]. To fulfil the needs of accurate classification, classifiers with better and accurate performance are the utmost need in the soil science domain. The next subsection discusses the motivation behind the present research work. The various abbreviations and their full description is given in Table 2 in appendix section.

Motivation
Within the agriculture sector, soil degradation is becoming a global threat due to various reasons like improper crop rotation, nutrient leaching, urbanization, removal of trees and vegetation, high grazing, and the most critical imbalance in fertilizers and other nutrient supplements [15]. Therefore, to handle the global problem of soil degradation and to achieve global food security targets, the assessment and management of soil quality or health are one of the mandatory need [16]. The managed soil health would lead to a reduction in the use of agrochemical supplements and the cost involved in various farming activities.

One of the reasons for soil degradation is the imbalanced use of fertilizers and nutrient supplements, which is basically driven by a lack of knowledge of farmers about the crop-specific needs of soils. Farmers in most of the developing countries are doing farming activities as per their traditional farming practices. They are generally unaware of the soil conditions of their farms, and in their desires to get higher yields, fertilizers are excessively applied, which not only degrades the soil conditions but also leads to harmful impacts on the environment. Therefore, awareness should be given to farmers about soil classification and assessment for better soil management. This leads the need to develop an accurate soil classifier that can answer the following questions.

a) How are the different soil types classified from a publically available soil dataset?

b) What are the widely used and popular ML-based classification models?

c) How a new classifier with higher classification accuracy, be modeled?

The next sections of the paper try to answer all of the above questions one by one. The next subsequent section shall explore the well-known classifiers and presents an overview of each respective method.

Materials and Methods

1.1. Dataset Collection.
The first and foremost requirement of this research work is collecting the soil dataset from the public domain repository. Therefore, an online data repository managed by the ICRISAT is accessed to get the soil dataset. ICRISAT is a non-political and non-profit organization that facilitates agricultural research for development in Asia and sub-Saharan Africa [5]. It also provides technical assistance to the Govt. of AP (Andhra Pradesh) in the development of primary sector. The dataset contains field-level data of soil health with spatial distribution in pilot sites in 13 districts of AP. The dataset consists of soil parameter details, i.e., chemical properties, previous crop, and location/farm owner details from different locations [17].

Tools and Computing Resources
All the implementation work and experimentation work is conducted using Intel® Core™ i3-6006U CPU (x64 based processor) @ 2.00 GHz notebooks with 8 GB RAM and 64-bit OS. An open source, simple, efficient and tool multiprocessing packages of python like scikit-learn and ‘pandas’ library is used for data processing, development of ensemble approach and analysis of results.

1.2. Overview of Classifiers
Three state of art and popular ML based classifiers have been chosen for preliminary classification of soil type from the soil properties as given in the dataset. Each classifier model is implemented and evaluated with same dataset for the performance evaluation and calculation of accuracy.

1.2.1. **K-Nearest Neighbours (KNN).** k-NN is one of the most popular supervised learning classifier for various applications [12], [18]. As evident from name It identifies the nearest neighbour k samples from the training data to the targeted sample point using Euclidean distance (ED) [19]. Let us assume that p and q are samples that may have ‘n’ dimensions in the feature space. The ED (p, q) in between the samples p and q is defined by equation (1).

\[
ED(p, q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}
\]  

1.2.2. **Decision Trees (DT).** DT is the potentially accurate classifier in many application contexts [20]. It utilizes different subsets of feature and decision rules at various instances of classification. DT secures information as a tree, which revamps many decision rules make it more explicit. As the name reveals, DT is a tree like structure with one root node, internal nodes, branches and many leaf nodes (LN) that represents the class designated to a sample [25]. Each internal node of a tree depicts a feature, and branches corresponds to conjunctions of elements which lead to classifications [21]. Based on the approach of tree construction, the DT model may vary. ID3 and CART are the most commonly used DT classifiers.

1.2.3. **Naïve Bayesian (NB).** NB is a probabilistic classifier that utilizes Bayes' theorem with strong independence (naive) assumptions and a small training dataset to estimate the target parameters [22]. The naive assumption supports the no-relationship theory between the presence or absence of each feature with other features [23], [25]. Within NB, the object ‘C’ given in equation (2) is the function of the prior probability \( P(C_j) \) of class \( C_j \) and the likelihood \( P(f_i|C_j) \) of class \( C_j \) w.r.t. feature \( f_i \) where \( P(f_i|C_j) \) is calculated from the ratio of count \((f_i, C_j)\) and count \((C_j)\) [24].

\[
C = \arg \max_{C_j} P(C_j) \prod_i P(f_i|C_j)
\]  

**Approach for Ensemble Classifier Development**

In any domain-oriented and data-driven research, getting the domain knowledge of targeted application and data collection is the pre-requisite. The study of many research papers helps in acquiring the domain knowledge of agriculture and research related to soil physio-chemical and nutrient properties and related applications. The soil dataset from the reliable source acts as a catalyst for the research work as getting a nutrient dataset is quite tricky. The physio-chemical properties of the soil include soil type, details of macro and micro nutrients, texture, previously grown crops. All of these parameters are useful in building intelligent models. Figure-1 depicts the methodological approach adopted for building the ensemble classifier (EC) model for soil data classification. The pictorial representation will help in easy understanding of entire methodology being used in development of classifier.
1.3. Dataset Preparation
Post collection of data, the next step is to prepare a dataset for pre-processing, which includes manual correction and transformation. The manual correction covers manual removal of irrelevant attributes and transformation of the dataset into a file format (xls or CSV) that is required by the analysis tool. In the collected dataset, the details of farmers (name, etc.) and address information (district, village) are removed as the analysis is more concentrated in numerical data analysis. Figure-2 depicts the various features in the processed dataset.

| Soil type | pH   | EC   | OC  | Avail-P | Exch-K | Avail-Ca | Avail-Mg | Avail-S | Avail-Zn | Avail-Fe | Avail-B | Avail-Cu | Avail-Mn | Classes |
|-----------|------|------|-----|---------|--------|----------|----------|--------|----------|----------|--------|----------|----------|---------|
| 1         | 6.19 | 0.07 | 0.18 | 7.13    | 41     | 587      | 101      | 5.16   | 0.3      | 0.17     | 0.49   | 0.51     | 15.24    | 1       |
| 2         | 8.4  | 0.33 | 0.31 | 10.34   | 102    | 811      | 261      | 9.91   | 0.16     | 0.57     | 3.24   | 0.44     | 6.9      | 1       |
| 1         | 7.1  | 0.11 | 0.17 | 8.46    | 46     | 582      | 48       | 3.77   | 0.37     | 0.19     | 5.54   | 0.42     | 8.34     | 1       |
| 2         | 7.78 | 0.14 | 0.21 | 5.68    | 44     | 5161     | 71       | 2.2    | 0.99     | 2.38     | 2.83   | 0.64     | 4.41     | 1       |
| 1         | 6.3  | 0.06 | 0.28 | 2.08    | 52     | 612      | 122      | 4.54   | 0.59     | 0.28     | 7.11   | 0.56     | 17.07    | 1       |

1.4. Data Pre-processing
Post data collection and preparation, the dataset is pre-processed for data cleaning purposes where data is cleaned via missing value imputation (MVI) and removal of outliers. Post cleaning a normalized dataset is obtained. For the cleaning portion, the classes which are not dataset are of no use. All such samples are removed. The ‘fillna()’ function of the pandas library is used to fill NA/NaN values. The missingness is handled using simple mean imputation method.
1.5. Dimensionality Reduction

The normalized dataset is processed for relationship analysis through the computation of correlation among different attributes of the dataset. The relationship analysis is followed by feature extraction (FE). FE involves the extraction of all the features that are required by the algorithm to classify the soil type. FE is followed by feature selection (FS). The FS is a data dimensionality reduction process that is generally used in ML techniques. It helps in holding the hidden irrelevant data and also suggests productive information mining. Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA) and Independent Component Analysis (ICA) are the popular strategies for attribute selection. The relationship or correlation is the most basic choice for element selection in the multi-attribute selection.

1.6. Data Splitting

This step involves the splitting of the processed dataset into 80:20 ratios of training dataset (80%) and the test dataset (20%) respectively.

1.7. Application of state of art classifiers

The next important step is to apply the classifiers for handling multi-attribute soil classification problems. This is achieved through the application of classifiers (k-NN, DT, and NB) on the dataset and comparative analysis of their performance measurement in terms of classification accuracy. The implementation of different algorithms is done using the open-source Python programming language. In DT, ‘entropy’ is the criterion, and 60 is the minimum samples for the leaf. In k-NN, five is taken as the no of nearest neighbours. On examined three classification techniques (Decision Tree Classifier, KNN, Naive Bayes) individually the models were giving an accuracy of (80.84%, 73.56% and, 72.90%) respectively.

1.8. Development and Evaluation of Ensemble Classifier (EC) Classifier

Post-implementation of different shortlisted methods, the hybrid or ensemble or fusion classifier is proposed and implemented, followed by an evaluation of its performance with the rest of the shortlisted methods. The implementation of the proposed classifier is done using the python language and its libraries like pandas, sklearn, and pydotplus. The dataset used in the development of the proposed model has a set of a "Class" associated with every "data point". In the fused model, the classification model is trained to identify "Class" of the sample. Every sample in the dataset is processed parallel and the majority voting in case of a tie, the DT classifier model, is given the upper hand.

1.9. Overall Classifier Performance Evaluation

Table-1 displays the accuracies of each classification model. The three classifiers (DT, KNN, NB), when applied individually to the agricultural dataset, results in an accuracy of 80.84%, 73.56% and, 72.90%, respectively. As shown, the DT algorithm obtained the highest accuracy. The most significant features among the attributes are the fifteen (15) selected soil nutrient features Table 1, including the crop type as the class. The proposed ensemble classifier (EC) gives the highest accuracy of 84.14% in classification of soil. Figure-3 depicts the variation of accuracy of the proposed ensemble classifier in comparison to other classifiers.

| Classifier        | Accuracy |
|-------------------|----------|
| DT                | 80.84%   |
| k-NN              | 73.56%   |
| NB                | 72.90%   |
| Ensemble Classifier | 84.14%   |
2. Results and Discussion
The results summary is tabulated in Table 1. From the experiments, the NB algorithm had the lowest accuracy of 72.90%, and the DT had the most outstanding accuracy of 80.84%. This is the individual performance of each algorithm when applied to the dataset separately. The fused or proposed algorithm, when applied to the same dataset, results in higher accuracy of 84.14% as compared to the remaining three classifiers. Therefore, it proves itself as the outstanding performer method for classification of soil type.

3. Conclusion and Future Scope
From the study, discussion and analysis of results given in this paper, firstly it is concluded that ML is an efficient approach for classification of soil. On analysing the results and performance, the proposed EC was the outstanding performer in terms of accuracy in comparison to the popular classifiers like DT, KNN and NB. For the studies related to agricultural soils, EC is quite capable and may be a better choice. The future work will focus on the inclusion of datasets with more soil quality indicators, usage of ML based missing value imputers and addition of more number of classifiers in fusion implemented for EC to further increase its accuracy. In addition, more no of classes of soil type in dataset will also help in making the classifier most robust.

Appendix
The various abbreviations and their full description is given in Table 2.

| Abbreviation | Full Description |
|--------------|-----------------|
| KNN: K-Nearest Neighbor | DT: Decision Trees, SVM: Support Vector Machine, RF: Random Forest, ANN: Artificial Neural Network, PCA: Principle Component Analysis, BP: Boosted Perceptron, MDT: Modified Decision Tree, NSC: Nearest Shrunken Centroid, MLR: Multinomial Logistic Regression, LMT: Logistic Model Trees, DSM: Digital Soil Mapping, GM: Geomorphic Modeling, CART: Classification and Regression Tree, USCS: Unified Soil Classification System, PER: Prediction Error Rate, BT: Bagged Trees, GKSVM: Gaussian kernel-based SVM, NN: Nearest Neighbors, NNet: Neural networks, GELM: Gaussian Extreme Learning Machine, ICRISAT: International Crops Research Institute for the Semi-Arid Tropics |
References

[1] El-Ramady HR, Alshaal TA, Amer M, Domokos-Szabolcsy É, Elhawat N, Prokisch J, Fári M. Soil quality and plant nutrition. In Sustainable Agriculture Reviews 14 2014; pp. 345-447. Springer, Cham.

[2] Karlen DL, Ditzler CA, Andrews SS. Soil quality: why and how?. Geoderma. 2003 Jun 1;114(3-4):145-56.

[3] Kovačević M, Bajat B, Gajić B. Soil type classification and estimation of soil properties using support vector machines. Geoderma. 2010 Jan 15;154(3-4):340-7.

[4] Hartemink AE. The use of soil classification in journal papers between 1975 and 2014. Geoderma Regional. 2015 Aug 1;1:127-39.

[5] https://www.icrisat.org/

[6] Padarian J, Minasny B, McBratney AB. Machine learning and soil sciences: A review aided by machine learning tools. Soil. 2020;6(1):35-52.

[7] Coopersmith EJ, Minsker BS, Wenzel CE, Gilmore BJ. Machine learning assessments of soil drying for agricultural planning. Computers and electronics in agriculture. 2014 Jun 1;104:93-104.

[8] Shastry KA, Sanjay HA, Kayva H. A novel data mining approach for soil classification. In 2014 9th International Conference on Computer Science & Education; 2014 Aug 22 (pp. 93-98). IEEE.

[9] Heung B, Ho HC, Zhang J, Knudby A, Bulmer CE, Schmidt MG. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma. 2016 Mar 1;265:62-77.

[10] Gambill DR, Wall WA, Fulton AJ, Howard HR. Predicting USCS soil classification from soil property variables using Random Forest. Journal of Terramechanics. 2016 Jun 1;65:85-92.

[11] Chandan, Thakur, R. An Intelligent Model for Indian Soil Classification using various Machine Learning Techniques. International Journal of Computational Engineering Research (IICER). 2018; Volume, 08, Issue, 9, Page 33-41.

[12] Laban N, Abdellatif B, Ebeid HM, Shedeed HA, Tolba MF. Machine Learning for Enhancement Land Cover and Crop Types Classification. In Machine Learning Paradigms: Theory and Application 2019; pp. 71-87. Springer, Cham.

[13] Rahman SA, Mitra KC, Islam SM. Soil classification using machine learning methods and crop suggestion based on soil series. In 2018 21st International Conference of Computer and Information Technology (ICICT) 2018; Dec 21 (pp. 1-4). IEEE.

[14] Sirsat MS, Cernadas E, Fernández-Delgado M, Khan R. Classification of agricultural soil parameters in India. Computers and electronics in agriculture. 2017; Apr 1;135:269-79.

[15] Karlen DL and Charles WR. Soil degradation: Will humankind ever learn?. 2015: 12490-12501.

[16] Lenka NK, Jaiswal SP, Thakur JK, Lenka S, Mandal A, Dwivedi AK, Lakaria BL, Biswas AK, Shukla AK, Yashona DS. Soil degradation effect on soil productivity, carbon pools and soil enzyme activity. CURRENT SCIENCE. 2017; Jun 25;112(12):2434.

[17] ICRISAT Development Center; Government of Andhra Pradesh, 2016, Soil health data - Andhra Pradesh Primary Sector Mission, https://doi.org/10.21421/D2/K3BPKW, ICRISAT Dataverse, V3.

[18] Shastry KA, Sanjay HA, Deexith G. Quadratic-radial-basis-function-kernel for classifying multi-class agricultural datasets with continuous attributes. Applied Soft Computing. 2017; Sep 1;58:65-74.

[19] Ertuğrul ÖF, Tağluk ME. A novel version of k nearest neighbor: Dependent nearest neighbor. Applied Soft Computing. 2017; Jun 1;55:480-90.

[20] Kingsford C, Salzberg SL. What are decision trees?. Nature biotechnology. 2008; Sep;26(9):1011-3.

[21] Che D, Liu Q, Rasheed K, Tao X. Decision tree and ensemble learning algorithms with their applications in bioinformatics. In Software tools and algorithms for biological systems. 2011;
[22] Al-Aidaroos KM, Bakar AA, Othman Z. Naive Bayes variants in classification learning. In 2010 International Conference on Information Retrieval & Knowledge Management (CAMP). 2010; Mar 17, pp. 276-281, IEEE.

[23] Bhargavi P, Jyothi S. Applying naive bayes data mining technique for classification of agricultural land soils. International journal of computer science and network security. 2009; Aug;9(8):117-22.

[24] Trovato G, Chrupała G, Takanishi A. Application of the naive bayes classifier for representation and use of heterogeneous and incomplete knowledge in social robotics. Robotics. 2016; Mar; 5(1):6.

[25] Abonyi J, Feil B, Abraham A. Computational intelligence in data mining. Informatica. 2005;29(1).

[26] Daryati D, Widiasanti I, Septiandini E, Ramadhan MA, Sambowo KA, Purnomo A. Soil characteristics analysis based on the unified soil classification system. In Journal of Physics: Conference Series 2019; Dec (Vol. 1402, No. 2, p. 022028). IOP Publishing.