Journal of Agricultural Sciences
(Tarim Bilimleri Dergisi)

J Agr Sci-Tarim Bili
e-ISSN: 2148-9297
jas.ankara.edu.tr

JAS

A Novel Machine Learning Approach: Soil Temperature Ordinal Classification (STOC)

Cansel KUCUK[a], **Derya BIRANT**[b], **Pelin YILDIRIM TASER**[c*]

[a]*Dokuz Eylul University, Graduate School of Natural and Applied Sciences, Izmir, TURKEY*
[b]*Dokuz Eylul University, Department of Computer Engineering, Izmir, TURKEY*
[c]*Izmir Bakircay University, Department of Computer Engineering, Izmir, TURKEY*

ARTICLE INFO
Research Article
Corresponding Author: Pelin YILDIRIM TASER, E-mail: pelin.taser@bakircay.edu.tr
Received: 21 January 2020; Revised: 17 November 2021, Accepted: 17 November 2021 / Online: 15 October 2022

Cite this article
KUCUK C, BIRANT D, TASER YILDIRIM P (2022). A Novel Machine Learning Approach: Soil Temperature Ordinal Classification (STOC). *Journal of Agricultural Sciences (Tarim Bilimleri Dergisi),* 28(4):635-649. DOI: 10.15832/ankutbd.866045

ABSTRACT

Soil temperature prediction is an important task since soil temperature plays an important role in agriculture and land use. Although some progress has been made in this area, the existing methods provide a regression or nominal classification task. However, ordinal classification is yet to be explored. To bridge the gap, this paper proposes a novel approach: *Soil Temperature Ordinal Classification* (STOC), which considers the relationships between the class labels during soil temperature level prediction. To demonstrate the effectiveness of the proposed approach, the STOC method using five different traditional machine learning methods (Decision Tree, Naive Bayes, K-Nearest Neighbors, Support Vector Machines, and Random Forest) was applied on daily values of meteorological and soil data obtained from 16 stations in three states (Utah, Alabama, and New Mexico) of United States at five soil depths (2, 4, 8, 20, and 40 inches) between the years of 2011 and 2020. The experiments show that the proposed STOC approach is an efficient method for soil temperature level (very low, low, medium, high, and very high) prediction. The applied STOC models (STOC.DT, STOC.NB, STOC.KNN, STOC.SVM, and STOC.RF) showed average accuracy rates of 90.95%, 77.09%, 90.84%, 89.94%, and 90.91% on the experimental datasets, respectively. It was observed from the experimental results that the STOC.DT method achieved the best soil temperature level prediction among the others.

Keywords: Agriculture, Classification, Decision tree, Machine learning, Random forest, Soil temperature level

1. Introduction

Soil temperature greatly influences plant growth and development, soil water and salt transport, soil carbon balance, and microbial activity and chemical reactions inside the soil (Onwuka & Mang 2018). Biochemical processes, such as germination of a seed, seedling emergence, uptake operation by roots, root growth, are realized under the suitable soil temperature. The high temperature in the soil can lead to a dramatic reduction of soil resistance to physical events such as erosion and subsidence. On the other hand, some biochemical activities can stop when the temperature drops dramatically and the soil freezes. In addition, a significant decrease in temperature changes the rate of organic matter decomposition and mineralization inside the soil. If the soil temperature is suitable, chemical activities in the soil continue smoothly and regularly. Soil temperature prediction is of great importance for this purpose. If further changes in soil temperature are predicted, new strategies can be developed to prevent undesired situations.

Machine learning has recently received much attention in the soil temperature prediction field (Abyaneh et al. 2016; Li et al. 2020a; Sattari et al. 2020). It discovers significant underlying patterns in raw data by constructing a model without any human intervention. The most widely studied machine learning technique is classification. Classification is the task of categorizing an input sample data into one of the predefined classes. In the literature, there are several soil temperature prediction studies that apply different machine learning techniques, such as regression (Abyaneh et al. 2016; Alizamir et al. 2020a; Alizamir et al. 2020b; Li et al. 2020b; Sattari et al. 2020) and time series (Mehdizadeh et al. 2020; Zeynoddin et al. 2020). For example, (Alizamir et al. 2020b) proposed a deep echo state network (Deep ESN) regression model for soil temperature prediction at 10 and 20 cm depths. In another study, a new time-series model, called fractionally autoregressive integrated moving average (FARIMA), was implemented for daily soil temperature prediction at four depths (5, 10, 50, and 100 cm) (Mehdizadeh et al. 2020).

In the previous machine learning-based soil temperature prediction studies, it is assumed that there is no order between the class labels of the dataset which will be predicted. However, the class attributes of some soil data have an inherent order. Due to this situation, this study proposes the ordinal classification technique for the first time, which considers natural order between class labels to solve the problem of soil temperature level prediction.

Ordinal classification is a special type of classification which assumes the class values of the attributes in the dataset are ordered and finite (Frank & Hall 2001). The ordinal classification aims to predict the target attribute of a given new sample by considering ranking relationships between the classes. For example, the soil temperature values can be categorized as "very low", "low", "medium", "high", and "very high". The order of these values is stated as "very low" < "low" < "medium" < "high" < "very high". The prediction performance of the machine learning models can be improved regarding the ranking relationship between the class labels of the used experimental dataset. Considering this motivation, a novel ordinal classification method, named STOC, was applied to real-world soil temperature data at five different soil depths (2, 4, 8, 20, and 40 inches) in this study.

Due to the nature of the problem of soil temperature level prediction, datasets are typically ordinal. Disregarding the orders among class labels, the problem is basically treated as a multi-class classification task, and it often results in a loss of performance. A possible way out is to develop a novel approach that makes use of order relations for samples during the classification task. Exploring such ranking information can usually help to increase the effectiveness of predictive classifiers. In this study, the order information was incorporated into the classification method, thus exploiting the implicit and relative knowledge simultaneously. More specifically, this study brings together notions from the fields of agriculture and machine learning for information fusion.

Soil temperature has a great impact on the growth and development of all types of plants. Because of this reason, soil temperature prediction plays an important role in the agriculture field. The machine learning-based soil temperature prediction studies in the literature generally apply regression (Kisi et al. 2015; Abyaneh et al. 2016; Citakoglu 2017; Mehdizadeh et al. 2018; Sanikhani et al. 2018; Xing et al. 2018; Delbari et al. 2019; Feng et al. 2019; Alizamir et al. 2020a; Alizamir et al. 2020b; Hao et al. 2020; Li et al. 2020a; Li et al. 2020b; Penghui et al. 2020; Sattari et al. 2020; Shamshirband et al. 2020; Tsai et al. 2020; Abimbola et al. 2021; Bayatvarkeshi et al. 2021; Wang et al. 2021) or time series (Bonakdari et al. 2019; Li et al. 2020c; Mehdizadeh et al. 2020; Nanda et al. 2020; Zeynoddin et al. 2020) methods.

In the regression studies, DT (Sanikhani et al. 2018; Sattari et al. 2020), Support Vector Regression (SVR) (Mehdizadeh et al. 2018; Xing et al. 2018; Delbari et al. 2019; Li et al. 2020a; Li et al. 2020b; Shamshirband et al. 2020), RF (Feng et al. 2019; Alizamir et al. 2020b; Tsai et al. 2020), NN (Kisi et al. 2015; Abyaneh et al. 2016; Citakoglu 2017; Hao et al. 2020; Penghui et al. 2020; Abimbola et al. 2021; Bayatvarkeshi et al. 2021; Wang et al. 2021), ELM (Alizamir et al. 2020a) algorithms have been preferred for predicting soil temperatures. In addition, some of the time-series studies have also applied NN (Bonakdari et al. 2019; Li et al. 2020c), ELM (Mehdizadeh et al. 2020; Zeynoddin et al. 2020), SVR (Nanda et al. 2020) algorithms for the prediction performance comparison.

Most of the soil temperature prediction studies are performed at multiple depths (Kisi et al. 2015; Sanikhani et al. 2018; Shamshirband et al. 2020). Sanikhani et al. (2018) applied extreme learning machine (ELM), neural network (NN), and M5 Model Tree (M5 Tree) models on the meteorological data obtained from two stations in Turkey for predicting soil temperatures at 5, 50, and 100 cm depths. The studies also estimate the features that affect the soil temperature prediction (Feng et al. 2019; Nanda et al. 2020). Nanda et al. (2020) discovered that while rainfall data does not affect the prediction performance, the soil moisture parameter improves the accuracy of the prediction problem.

The present studies assumed that there is no order between the target attribute of the soil dataset, which will be predicted. However, the target attribute values have a natural order, such as very high, high, medium, low, very low, and it can affect the prediction performance of the soil temperature level. Because of this reason, the STOC approach, which takes into account the inherent order of class labels, is proposed to classify the soil temperature levels of the experimental dataset.

Table 1 shows the comparison of this study with the existing studies. This method differs from the existing methods in three respects. First, they performed the prediction of numeric target values, while this paper focuses on the classification of ordinal and categorical target values. Second, this study uses different methods such as KNN and NB. Third, in this study, the performance was tested with some metrics different from those used by aforementioned studies.

Table 1- Comparison of this study with the existing studies.

Reference	*Year*	*Algorithm*							*Task*			*MD*	*Country*
		DT	*SVR*	*NB*	*KNN*	*RF*	*NN*	*ELM*	*R*	*T*	*O*		
Abimbola et al. 2021	2021						√		√			√	USA
Bayatvarkeshi et al. 2021	2021						√		√			√	Iran
Wang et al. 2021	2021						√		√			√	Switzerland
Alizamir et al. 2020a	2020	√					√	√	√			√	Turkey
Alizamir et al. 2020b	2020	√				√	√		√			√	USA
Hao et al. 2020	2020						√		√			√	Switzerland
Sattari et al. 2020	2020	√							√			√	Turkey
Li et al. 2020a	2020		√			√	√		√				USA
Li et al. 2020b	2020		√				√	√	√				China
Li et al. 2020c	2020						√			√		√	Switzerland
Zeynoddin et al. 2020	2020							√		√		√	Iran
Mehdizadeh et al. 2020	2020						√			√		√	Iran
Nanda et al. 2020	2020		√			√	√			√			India
Shamshirband et al. 2020	2020		√				√		√			√	Iran
Penghui et al. 2020	2020						√		√				USA
Tsai et al. 2020	2020					√			√				Taiwan
Bonakdari et al. 2019	2019						√			√		√	USA
Delbari et al. 2019	2019		√				√		√			√	Iran
Feng et al. 2019	2019					√	√	√	√				China
Mehdizadeh et al. 2018	2018		√						√			√	Iran
Sanikhani et al. 2018	2018	√					√	√	√			√	Turkey
Xing et al. 2018	2018		√						√			√	USA
Citakoglu 2017	2017						√		√			√	Turkey
Abyaneh et al. 2016	2016						√		√			√	Iran
Kisi et al. 2015	2015						√		√			√	Turkey
Proposed approach		√	√	√	√	√	√				√	√	USA

(R=Regression, T=Time series, O=Ordinal Classification, MD=Multiple-Depth)

The novelty and main contributions of this work are as follows. (i) It is the first attempt to apply ordinal classification for soil temperature level prediction. (ii) This paper proposes a novel method, called STOC, which takes into account the relationships between the class labels during soil temperature level prediction. (iii) This study is also original in that it compares alternative base learners in conjunction with the proposed method, including decision tree (DT), Naive Bayes (NB), k-nearest neighbors (KNN), support vector machines (SVM), and random forest (RF). (iv) This is the first study using ordinal classification to predict soil temperature levels at five different soil depths. The main motivation of this study is to improve the performance of the soil temperature level prediction of the models by considering the ranking among the class labels.

2. Material and Methods

2.1. Dataset description

In this study, 16 different experimental datasets which contain daily values of meteorological and soil data obtained from 16 stations in three states (Utah, Alabama, and New Mexico) of United States at five soil depths at the interval of 01.01.2011 and 31.05.2020. Figure 1 presents the map of the stations in each state of the United States.

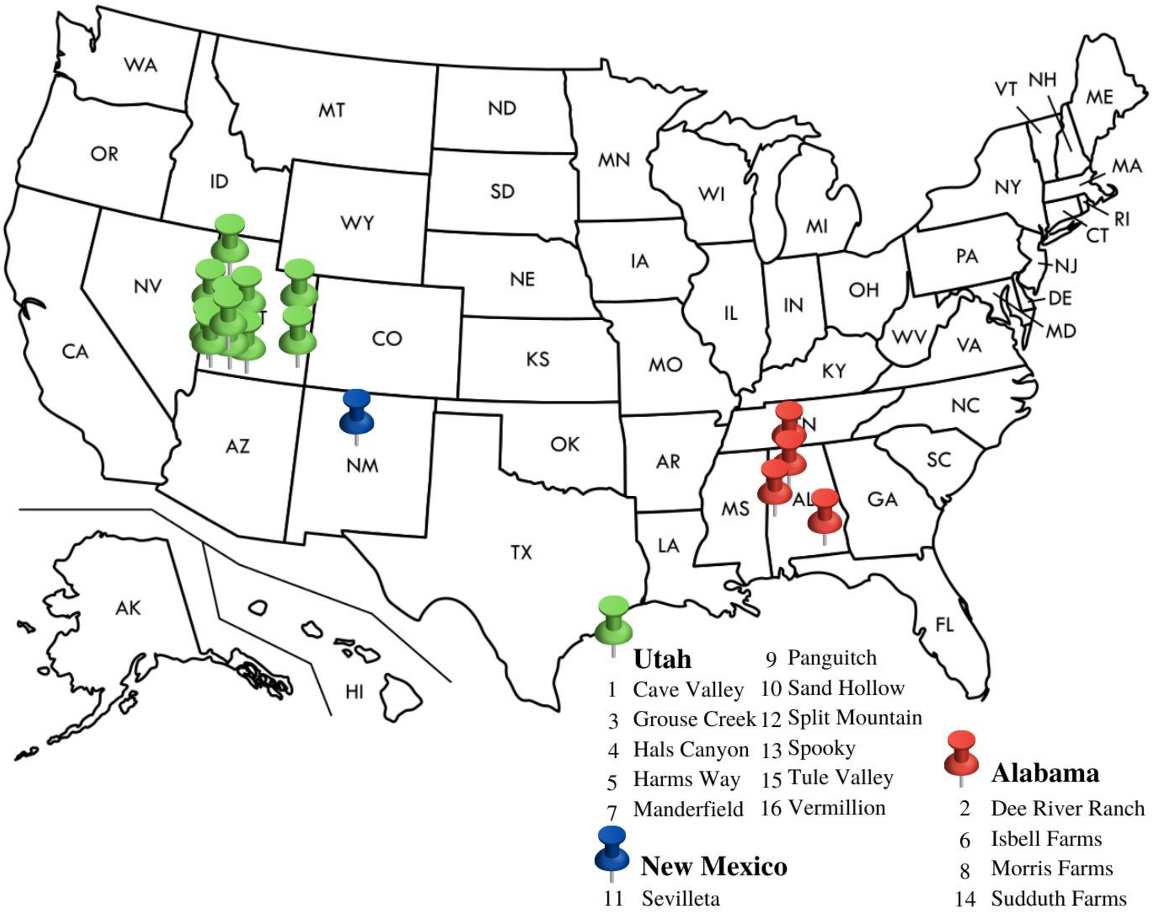


Figure 1- The map of the stations used in this study

Each data set in this study includes meteorological parameters such as air temperature, precipitation, relative humidity, solar radiation, wind speed, and vapor pressure, and soil parameters such as moisture and temperature. These real-world and publicly available datasets were gathered from the website (https://www.wcc.nrcs.usda.gov) of the National Water and Climate Center (NWCC) in the National Resources Conservation Service (NRCS) under the United States Department of Agriculture. Table 2 presents the details of meteorological and soil parameters for each dataset.

Table 2- Parameters of the datasets

Type	*Parameters*	*Unit*
Meteorological Parameters	Air Temperature Maximum	degF
	Air Temperature Minimum	degF
	Precipitation Increment	in
	Relative Humidity	pct
	Solar Radiation Average	watt/m^2
	Solar Radiation/Langley Total	langley
	Wind Speed Maximum	mph
	Wind Speed Average	mph
	Vapor Pressure – Partial	in Hg
	Vapor Pressure – Saturated	in Hg
Soil Parameters	Soil Moisture Percent -2in	pct
	Soil Moisture Percent -4in	pct
	Soil Moisture Percent -8in	pct
	Soil Moisture Percent -20in	pct
	Soil Moisture Percent -40in	pct
	Soil Temperature Observed -2in	degF
	Soil Temperature Observed -4in	degF
	Soil Temperature Observed -8in	degF
	Soil Temperature Observed -20in	degF
	Soil Temperature Observed -40in	degF

Figures 2 and 3 present the annual and seasonal variations of maximum (max) and minimum (min) soil temperature values by soil depth between 2011 and 2019, respectively. The data for 2020 are not included because they only contain soil temperature values for the first five months of the year. In Figure 2, it is clearly seen that as the soil depth increases, the minimum soil temperature increases, while the maximum soil temperature decreases for each year. Figure 3 reveals that when the air temperatures rise, especially in the spring and summer seasons, an increase in soil temperature values is also observed. In other seasons, the soil temperature decreases in direct proportion to the air temperature.

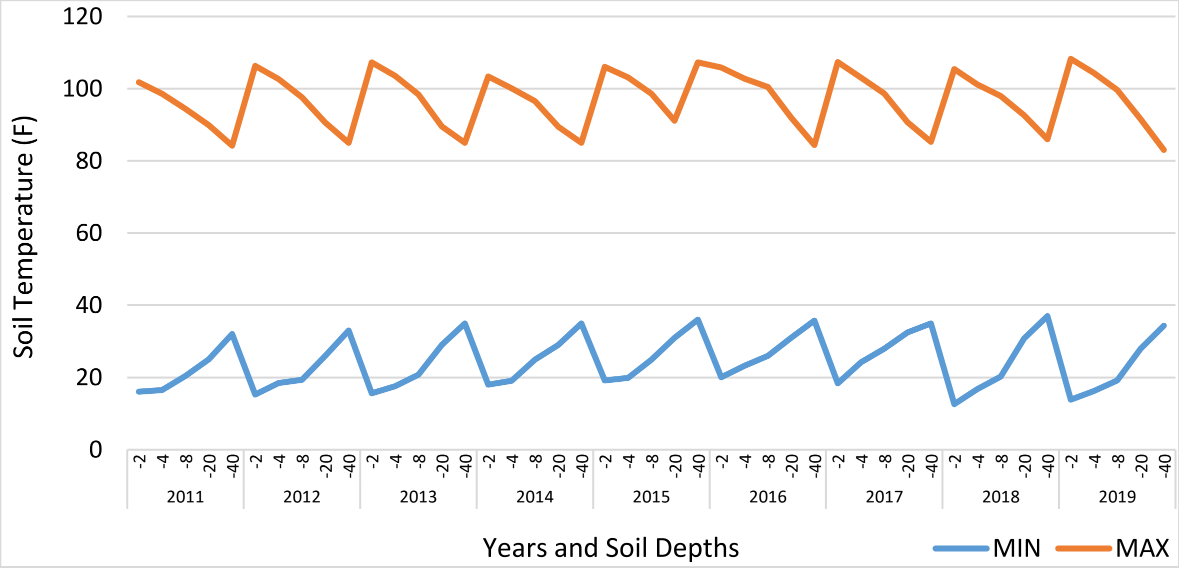


Figure 2- The annual variations of max and min soil temperature values by soil depth

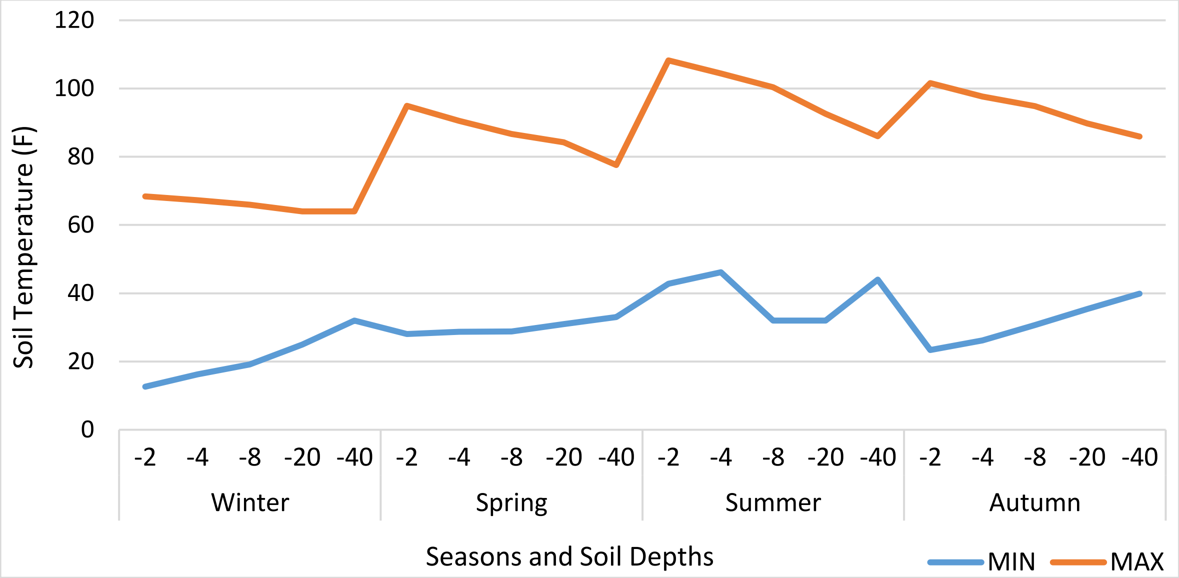


Figure 3- The seasonal variations of max and min soil temperature values by soil depth

The datasets considered in this study have passed through data preprocessing steps using the Python Scikit-learn library for the implementation of the STOC method. First, "date" and "station ID" parameters were extracted from each dataset because they are not correlated with the target parameter (soil temperature observed). Then, the hourly collected data were converted into daily data and the instances which have no soil temperature value were eliminated. Furthermore, the proposed STOC method cannot handle continuous target parameters, so, the first, soil temperature observed parameter for each soil depths in Fahrenheit degree were converted to Celsius degree, and then they were discretized into five different levels such as "very low", "low", "medium", "high", and "very high" according to specific value ranges as shown in Table 3.

Table 3- Discretization of soil temperature observed parameter

Continuous Values (Celsius)	*Continuous Values (Fahrenheit)*	*Categorical Values*
[..., 5]	[..., 41]	Very low
(5, 10]	(41, 50]	Low
(10, 15]	(50, 59]	Medium
(15, 20]	(59, 68]	High
(20, …]	(68, …]	Very high

2.2. Proposed approach

In this study, the authors propose a novel approach: STOC. This approach classifies the soil temperature levels of meteorological and soil data at different soil depths by considering the natural order of class labels. In this approach, the ordinal class classifier algorithm (Frank & Hall 2001), which converts an ordinal classification problem including multiple class labels to a binary classification problem was preferred. According to this algorithm, *k* different ordinal class labels are converted to binary values considering their inherent order. For example, consider that there are five different class labels (C_1, C_2, C_3, C_4, and C_5) and the order between the classes is as follows: $C_5 > C_4 > C_3 > C_2 > C_1$. In the first phase, while the class values that are higher than C_1 (C_2, C_3, C_4, and C_5) are labeled as 1, the rest of them (C_1) are labeled as 0. Then, the same processes are applied to all class labels, so the ordinal classification approach is transformed into a binary classification problem. Besides, five different traditional classification algorithms (DT, NB, KNN, SVM, and RF) were chosen as base learners for the ordinal class classifier algorithm in the STOC method.

Figure 4 shows a general overview of the proposed approach. First, in the data acquisition step, meteorological (air temperature, air pressure, precipitation, relative humidity, and solar radiation) and soil (soil temperature and soil moisture) data is obtained from the sensors (i.e. solar radiation sensor, wind speed, and direction sensor, etc.) of the Soil Climate Analysis Network (SCAN) in 16 stations of United States. These data are stored in a cloud platform. In the next step, first, the datasets are passed through a data preprocessing step (discretization and missing data elimination) and then they are converted to binary datasets to make them ready for the implementation of the STOC method. In the training phase, the STOC approach is applied to the meteorological and soil data using DT, NB, KNN, SVM, and RF as base learners. After that, the performance of the soil temperature level prediction of these models are evaluated using the *n*-fold cross-validation technique selecting *n* as 10. Then, in the prediction phase, a new sample is predicted using the STOC method. Finally, the obtained accuracy rate and F-measure results from these models are represented in the presentation step.

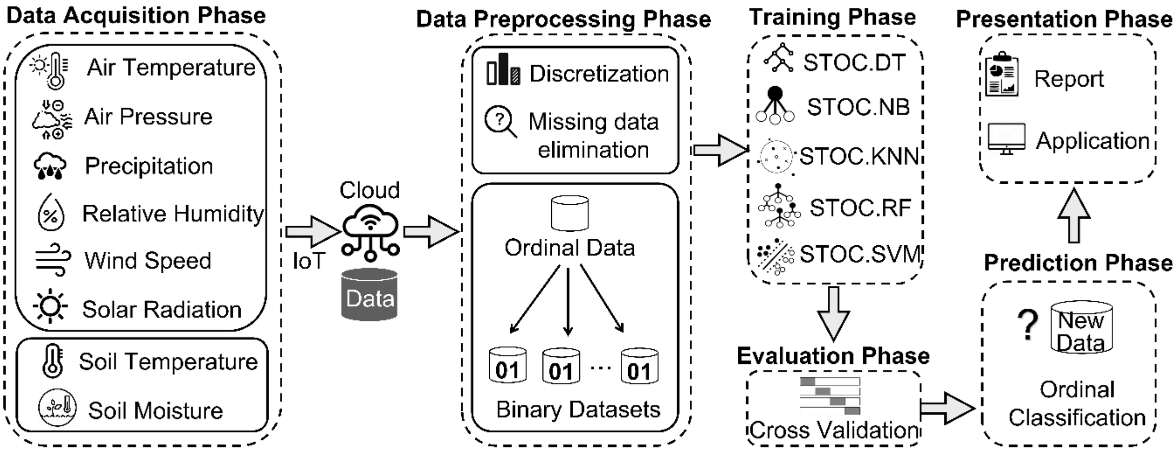


Figure 4- The general overview of the proposed STOC approach

In addition to the general overview of the proposed STOC approach, the flowchart of it is also presented in Figure 5. In the training phase, the original raw data passes through data preprocessing steps and then the ordinal values are converted to binary values using the binary decomposition method. After that, the experimental dataset is divided into training and test data. The selected ordinal classification algorithm is applied to the training data to construct the prediction model. The constructed model is then evaluated using the test data in terms of an accuracy metric. Finally, in the prediction phase, the class label of new unseen data is predicted using the constructed classification model.

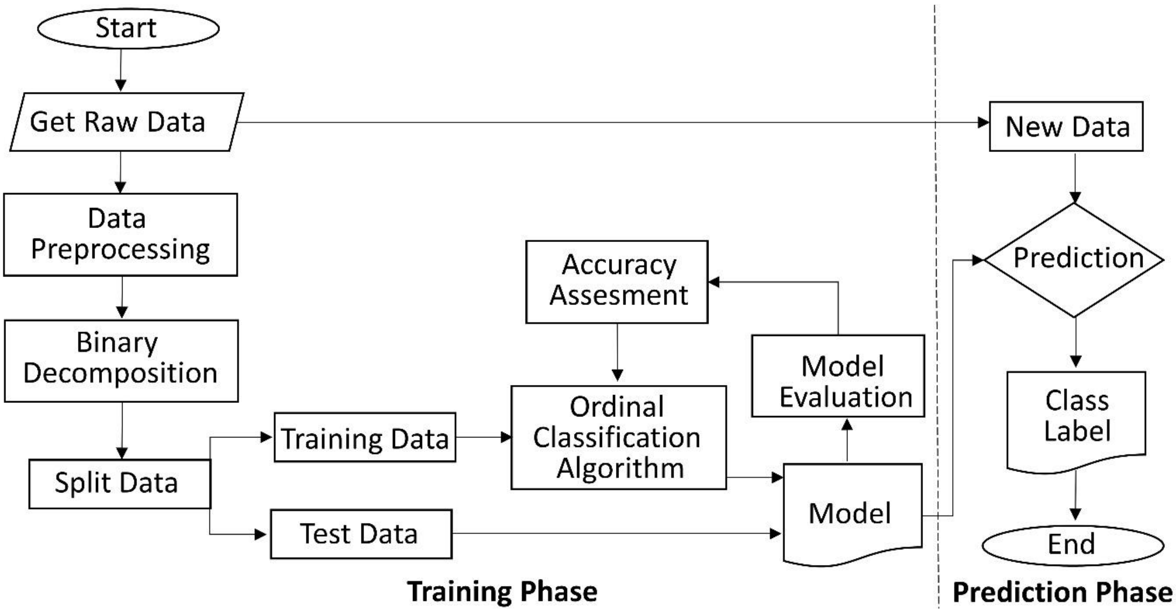
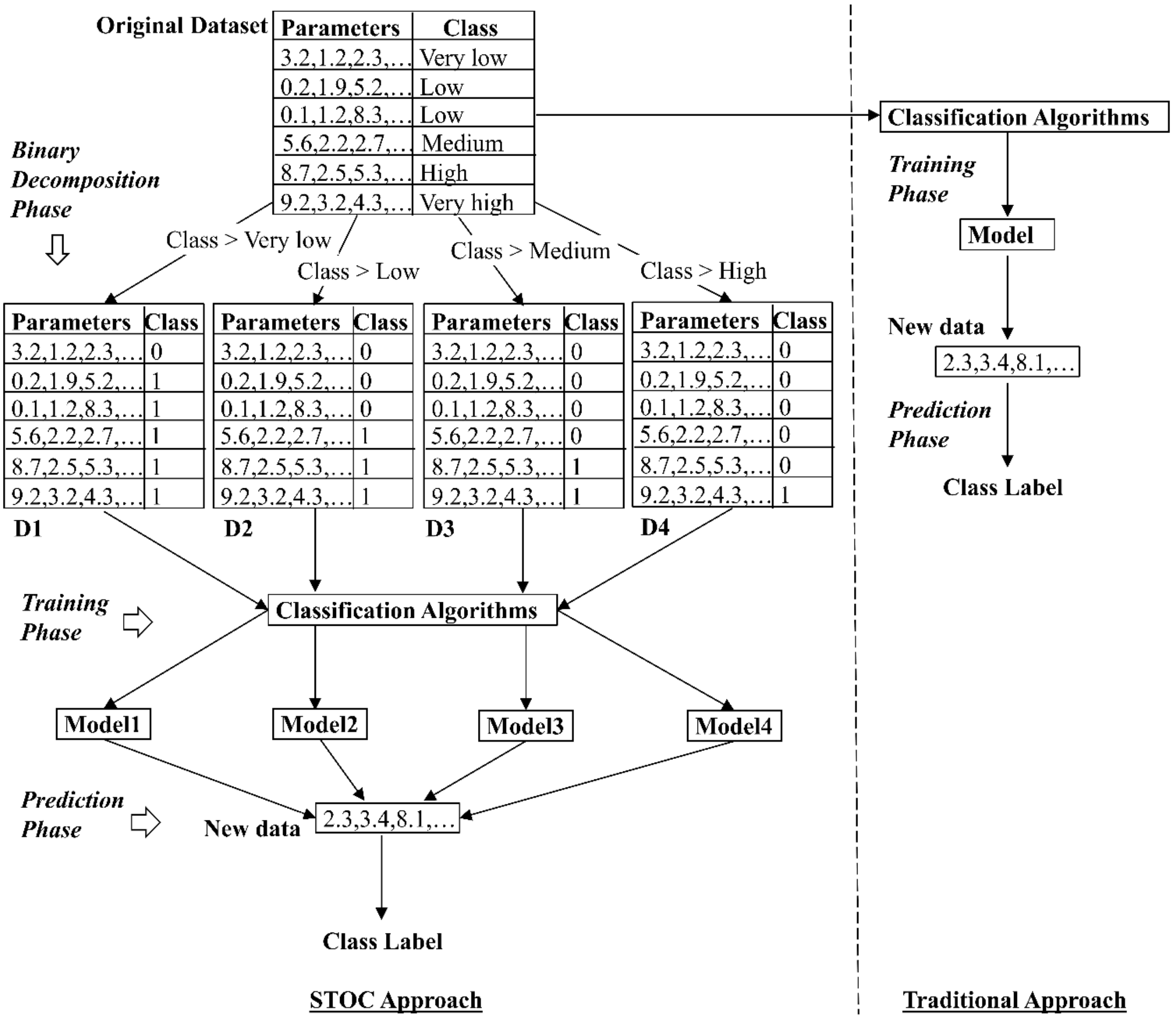


Figure 5- The flowchart of the proposed STOC approach

A tangible example of the proposed STOC approach and its differences from the traditional approach is illustrated in Figure 6. Assume that each instance in the example dataset, including daily values of meteorological and soil data, has an ordinal class attribute with five values: very low, low, medium, high, and very high. These values are on an ordinal scale, such as low < very low < medium < high < very high. In the proposed approach, first, these values are converted to binary values considering their ordinal scales. For example, the samples whose class values are higher than “very low” are labeled as 1 and the others are labeled as 0. Thus, the first dataset (D1) is obtained. In this manner, four different transformed datasets (D1, D2, D3, and D4) are constructed. Then, the base classifiers of the STOC approach (DT, NB, KNN, SVM, and RF) are applied to these datasets and multiple models are generated. The proposed application was developed by using the WEKA machine learning library (Witten et al. 2016) in the Java programming language. In the prediction step, the class label of the new data is predicted using these models by choosing the class with the highest probability. However, in the traditional approaches, the orders among class labels are disregarded. Thus, the classification algorithms are directly applied to the original dataset without any conversion process and the model is constructed. Finally, similar to the STOC approach, new data is classified with one of the predefined classes using the model.



Original Dataset

Parameters	Class
3.2,1.2,2.3,…	Very low
0.2,1.9,5.2,…	Low
0.1,1.2,8.3,…	Low
5.6,2.2,2.7,…	Medium
8.7,2.5,5.3,…	High
9.2,3.2,4.3,…	Very high

D1

Parameters	Class
3.2,1.2,2.3,…	0
0.2,1.9,5.2,…	1
0.1,1.2,8.3,…	1
5.6,2.2,2.7,…	1
8.7,2.5,5.3,…	1
9.2,3.2,4.3,…	1

D2

Parameters	Class
3.2,1.2,2.3,…	0
0.2,1.9,5.2,…	0
0.1,1.2,8.3,…	0
5.6,2.2,2.7,…	1
8.7,2.5,5.3,…	1
9.2,3.2,4.3,…	1

D3

Parameters	Class
3.2,1.2,2.3,…	0
0.2,1.9,5.2,…	0
0.1,1.2,8.3,…	0
5.6,2.2,2.7,…	0
8.7,2.5,5.3,…	1
9.2,3.2,4.3,…	1

D4

Parameters	Class
3.2,1.2,2.3,…	0
0.2,1.9,5.2,…	0
0.1,1.2,8.3,…	0
5.6,2.2,2.7,…	0
8.7,2.5,5.3,…	0
9.2,3.2,4.3,…	1

Figure 6- The comparison of the proposed STOC and traditional approaches

The main benefits of the proposed STOC approach can be summarized as follows:

- Using the meteorological data, future soil temperature levels for a region can be predicted.
- An intelligent soil management system can be designed using the predicted values.
- The crop yield in agriculture can be increased by using the predicted soil temperature levels.
- The proposed approach can play a role in reducing unnecessary resource consumption in agriculture, such as water, sensor, and pesticide.

2.3. Formal definition of the proposed method

Definition 1. STOC refers to the problem that uses ordinal data obtained from sensors and aims to construct a prediction model by considering the orders among class labels to correctly predict the level of the soil temperature according to the methodological data.

Let D be an ordinal meteorological and soil data D=$\{(x_i, y_i)_{i=1}^{N} \in X\ x\ Y\}$, where $X \in R^d$ and its related class label y_i belongs to the output space Y=$\{C_1, C_2, C_3, \ldots, C_k\}$ having k classes with a data ranking structure $C_1 < C_2 < C_3 < \ldots < C_k$, where < denotes the linear order relation. In this method, five different traditional classification algorithms (DT, NB, KNN, SVM, and RF) were implemented as base learners. The aim of classification is to categorize an input vector x with one of the k class labels by selecting the maximum probability value among the evaluated probabilities for each label using a binary classification method. In this approach, the probabilities for each k class are computed as follows:

$$\mathrm{P}(\mathrm{C}_1 |\ \mathrm{x}\ ) = 1 - \mathrm{P}\ (\mathrm{Class} > \mathrm{C}_1 |\ \mathrm{x}) \tag{1}$$

$$\mathrm{P}(\mathrm{C_j}|\ \mathrm{x}\ ) = \mathrm{P}(\mathrm{Class} > \mathrm{C_{j-1}}|\mathrm{x}) - \mathrm{P}\ (\mathrm{Class} > \mathrm{C_j}|\ \mathrm{x})$$
$$\mathrm{P}(\mathrm{C_k}|\ \mathrm{x}\ ) = \mathrm{P}(\mathrm{Class} > \mathrm{C_{k-1}}|\mathrm{x})$$

Where; j={2,3,….,k}.

2.4. Proposed algorithm

Algorithm 1 presents the pseudocode of the STOC method. In the first step of this algorithm, the original dataset is passed through a binary decomposition process for converting ordinal data D to binary datasets $\{D_i\}_{i=1}^{k-1}$ to encode the ranking relationship among the classes. Here, the label y_j associated with the sample x_j is replaced with $y_j = 1$ for all $y_j > C_i$, or, $y_j = 0$ for all $y_j <= C_i$. In other words, for a particular class C_i, the values lower than or equal to C_i are labeled as 0 and the others are labeled as 1. In this way, the algorithm converts an ordinal classification problem involving k classes into k-1 binary classification problems. The dataset D^i is created for each $i \in \{1, 2, ..., k\text{-}1\}$ by using classes $\{C_1, C_2, ..., C_i\}$ against $\{C_{i+1}, C_{i+2}, ..., C_k\}$. After that, a model M_i is constructed for each binary dataset D_i using one of the classification algorithms (DT, NB, KNN, SVM, and RF) as a base learner. In other words, the first model M_1 is constructed to predict what is the probability for a given instance belonging to any class which is located higher than C_1, the second model M_2 is built to predict the probability of belonging to any class that is ordered higher than C_2, and so on. In the final step, the class label of each instance x in the test set T is predicted by choosing the maximum probability value among the evaluated probabilities for each class label using the constructed classification models M^*. In this process, the probability of the first class P(C_1) is computed by subtracting the probability of the upward union of classes P($L_x > C_1$) from 1. The probability of the last class P(C_k) is computed by directly considering the probability of the upward union of classes P(L_x $\mathrm{C_{k-1}}$). In other cases (for the intermediate classes, where $2 \leq j \leq k$-1), the probabilities are computed by considering the probabilities of both upward and downward unions of classes. In the end, the class label with the highest probability (MAX) is predicted as the final class y for the test query x.

Algorithm 1: STOC

```
Inputs:
  D: ordinal dataset D={(x_1,y_1), (x_2,y_2),…., (x_l,y_l)} with n instances
  T: test set to be predicted
  Y: ordinal class labels y ϵ {C_1,C_2,….,C_k} with an order C_1 < C_2 < ... < C_k
  k: the number of classes
 Output:
 M*: Ordinal classifiers
 Ŷ : Predicted class labels
Begin:
   // Step 1 – Binary decomposition
   for i = 1 to k-1 do
         foreach (x_j,y_j) in D
               if (y_j <= C_i)
                     D_i.Add(x_j,0)
               else
                     D_i.Add(x_j,1)
               end if
         end foreach
   end for
   // Step 2 – Model construction
    for i = 1 to k-1 do
         M_i=Train(D_i)
         M* = M* ∪ M_i
    end for
   // Step 3 – Classification
   foreach x in T
         y = M*(x) = MAX (
               P(C_1) = 1 − P(L_x > C_1)
               for j = 2 to k-1 do
                     P(C_j) = P(L_x > C_{j−1}) - P(L_x > C_j)
               end for
               P(C_k) = P(L_x > C_{k−1}) )
               Ŷ = Ŷ ∪ y
   end foreach
End Algorithm
```

2.5. Evaluation metrics

The proposed STOC models using five different base classifiers were compared by using the *n*-fold cross-validation technique selecting n as 10. The prediction performance of each model was tested in terms of accuracy rate and F-measure metric, and the obtained results were presented via tables and graphs in this study.

The accuracy rate gives a ratio of correctly classified instances to all instances in the dataset, as shown in Equation (2).

$$\text{Accuracy rate} = \frac{\#\text{ of correctly classified instances}}{\#\text{ of total instances}} \tag{2}$$

F-measure is a useful metric for presenting the classification performance of any model, which is calculated as a harmonic mean of precision and recall values as shown in Equation (3).

$$\text{F} - \text{measure} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

3. Results

In the experimental studies, the proposed STOC method was executed on real-world meteorological and soil data from 16 stations for predicting soil temperature levels at five soil depths. In this study, five different traditional classification algorithms (DT (C4.5 algorithm), NB, KNN, SVM, and RF) were chosen as base learners for the STOC method. Except for the KNN algorithm, all other base learners were implemented with their default parameter values. In each experiment, the *k* value of the KNN algorithm was selected as $log_2(n)$(n is the number of instances in each dataset). To test the success of the proposed STOC method on the datasets, accuracy rate and F-measure values were evaluated.

Tables 4, 5, 6, 7, and 8 present the accuracy rate results of the applied STOC models using the traditional classification algorithms as base learners (STOC.DT, STOC.NB, STOC.KNN, STOC.SVM, and STOC.RF) on the datasets obtained from 16 stations in three states of United States at five soil depths (D). The results obtained from these tables indicate that the STOC models using five classification algorithms show different classification performances on these datasets. For example, STOC.KNN and STOC.SVM algorithms achieved the highest accuracy rate value (96.3%) on the dataset with 4 inches depth of the Sudduth Farms and Sevilleta stations, respectively. Also, when the accuracy rate results are considered in general, it is observed that STOC.NB showed a worse classification ability than the other STOC models. However, the applied STOC models, except for STOC.NB, show over 75.5% performance of the soil temperature level prediction.

Table 4- The accuracy rates (%) of the applied STOC.DT model on 16 stations at five soil depths (inch)

	Station ID															
D	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***	***14***	***15***	***16***
2	90.9	90.4	92.9	89	92.6	92.6	88.6	95.3	95.6	94.6	94.8	93.6	93.4	92.6	92.3	92
4	95.3	93.2	93.5	91.1	93.4	92.1	92.6	94	95.4	94.5	94.9	94	94	95.2	92.6	92.7
8	93.9	92.9	92.4	89.1	90.4	90.1	92.9	84	95.4	92.8	91.1	92.5	94.1	94.7	92	91.1
20	87.3	89.5	88.8	88.3	88.5	84.8	88.7	89.6	91.8	88.9	87.3	91	90.3	90	90.3	87.7
40	84.6	89.4	90.5	87.8	88.4	82	90.4	84.7	93.7	86.7	88.6	87.7	89.1	85.9	85.8	89.7

Table 5- The accuracy rates (%) of the applied STOC.NB model on 16 stations at five soil depths (inch)

	Station ID															
D	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***	***14***	***15***	***16***
2	81.1	77	84.3	84	83.2	77	80.3	81.8	85.3	79.9	83.3	89.4	87.1	79.3	85.6	81.4
4	85.1	79.7	84.5	85.3	85.5	76.3	81.5	82.8	78.5	81.7	84.9	90.7	88.3	80.2	88.1	81
8	85.9	79.5	82.5	83.2	82.6	81.2	80.8	74.4	89.1	80.8	80.8	88.1	87.5	79.3	85.9	80.9
20	75.4	78	72.6	71.2	75.3	73.4	68.4	80.8	77.3	73.8	75	76.3	75.4	76.5	78.6	66.7
40	63.5	75.9	59	55.7	51.7	63.3	65.6	72.6	69.5	70.5	58.6	58.3	64	67.1	60.5	71

Table 6- The accuracy rates (%) of the applied STOC.KNN model on 16 stations at five soil depths (inch)

	Station ID															
D	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***	***14***	***15***	***16***
2	91.9	91.2	92	88.6	93.2	92.1	88.8	95.1	95.7	94.3	94.9	93.4	93.6	93.4	91.2	92.7
4	95	94	94.7	90.9	93.7	93.2	92.4	95	95.7	95.5	96	93.1	94.9	96.3	92.8	93.4
8	95	93.7	92.9	89.6	91.5	90.3	93.5	83.4	95.5	93.8	91.4	93.6	94.5	95.4	92.2	92.2
20	87.4	88.7	88.3	86.8	87	87.3	86.7	90.4	88.1	88.2	87.4	89.6	89	89.7	89.9	86.3
40	83.5	87.3	86.9	82	82.9	82.4	84.5	85.8	89.6	85.9	85.8	84.3	85.7	86.1	81.6	86.4

Table 7- The accuracy rates (%) of the applied STOC.SVM model on 16 stations at five soil depths (inch)

	Station ID															
D	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***	***14***	***15***	***16***
2	92.1	90.4	93.1	90.2	93.6	92.3	90.4	95.4	95.8	94.9	95.6	94.1	93.9	93.5	92.2	93.3
4	95.2	93.9	94.8	92.6	94.5	92.6	92.9	95.2	95.9	95.4	96.3	93.9	95	95.9	93.6	93.8
8	95.3	93.8	93	89.8	91.5	89.3	93.2	83.4	95.6	94	91.3	94.1	94.8	95.6	93.4	92.3
20	86.6	85	87.7	86.5	87.4	86.1	84.9	89.7	87.8	87.9	85.3	90.1	89.5	88.7	90.4	86.3
40	80.2	85.7	85.3	75.5	83.1	80.8	83.4	83.9	81.8	83.9	80.8	79	81.8	85.5	77	84.7

Table 8- The accuracy rates (%) of the applied STOC.RF model on 16 stations at five soil depths (inch)

	Station ID															
D	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***	***14***	***15***	***16***
2	92.5	92.2	93.2	90.2	93.2	93.6	89.7	95.4	96	95.3	95.7	94.1	94.3	93.5	92.8	92.9
4	95.5	94.1	94.1	91.8	94	92.6	92.6	95.3	95.8	95.4	95.7	94.4	95	96.1	93.5	93.2
8	94.8	94	92.2	90.3	90.8	89.5	93.6	84.5	95.2	93.7	91.8	93.6	94.6	95.9	93.3	92.1
20	87.6	87	87.6	88.7	87.7	87	85.4	90.7	87.7	88.4	87.2	89.5	89.7	90.5	89.7	84.9
40	83.3	87.9	86	85.7	85.4	81.1	86.9	84.2	87.7	85.9	86.5	82.4	86.5	85.1	83.6	85.5

The average accuracy rates of the STOC models on these datasets were calculated to reveal which model is more successful in soil temperature level prediction. The obtained values are illustrated in the graph given in Figure 7. The results from this graph indicate that STOC.DT achieved the best performance of the soil temperature level prediction with an accuracy rate of 90.95%. This is probably because of the fact that the DT algorithm ignores unrelated parameters of the experimental dataset through information gain and defines logic for the split branches, and as a result, it usually provides accurate results. It is the most successful predictive model since it provides an exhaustive analysis of the consequences of each possible decision. The decision tree method divides the dataset at a much deeper level which is not as easily accomplished with other classifiers such as support of vector machines (Mim et al. 2018). According to the results given in Figure 7, it is also possible to say that STOC.RF and STOC.KNN reached higher classification abilities with quite close accuracy rates of 90.91% and 90.84%, respectively.

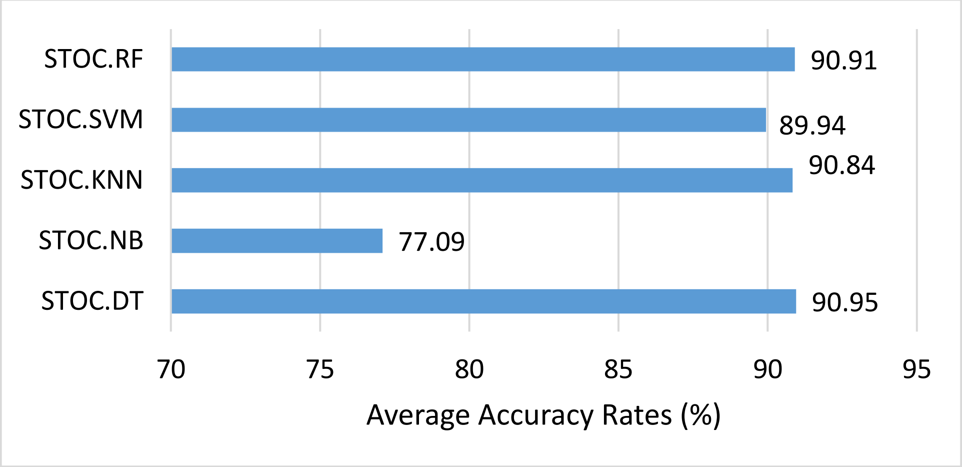


Figure 7- The average accuracy rates (%) of the applied STOC models on the 16 different stations

Furthermore, F-measure values of the applied STOC models using the traditional classification algorithms on the 16 datasets obtained in three of the United States at five soil depths were evaluated. In Figures 8, 9, and 10, the average F-measure values of the STOC models for each state were illustrated, respectively. The fact that the F-measure value is close to 1 presents that the model offers a successful classification ability. These graphs in the figures presented that the STOC.NB model gives the worst prediction performance among the others. Because the NB classifier assumes that the parameters of the dataset are independent of each other. However, some of the parameters often contain similar information and they depend on one another (Jiang et al. 2020). This may cause the NB classifier to yield inaccurate prediction results. The graphs also indicated that all the STOC models, except for STOC.NB, provide over 0.8 F-measure value. Therefore, it is possible to say that the proposed STOC approach is a successful method for soil temperature level prediction of ordinal data.

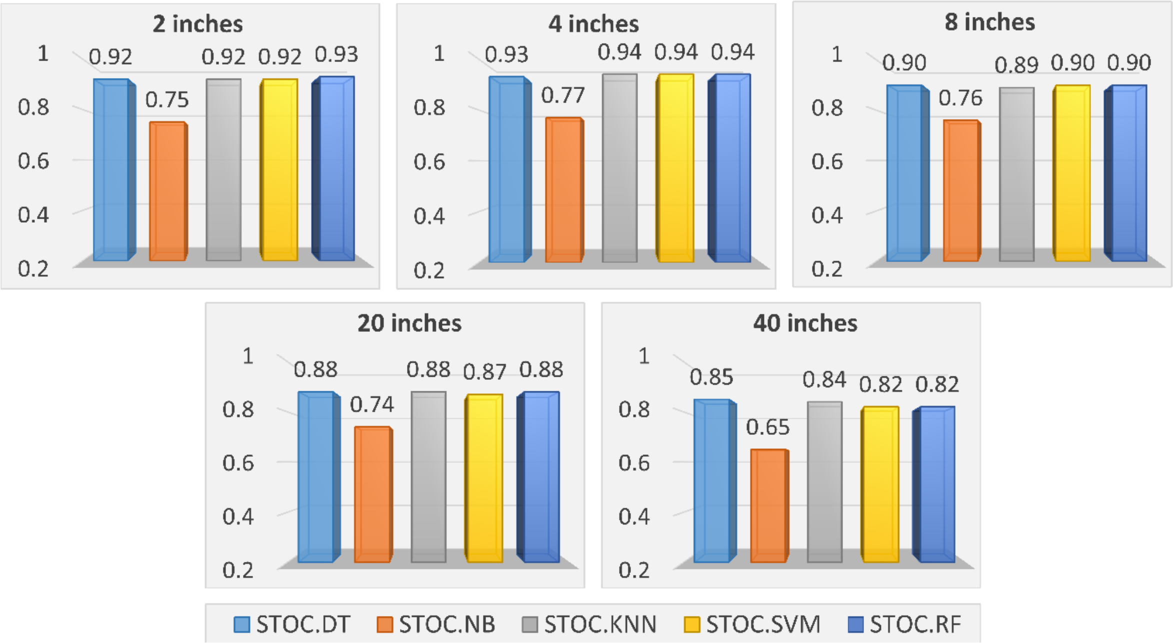


Figure 8- The average F-measure values of the applied STOC models on the Alabama state stations

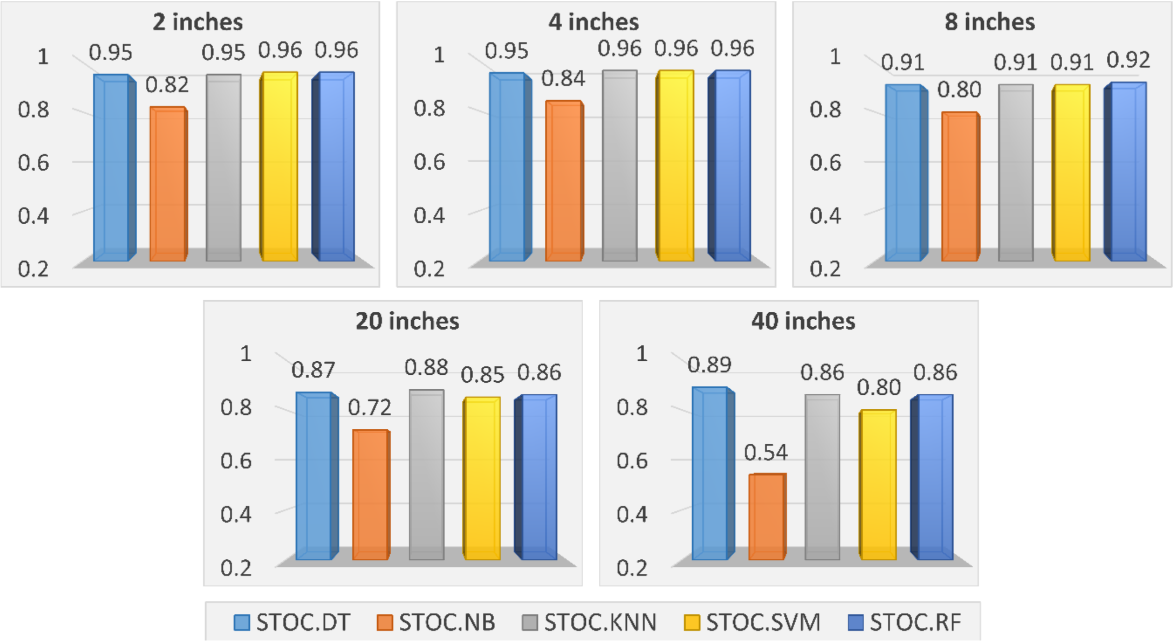


Figure 9- The average F-measure values of the applied STOC models on the Mexico state stations

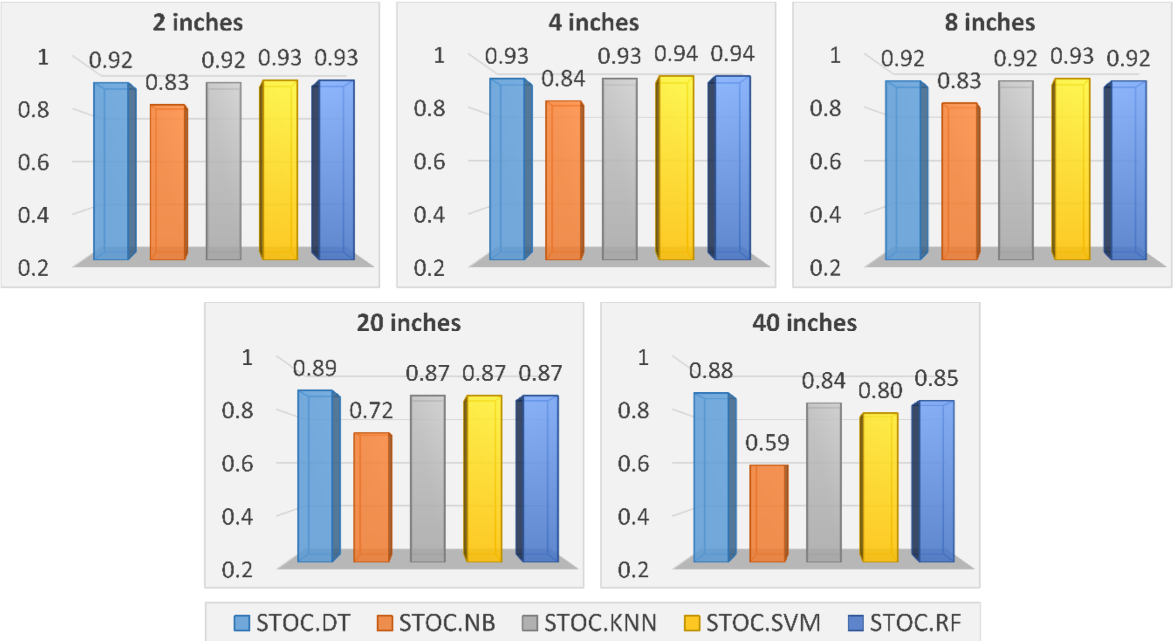


Figure 10- The average F-measure values of the applied STOC models on the Utah state stations

4. Discussion

Differently from the existing machine learning-based soil prediction studies, this study focuses on the classification of ordinal and categorical target values. It also applies different methods from the present studies, such as KNN and NB. Finally, because the target attribute value is categorical, the performances of the applied algorithms were tested with accuracy and F-measure metrics. The detailed comparison of this study with the existing studies is given in Table 1.

The main findings of this study can be summarized as follows:

- The proposed STOC method achieved high accuracy (>90%) on the prediction of soil temperature level. This means that it has a high capability of temperature prediction for new samples in the presence of ordinal historical data.

- The STOC method can be applied to any ordinal soil dataset without any prior information or specific assumptions about the given data.

- When the STOC approach was tested in combination with different classification algorithms (DT, NB, KNN, SVM, and RF), the SSOC.DT method achieved higher accuracy (90.95%) than the rest on average.

- Daily soil temperature level prediction at different depths is necessary for real-life agricultural management systems. The accuracy of prediction varies as the soil depth changes for many locations. STOC usually performed over 0.8 of F-measure value, which means that the proposed method has a good generalization ability at all soil depths. Since STOC covers multiple soil depths, it enables numerous agricultural applications, and thus it expands the use of machine learning in the field of agriculture.

- Since soil temperature is highly dependent on various parameters (i.e., meteorology, soil type, vegetation, lithology, slope angle, and background concentration), this study proposes station-based models. In other words, a separate model is built for each measurement location (monitoring node) by using the station-based soil data.

- Based on the experimental results reported in this study, it is possible to say that the proposed STOC approach is a successful method for making a prediction on ordinal soil temperature data.

- It was observed that STOC has many advantages for predicting soil temperature level, which can provide increasing crop yield, reducing unnecessary resource consumption (i.e., water, sensor, and pesticide), and providing additional information when ordinal data is available.

5. Conclusions

In this study, a novel STOC approach is proposed. The proposed approach categorizes the soil temperature levels of meteorological and soil data at different soil depths by taking into account the inherent order of class labels. While the previous soil temperature prediction studies do not consider the natural ranking between the class labels, this study considers them by using the binary decomposition technique in ordinal classification. Thus, this study is the first attempt to implement ordinal classification for soil temperature level prediction. This study is also original in that it compares alternative base learners (DT, NB, KNN, SVM, and RF) in conjunction with the proposed method.

In the experiments, the proposed STOC method with different classifiers (STOC.DT, STOC.NB, STOC.KNN, STOC.SVM, and STOC.RF) was applied on daily values of meteorological and soil data obtained from 16 stations at five soil depths and compared with each other in terms of F-measure and accuracy rate metrics. The results show that the proposed STOC method provides accurate soil temperature classification performance. The constructed STOC models, except for STOC.NB, achieved over 75.5% ability of the soil temperature level prediction. Also, it is clearly understood from the experimental result that the STOC.DT shows the best performance of the soil temperature level prediction with an average accuracy rate of 90.95%.

As a future study, a multi-output regression technique can be implemented on the same meteorological and soil data for predicting continuous soil temperature values. In addition, an improved time-series study that solves the soil temperature prediction problem using a supervised learning technique can be performed.

Acknowledgment

We thank Volkan KIRANOGLU and Goksu TUYSUZOGLU for supporting the preparation of the dataset used in this study.

References

Abimbola O P, Meyer G E, Mittlstet A R, Rudnick D R & Franz T E (2021). Knowledge-guided machine learning for improving daily soil temperature prediction across the United States. *Vadose Zone Journal* e20151: 1-18. https://doi.org/10.1002/vzj2.20151

Abyaneh H Z, Varkeshi M B, Golmohammadi G & Mohammadi K (2016). Soil temperature estimation using an artificial neural network and co-active neuro-fuzzy inference system in two different climates. *Arabian Journal of Geosciences* 9(2016): 1-9. https://doi.org/10.1007/s12517-016-2388-8

Alizamir M, Kisi O, Ahmed A N, Mert C, Fai C M, Kim S, Kim N W & Shafie A E (2020a). Advanced machine learning model for better prediction accuracy of soil temperature at different depths. PLOS ONE 15(4): 1-25. https://doi.org/10.1371/journal.pone.0231055

Alizamir M, Kim S, Kermani M Z, Heddam S, Shahrabadi A H & Gharabaghi B (2020b). Modelling daily soil temperature by hydro-meteorological data at different depths using a novel data-intelligence model: deep echo state network model. Artificial Intelligence Review 102: 1-28. https://doi.org/10.1007/s10462-020-09915-5

Bayatvarkeshi M, Bhagat S K, Mohammadi K, Kisi O, Farahani M, Hasani A, Deo R & Yaseen Z M (2021). Modeling soil temperature using air temperature features in diverse climatic conditions with complementary machine learning models. Computers and Electronics in Agriculture 185: 1-15. https://doi.org/10.1016/j.compag.2021.106158

Bonakdari H, Moeeni H, Ebtehaj I, Zeynoddin M, Mahoammadian A & Gharabaghi B (2019). New insights into soil temperature time series modeling: linear or nonlinear? Theoretical and Applied Climatology 135: 1157-1177. https://doi.org/10.1007/s00704-018-2436-2

Citakoglu H (2017). Comparison of artificial intelligence techniques for prediction of soil temperatures in Turkey. Theoretical and Applied Climatology 130: 545-556. https://doi.org/10.1007/s00704-016-1914-7

Delbari M, Sharifazari S & Mohammadi E (2019). Modeling daily soil temperature over diverse climate conditions in Iran-a comparison of multiple linear regression and support vector regression techniques. Theoretical and Applied Climatology 135: 991-1001. https://doi.org/10.1007/s00704-018-2370-3

Feng Y, Cui N, Hao W, Gao L & Gong D (2019). Estimation of soil temperature from meteorological data using different machine learning models. Geoderma 338(2019): 67-77. https://doi.org/10.1016/j.geoderma.2018.11.044

Frank E & Hall M (2001). A simple approach to ordinal classification. In: *European Conference on Machine Learning* 3-5 September, Freiburg, Germany, pp. 145-156. https://doi.org/10.1007/3-540-44795-4_13

Hao H, Yu F & Li Q (2020). Soil temperature prediction using convolutional neural network based on ensemble empirical mode decomposition. IEEE Access 9: 4084-4096. https://doi.org/10.1109/ACCESS.2020.3048028

Jiang L, Yuan P, Zhang Q & Liu Q (2020). A study of the Naive Bayes classification based on the Laplacian matrix. *International Journal of Computer Science* 47:4 713-722

Kisi O, Tombul M & Kermani M Z (2015). Modeling soil temperatures at different depths by using three different neural computing techniques. Theoretical and Applied Climatology 121: 377-387. https://doi.org/10.1007/s00704-014-1232-x

Li C, Zhang Y & Ren X (2020a). Modeling hourly soil temperature using deep BiLSTM neural network. Algorithms 13(7): 173-187. https://doi.org/10.3390/a13070173

Li Q, Hao H, Zhao Y, Ceng Q, Liu G, Zhang Y & Yu F (2020b). GANs-LSTM model for soil temperature estimation from meteorological: A new approach. IEEE Access 9: 59427-59443. https://doi.org/10.1109/ACCESS.2020.298299

Li Q, Zhao Y & Yu F (2020c). A novel multichannel long short-term memory method with time series for soil temperature modeling. IEEE Access 8: 182026-182043. https://doi.org/10.1109/ACCESS.2020.3028995

Mehdizadeh S, Behmanesh J & Khalili K (2018). Comprehensive modeling of monthly mean soil temperature using multivariate adaptive regression splines and support vector machine. Theoretical and Applied Climatology 133: 911-924. https://doi.org/10.1007/s00704-017-2227-1

Mehdizadeh S, Fathian F, Safari M J & Khosravi A (2020). Developing novel hybrid models for estimation of daily soil temperature at various depths. Soil & Tillage Research 197(2020): 1-12. https://doi.org/10.1016/j.still.2019.104513

Mim F S, Galib S M, Hasan M F & Jerin S A (2018). Automatic detection of mango ripening stages – An application of information technology to botany. Scientia Horticulturae 237: 156-163. https://doi.org/10.1016/j.scienta.2018.03.057

Nanda A, Sen S, Sharma A N & Sudheer K P (2020). Soil temperature dynamics at Hillslope scale-field observation and machine learning-based approach. Water 12(2020): 713-734. https://doi.org/10.3390/w12030713

Onwuka B & Mang B (2018). Effects of soil temperature on some soil properties and plant growth. Advances in Plants & Agriculture Research 8(1): 34-37. https://doi.org/10.15406/apar.2018.08.00288

Penghui L, Ewees A A, Beyaztas B H, Qi C, Salih S Q, Al-Ansari N, Bhagat S K, Yaseen Z M & Singh V P (2020). Metaheuristic optimization algorithms hybridized with artificial intelligence model for soil temperature prediction: Novel model. IEEE Access 8: 51884-51904. https://doi.org/10.1109/ACCESS.2020.2979822

Sanikhani H, Deo R C, Yaseen Z M, Eray O & Kisi O (2018). Non-tuned data intelligent model for soil temperature estimation: A new approach. Geoderma 330(2016): 52-64. https://doi.org/10.1016/j.geoderma.2018.05.030

Sattari M T, Avram A, Apaydin H & Matei O (2020). Soil temperature estimation with meteorological parameters by using tree-based hybrid data mining models. Mathematics 8(9): 1-21. https://doi.org/10.3390/math8091407

Shamshirband S, Esmaeilbeiki F, Zarehaghi D, Neyshabouri M, Samadianfard S, Ghorbani M A, Mosavi A, Nabipour N & Chau K W (2020). Comparative analysis of hybrid models of firefly optimization algorithm with support vector machines and multilayer perceptron for predicting soil temperature at different depths. Engineering Applications of Computational Fluid Mechanics 14(1): 939-953. https://doi.org/10.1080/19942060.2020.1788644

Tsai Y Z, Hsu K S, Wu H Y, Lin S I, Yu H L, Huang K T, Hu M C & Hsu S Y (2020). Application of random forest and ICON models combined with weather forecasts to predict soil temperature and water content in a greenhouse. Water 12(4): 1-23. https://doi.org/10.3390/w12041176

Wang X, Li W & Li Q (2021). A new embedded estimation model for soil temperature prediction. Scientific Programming 2021: 1-16. https://doi.org/10.1155/2021/5881018

Witten I H, Frank E, Hall M A & Pal C J (2016). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers, Cambridge, MA, USA https://doi.org/10.1016/C2009-0-19715-5

Xing L, Li L, Gong J, Ren C, Liu J & Chen H (2018). Daily soil temperatures predictions for various climates in United States using data-driven model. Energy 160(2018): 430-440. https://doi.org/10.1016/j.energy.2018.07.004

Zeynoddin M, Ebtehaj I & Bonakdari H (2020). Development of a linear based stochastic model for daily soil temperature prediction: One step forward to sustainable agriculture. Computers and Electronics in Agriculture 176(2020): 1-24. https://doi.org/10.1016/j.compag.2020.105636