DOI: 10.12731/2658-6649-2022-14-1-262-285
UDC 631.4

BIG DATA APPLICATION IN THE NEYMAN-PEARSON REGRESSION AND DEEP BERNOULLI AND BOLTZMANN FOR IOT BASED SOIL QUALITY PREDICTION

G. Balaji, P. Vijayakumar

Purpose. *The aim of the work is to enhance the soil quality prediction accuracy and time by using feature selection and classification-based model.*

Background. *Soil quality analysis was handled based on the farmer's first-hand data mining competence and with the world population expected to increase exponentially (i.e., big data), erratic changes in climate have started influencing soil capitulates incorrectly. Big data is employed used to examine the large amount of dataset for soil quality analysis. It is helps to address a lot of new and significant farming decisions and issue. Soil quality analysis depends on fertility of the soil. By soil quality analysis accuracy prediction is very crucial for practical utilization of resources. But, the existing data mining techniques failed to select the correct crop based on the soil and environmental features.*

Material and methods. *The study presents the data mining techniques based on smart and efficient soil quality prediction for agriculture development. In our work, first, linear regression and the Neyman-Pearson correlation-based feature selection model is employed to obtain the computationally efficient and relevant features. Next, an enhanced deep learning model called deep Bernoulli and Boltzmann IoT-based soil quality prediction is designed to classify the complex soil features with better sensitivity and specificity.*

Results. *Experimental results obtained confirmed the performance and reliability of the proposed method. The result evaluations are carried out on the basis of the prediction accuracy, prediction time, sensitivity and specificity.*

Conclusion. *The result shows that the NPR-DBB method achieves better results than the state-of-the-art methods.*

Keywords: *big data; Internet of things; linear regression; Neyman-Pearson correlation; feature selection; deep Bernoulli; Boltzmann soil quality prediction*

For citation. *Balaji G., Vijayakumar P. Big Data Application in the Neyman-Pearson Regression and Deep Bernoulli and Boltzmann for IOT Based Soil Quality Prediction. Siberian Journal of Life Sciences and Agriculture, 2022, vol. 14, no. 1, pp. 262-285. DOI: 10.12731/2658-6649-2022-14-1-262-285*

Introduction

The main objective of precision agriculture remains in increasing the productivity and permitting rational input utilization therefore minimizing the environmental factors due to improper agricultural enactments. By means of it, the inputs in turn can be utilized in an arbitrary manner with the purpose of conclave the distinct requirements of each location's soil, that produce optimized production result. Hence, it becomes necessity to distinguish the soil chemical volatility and physical features via representative sampling. The employment of revolutionary sensors to evaluate soil composition and plant requires an inclination in precision agriculture. In any case, soil prediction models are constructed by employing machine learning algorithms. The objective remains in make farming more significant and productive with minimal influence on the environment.

An architectural model that estimates the soil fertility and productivity via context history with partial least squares regression was proposed in [1]. The partial least squares regression as the multivariate technique was made on the basis of linear regression and to eliminate outliers, spectral region selection technique was applied. The partial least squares regression includes the spectral data such as soil, food, medicine, fuel, etc. With this the prediction accuracy was improved with minimum error.

Classification and prediction of soil parameters on the basis of village assists in minimizing the extravagant disbursement on fertilizer inputs, growing profitability, minimize the laborious time involved in chemical soil analysis and so on. These five classification issues were addressed with the aid of fast learning classification method called extreme learning machine (ELM) using distinct activation functions involving Gaussian radial basis, sine-squared, hyperbolic tangent and triangular basis. With this, the precision, recall and accuracy were improved significantly.

Fertilization is one of the most predominant characteristics of agriculture to acquire paramount capitulates. However, fertilization has often been accomplished on the basis of the circumstance and inclination of farmers without accurate on the necessary of plants. Hence, an instrument and service arrangement framework are required to bestow fertilization recommendations that can help farmers in enhancing yields, minimizing production costs, and putting a stop to environmental pollution from imprudent fertilizer.

New methods concerning smart farming (SF) where the work concentrated on data gathering, transmission, storage, analysis, and moreover appropriate solutions were provided in [3]. The smart farming is a promising idea that refers to

deal with farms that benefit the agricultural industry. The smart farming comprises the advanced technology big data, cloud, and IoT for improving the quantity and quality of products. Here, IoT being indispensable poles in smart systems, agriculture management was improved via smart farming. One of the paramount issues facing by the IoT technology is the management of data. The increase in the number of sensors will obviously results in the increase in the obtained data or big data. As a result, data analysis, processing will naturally become a complicated task. To address this issue, big data management for IoT was discussed in [4].

One of the most significant factors of agriculture is concerned to fertilization so that it results in maximum profit. But, in the recent years, fertilization has been performed on the basis of the experience and instinct of farmers without accurate information on the nutritional plant necessitates. Hence, platforms are required to bestow recommendations on fertilization aspects that can aid farmers in enhancing the yields, minimizing production costs, and eliminating environmental pollution from unrestricted fertilizer.

An integrated service system platform using JMeter software to provide fertilizer recommendations was discussed in detail in [5].

A low-cost smart agriculture framework for smart agriculture employing IoT featuring sensor for an extensive range of temperature and phosphate detection was presented in [6]. However, the energy consumption involved in the analysis was not discussed. To address this aspect, a zoning irrigation system was designed in [7] based on IoT and fuzzy control system was proposed. Here, with the aid of fuzzy control technology optimized usage of water and energy were said to be ensured. Yet another method addressing the energy consumption aspect concentrating on the energy for estimating drought severity level and accordingly prediction was proposed in [8].

A review of big data for agri-food supply chains was investigated in [9]. In the recent few years, sprinklers are monitored manually that in turn results in the continuous sprinkler monitoring. Moreover, occurrence of water wastage at certain stages in the sprinkler system has to be eliminated. The monitoring of sprinklers in a continuous manner is necessitated by means of surveillance with the purpose of recognizing drawbacks but still necessitates human intervention to eliminate certain issues. Therefore, to eliminate continuous monitoring and also to minimizing the human efforts a novel method utilizing Internet of things (IoT) by means of proximity sensor, temperature of soil sensor, on the basis of the water flow and land wetness were proposed in [10].

Blockchain swiftly became an indispensable technique in numerous applications of precision agriculture discipline. The requirement to design smart

P2P systems potential of validating, securing, monitoring, and agricultural data analysis is nearing to conceptualize about constructing blockchain-based IoT systems in designing precision agriculture.

An elaborate survey on the significance of combining blockchain and IoT in designing smart applications for precision agriculture was investigated in [11]. Also, the major advantages and disadvantages in managing numerous sub-sections in precision agriculture were also detailed in brief. Materializing the upcoming transpose tendency of soil fertility has substantial importance in enhancing the quality of soil, attaining high-quality crop production and finally, sustainable development in the field of agriculture.

Stochastic petri net was utilized in [12] for predicting soil fertility. Machine learning algorithms were employed in [13] for maize prediction in Eastern and Southern Africa. The precise soil nutrient estimation specifically paramount owing to the fact that its' influence on plant growth and forest regeneration have a positive effect. With the objective of analyzing the soil nutrient content and natural regeneration quality of Dacrydium pectinatum communities in China, advanced and precise estimation measures are of primary importance.

Machine learning techniques were employed in [14] to estimate soil nutrient content. Here, soil nutrient evaluation mechanisms were constructed by employing six distinct support vector machines and four artificial neural networks. With this the error involved in analyzing soil nutrient content were reduced with significant improvement in the accuracy level. Artificial intelligence system was employed in [15] for aiding soil classification.

Soil organic matter prediction was performed in [16] by employing multilayer perceptron and two convolutional neural networks. With this best prediction results were achieved for predicting soil organic matter content. In [17], heuristic algorithms were applied for perishable agricultural products employing two distinct types of harvesting, therefore contributing to both cost and time. In [18] a review of machine learning algorithms was applied to obtain soil sciences for moisture prediction. Soil property prediction using multi-target stacked generalization was performed in [19] to minimize average error rate for assessing soil quality. In [20], deep reinforcement model was applied for crop yield prediction. Spectrophotometric methods were introduced in [21] to improve the quality of natural sources. Environmental and economic efficiency protective afforestation was introduced in [22] to achieve the quality of soil and climatic conditions. Water resource was developed in [22] for enhancing the efficiency.

Motivated by the above two works and also the state-of-the-art works presented in the literature sections, in this work with the objective of improving

prediction accuracy, sensitivity, specificity and reducing the prediction time involved in mining and analyzing of soil quality, of the Neyman-Pearson regression and deep Bernoulli and Boltzmann (NPR-DBB) IoT-based soil quality prediction is proposed.

To summarize the contributions of this work are listed below.

- To design an efficient method for predicting data mining and IoT-based soil moisture quality involving huge samples.
- To find the relevant features, linear regression and the Neyman-Pearson correlation-based feature selection model is applied to the raw sample soil data instances.
- To get better prediction accuracy with minimum time and sensitivity rate by selecting the computationally efficient and relevant features using deep Bernoulli and Boltzmann IoT-based soil quality prediction algorithm.
- To propose a new IoT-based soil quality prediction method to forecast the soil moisture and accordingly perform irrigation based on soil humidity.
- The performance of the proposed NPR-DBB based soil quality prediction method is compared with the state-of-the-art methods.
- Finally, a series of experiments were conducted using soil moisture prediction dataset with different performance metrics. The metric assessment results like, prediction accuracy, prediction time, sensitivity and specificity are utilized to find the performance improvement of the NPR-DBB method over the existing works.

Materials and research methods

Soil quality water pollution can be done based on several measures, like, soil fertility, classification based on states and districts, soil moisture control and so on. It is a significant challenge for farmers and agriculturists to predict the soil quality. This work defines the prediction of soil quality using machine and deep learning techniques. The main objective of this work is to improve the soil quality prediction accuracy and time by feature selection and classification-based model. The workflow of the Neyman-Pearson regression and deep Bernoulli and Boltzmann (NPR-DBB) IoT-based soil quality prediction is illustrated (Fig. 1).

As illustrated in the above figure, the Neyman-Pearson regression and deep Bernoulli and Boltzmann (NPR-DBB) IoT-based soil quality prediction is split into three stages. In the first stage, the details regarding the soil moisture prediction big data dataset are provided. Second, feature selection using linear regression and the Neyman-Pearson correlation is designed. Finally, the soil quality prediction using deep Bernoulli and Boltzmann model is presented.

The elaborate description of the proposed NPR-DBB is provided in the following sections.

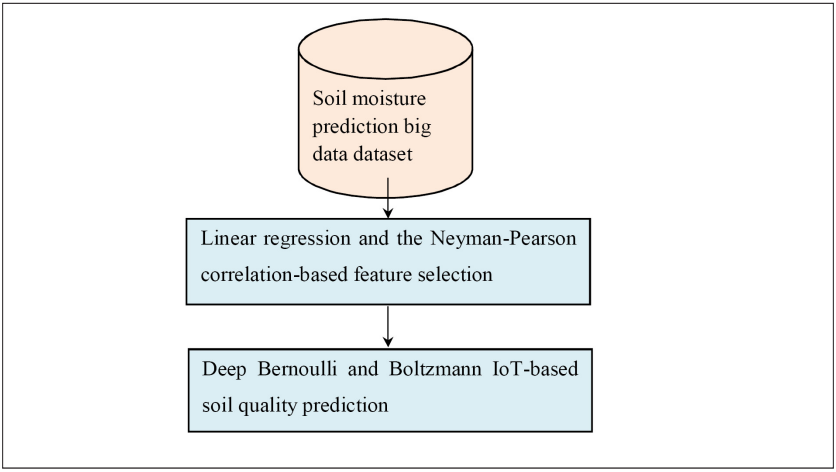

Fig. 1. Block diagram of the Neyman-Pearson regression and deep Bernoulli and Boltzmann (NPR-DBB) IoT-based soil quality prediction

- **Soil moisture prediction big data dataset**

The data used for the experimentation has been collected from [24]. This collection of the dataset includes a combination of distinctive soil data that is a measure of predicting soil humidity in their fields that allow farmers to prepare their irrigation schedules in an optimal and efficient manner using a wide range of data. Sensor-based irrigation, machine and deep learning algorithms can assist farmers with a solution to manage water utilization in a more effective manner. In this work, a soil moisture prediction dataset has been obtained through various IoT sensor devices and the IoT soil moisture sensors were set up in each of the fields followed by which an IoT weather station was positioned near the fields. These IoT devices transmitted the following data in five-minute interval gap. They are

- Soil humidity
- Air temperature
- Air humidity
- Pressure
- Wind speed
- Wind gust
- Wind direction

A sensor being a device detects and responds to certain input acquired from the physical environment (i.e., soil). Three distinct types of irrigation schedules were analyzed. They are:

1. usual irrigation
2. less than crop needs
3. based on water loss

The dataset comprises numerous data and with this, significant and pertinent features are selected for further soil quality prediction. Among numerous data, data consumed in the range of 2800 to 28000 are used. Also, with the purpose of predicting the class of soil fertility, a target vector feature column referring to 'soil humidity' is utilized ranging between 0 and 1. On the basis of the resultant soil humidity value soil moisture prediction can be made and accordingly the irrigation analysis for agriculture data for farmers is made by means of the deep learning model.

- **Linear regression and the Neyman-Pearson correlation-based feature selection**

Feature selection is distinguished as the course of action in connection with eliminating the imprudent and dispensable features or attributes utilizing statistical measures from a raw soil quality dataset, to enhance the learning algorithm. The central objective of feature selection is to attain a significant subgroup of features for elucidating and mining a dataset. With the aid of machine learning, feature selection model gives us a method for minimizing execution time involved in predicting soil quality, improving the soil quality prediction accuracy. To be more specific, feature selection is an enormously utilized pre-processing and data mining model for higher-dimensional data or big data. In this work linear regression and the Neyman-Pearson correlation-based feature selection for agriculture data to analyze the soil moisture and accordingly perform the irrigation is designed. In machine learning algorithm, the linear regression is one of the most available and popular regression. It is a numerical method that is employed for predictive analysis. Linear regression is applied for measuring the linear relationship among a dependent and one or more independent variables. The Neyman-Pearson correlation-based feature selection is utilized to automatically identify the relevant features in a very huge dataset. The structure of linear regression and the Neyman-Pearson correlation-based feature selection model is showed (Fig. 2).

Fig 2 explains the structure of linear regression and the Neyman-Pearson correlation-based feature selection model. Initially, the soil moisture prediction big data dataset is considered. Next, the linear regression is applied to perform

the feature selection. Followed by, the Neyman-Pearson correlation is utilized to measure the correlation between perceived and mean feature and evaluate the hypothesis for selecting the computationally efficient and correlated features.

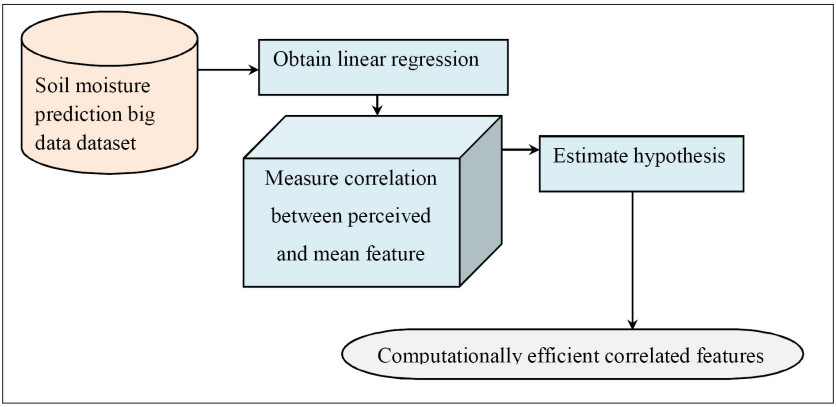

Fig. 2. Structure of linear regression and the Neyman-Pearson correlation-based feature selection model

As shown in the above figure, let us consider a soil quality big data dataset '*DS*' where '*m*' denotes the number of soil samples and '*n*' represents the number of features. Then the feature set is represented as '$F = \{F_1, F_2, F_3, \ldots, F_{\text{吃}}\}$' and the class label set '*C*' is represented as 'C = {–1,+1}'. A predictable feature 'F_i' is discerned to be relevant if and only if there persuade certain probability '$Prob(F_i)$' and '*Y*' such that ''$Prob(F_i = fi)>0$' and linearly regression with each other as given in the below equation.

$$Prob\ (Y = y|F_i = f_i) \neq Prob\ (Y = y) \quad (1)$$

$$y_i = \alpha_0 + \alpha_1 f_{i1} + \alpha_2 f_{i2} + \cdots + \alpha_n f_{in} \quad (2)$$

Also, in addition to the removal of irrelevant features, redundant features should also be eliminated as well. So that time and accuracy can be improved to a great extent. A feature is perceived as redundant if it is tremendously or immensely correlated with one or more other features. In addition, this emerged in a hypothesis for feature selection that embodies mining of features that are notably correlated with class but disjoint with one other.

Let us consider with a given feature set '$F = \{F_1, F_2, F_3, \ldots, F_n\}$', (soil humidity, air temperature – minimum temperature and maximum temperature, air humidity – relative humidity, pressure, wind speed, wind gust, wind direction, solar irradiance, sun, coefficient cultural, evapotranspiration measured, evapo-

transpiration reference, rainfall, water need) if the correlation between the discrete feature and a foreign variable is identified and the correlation between every other feature pair is provided, then the correlation between the complex test consisting of the comprehensive features and the foreign variable is mathematically formulated as given below.

$$PCC = \frac{\sum_{i=1}^{n}(f_i - f')(y_i - y')}{\sqrt{\sum_{i=1}^{n}(f_i - f')^2}\sqrt{\sum_{i=1}^{n}(y_i - y')^2}} \quad (3)$$

From the above equation (3), the pearson correlation coefficient '*PCC*' to obtain the correlation between the discrete feature and a foreign variable is formulated based on the perceived 'f_i' and mean values of the feature 'f'', perceived 'y_i' and mean 'y'' values of the dataset classification respectively. Then, the piled correlation coefficient between the features and output variables are mathematically formulated as given below.

$$A_{ny} = Prob\,(F_n, y) \rightarrow H_o: obtaining\ relevant\ feature \quad (4)$$

Second, the piled correlation coefficient between diversified features is mathematically formulated as given below.

$$A_{nn} = Prob\,(F_n, F_n) \rightarrow H_1: eliminating\ redundant\ feature \quad (5)$$

Next, the classified correlation coefficient selecting the relevant feature is mathematically formulated as given below.

$$J(F_n, y) = \frac{nA_{ny}}{\sqrt{n+(n-1)A_{n\text{捍}}}} \quad (6)$$

Finally, the Neyman-Pearson correlation is a method to determine hypothesis test. The Neyman-Pearson correlation is utilized to choose the important correlated features by using Pearson correlation coefficient. Then, the Neyman-Pearson detector function is employed to create two distinct hypotheses and reduce the redundant features. The Neyman-Pearson correlated probability distribution to obtain the computationally efficient relevant features, eliminating the redundant features is mathematically formulated as given below.

$$\frac{Prob(A_{ny}=RF|H_1)}{Prob\,(A_{ny}=RF|H_0)} \quad (7)$$

From the above hypothesized results obtained from 'H_0' and 'H_1', it is inferred that the correlation between a group and a foreign feature is an operation of the total number of discrete distinguishing features in the group. The formula from (7) is obtained from the Neyman-Pearson correlated probability distribution by normalizing all the variables in the feature subset. The pseudo code representation of the Neyman-Pearson correlation-based feature selection is given below.

Algorithm 1.

The Neyman-Pearson correlation-based feature selection

Input: Dataset '*DS*', Feature set '$F = F_1,F_2,F_3,...,F_n$'
Output: Computationally efficient and relevant features
Step 1: **Initialize** number of soil samples '*m*', number of features '*n*' Step 2: **Begin** Step 3: **For** each Dataset '*DS*' with Features '*F*' Step 4: Formulate linear regression function as in equation (2) Step 5: Estimate pearson correlation coefficient '*PCC*' to obtain correlation between discrete feature and foreign variable using equation (3) Step 6: Estimate piled correlation coefficient between features and output variables as in equation (4) Step 7: Estimate piled correlation coefficient between diversified features as in equation (5) Step 8: Evaluate classified correlation coefficient as in equation (6) Step 9: Estimate the Neyman-Pearson correlated probability distribution as in equation (7) Step 10: **Return** (feature selected '*FS*') Step 11: **End for** Step 12: **End**

As given in the above the Neyman-Pearson correlation-based feature selection algorithm with the objective of obtaining computationally efficient and relevant features (soil humidity, air temperature, air humidity, water need) higher correlated features are first selected using Pearson Correlation Coefficient. This is due to the reason that higher the correlation between the perceived and mean variable, the higher is the correlation between the piled and extrinsic variables. Followed by which, two distinct hypotheses are generated by employing the Neyman-Pearson detector function. This is owing to the reason that redundant features associated with other features have to be eliminated. This is performed using the Neyman-Pearson detector. With this computationally efficient and relevant features are selected with minimum time and maximum accuracy.

- **Deep Bernoulli and Boltzmann IoT-based soil quality prediction**

With the computationally efficient and relevant features selected using the Neyman-Pearson correlation-based feature selection algorithm, in this section, a deep learning classification model is presented. The functions involved in the proposed deep Bernoulli and Boltzmann IoT-based soil quality prediction model are presented in this section. The soil quality depends on soil PH and the amount of the soil macronutrients (nitrogen, phosphorus and potassium).

This is aids to predict the quality of soil by using deep learning classification model. Followed by, the classification model is helps to save the time of soil analysis for farmer. The structure of deep Bernoulli and Boltzmann IoT-based soil quality prediction model is demonstrated (Fig. 3).

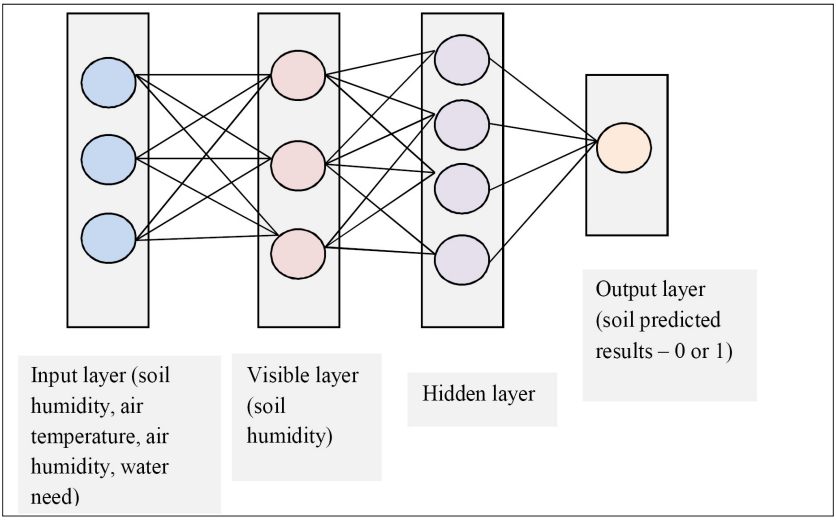

Fig. 3. Structure of deep Bernoulli and Boltzmann IoT-based soil quality prediction model

As illustrated in the above figure, the proposed deep Bernoulli and Boltzmann IoT-based soil quality prediction model consists of two distinct layers, namely, Visual Layer or visible layer '*VL*' and the Hidden Layer '*HL*'. The selected features '*FS*' in these two layers are said to be unanimously self-sufficient to each other and the selected features are modeled as '$Prob(HL|VL) = Prob(HL_1|VL_1)$, $Prob\ (HL_2|VL_2), \ldots Prob(HL_n|VL_n)$'. Moreover, the visual layer and the hidden layer are associated with the Weight '*W*' respectively. Also, '*BHL*' and '*BVL*' forms the '*HL*' and '*VL*' biases with '*m*', '*n*' denoting the number of visible and hidden instances from the selected features (soil humidity, air temperature, air humidity, water need) respectively.

For each '*HL*' and '*VL*' let us consider two vectors, hidden layer vector '$hl = \{hl_1,hl_2,hl_3,...,hl_m\}$' and visual layer vector '$vl = \{vl_1,vl_2,vl_3,...,vl_m\}$'. Then, the pair of Boolean vectors for mining hidden and visual layer vectors data is mathematically represented as given below.

$$EY(vl,hl,\beta) = -\sum_{j=1}^{n} BHL_j vl_j - \sum_{i=1}^{m} BVL_i hl_i - \sum_{i=1}^{m}\sum_{j=1}^{n} BHL_j BVL_i W_{ij} \quad (8)$$

From the above equation (8), '$\beta = \left(W_{ij}, F\sim_i, y_i\right)$' forms the Boltzmann factor, 'W_{ij}' being the weight associating between the '*ith*' feature instance in the current layer and the '*jth*' feature instance in the previous layer. Moreover, the hinge probability distribution for mining '(*vl,hl*)' data is mathematically formulated as given below.

$$HPD = \frac{e^{-EY(vl,hl,\beta)}}{Stand(\beta)} = Prob\,(vl, hl|\beta) \quad (9)$$

From the above equation (9), '*Stand*(β)' forms the standardized criterion and the negligible classification function that includes irrigation variables (i.e., irrigation variable is set to 1 when the irrigation is turned on and the irrigation variable is set to 0 when the irrigation is turned off) associated with each of the four fields (i.e., maize [less water irrigation], peanuts [irrigation based on water loss], peanuts [less water irrigation] and peanuts [normal irrigation]) using Bernoulli distribution is mathematically formulated as given below.

$$Prob\left(vl_i^k = 1\}hl\right) = \frac{\exp\left(BVL_i^k + \sum W_{ij}^k HL_j\right)}{\sum \exp\left(BV\,\%_{oo}i^k + \sum W_{ij}^k HL_j\right)} \quad (10)$$

Then, the final soil moisture predicted value using the soil humidity is formulated according to the hidden layer (i.e., mining and classification results) and visual layer (i.e., soil humidity) functional values. The activation probability of '*jth*' element with known '*vl*' units is formulated as given below.

$$Prob\left(hl_j = 1|vl,\beta\right) = \sigma\left(BHL_j + \sum \mathcal{U}l_j W_{ij}\right) \quad (11)$$

In a similar manner, the activation probability of '*ith*' element with known '*hl*' units is formulated as given below.

$$Prob\left(vl_j = 1|hl,\beta\right) = \sigma\left(BVL_j + \sum W_{ij} hl_j\right) \quad (12)$$

On the basis of the resultant values in the visible units (i.e., soil humidity values), the soil moisture prediction is made and accordingly, irrigation process is carried out. With the Bernoulli logistic function, both the sensitivity and specificity rate of soil moisture prediction for agriculture data are said to be improved. The pseudo code representation of deep Bernoulli and Boltzmann IoT-based soil quality prediction is given below.

Algorithm 2.

Deep Bernoulli and Boltzmann IoT-based soil quality prediction

Input: Dataset '*DS*', Feature set '$F = F_1,F_2,F_3,...,F_n$'
Output: Robust soil quality prediction
Step 1: **Initialize** feature selected '*FS*' Step 2: **Begin** Step 3: **For** each Dataset '*DS*' with Features '*F*' Step 4: Formulate pair of Boolean vectors for hidden and visual layer vectors as in equation (8) Step 5: Estimate hinge probability distribution as in equation (9) Step 6: Evaluate logistic function to visible unit based on Bernoulli distribution as in equation (10) Step 7: Formulate activation probability of '*jth*' element with known '*vl*' units as in equation (11) Step 8: Formulate activation probability of '*ith*' element with known '*hl*' units as in equation (12) Step 9: **If** 'vl_j(*i.e.,soil humidity*) = 1' Step 10: **Then** irrigation is set to on Step 11: **End if** Step 12: **If** 'vl_j(*i.e.,soil humidity*) = 0' Step 13: **Then** irrigation is set to off Step 14: **End if** Step 15: **End for** Step 16: **End**

As given in the deep Bernoulli and Boltzmann IoT-based soil quality prediction algorithm for soil quality prediction, with the objective of improving the sensitivity and specificity rate involved in the classification process, a duality problem is formulated. First with the relevant feature selected as input, a pair of Boolean vectors is modeled for predicting soil quality. Second, a logistic function based on Bernoulli distribution is formulated for efficient differentiation between the soil quality parameters. Finally, activation function with respect to visual and hidden layers is modeled to arrive at the result. Finally, the predicted results of soil quality (i.e., either normal or contaminated results) are obtained.

Results and discussion

In this section, simulations of the proposed the Neyman-Pearson regression and deep Bernoulli and Boltzmann (NPR-DBB) IoT-based soil quality prediction method, partial least square regression [1] and extreme learning machine (ELM) [2] are developed in Python by soil moisture prediction dataset. The

proposed IoT-based deep learning method evaluates the experimental agriculture land based on the considered attributes for predicting the soil moisture. This dataset was obtained as a portion of an experiment performed by Wazihub utilizing low-cost internet of things (IoT) sensors for a period of 4 months in 4 fields cultivating maize and peanuts in Senegal.

An IoT sensor was first positioned in four different plots of land that were planted with either maize or peanuts with an assumption or hypothesis that similar amount of maize was sown in maize plot and for peanuts. Also, the plots were positioned right next to each other, separated by a one-meter perimeter. The collected agriculture dataset here is applied to the machine and deep learning algorithm for training purpose. The performance of the proposed method is compared with the results obtained from state-of-the-art methods. In the proposed method, different performance metrics are utilized for predicting soil quality. Experimental evaluation is carried out on factors such as prediction accuracy, prediction time, sensitivity and specificity with respect to distinct samples. The dataset description is provided in Table 1.

Table 1.

Dataset description

S. No	Feature name	Description
1	Timestamp	Time of recording
2	Soil_humidity_1	Soil humidity of field 1
3	Irrigation_field_1	1 = irrigation on; 0 = irrigation off
4	Soil_humidity_2	Soil humidity of field 2
5	Irrigation_field_2	1 = irrigation on; 0 = irrigation off
6	Soil_humidity_3	Soil humidity of field 3
7	Irrigation_field_3	1 = irrigation on; 0 = irrigation off
8	Soil_humidity_4	Soil humidity of field 4
9	Irrigation_field_4	1 = irrigation on; 0 = irrigation off
10	Air_temperature	Temperature of the air
11	Air_humidity	Temperature of the humidity
12	Air_pressure	Temperature of the pressure
13	Wind_speed	Speed of the wind
14	Wind_gust	Speed of the gust of the wind
15	Wind_direction	Direction of the wind
16	Solar_irradiance	Power per unit area
17	Sun	Radiant energy emitted by sun
18	Kc	Crop coefficient

The end of a table

19	ETc	Evapotranspiration rate (testing crop)
20	ETo	Evapostranspirate rate (training crop)
21	Rainfall	Rainfall per day
22	Water_need_1day	Water needs for one day
23	Water_need_2days	Water needs for two days
24	Water_need_3days	Water needs for three days

Prediction accuracy: The first significant parameter concerning soil quality prediction is the prediction accuracy. This is due to the reason that the prediction accuracy analyzes how far the soil quality is assessed. This is mathematically formulated as given below:

$$P_{acc} = \sum_{i=1}^{n} \frac{S_{AP}}{S_i} * 100 \quad (13)$$

From the above equation (13), the prediction accuracy 'P_{acc}' is measured based on the samples involved in the simulation process 'S_i' and samples accurately predicted 'S_{AP}'. It is measured in terms of percentage.

First, the results of prediction accuracy using NPR-DBB, partial least square regression [1] and ELM [2] are provided in Table 2. Comparison made with the proposed NPR-DBB method has produced better results than soil quality prediction with partial least square regression [1] and ELM [2] respectively.

Table 2.

Comparative analysis for prediction accuracy using NPR-DBB, partial least square regression [1] and ELM [2]

Samples	Prediction accuracy (%)		
	NPR-DBB	partial least square regression	ELM
2800	96.42	94.1	92.85
5600	94.15	92.15	89.15
8400	93	90	85.35
11200	92.15	87.35	82
14000	90	85.45	80.45
16800	89.35	83	78
19600	86	82	75.35
22400	85.25	79	74
25200	85.15	77.15	73.15
28000	85	75	71

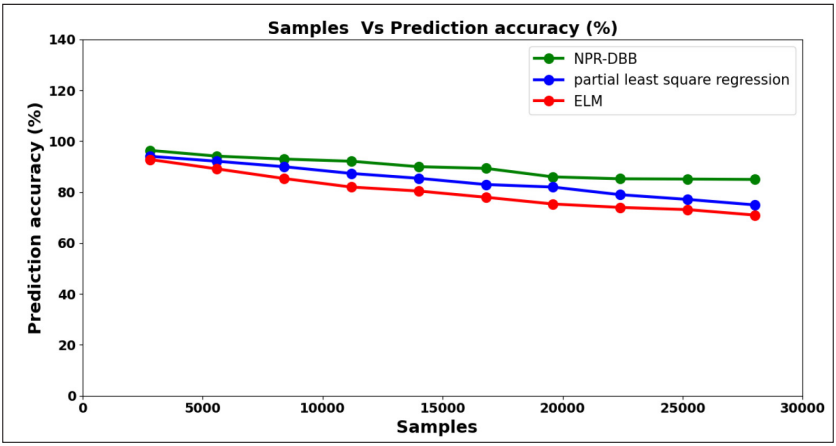

Fig. 4. Graphical representation of prediction accuracy

Fig. 4 given above illustrates the prediction accuracy with respect to 28000 different sample soil data instances acquired by means of soil moisture prediction dataset. As illustrated in the above graphical representation a linear decrease in the curve is inferred increasing the sample soil data instances provided as simulation process. This is due to the reason that increasing the sample soil data obtained for sample processing higher is the water soil data instances collected in the process for monitoring the soil moisture prediction. Hence, linearity is said to be observed between the sample soil data instances and prediction accuracy. However, with simulations performed with 2800 sample soil data instances, the prediction accuracy was observed to be 0.96 using NPR-DBB, 0.94 using [1] and 0.93 using [2] respectively. From this the prediction accuracy was found to be improved using the proposed BG-SCDP method when compared to [1] and [2]. This is because of the application of the Neyman-Pearson correlation-based feature selection. By applyng this algorithm, significant correlated features are initially selected with the aid of the Pearson correlation coefficient. Next, two distinct hypotheses were generated using the Neyman-Pearson detector function, therefore eliminating the redundant features associated. Due to this the prediction accuracy using BG-SCDP method was found to be comparatively better by 6% compared to [1] and 12% compared to [2].

Prediction time: The second significant parameter involved in the analysis of soil quality prediction is the prediction time. Prediction time here refers to

the time consumed in predicting the soil quality. This is mathematically formulated as given below.

$$P_{time} = \sum_{i=1}^{n} S_i * Time\ [SQP] \quad (14)$$

From the above equation (14), prediction time 'P_{time}' is measured based on the samples employed for simulation 'S_i' and the time consumed in soil quality prediction '*Time*[*WQP*]'. It is measured in terms of milliseconds (ms).

According to results obtained from (14), the prediction time is as shown in Table 3. As provided from the results of experiments in Table 3, the proposed NPR-DBB method has produced better results compared to partial least square regression [1] and ELM [2].

Table 3.

Comparative analysis for prediction time using NPR-DBB, partial least square regression [1] and ELM [2]

Samples	Prediction time (ms)		
	NPR-DBB	partial least square regression	ELM
2800	2380	3220	5180
5600	2550	3355	5855
8400	2915	3555	6135
11200	3135	4215	6355
14000	3255	4925	6585
16800	4155	5215	7025
19600	5325	6855	7935
22400	6145	7535	8325
25200	7255	8215	8815
28000	7500	9350	9955

Fig. 5 given above shows the prediction time with respect to 28000 sample soil data instances utilized for soil fertility or soil quality analysis. In the above figure the horizontal axis represents the number of sample soil data instances collected for simulation and the vertical axis represents the prediction time. From graph, the prediction time of NPR-DBB is comparatively lower than that of partial least square regression [1] and ELM [2] for soil quality prediction. Also, the soil quality prediction time is said to be increasingly proportional to the number of sample soil data instances involved in the process of simulation for analyzing soil quality. Moreover, with the simulations conducted using 28000 sample soil data instances, the time consumed in monitoring the soil quality were found to be 2380 ms using NPR-DBB method, 3220 ms using [1]

and 5180 ms using [2]. With this result, the prediction time was observed to be better. This is because of application of the linear regression machine learning prior to identification of highly correlated features. With this the prediction feature is said to be relevant if and only if there exists certain probability measure where the features are identified for being linearly regression with each other. This in turn only returned the computationally efficient relevant feature for further processing. Therefore, the prediction time of NPR-DBB is reduced by 22% compared to [1] and 41% compared to [2].

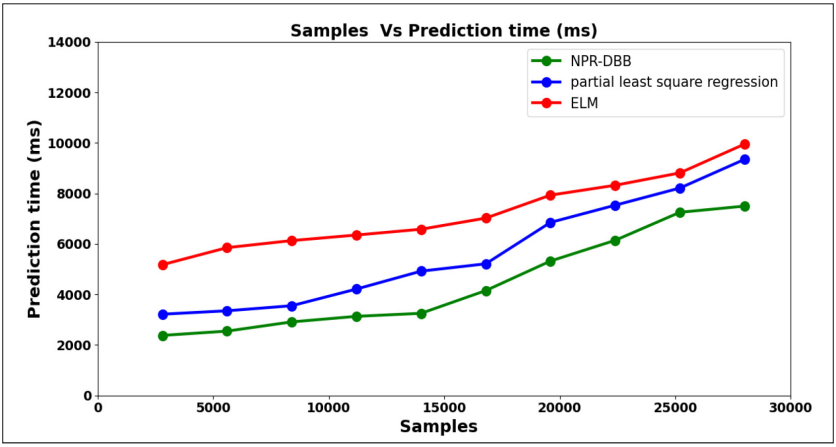

Fig. 5. Graphical representation of prediction time

Sensitivity: Sensitivity rate is defined as the rate that the method provides precise results of positive soil quality predictions. This is mathematically formulated as given below.

$$P_{time} = \sum_{i=1}^{n} S_i * Time\ [SQP] \tag{15}$$

From the above equation (15), sensitivity rate '*Sen*' is measured on the basis of the true positive rate '*TP*', (i.e., soil samples correctly predicted as it is) and the false negative rate '*FN*', (i.e., soil samples wrongly predicted) respectively.

Table 4 given below shows the sensitivity rate using the three methods, NPR-DBB, partial least square regression [1] and ELM [2] respectively for analyzing soil fertility or soil quality analysis to monitor the agricultural land suitability for different irrigation types with samples collected in the range of 2800 to 28000.

Table 4.

Comparative analysis for sensitivity using NPR-DBB, partial least square regression [1] and ELM [2]

Samples	Sensitivity (%)		
	NPR-DBB	partial least square regression	ELM
2800	0.98	0.96	0.95
5600	0.96	0.93	0.9
8400	0.95	0.9	0.88
11200	0.93	0.87	0.84
14000	0.91	0.85	0.81
16800	0.89	0.82	0.79
19600	0.86	0.8	0.75
22400	0.84	0.77	0.71
25200	0.82	0.74	0.69
28000	0.79	0.72	0.65

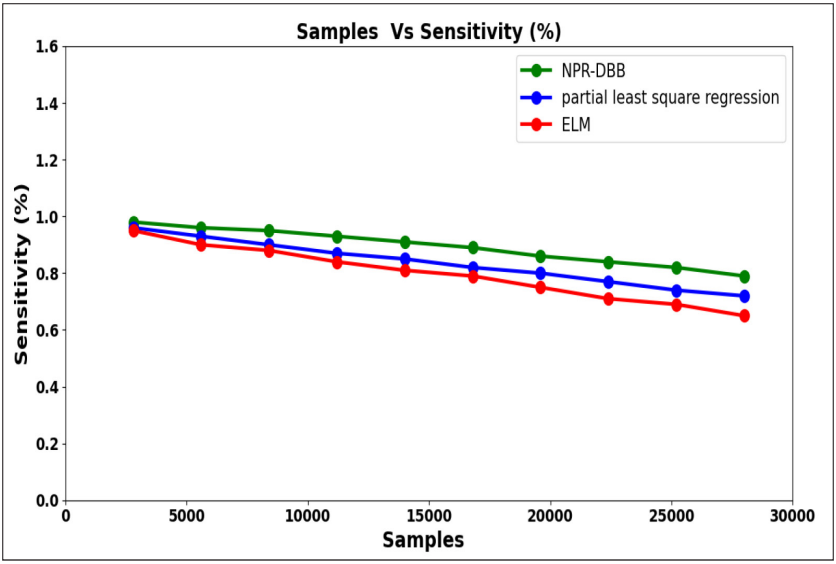

Fig. 6. Graphical representation of sensitivity

Fig. 6 given above show the sensitivity rate of NPR-DBB, Partial least square regression [1] and ELM [2]. It can be inferred from the figure that the NPR-DBB method is comparatively better to the [1] and [2] and its sensitiv-

ity rate is always the highest. When the number of sample soil data instance for soil quality analysis is small, the advantage of the sensitivity rate on the NPR-DBB method is not obvious. However, with the increase in the number of sample soil data instance, the sensitivity rate is found to be obvious. This is due to the reason that with the number of sample soil data instances being small, the moisture factors influenced in monitoring the soil quality is not found to be dense.

On the other hand, with the increase in the number of sample soil data instances the soil quality analysis indicators is also found to be dense and hence a decrease in the sensitivity rate is identified with higher number of sample data points. With 2800 sample soil data instances involved in simulation and 2755 samples being true positive using NPR-DBB method, 2715 samples being true positive using [1] and 2685 samples being true positive using [2], the overall sensitivity rate using the three methods were found to be 0.98, 0.96 and 0.95 respectively. With the application of Boltzmann factor and weight associating between feature instance in the current layer and feature instance in the previous layer that plays an important role in identifying the soil humidity owing to the reason that it being a robust parameter, designs for estimating the irrigation type. In other words, initially it filters the relevant features using hinge probability distribution and returns data mine results. Next, the final block with the aid of previous layer's result make a final soil moisture prediction on the basis of soil humidity, air temperature, air humidity and water need respectively. As a result, the sensitivity rate using NPR-DBB method is found to be better by 7% compared to [1] and 13% compared to [2].

Specificity: Specificity rate is also a significant metric to be estimated. This is obtained on the basis of the number of negative classifications in a class. This is mathematically formulated as given below.

$$Sen = \frac{TP}{TP+FN} \tag{16}$$

From the above equation (16), specificity rate '*Spe*' is measured on the basis of the true negative rate '*TN*', (i.e., soil samples correctly predicted negative class) and the false positive rate '*FP*', (i.e., soil samples falsified though samples are true) respectively.

Finally, the specificity rate is described in Table 5 for analyzing soil quality to monitor agricultural land with the comparison of NPR-DBB, partial least square regression [1] and ELM [2].

Table 5.

Comparative analysis for specificity using NPR-DBB, partial least square regression [1] and ELM [2]

Samples	Specificity (%)		
	NPR-DBB	partial least square regression	ELM
2800	0.96	0.94	0.93
5600	0.93	0.91	0.88
8400	0.92	0.88	0.85
11200	0.9	0.75	0.73
14000	0.89	0.73	0.7
16800	0.87	0.71	0.68
19600	0.85	0.69	0.63
22400	0.83	0.67	0.61
25200	0.8	0.65	0.6
28000	0.78	0.63	0.58

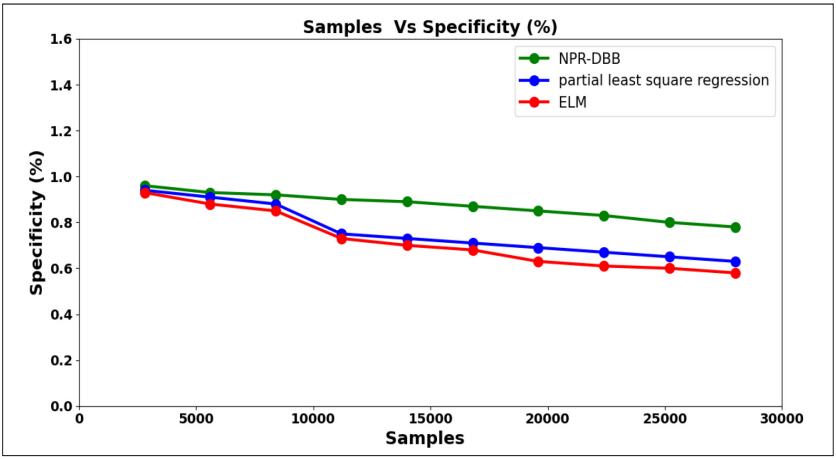

Fig. 7. Graphical representation of specificity

Finally, Fig. 7 given above shows the graphical representation of specificity for 2800 distinct sample soil data instances. From the above graphical representation, the specificity rate is found to be inversely proportional to each other. In other words, increasing the samples causes a decrease in the specificity rate. However, comparative analysis with the existing two methods, [1] and [2], shows improved specificity rate using NPR-DBB method. The reason behind the improvement was due to the application of deep Bernoulli and Boltzmann IoT-based soil quality prediction algorithm. By applying this algorithm, initially with the relevant

feature utilized as input, Boolean pair vectors were modeled. Second, logistic function based on Bernoulli distribution was formulated for significant and distinct differentiation between soil quality parameters. Finally, activation function with respect to visual based on soil humidity and hidden layers are formulated to arrive at the result, therefore reducing the false positive rate to a greater extent. With this the specificity rate was said to be improved using NPR-DBB method by 17% compared to [1] and 23% compared to [2] respectively.

Conclusion

The proposed deep learning technique called Neyman-Pearson regression and deep Bernoulli and Boltzmann (NPR-DBB) IoT-based soil quality prediction by using soil moisture dataset. First, with each sample soil data instances considered as input computationally efficient and relevant features are selected using the Neyman-Pearson correlation-based feature selection algorithm. Deep Bernoulli and Boltzmann IoT-based soil quality prediction algorithm is utilized to robust classified results. Simulation results demonstrate the efficient performance of the proposed NPR-DBB method. Also, comparison simulation results disclosed that the proposed method outperforms current state-of-the-art soil moisture prediction for agriculture in analyzing the irrigation level in terms of prediction accuracy and sensitivity. Also, it is shown that the prediction time of NPR-DBB method is minimal with respect to the optimal solution showing greater specificity, ensuring smooth soil quality analysis.

References

1. Helfera G.A., Barbosa J.L.V., Santos R. et al. A computational model for soil fertility prediction in ubiquitous agriculture. *Computers and Electronics in Agriculture*, Elsevier, 2020, vol. 175, pp. 105602. https://doi.org/10.1016/j.compag.2020.105602
2. Suchithra M.S., Pai M. L. Improving the prediction accuracy of soil nutrient classification by optimizing extreme learning machine parameters. *Information Processing in Agriculture,* Elsevier, 2019, vol. 7, no. 1, pp. 72-82. https://doi.org/10.1016/j.inpa.2019.05.003
3. Mohamed E. S., Belal AA., Abd-Elmabod S.K. et al. Smart farming for improving agricultural management. *The Egyptian Journal of Remote Sensing and Space Sciences,* Elsevier, 2021, vol. 24, no. 3, pp. 971-981. https://doi.org/10.1016/j.ejrs.2021.08.007
4. Sawalha S., Al-Naymat G. Towards an efficient big data management schema for IoT. *Journal of King Saud University – Computer and Information Sciences,* Elsevier, 2021. https://doi.org/10.1016/j.jksuci.2021.09.013

5. Andrianto H., Suhardi, Faizal A. et al. Performance evaluation of IoT-based service system for monitoring nutritional deficiencies in plants. *Information Processing in Agriculture,* Elsevier, 2021. https://doi.org/10.1016/j.inpa.2021.10.001
6. Akhtera F., Siddiqueia H.R., Alahib E.E., Mukhopadhyay S.C. Design and development of an IoT-enabled portable phosphatedetection system in water for smart agriculture. *Sensors and Actuators A: Physical,* Elsevier, 2021, vol. 330, pp. 1-11. https://doi.org/10.1016/j.sna.2021.112861
7. Benyezza H., Bouhedda M., Rebouh S. Zoning irrigation smart system based on fuzzy control technology and IoT for water and energy saving. *Journal of Cleaner Production,* Elsevier, 2021, vol. 132, pp. 1-15. https://doi.org/10.1016/j.jclepro.2021.127001
8. Kaur A., Sood S.K. Energy efficient cloud-assisted IoT-enabled architectural paradigm for drought prediction. *Sustainable Computing: Informatics and Systems,* Elsevier, 2020, vol. 30, pp. 1-15. https://doi.org/10.1016/j.suscom.2020.100496
9. Rejeb A., Rejeb K., Zailani S. Big data for sustainable agri-food supply chains: a review and future research perspectives. *Journal of Data, Information and Management,* Springer, 2021, vol. 3, no. 3, pp. 167–182. https://doi.org/10.1007/s42488-021-00045-3
10. Santhi M.V.B.T., Raju S. H., Krishna P.S.R. et al. Full smart sprinklers: Monitoring of sprinkler watering using IoT. *Materials Today: Proceedings,* Elsevier, 2020, vol. 10, no. 339, pp. 1-6. https://doi.org/10.1016/j.matpr.2020.12.399
11. Torky M., Hassanein A.E. Integrating blockchain and the internet of things in precision agriculture: Analysis, opportunities, and challenges. *Computers and Electronics in Agriculture,* Elsevier, 2020, vol. 178, pp. 1-23. https://doi.org/10.1016/j.compag.2020.105476
12. Geng X., Zhu C., Zhang J., Xiong Z. Prediction of Soil Fertility Change Trend Using a Stochastic Petri Net. *Journal of Signal Processing Systems,* Springer, 2020, vol. 93, no. 3, pp. 1-13. https://doi.org/10.1007/s11265-020-01594-3
13. Mupangwa W., Chipindu L., Nyagumbo I. et al. Evaluating machine learning algorithms for predicting maize yield under conservation agriculture in Eastern and Southern Africa. *Springer Nature*, 2020, vol. 2, no. 952, pp. 1-14. https://doi.org/10.1007/s42452-020-2711-6
14. Wu C., Chen Y., Hong X. et al. Evaluating soil nutrients of Dacrydium pectinatum in China using machine learning techniques. *Forest Ecosystems,* Springer, 2020, vol. 7, no. 30, pp. 1-14. https://doi.org/10.1186/s40663-020-00232-5
15. Inazumi S., Intui S., Jotisankasa A. et al. Artificial intelligence system for supporting soil classification. *Results in Engineering,* Elsevier, 2020, vol. 8, pp. 1-9. https://doi.org/10.1016/j.rineng.2020.100188
16. Xu Z., Zhao X., Guo X. Deep Learning Application for Predicting Soil Organic Matter Content by VIS-NIR Spectroscopy. *Computational Intelligence and Neuroscience,* Hindawi, 2019, vol. 2019, pp. 1-11. https://doi.org/10.1155/2019/3563761

17. Li J., Gao X., Guo B., Wu M. Production plan for perishable agricultural products with two types of harvesting. *Information Processing in Agriculture,* Elsevier, 2019, vol. 7, no. 1, pp. 83-92. https://doi.org/10.1016/j.inpa.2019.05.001
18. Padarian J., Minasny B., McBratney A.B. Machine learning and soil sciences: a review aided by machine learning tools. *Soil, European Geosciences Union*, 2020, vol. 6, no. 1, pp. 35-52. https://doi.org/10.5194/soil-6-35-2020
19. Santana E. J., Santos F. R., Mastelini S. M. et al. Improved prediction of soil properties with Multi-target Stacked Generalisation on EDXRF spectra. Elsevier, 2019, vol. 209, pp. 1-12. https://doi.org/10.1016/j.chemolab.2020.104231
20. Elavarasan D., Vincent P. M. D. Crop Yield Prediction Using Deep Reinforcement Learning Model for Sustainable Agrarian Applications. *IEEE Access,* 2020, vol. 8, pp. 86886-86901. https://doi.org/10.1109/ACCESS.2020.2992480
21. Anteh J.D., Timofeeva O.A., Mostyakova A.A. Assessment of Mineral Nutrient Impact on Metabolites Accumulation In Kale (Brassica Oleracea Var. Sabellica). *Siberian Journal of Life Sciences and Agriculture,* 2021, vol. 13, no. 3, pp. 208-224. https://doi.org/10.12731/2658-6649-2021-13-3-208-224
22. Suchkov D.K**.** Environmental and Economic Efficiency Protective Afforestation in the Arid Zoned. *Siberian Journal of Life Sciences and Agriculture,* 2021, vol. 13, no. 3, pp. 119-138. https://doi.org/10.12731/2658-6649-2021-13-3-119-138
23. Bondarenko V.L., Senova E.A., Gurina I.V., Aliferov A.V. Fundamental of technology convergence when water resource in agriculture production. *Siberian Journal of Life Sciences and Agriculture,* 2017, vol. 9, no. 1, pp. 100-114. https://doi.org/10.12731/wsd-2017-1-100-114
24. Wazihub Soil Moisture Prediction Challenge. URL: https://zindi.africa/competitions/wazihub-soil-moisture-prediction-challenge/data (accessed 20.12.2021)

DATA ABOUT THE AUTHORS

G. Balaji, PhD Research Scholar, Department of Computer Science
Sri Jayendra Saraswathy Maha Vidyalaya College of Arts and Science
182, SIHS Colony Road, Singanallur, Coimbatore 641005, Tamil Nadu, India
balajiphd08@gmail.com

P. Vijayakumar, Associate Professor & Head, Department of Computer Applications
Sri Jayendra Saraswathy Maha Vidyalaya College of Arts and Science
182, SIHS Colony Road, Singanallur, Coimbatore 641005, Tamil Nadu, India

Поступила 13.12.2021 Received 13.12.2021
После рецензирования 11.01.2022 Revised 11.01.2022
Принята 27.01.2022 Accepted 27.01.2022