International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

A Comparative Analysis of Machine Learning Models for Soil Health Prediction and Crop Selection

[1]Yogesh Mali, [2]Vijay U. Rathod, [3]Masira M. S. Kulkarni, [4]Pranita Mokal, [5]Sarita Patil, [6]Vidya Dhamdhere, [7]Dipika R. Birari

Submitted: 26/05/2023 **Revised**: 07/07/2023 **Accepted**: 23/07/2023

Abstract: This research paper explores the concept of soil health intelligence and crop recommendation using image analysis techniques. The proposed work focuses on predicting the soil type, pH, suitable crops that can be produced in that soil, and nutrients present in the soil. A soil health intelligence system is proposed in this work by combining machine and deep learning algorithms with image processing techniques. The system is trained with a dataset of soil images and another dataset containing the RGB (Red, Green, and Blue) values along with the pH of an image. The proposed model's ability to predict soil type and properties was assessed through the use of a test dataset, and the findings suggest that the system's accuracy is high, with minimal error. These results highlight the potential of image analysis as a real-world and competent approach for determining the properties of soil in both agriculture and soil science. The image dataset was trained with Convolutional Neural Networks to predict soil type, while the pH-recognition dataset was trained with many regression models, of which the XGBoost Regressor performed the best. Potential benefits of the system include giving farmers, agronomists, and researcher's vital information on soil management and crop productivity so they may make informed decisions. The findings have major implications for enhancing soil health and promoting sustainable agriculture.

Keywords: *Soil, pH, Predictions, Image processing, Machine learning.*

I. Introduction

Agriculture is a vital part of many countries' economies. India is one of the largest countries that depend solely on agriculture and related products. Agriculture employs more than 50% of the working population across the world and is one of the key drivers of the industrial revolution and a specific region's economy. Soil management plays a critical role to ensuring agricultural sustainability, and it is critical to understand the long-term repercussions of the various ways of managing soil and to pay close attention to soil quality. Like air and water, soil is considered a vital natural resource that offers a diverse range of paybacks to human beings in the form of commodities and services provided by various ecosystems. In agriculture, the soil is the most important and fundamental thing [1].

The quality of the soil is the single most significant factor in crop production in any form. The term "soil" can be understood in a variety of ways depending on who you ask. To a geologist, the term "soil" refers to the products of past surface products. It is significant to a penologist because it depicts the on-going physical and chemical processes that are taking place. According to the civil engineering perspective, soil is considered as the solid material that serves as a base for constructing the foundations of various structures such as buildings, roads, and other infrastructure. When referring to the rocks or minerals that are being extracted or mined, a mining engineer will refer to the soil as the dispersed particles of debris or residues that surround the target material. When it comes to highway engineering, the soil is the material that will be used as the base for the track bed.

It is possible for different people, like geologists, penologists, civil engineers, mining engineers, and highway engineers, for their own unique reasons, to define soils in a variety of distinct ways. The term "environmental medium" refers to the role that soil plays in agriculture from the perspective of an agriculturist. The study of soil involves the identification of externally observable patterns present in the soil. The study of soil characteristics is essential for understanding and managing the natural resources of a region. There are many different kinds of soil, and not all of them are ideal for growing every kind of crop. Because different types of soil each contain their own unique set of qualities that make them ideal for growing specific types of crops. Take sandy soil as an example; it requires a significant

[1], [2], [3], [4],[5],[6]G H Raisoni College of Engineering & Management Wagholi, Pune, Maharashtra, India
[7]Department of Information Technology, Army Institute of Technology, Pune, Maharashtra, India
[1]yogeshmali3350@gmail.com,[2]vijay.rathod25bel@gmail.com, [3]masirashaikh96@gmail.com,[4]pranitamokal62@gmail.com, [5]saritapatil555@gmail.com, [6]vidya.dhamdhere@gmail.com, [7]dipikabirari001@gmail.com, [7]Dipika R. Birari

amount of water. Clay soil, on the other hand, has a greater capacity to store water; hence, it has a lower overall water requirement. The study of soil entails becoming familiar with the distinct patterns that can be observed on soil. The soil classification system is utilized to categorize soil types into groups according to their respective soil properties [2]. The process of soil classification encompasses various stages, including analysing the composition and structure of the soil, determining its classification levels, and ultimately applying this knowledge in practical settings.

In recent years, there has been a rise in interest in the study of soil texture and colour classification by utilizing digital methods on photographs of soil. The classification of soils is an especially fundamental component of any viable agricultural enterprise. Studies classify the soil based on the colour of the soil and the texture of the soil.

The classification of soil can be carried out using a variety of standard approaches, both in the lab and in the field. Chemical analysis and image analysis are the two methods that can be utilized to determine the type of soil. Chemical analysis is often carried out in a laboratory using a variety of chemicals, a process that is not only expensive but also time-consuming, making it difficult for average farmers to access. [3, 4, 5] Image analysis-in most cases, engineers will categorize soils based on the engineering properties of the soil. The latest categorizations enable a seamless progression from on-site examination to initial projections of the characteristics and performance of soil mechanics. Soil sections captured with conventional cameras, microscopes, or scanners and observed under polarized light can be digitized to generate three-dimensional models that exhibit diverse geometric properties. The perspectives that soil is a resource and that soil itself is a material are both useful places to start when attempting to classify soil.

The pH value of the soil is the most important element to be examined before beginning the cultivation of any crop [6]. The pH level of soil is a crucial determinant of crop health and productivity, as well as the general well-being of an ecosystem. The determination of soil pH is crucial in identifying its inherent acidity or alkalinity, and optimal plant growth is contingent upon achieving the appropriate pH equilibrium, which can be ascertained via testing. The pH value of soil indicates its acidity or basicity [7]. Table 1 shows the pH value and its meaning. Effective crop management and sustainable agriculture necessitate accurate and efficient soil pH measurement and prediction systems.

Table 1: The pH value of the soil and meaning

pH value	7	<7	>7	>=5.5 and <7.0
Meaning	Neutral	Acidity	Basicity	Ideal for cultivation

Farmers usually provide soil samples to specialized laboratories for the purpose of evaluating soil pH or referring to soil pH colour charts. An expert periodically aids agricultural producers in ascertaining the soil's pH level. Nevertheless, obtaining proficient perspectives may not always be feasible in every circumstance. Each of the aforementioned alternatives necessitates a certain amount of time, exertion, and specialized expertise. The utilization of a soil pH chart as a sole means of assessing soil pH is deemed inadequate due to its reliance on human perception and the expertise of a trained specialist. To determine the pH level of soil in a laboratory setting, it is necessary to employ a soil pH meter and a soil colour pH card. The procedure of utilizing a pH meter on soil for a comparatively uncomplicated soil sample necessitated over an hour. The proliferation of technological advancements and greater computer usage has led to an increased prevalence of automation in daily life. Consequently, the acceleration of the process is accompanied by a reduction in the susceptibility of the final product to inaccuracies. Two techniques, namely image processing and regression, will be utilized to attain the objective of estimating the soil's pH [7].

In recent years, the use of machine learning and deep learning techniques has become increasingly popular in the field of soil science, as they offer a powerful tool for characterizing and predicting soil properties. On soil images, by applying image processing and machine learning approaches, soil pH can be forecasted. Regression analysis is a statistical method employed to establish the correlation between a dependent variable (soil pH) and one or more independent variables (image features). The objective of this research paper is to investigate the application of machine learning and deep learning methodologies in the categorization of soil varieties and the estimation of soil pH. This research aims to present an analysis of the current state of the art in soil characterization through the utilization of machine learning and deep learning. Additionally, the study will showcase the outcomes of a case study that highlights the efficacy of these techniques in soil type classification and pH prediction.

II. Related Works

The hue of the soil plays a crucial role in distinguishing diagnostic strata and features, which are utilized for the taxonomic categorization of the soil, such as mollic and umbric epipedons [8]. The absence of well-defined standards makes the task of ascertaining the hue of the soil a challenging endeavour. As per the American Standards Association [9, 10], the Munsell Soil Colour Chart [11] is solely employed for visual characterization purposes. According to Kirillova et al., the measurement of colour is typically conducted through the use of a spectrophotometer. The MSCC has been the preferred technique for soil colour determination in the past due to its comparative simplicity.

The colour chips of MSCC are generally expected to correspond with the colour of the soil. However, the degree of similarity between the two may be contingent upon the nature of the illumination and the dispersion of the lamp spectrum. Furthermore, the MSCC exhibits a high degree of subjectivity to human perception, thereby rendering the presentation of SOM prediction utilizing MSCC notation a challenging undertaking.

Diffuse reflectance spectroscopy (DRS) has been identified as a promising alternative approach for measuring soil organic carbon (SOC) and soil organic matter (SOM) due to the strong correlations observed between SOM and reflectance values in the visible and near infrared or mid-infrared range. This has been demonstrated in previous studies [12, 13, 14].

Melo and colleagues proposed a method of multivariate image analysis that utilizes picture segmentation to differentiate between clay and sandy soil types. This approach has been documented in several studies [3, 4 15]. Soil micromorphology-based image analysis was utilized for soil classification [16]. The authors Liu et al. [17] proposed a categorization system for urban soil utilizing a support vector machine (SVM) approach.

B. Bhattacharya et al. [18] employ segmentation, feature extraction, and classification techniques in their study. Segmentation algorithms are utilized to segregate measured signals. By means of the boundary energy method, characteristics are derived from the input data. Based on these identified characteristics, classifiers such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), and decision trees are employed to generate desirable outcomes.

Chung et al. [19] investigated the soil texture classification system based on RGB form images. Each segmented soil sample was subjected to image acquisition using a CCD camera that had been reduced in size, resulting in the acquisition of four surface pictures. The utilization of the pipette method was also employed in the analysis of the texture fractions. The results indicate that 48% of the soil samples yielded identical findings when comparing the in-situ image processing method and the laboratory method. The previously mentioned standards were employed to categorise the soil.

Shenbagavalli and Ramar [20] devised an algorithm for classifying soil based on its texture that makes use of mask convolution. In order to investigate the soil photos, feature extraction 3x3 Law's mask convolution was utilized. In order to construct the feature vector that will be used in the subsequent operations, the absolute mean, mean, skewness, kurtosis, and standard deviation of the soil picture were computed.

Bose Chaudhuri Hocquenghem codes (BCH) used multiclass support vector machines (SVM). The decoding process of a standard method remains constant irrespective of the Signal-to-Noise Ratio (SNR) conditions. The absence of local minima in Thaw, coupled with the high decoding capability of outlier resilient SVM for BCH codes [21], is noteworthy. ABDF is a software tool that offers a wide range of features and practical applications. Additionally, it presents a graphical user interface that facilitates the exploration of extensive data sets [22].

Zhongjie Zhang et al. have reported that a degree of uncertainty may exist between the correlation of soil composition and the mechanical behaviour of soil when obtained through the use of CPT. The lack of clarity leads to the convergence of numerous categories of soil. The present methodology involves the utilization of point and region estimation. In this context, the writer presents a novel fuzzy approach that is independent of CPT, as documented in reference [23].

I.T. Young et al. [24] present a notion for bending energy-based analysis of biological shape techniques. It also explains sampling theorem for connected and closed contours and provides a quick approach for determining bending energy. The utilization of RGB image processing in order to detect iron and carbon from soil images has been documented in references [25, 26]. Vibhute et al[27] have devised a methodology for categorising soil utilizing hyperspectral image data. Furthermore, the hyper spectral image was utilized in [28] and [29]. Nevertheless, the cost of a hyper-spectral imaging camera exceeds that of a smartphone camera. Stiglitz and colleagues utilized the Cyan, Magenta, Yellow, and Black colour conversion method derived from Munsell colour [25] to develop software for instantaneous identification of soil types [30].

The CIEL*a*b* mode was utilized to determine the soil moisture content in [31], while the L*a*b* mode was employed to evaluate the soil depth. In reference [32], a mobile application utilizing RGB image processing techniques has been developed for the purpose of identifying soil colour classification. The images were converted from the RGB colour space to the CIExyz and Munsell HVC colour models.

The classification of dry and moist sample photos obtained by GPS-enabled devices was the subject of a second investigation [33]. They also processed images in RGB, CMYK, CIElab, and XYZ modes.

An image was subjected to a computer vision-based texture analysis, as suggested by Sofou et al. [26]. The morphological partial differential equation-based method, which depended on the contrast of the picture, was utilized for the segmentation process. It was proposed to use variations in surface texture images as a local modulation component for the purpose of texture analysis. The in-field method that Breul and Gourves [34] presented for characterizing soil on the basis of its textural properties using third-order moment is described below. Textural investigation in the field using spectral methods applied to sub-surface soil pictures is one of the proposed investigations. This method seeks to rapidly distinguish fine material from coarse material and characterize a large percentage of materials with a grain size of 80 microns or finer. A proposal for a soil image retrieval system was made by Shenbagavalli and Ramar [28]. As a result of the findings, we concluded that the proposed retrieval method is effective.

The pH of the soil was predicted using the RGB color space of soil images [7, 35, 36, 37, 38, 39, 40]. The soil is photographed using a digital camera, and the following characteristics shown in equation 1 were utilized to forecast the pH of the soil.

$$Soil\ feature = Red\ /\ Green\ /\ Blue \quad (1)$$

Abu et al. [41] devised an expert system utilizing fuzzy logic to control soil pH. The method involved the correction of soil pH levels to facilitate the replacement of fertilizer by farmers and to ensure optimal plant quality. The physical properties of soil were demonstrated by Babu et al. [37] with a presented methodology. LabView was employed to construct the technology for fractal dimension measurement. The equation for fractal dimension is depicted in equation 2.

$$Fractal\ dimension = \frac{log\ (y(f))}{log\ (\frac{1}{f})} \quad (2)$$

24-bit color photographs are employed as input to assess the model's performance subsequently transformed into 8-bit, and the features are extracted utilizing an equation suggested by Kumar et al. [39]. In their study, Aziz et al. [35] employed the RGB values of images as provided by Kumar et al. [39] to train and test a neural network. The authors achieved an accuracy rate of 80% by utilizing a hidden layer comprising of 10 neurons. The pH value of the soil was determined through the computation of the mean RGB values of soil images by Garibashvili and Mahantesh S.D. et al. [38]. The mean RGB values were contrasted against both the factual soil pH and their projected pH. The authors Barman et al. [7] have proposed a technique for soil pH estimation that involves the utilization of HSV color image processing and regression methodologies, including linear, logarithmic, exponential, and quadratic models. The approach also involves the computation of hue, saturation, and value of soil images.

III. Proposed System

A. Dataset description

In the current work two different datasets are being used, the first being an image dataset for classifying the type of soil, consisting of 235 images (approximately 30 per class) which has been manually created by the authors. Soil is classified into eight types: "alluvial soil," "arid and desert soil," "black soil," "cinder soil," "laterite soil," "peat soil," "red soil," and yellow soil. Figure 1 shows the sample images of the dataset from each class. The second dataset utilized in this study is pH-recognition, a Comma Separated Value file, collected from Kaggle [42], which contains pH values for a range of R, G, and B combinations within an image, i.e., between 0 and 255. The dataset consists of four attributes, specifically blue, green, red, and label (pH). Table 2 shows the sample RGB values of the data set. Figures 2, 3, and 4 display scatter plots that illustrate the relationship between red, green, and blue values, ranging from 0 to 255, and their corresponding pH values, ranging from 0 to 14, within the dataset. It is evident from the plots that the data points are not linearly distributed, but rather exhibit a wide dispersion across the plot's dimensions.

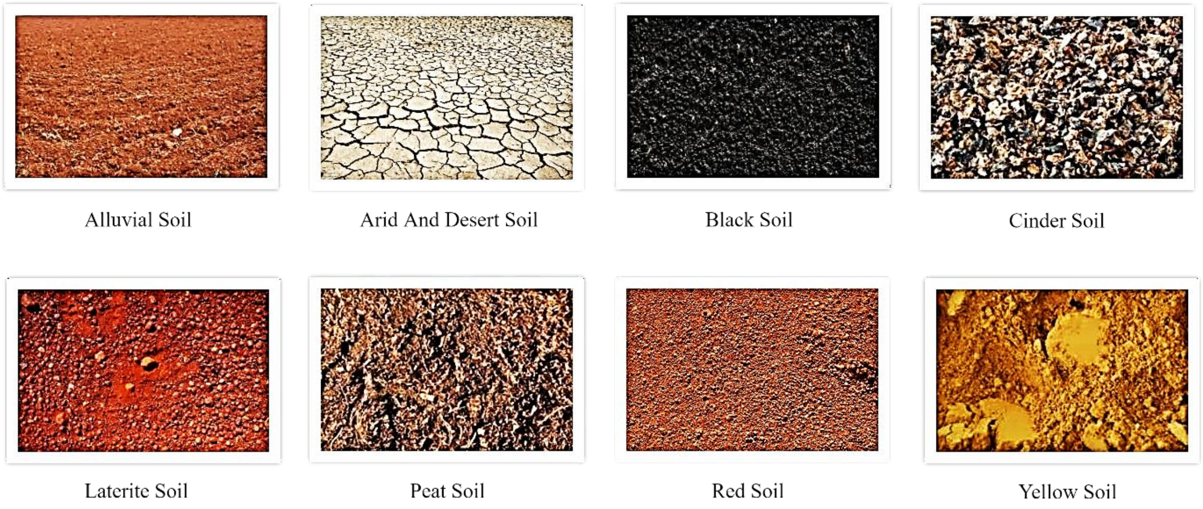

Fig 1: Sample images of the dataset from each class types of soils

Table 2: Sample RGB values of the dataset

Red (R)	25	55	67	67	104	183	207	166
Green (G)	212	197	185	170	181	185	143	82
Blue(B)	180	130	72	143	16	2	65	54

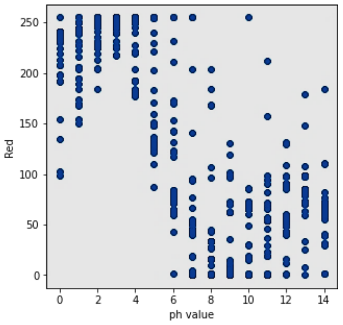

Fig 2: Plot of all the values of red w.r.t pH

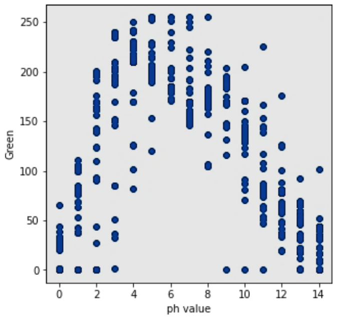

Fig3: Plot of all the values of green w.r.t pH

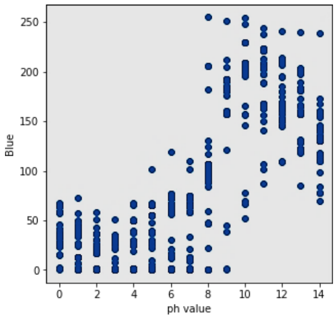

Fig 4: Plot of all the values of blue w.r.t pH

IV. Methodology

1. Linear regression

Linear regression is a statistical methodology employed to establish a linear correlation between a reliant variable and one or more autonomous variables. The objective is to formulate a mathematical equation thatcan forecast the dependent variable's value by considering the independent variables' values. Linear regression is a statistical technique utilized to construct a model that describes the association between two variables by employing a linear equation to the gathered data.

The fundamental formula for simple linear regression with a single predictor variable (X) is given 3.

$$Y = \beta 0 + \beta 1X + \varepsilon \quad (3)$$

Where:

Y is the dependent variable or outcome we want to predict. β0 represents an intercept or constant term.

β1 is the slope coefficient.

X is the variable of interest or predictor

ε is the random noise or error component t

The primary aim of linear regression is to ascertain the optimal values for β0 and β1 that minimize the sum of squared errors between the predicted and actual Y values. The conventional approach for this task involves the utilization of either the least squares technique or the gradient descent optimization algorithm.

2. Random Forest regression

The Random Forest Regressor is a form of ensemble learning technique that integrates numerous decision trees to generate forecasts. The concept underlying random forest regression involves the construction of numerous decision trees utilizing distinct subsets of both the data and features. Subsequently, the predictions of these trees are averaged to yield a more precise and resilient outcome.

The construction of each decision tree within a random forest involves the utilization of a random subset of the available features and training data. The implementation of this technique aids in mitigating over fitting while simultaneously augmenting the heterogeneity of the trees within the forest. In the course of the training procedure, every decision tree within the random forest is constructed via a recursive partitioning of the data into increasingly smaller subsets, utilizing the features that yield the highest information gain. After constructing the trees, they can be employed to generate forecasts on novel data by transmitting it through every tree and computing the mean outcomes.

The algorithm can be mathematically represented as follows:

1. Let X represents the matrix of input data, where each row represents an observation, and each column represents a feature.

2. Consider y to be the variable-vector target.

3. Let T represent the total number of decision trees within the forest.

4. For every tree from t = 1 to T:

4.1. A bootstrap sample is a subset of the data selected at random with replacement. Let X_t and y_t represents the vectors of data input and variable target for this sample.

4.2. Select at random a feature subset, often known as a feature subset. Let F_t represents the set of chosen features.

4.3. Train a decision tree using only the features in F_t and the data X_t and y_t.

4.4. The forest is where the decision tree is stored.

5. To make a prediction for a new observation x, compute and average the prediction for each decision tree in the forest, shown as in equation 4.

$$y\_hat(x) = 1/T * sum(y\_t(x)) \quad (4)$$

Where, y_t(x) is the output of the decision tree t for the input x.

V. LASSO Regression

The Least Absolute Shrinkage and Selection Operator (LASSO) regression is a linear regression approach that is employed to identify a subset of features for utilization in the model. This is accomplished by reducing the coefficients of irrelevant features to zero. The Lasso regression technique operates by incorporating a penalty term into the cost function of least squares, thereby imposing a restriction on the magnitude of the coefficients. The penalty term in question exhibits proportionality to the absolute value of the coefficients. This property results in the tendency of small coefficients to approach zero, leading to the exclusion of the corresponding features from the model.

Following is the mathematical formulation of LASSO regression:

1. The matrix of input data can be represented by X, where each row denotes an observation, and each column denotes an independent variable.
2. Consider y to be the variable-vector target.
3. Let β = (β0, β1, ..., βp) be the estimated vector of coefficients, where β0 is the intercept term.
4. The LASSO regression problem may be expressed as shown is equation 5.

$$\text{minimize } \left|\left|y - X\beta\right|\right|2 + \lambda * |\beta||\_1 \quad (5)$$

Where ||.||_1 is the L1 norm and λ is a hyper parameter that controls the strength of the penalty term. The larger the values of λ, the more coefficients are pushed towards zero, resulting in sparser solutions.

5. The norm of in L1 is defined as:

$$||\beta||\_1 = |\beta1| + |\beta2| + ... + |\beta p|$$

6. Optimization techniques such as coordinate descent and gradient descent can be used to deal with the optimization problem.

Coordinate descent updates is given by equation 6.

$$\beta\_j = S(z\_j, \lambda/2) / (X\_jT X\_j) \quad (6)$$

Where $S(z, \lambda/2)$ is the soft-thresholding operator defined as given in equation 7.

$$S(z, \lambda/2) = sign(z) * max(0, |z| - \lambda/2) \quad (7)$$

And $z\_j = X\_j\text{^}T (y - X\_\{-j\} \beta\_\{-j\})$ is the residual, where $X\_\{-j\}$ and $\beta\_\{-j\}$ are the matrices and vectors without the j-th column.

Gradient descent updates are shown in equation 8.

$$\beta = \beta - \eta * \nabla(||y - X\beta||2 + \lambda * ||\beta||\_1) \quad (8)$$

Where η is the rate of learning and (.) is the gradient operator.

3. XGB Regression

Extreme Gradient Boosting (XGBoost) The regression algorithm is a widely used machine learning technique for performing regression tasks. The XGBoost algorithm is a boosting technique that leverages an ensemble of weak prediction models, specifically decision trees, to construct a robust model for predictive tasks. The XGBoost algorithm operates through a process of iteratively augmenting the ensemble model with decision trees, wherein each subsequent tree is designed to minimize the residual errors of its predecessor. The aforementioned procedure is iterated until either a predetermined limit of trees is attained or until no additional enhancements can be achieved.

In order to mitigate the issue of over fitting, XGBoost employs regularization methods, such as L1 and L2 regularization, as well as early stopping, which enable the algorithm to halt the addition of new trees when the validation loss fails to demonstrate improvement beyond a specified number of iterations.

The XGBoost regression algorithm is mathematically described as follows:

1. The matrix of input data can be represented by X, where each row denotes an observation, and each column denotes an independent variable.

2. Consider y to be the variable-vector target.

3. Let h_i(x) represent the i-th decision tree's forecast.

4. The objective function of XGBoost is specified as shown in equation 9.

$$Obj = 1/2 * sum\_in [y\_i - sum\_kM f\_k(x\_i)]2 + gamma * sum\_kM Omega(f\_k) \quad (9)$$

Where n represents the number of observations, M represents the number of decision trees, f_k represents the prediction of the k-th tree, gamma controls the complexity of the trees, and Omega(f_k) penalises the complexity of the k-th tree.

5. Optimizing the objective function using gradient descent.

Gradient updates are given in equation 10.

$$g\_i = \partial(y\_i - sum\_kM f\_k(x\_i)) / \partial f\_M(x\_i) \ h\_i = \partial 2(y\_i - sum\_kM f\_k(x\_i)) / \partial f\_M(x\_i)2 \ f\_k = f\_k - \eta * [sum\_i g\_i / (sum\_i h\_i + lambda)] \quad (10)$$

Where g_i is the first derivative and h_i is the second derivative of the loss function with respect to the prediction, respectively, η is the learning rate, and lambda is the L2 regularization parameter.

6. When a maximum number of trees are achieved or when the improvement in the goal function falls below a specific level, the algorithm terminates.

4. Ridge regression

Ridge Regression is a linear regression methodology that is used to examine data that exhibits multicollinearity, a phenomenon characterized by the existence of independent variables that are highly correlated. The conventional approach to linear regression involves minimizing the sum of squared residuals. However, Ridge Regression introduces an extra penalty term to the cost function. The penalty term is a mathematical function that is dependent on the squared magnitude of the coefficients and is subsequently multiplied by a parameter referred to as lambda, which is also recognized as the regularization parameter. The inclusion of a penalty term aids in the contraction of coefficients towards zero, thereby diminishing their variance and subsequently mitigating the impact of multicollinearity. Consequently, Ridge Regression is more appropriate for datasets exhibiting a high degree of correlation among independent variables. The determination of the optimal lambda value can be achieved through various methods, including cross-validation. It is worth noting that an increase in lambda values leads to a greater degree of shrinkage of the coefficients.

Following is the mathematical formulation of ridge regression:

1. The matrix of input data can be represented by U, where each row denotes an observation, and each column denotes an independent variable.

2. Consider y to be the variable-vector target.

3. Let β = (β0, β1, ..., βp) be the estimated vector of coefficients, where β 0 is the intercept term.

4. The ridge regression problem is stated in equation 11.

$$\text{minimize } ||y - U\beta||2 + \lambda * ||\beta||2 \quad (11)$$

Where ||.|| is the L2 norm and λ is a hyper parameter that controls the strength of the penalty term. The larger the value of λ, the smaller the coefficients become, resulting in a more stable and less flexible model.

5. The L2 norm of is defined as follows:

$$||\beta||2 = \beta12 + \beta22 + ... + \beta p2$$

6. Using linear algebra techniques such as matrix inversion or singular value decomposition, the optimization problem can be solved.

The answer to the problem of ridge regression is given by equation 12.

$$\beta = (UT\ U + \lambda I)(-1)\ UT\ y \quad (12)$$

Where I represent the p x p identity matrix.

5. Elastic net

Elastic net linear regression utilizes regularization techniques from both the lasso and ridge methodologies to impose penalties on regression models. The proposed approach integrates the lasso and ridge regression techniques, leveraging their respective limitations to enhance the efficacy of statistical model regularization.

The elastic net approach addresses the limitations of the lasso method, which is restricted to a small number of samples in the context of high-dimensional data. The elastic net methodology permits the inclusion of "n" predictors until reaching saturation. In cases where the variables exhibit high interconnectivity, the lasso method tends to favors the selection of a single variable from each group while disregarding the remaining variables.

The elastic net is a regularization technique that addresses the limitations of lasso by incorporating a quadratic term (||β||2) in the penalty function. When utilized independently, this term corresponds to ridge regression. The quadratic expression of the penalty term induces convexity in the loss function. The elastic net model integrates the favorable attributes of both lasso and ridge regression techniques.

The elastic net method's estimate is determined through a two-stage procedure that involves the utilization of the lasso and regression approaches. The initial step involves the identification of the coefficients for ridge regression, followed by a subsequent stage wherein the coefficients are subjected to shrinkage via lasso.

Consequently, the coefficients undergo dual forms of shrinkage through the utilization of this approach. The utilization of the naive form of the elastic net leads to a dual shrinkage phenomenon that yields diminished predictability and notable bias. The rescaling of coefficients is performed by multiplying them with (1+λ2) in order to accommodate the aforementioned impacts.

The mathematical formulation of elastic net is as follows:

1. The matrix of input data can be represented by X, where each row denotes an observation, and each column denotes an independent variable.

2. Consider y to be the variable-vector target.

3. Let β = (β0, β1, ..., βp) be the estimated vector of coefficients, where β0 is the intercept term.

4. The elastic net problem can be expressed as shown in equation 13.

$$Minimize\ ||y - X\beta||2 + \lambda 1 * ||\beta||\_1 + \lambda 2 *||\beta||2 \quad (13)$$

Where ||.||_1 is the L1 norm and ||.||2 is the L2 norm, λ1 and λ2 are hyper parameters that determine the penalty term strengths. The greater the values of λ1 and λ2, the lower the coefficients, resulting in a model that is more stable and less flexible.

5. The norm of β in L1 is defined as: ||β||_1 = |β1| + |β2| + ... + |βp|

6. It is possible to tackle the optimization problem using iterative techniques such as coordinate descent or gradient descent.

The following updating rules shown in equation 14 apply to the elastic net algorithm.

$$\beta j = S(\Sigma\_i=1\hat{}n(xij)(yi - \Sigma\_k\neq j\ xik * \beta k), \lambda 1) / (\Sigma\_i=1\hat{}n\ (xij)\hat{}2 + \lambda 2) \quad (14)$$

Where S represents the soft-thresholding operator and is defined as shown in equation 15.

$$S(z, \gamma) = sign(z) * max(0, |z| - \gamma) \quad (15)$$

6. Polynomial regression

The polynomial regression technique is utilized to model the association between a dependent variable y and an independent variable c as a polynomial of nth degree. Polynomial Regression is a type of Linear Regression that involves modeling a non-linear relationship between the independent and dependent variables using a polynomial equation. This approach is used when there is a curvilinear pattern in the datapoints.

Polynomial regression can model relationships between variables of any degree, as contrast to linear regression, which only models a linear relationship between the input variables and the output. Finding the optimal coefficients to minimize the discrepancy between the expected and actual values entails fitting a polynomial equation to the data.

Following is the mathematical formulation of polynomial regression:

1. The matrix of input data can be represented by C, where each row denotes an observation, and each column denotes an independent variable.

2. Consider y to be the target variable-vector.

3. Let n indicate the degree of the polynomial function.

4. The following expression describes the polynomial regression model:

$$y = \beta 0 + \beta 1c + \beta 2c2 + \dots + \beta n{*}cn + \varepsilon$$

Where 0, 1,..., c are the estimated coefficients, ci is c raised to the power of i, and is the error term.

5. The objective of OLS regression is to minimize the sum of squared errors, as in equation 16.

$$SSE = \Sigma(yi - f(ci))^2 \tag{16}$$

Where (ci) is the expected value of y from the ith observation.

6. Minimizing the SSE yields the coefficient OLS estimator as given in equation 17.

$$\beta = (C^T\, C)^{(-1)}\, C^T\, y \tag{17}$$

Where CT represents the transposition of C and (CT C)$^{(-1)}$ represents the inverse of the matrix product CT C.

7. AdaBoost Regressor

The AdaBoost Regressor is a meta-estimator that initiates the modeling process by fitting a Regressor on the initial dataset. Subsequently, the dataset is subjected to fitting of additional Regressor copies, wherein the weights of instances are modified based on the error of the present prediction. Subsequently, the AdaBoost Regressor is trained on the initial dataset once more. Consequently, as a result of this phenomenon, subsequent Regressor tends to priorities cases that present greater difficulties. The current study utilized multiple models as base estimators, with the specific choice of base estimator being the Random Forest Regressor and Decision Tree Regressor for prediction purposes.

Following is the mathematical formulation of the AdaBoost Regressor:

1. The matrix of input data can be represented by R, where each row denotes an observation, and each column denotes an independent variable.

2. Consider y to be the variable-vector target.

3. Let h(r) be a model of weak regression with a maximum depth of 1.

4. Let E be the number of weak learners, also known as the number of iterations.

5. Initialize the weights a_h associated with each observation as follows:

$$a\_h = 1/o$$

Where n represents the number of observations.

6. For every iteration m between 1 and E:

a. Fit a weak regression model z_e(r) using the current weights a to the training data.

b. Determine the weak learner's weighted mean absolute error as per equation 18.

$$\varepsilon\_e = \Sigma\_i{=}1\hat{}o\ a\_h\ |y\_h - m\_e(r\_h)| \,/\, \Sigma\_h{=}1\hat{}o\ a\_h \tag{18}$$

c. Calculate the weak learner’s weight as given in equation 19.

$$\alpha\_e = \log((1-\varepsilon\_e)/\varepsilon\_e) \tag{19}$$

d. Update each observation's weights w_i as shown in equation 20.

$$a\_h = a\_h * \exp(\alpha\_e * |y\_h - z\_e(r\_h)|) \tag{20}$$

e. Normalize the weights so their sum equals 1 as in equation 21.

$$a = a \,/\, \Sigma\_h{=}1\hat{}o\ a\_h \tag{21}$$

7. The conclusive prediction is the weighted total of the weak learners is given by equation 22.

$$f(r) = \Sigma\_e{=}1\hat{}E\ \alpha\_e\ z\_e(r) \tag{22}$$

8. Decision Tree Regressor

A Decision Tree (DT) Regressor creates a tree-like structure to depict the relationship between a dependent variable and one or more independent variables. Each node in the tree represents a decision or branch depending on the value of a certain independent variable, and each branch reflects the possible outcomes of that decision.

The final predictions are made by tracing the tree's branches to the correct leaf node. Finding the splits that minimize the variance of the dependent variable is how the algorithm operates. The ultimate result of the tree is the mean target value of the training instances located in the corresponding terminal node. The decision tree Regressor model is characterized by its ease of comprehension and interpretability, as well asits capacity to handle both linear and non-linear associations between features and target variables.

Following is the mathematical description of decision tree regression:

1. The matrix of input data can be represented by M, where each row denotes an observation, and each column denotes an independent variable.

2. Consider y to be the variable-vector target.

3. Let T represent a DT with internal nodes. Each internal node represents a feature-based split, whereas each leaf node represents a prediction.

4. For each internal node j, let f_j represent the feature upon which the split is based and let s_j representthe split threshold value. The division can then be stated as:

If x_fj s_j, then proceed to the left child; otherwise, proceed to the right child.

5. Let c_k represent the anticipated value for the observations that fall under each leaf node k.

6. The objective is to identify the tree T that minimizes the variance of the dependent variable across all leaf nodes. The definition of the variance of the dependent variable at a leaf node k is given by equation 23.

$$Var(y\_q) = 1/g\_q\ \Sigma\_i=1\hat{}g\_k\ (y\_i - v\_q)^2 \qquad (23)$$

Where g_q is the number of observations that fall under node q and v_q represents the expected value forthose data.

7. The criterion for dividing each internal node is the variance reduction that results from the split. Thevariance reduction for split j is defined as shown in equation 24.

$$\Delta Var(p) = Var(y) - (Var(y\_left) + Var(y\_right)) / 2 \qquad (24)$$

Where Var(y) is the variance of the dependent variable in the parent node, Var(y_left) is the variance of the dependent variable in the child node that is towards the left, and Var(y_right) is the variance of the dependent variable in child node that is towards the right.

8. The tree grows recursively by selecting the feature and threshold that yields the largest variance reductionand continuing the process on the subsequent subsets until a stopping requirement is reached.

9. MLP Regressor

The MLP Regressor, also referred to as the Multi-Layer Perceptron Regressor, is a neural network utilizedfor the purpose of addressing regression-related problems. The system is composed of multiple layers of synthetic neurons that analyze and alter the input information in order to anticipate the desired outcome. Neurons receive inputs, undergo a non-linear activation process, and subsequently transmit the modified outputs to the succeeding layer. The intermediate layers are utilized to obtain complex representations of the input data. The algorithm is trained through the utilization of the gradient descent method, with the objective of minimizing the difference between the predicted and actual target values. The MLP Regressor exhibits the ability to effectively manage intricate and non-linear connections between the input and outputvariables.

The MLP Regressor can be mathematically represented as follows:

1. The matrix of input data can be represented by S, where each row denotes an observation, and each column denotes an independent variable.

2. Consider y to be the variable-vector target.

3. Let D and h represent, respectively, the weight matrix and bias vector for each neuron in the MLP Regressor.

4. Let f represent the activation function applied to the weighted sum of each neuron's inputs.

5. The MLP Regressor output for a given input s is given by equation 25.

$$y\_pred(s) = f(D\_2 * f(D\_1 * x + h\_1) + h\_2) \qquad (25)$$

Where D_1 and h_1 represent the weight matrix and bias vector for the first layer of neurons, D_2 and h_2 represent the weight matrix and bias vector for the second (output) layer of neurons, and f represents the activation function.

6. Determining the weight matrices D_1 and D_2 and bias vectors h_1 and h_2 that minimise the mean squared error (MSE) between the predicted values y_pred(s) and the actual target values y in the training set is the goal as in equation 26.

$$MSE = 1/m * \Sigma\_i=1\hat{}m\ (y\_i - y\_pred(s\_i))^2 \qquad (26)$$

Where m is the number of training set observations.

7. Using gradient descent, the optimization problem can be handled by updating the weights and biases inthe direction of the negative gradient of the MSE with respect

to the weights and biases as in equation 27.

$$D\_1 = D\_1 - \alpha * \partial MSE/\partial D\_1 h\_1 = h\_1 - \alpha * \partial MSE/\partial h\_1$$
$$D\_2 = D\_2 - \alpha * \partial MSE/\partial D\_2$$
$$h\_2 = h\_2 - \alpha * \partial MSE/\partial h\_2 \quad (27)$$

where α is the learning rate, which controls the size of the weight updates, and ∂MSE/∂D and ∂MSE/∂h are the partial derivatives of the MSE with respect to the weights and biases, respectively.

8. The partial derivatives can be calculated using the back propagation technique, which propagates the error from the output layer back through the network and uses the chain rule to calculate the derivatives.

The Mean Squared Error (MSE) is a statistical metric utilized for assessing the accuracy of a predictive or estimation model as shown in equation 28. MSE value of zero signifies an exact correspondence between the anticipated and factual values. Conversely, higher MSE values indicate more substantial deviations between the predicted and actual values.

$$MSE = 1/h * \sum ((i) - \hat{a}(i))\ 2 \quad (28)$$

Where,

h: The count of number of data points,

a (i) : The actual value of the data point and

a ̆(i) : The corresponding predicted value of the data point

Root Mean Squared Error

The Root Mean Square Error (RMSE), a widely employed statistical metric as shown in equation 29, utilized for assessing the accuracy of a prediction or forecasting model. RMSE serves as an indicator of the degree of deviation between the predictions and actual values. A reduced RMSE value is indicative of a superior model fit to the data, as it implies a decreased discrepancy between the projected and observed values.

$$\mathrm{RMSE} = \frac{\sqrt{\sum ||a(i) - \hat{a}(i)||\ 2\ F\ i=1 \div F}}{} \quad (29)$$

Where,

F: The count of number of data points,

a(i) : The actual value of the datapoint and

a ̆(i) : The corresponding predicted value of the datapoint.

Mean Absolute Error

Mean Absolute Error (MAE) is a performance metric for regression models. The MAE is computed by accumulating the absolute value of the difference between the predicted and actual values and dividing by the total number of observations as per equation 30. The resultant value represents the average deviation between the predicted and actual values, with smaller values indicating superior model performance.

$$MAE = |(si\text{-}sj)\ |\ /\ j \quad (30)$$

Where,

j : Number of data points,

si : Actual value and

sj : The predicted value.

Coefficient of Determination (R^2)

R^2 score as per equation 31 is a statistical indicator of the proportion of variance in the dependent variable (target variable) that can be explained by the independent variables (features) in a regression model. The correlation is expressed as a numeric value between 0.0 and 1.0, where 1.0 represents a perfect fit, which is highly reliable for making future predictions. A value of 0.0, on the other hand, indicates that the model is erroneous and fails to model the data effectively.

$$R = m\left(\sum cd - (\sum c)(\sum d)\right)/\sqrt{}[m\sum c2 - (\sum c)2][m\sum d2 - (\sum d)2] \quad (31)$$

Where,

m = Total number of observations Σc = Sum of values of variable c Σd = Sum of values of variable d

Σcd = Sum of the Product of variables c and d Σc^2 = Sum of Squares of the variable c

Σd^2 = Sum of Squares of the variable d

Coefficient of determination =Square of correlation coefficient = R^2

Where,

m = Total number of observations,

Σc = Sum of values of variable c

Σd = Sum of values of variable d

Σcd = Sum of the Product of variables c and d

Σc^2 = Sum of Squares of the variable c

Σd^2 = Sum of Squares of the variable d

Coefficient of determination =Square of correlation coefficient = R^2

Process flow and Architecture of the Proposed work:

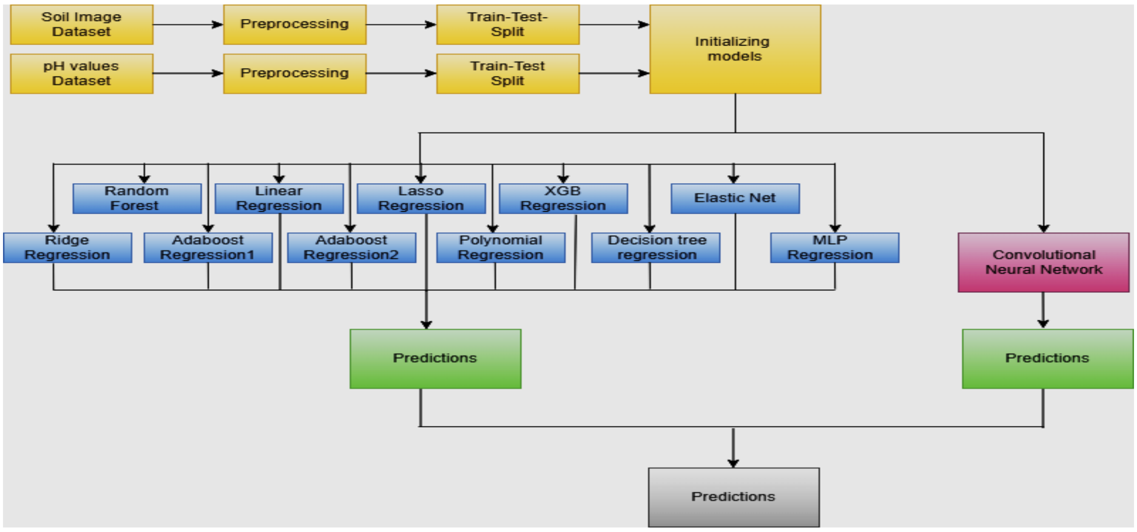

Fig 5: Process flow of the proposed work

Convolutional Neural Network (CNN)

CNN is a class of deep learning models that are commonly employed in tasks related to the recognition of images and videos. The convolutional layer is a fundamental element of a Convolutional Neural Network (CNN) and executes a mathematical function known as convolution.

The convolution operation takes in an input (image), a set of filters (also known as kernels or weights) and applies the dot product of each filter with the overlapping regions of the input. The result is a new feature map that summarizes the local features in the input. By repeating this operation multiple times, CNN is ableto extract high-level features from the input image.

The key mathematics behind convolutions includes linear algebra and calculus. Specifically, the dot product and matrix multiplication are used to compute the convolution, while gradient descent and back propagation are used to update the network weights during training.

Convolution: It is a mathematical operation that computes the sum of element-wise multiplications between the input and a small, fixed filter.

Pooling: Pooling is a technique used in deep learning that involves down sampling the input data to reduceits spatial dimensions. This process is designed to retain critical features while discarding non-essential information. Max pooling is the most frequently used pooling operation.

Activation Functions: Activation functions are mathematical functions that are utilized to introduce non- linearity into a neural network. This non-linearity enables the network to acquire intricate representations of the data. The rectified linear unit (ReLU), sigmoid, and hyperbolic tangent (tanh) are among the frequently employed activation functions in CNNs.

Matrix Multiplication: It is a fundamental operation in linear algebra that is used in the fully connected layers of a CNN to make the final predictions.

Back propagation: Back propagation is an algorithm that is commonly utilized to calculate the gradients ofthe loss function in relation to the network weights. These gradients are subsequently used to modify the weights through the application of gradient descent.

Pooling layers are frequently employed in CNNs to perform down sampling of the feature maps, thereby decreasing their dimensionality. The process is commonly executed via operations such as max pooling, wherein the maximum value within each region of the feature map is selected.

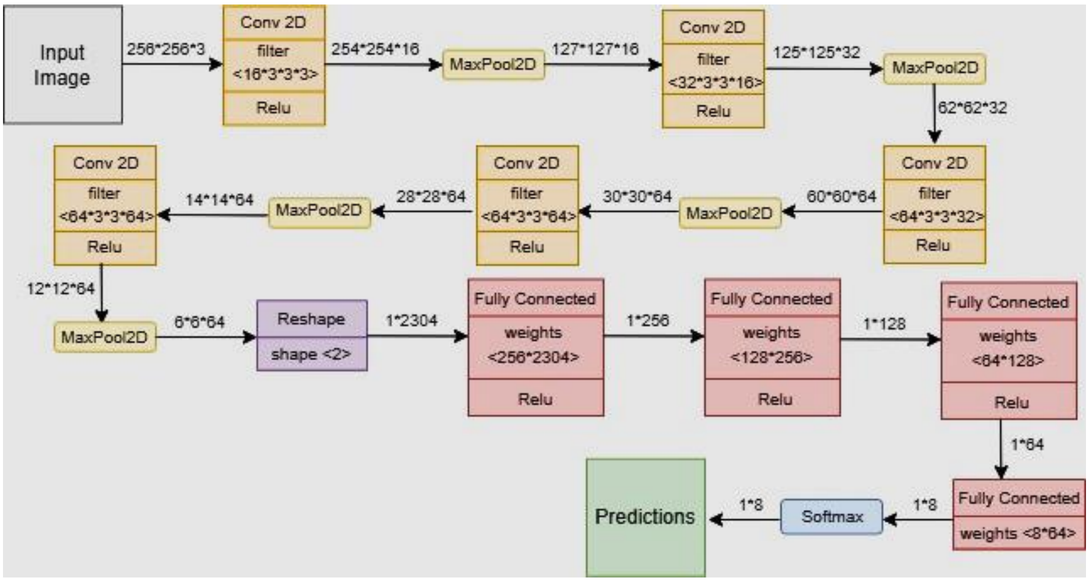

Fig 6: CNN architecture

The 235-image dataset was imported, minimal preprocessing has been done and then splitted into train, testand validation datasets in the ratio of 0.8:0.1:0.1(i.e., 80% data for train set and 10% each for test and validation sets) and has been ran for 100 epochs. As the dataset consists of images, Convolutional Neural Networks (CNN) has been chosen for the current work. An image size of 256*256 has been set and passedto a CNN consisting of 6 convolutional layers and 6 max pooling layers alternatively placed followed by a flatten and couple of dense layers. Sparse categorical cross entropy has been used as the loss function as the current work is a multiclass classification and optimizer being set to Adam who has been clearly depicted in figure 6. Results with graphical interpretation have been discussed in the section below. The current dataset would predict the type of soil. The pH-recognition dataset [42] contains 653 rows and fourattributes namely blue, green, red, and label (pH), in which 80% of the dataset being used for training and 20% used for testing as the dataset appeared to be flawless and did not contain any null values or superfluous attributes, a significant amount of the pre-processing of the data was skipped. Both the quality and quantity of the data that were acquired are considered in the analysis. The values that were determined are also presented in graphical form.

The predictions were done using multiple regression models namely Decision Tree Regressor, XGB Regressor, Lasso Regressor, Elastic Net, Linear Regressor, Ridge Regressor AdaBoost Regressor(base_estimator=Random Forest Regressor), Random Forest Regressor, AdaBoost regressor(base_estimator=Decision Tree Regressor), Polynomial Regressor, and MLP Regressor.

Various parameters have been customized during the training of the model which includes setting degree to4 for polynomial regression, maximum depth of two for the Random Forest Regressor, regularization hyper parameter being 0.0001 for elastic net, lasso and ridge regressors, learning rate of 0.5 and 220 estimators for XGB Regressor, maximum iterations were set to 50, early stopping being true and solver being 'Limited-memory BFGS'(Broyden–Fletcher–Goldfarb–Shanno) for MLP Regressor.Adaboost regressor with base estimator being decision tree regressor and Decision Tree Regressor with default parameters have been used and no customizations have been done. Figure 5 shows the process flow of theproposed work.

VI. Results and discussions

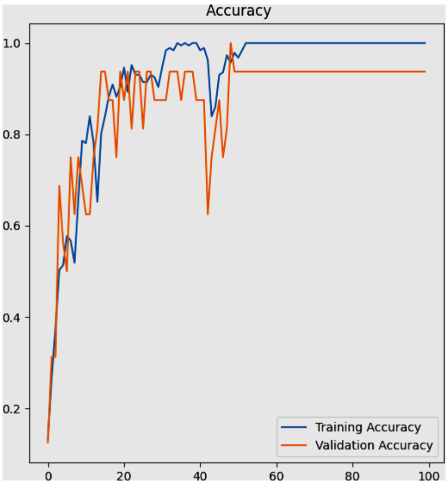

Fig 7 : Test and train accuracies of proposed CNN model

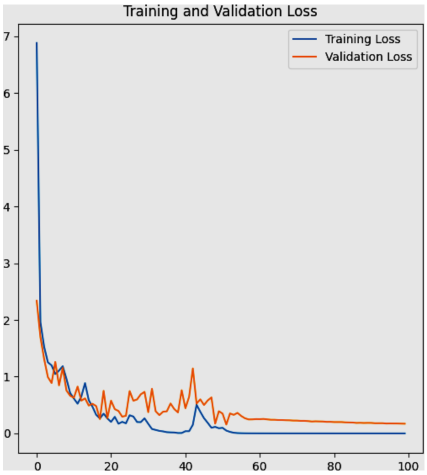

Fig 8 : Test and train losses of proposed CNN model

Figures 7 and 8 depict the performance of CNN model where our test and train accuracies have been 90.62%and 91.66%, followed by loss and validation loss being 0.27 and 0.17.

MAE, MSE, RMSE and R^2 score values of various models are shown in Table 3. From Table 3, XGBoost Regressor has been the best among all the regressors with the highest R^2 score of 98% followed by Adaboost Regression (be=Decision Tree Regressor) and Polynomial Regression(d=4) with R^2 scores of 97% and 95%respectively.

Table 3: MAE, MSE, RMSE and R^2 score values of machine learning models.

S.No	Model	MAE	MSE	RMSE	R^2
1.	XGB Regression	0.38	0.38	0.61	0.98
2.	Adaboost Regression(be=Decision Tree Regressor)	0.36	0.56	0.75	0.97
3.	Polynomial Regression(d=4)	0.65	0.87	0.93	0.95
4.	Decision Tree Regression	0.46	1.02	1.01	0.94
5.	Adaboost Regression(be=Random Forest Regressor)	0.90	1.25	1.11	0.93
6.	Random Forest Regression	0.94	1.43	1.19	0.92
7.	MLP Regressor	0.86	2.13	1.46	0.89
8.	Elastic Net	1.79	4.90	2.21	0.75
9.	Lasso Regression	1.79	4.90	2.21	0.75
10.	Linear Regression	1.79	4.90	2.21	0.75
11.	Ridge Regression	1.79	4.90	2.21	0.75

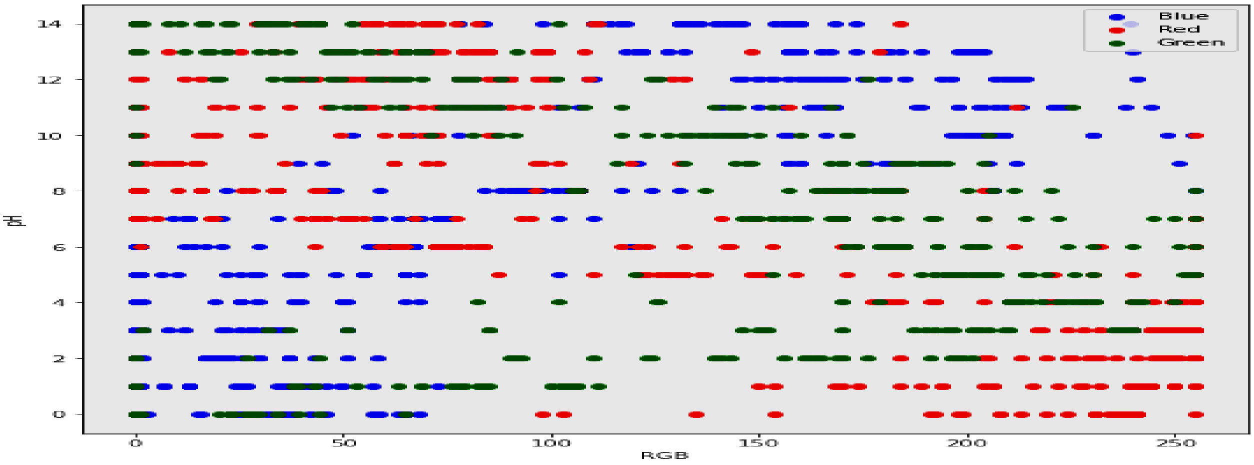

Fig 9: Graph of R, G, B values and their corresponding values of pH

Based on the outcomes of models such as Decision Tree Regression and Random Forest Regression outperformed Linear Regression due to the non-linear nature of the data, which can be readily deduced from Figure 9. Since the data were widely dispersed, these algorithms outperformed linear regression.

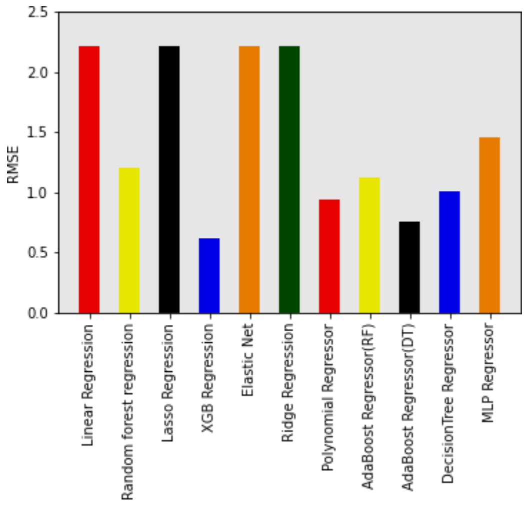

Fig 10 : RMSE of proposed models.

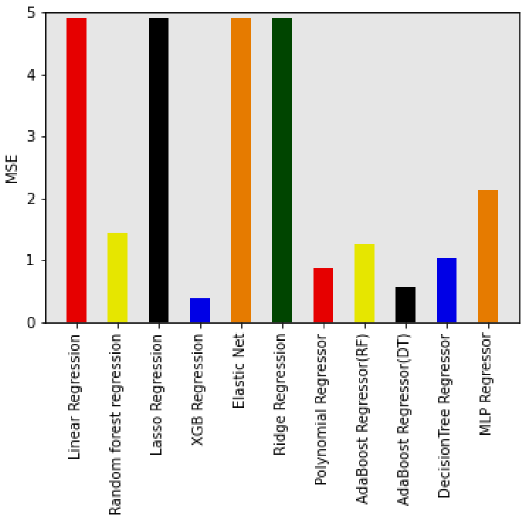

Fig 11 : MSE of proposed models.

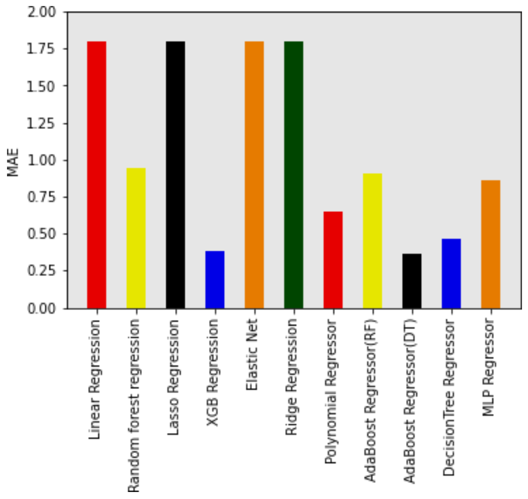

Fig 12: MAE of proposed models.

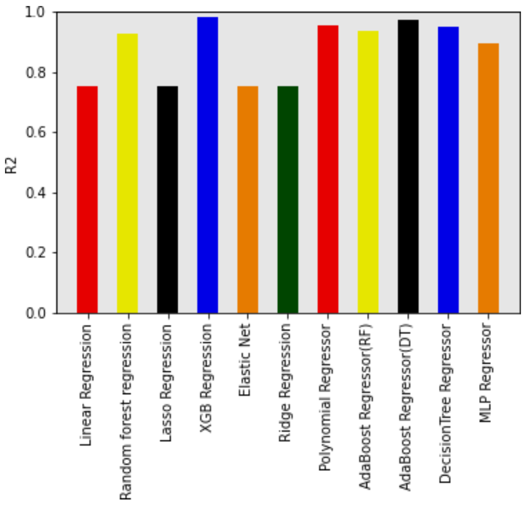

Fig 13: R^2 of proposed models.

Figures 10 – 13 show us the comparison in terms of performance between all 11 regression models in termsof

MAE, MSE, RMSE and R^2 score respectively.

The input image has been separated into 3 layers (i.e., R, G, B) and the average of red, green, and blue values have been calculated after parsing through each and every pixel of the image. Then these RGB valuesare sent to our model for pH predictions, and it would also predict the chemical characteristics, deficient nutrients and suitable crops based on pH as depicted in Table 4 where the list of deficient nutrients abbreviated as follows: Ca (Calcium), Mg (Magnesium), N(Nitrogen), P(Phosphorus), K(Potassium), B(Boron), Mo (Molybdenum), S(Sulphur), Fe (Iron), Mn (Manganese).

Table 4 : Predictions of the model with respect to corresponding pH range

S.No	pH Range	Chemical characteristics	DeficientNutrients	Suitable Crops
1	Below 4.0	Strongly acidic	All Nutrients	Not suitable for any crop
2	4.0-6.0	ModeratelyAcidic	B, Ca, K, Mg,Mo, N, P	Wheat, Soyabean, Rice, Potato, Pea,Peanut
3	6.0-6.5	Slightly Acidic	Ca, Mg, N, P, S	Wheat, Soyabean, Rice, Sweet Potato,Corn, Beetroot
4	6.5-7.5	Neutral	No deficientnutrients	Wheat, Soyabean, Barley, Vegetables, Oilseeds, Mushrooms, Oats, Cotton
5	7.5-8.5	Slightly Alkaline	Fe, Mn, N, P	Vegetables, Oilseeds, Mushrooms,Oats, Cotton, Cucumber, Garlic
6	Above 8.5	Strongly Alkaline	All Nutrients	Not suitable for any crops

VII. Conclusion

The objective of the proposed work is to classify different soil types and forecast their pH levels using a combination of machine learning algorithms and image processing techniques to create a soil health intelligence system that can suggest suitable crops. The system is trained on two datasets consisting of soil images, and Red, Green, Blue (RGB) values. From the results, the XGBoost Regressor has effectively predicted soil pH and its characteristics with a very good R2 score of 98%, and Convolutional Neural Networks have also performed well, with test and training accuracy of 90.62% and 91.66%, followed by train loss and validation loss of 0.27 and 0.17, respectively. The proposed approach examines andcategorizes the data, resulting in a stable and effective method for soil classification and pH prediction. This research work has major implications for agriculture and soil sciences since it provides a rapid, inexpensive, and non-invasive technique for soil assessment. In addition, the proposed method can be extended in future work to predict other soil parameters, such as nutrient levels, organic matter content, and water retention capacity, by applying machine learning and deep learning techniques to enhance our understanding of soil parameters and recommended farming.

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The first dataset of images would be sent if requested, while the second dataset is publicly availablein Kaggle [42].

References

[1] Rahman, Sk & Mitra, Kaushik & Islam, S.M. (2018). Soil Classification Using Machine Learning Methods and Crop Suggestion Based on Soil Series. 1-4. 10.1109/ICCITECHN.2018.8631943.

[2] Park, J., and Santamarina, J. C., “Revised soil classification system for coarse-fine mixtures,” Journal of Geotechnical and Geoenvironmental Engineering, Vol. 143, No. 8, 2017.

[3] Morais, P. A. O., de Souza, D. M., Madari, B. E., Soares, A. S., & deOliveira, A. E. (2019). Using image analysis to estimate the soil organic carbon content. Microchemical Journal, 147, 775e781.

[4] Morais, Pedro Augusto & Souza, Diego & Carvalho, Márcia Thaís & Madari, Beáta & Oliveira, Anselmo. (2019). Predicting Soil Texture Using Image Analysis. Microchemical Journal. 146. 10.1016/j.microc.2019.01.009.

[5] Kovačević, Miloš & Bajat, Branislav & Gajic, Bosko. (2010). Soil type classification and estimation of soil properties using support vector machines. Geoderma. 154. 340-347. 10.1016/j.geoderma.2009.11.005.

[6] Barman, Utpal. (2019). Prediction of Soil pH using Smartphone based Digital Image Processing and Prediction Algorithm. JOURNAL OF MECHANICS OF CONTINUA AND MATHEMATICAL SCIENCES. 14. 10.26782/jmcms.2019.04.00019.

[7] Barman, Utpal & Choudhury, Ridip & Talukdar, Niyar & Deka, Prashant & Kalita, Indrajit &

Rahman, Naseem. (2018). Predication of soil pH using HSI color image processing and regression over Guwahati, Assam, India. Journal of Applied and Natural Science. 10. 805-809. 10.31018/jans.v10i2.1701.

[8] Soil Survey Staff. (2014). Keys to soil taxonomy (12th ed.).Washington, D.C: USDA-Natural Resources ConservationService, U.S. Government Print Office.

[9] Kirillova, Nataliya & Grauer-Gray, Jenna & Sileva, T. & Artemyeva, Zinaida & Burova, E. (2018). New perspectives to use Munsell color charts with electronic devices. Computers and Electronics in Agriculture. 155. 373-385. 10.1016/j.compag.2018.10.028.

[10] Marqués-Mateu, Angel & Moreno-Ramón, Héctor & Balasch, Sebastià & Ibáñez-Asensio, Sara. (2018). Quantifying the uncertainty of soil colour measurements with Munsell charts using a modified attribute agreement analysis. Catena. 171. 10.1016/j.catena.2018.06.027.

[11] Munsell Color. (2000). Munsell soil color charts: Year 2000 revised washable edition. New York: Greta Macbeth.

[12] Chang, Cheng-Wen & Laird, David & Mausbach, Maurice & Hurburgh, Charles. (2001). Near-Infrared Reflectance Spectroscopy–Principal Components Regression Analyses of Soil Properties. Soil Science Society of America Journal. 65. 480-490. 10.2136/sssaj2001.652480x.

[13] Viscarra Rossel, Raphael & Walvoort, D.J.J. & Mcbratney, Alex & Janik, L. & Skjemstad, J.O. (2006). Visible, Near Infrared, Mid Infrared or Combined Diffuse Reflectance Spectroscopy for Simultaneous Assessment of Various Soil Properties. Geoderma. 131. 59-75. 10.1016/j.geoderma.2005.03.007.

[14] Nocita, Marco & Stevens, Antoine & Tóth, Gergely & Panagos, Panos & Wesemael, Bas & Montanarella, Luca. (2013). Prediction of Soil Organic Carbon Content by Diffuse Reflectance Spectroscopy Using a Local Partial Least Square Regression Approach. Soil Biology and Biochemistry. 10.1016/j.soilbio.2013.10.022.

[15] Heung, Brandon & Ho, Hung Chak (Derrick) & Zhang, Jin & Knudby, Anders & Bulmer, Chuck & Schmidt, Margaret. (2016). An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma. 265. 62-77. 10.1016/j.geoderma.2015.11.014.

[16] Sofou, Anastasia & Evangelopoulos, Georgios & Maragos, Petros. (2005). Soil Image Segmentation and Texture Analysis: A Computer Vision Approach. Geoscience and Remote Sensing Letters, IEEE. 2. 394- 398. 10.1109/LGRS.2005.851752.

[17] Liu, Yong & Wang, Huifeng & Zhang, Hong & Liber, Karsten. (2016). A comprehensive support vector machine-based classification model for soil quality assessment. Soil and Tillage Research. 155. 19-26. 10.1016/j.still.2015.07.006.

[18] B. Bhattacharya, and D.P. Solomatine "An algorithm for clusteringand classification of series data with constraint of contiguity",Proc. 3T"d nt. Conf: on Hybrid and Intelligent Systems,Melboume, Australia, 2003, pp. 489-498.

[19] Chung, Sun-Ok & Cho, Ki-Hyun & Cho, Jin-Woong & Jung, Ki-Youl & Yamakawa, Takeo. (2012). Soil Texture Classification Algorithm Using RGB Characteristics of Soil Images. Journal- Faculty of Agriculture Kyushu University. 57. 393-397. 10.5109/25196.

[20] Shenbagavalli, R. (2011). Classification of Soil Textures Based on Laws Features Extracted from Preprocessing Images on Sequential and Random Windows. Bonfring International Journal of Advances in Image Processing. 1. 15-18. 10.9756/BIJAIP.1004.

[21] Sudharsan, V. and B. Yamuna. “Support vector machine based decoding algorithm for BCH codes.” Journal of telecommunications and information technology (2016).

[22] Unmesha Sreeveni.U .B, Shiju Sathyadevan “ADBF Integratable Machine Learning Algorithms –Map reduce Implementation” Second International Symposium on computer vision and the Internet(VisionNet’15).

[23] Zhang, Z., & Tumay, M.T. (2000). Statistical to Fuzzy Approach toward CPT Soil Classification. Journal of Geotechnical and Geoenvironmental Engineering, 125, 179-186.

[24] I.T. Young, and T.W. Calvert, "An analysis technique for biological shape", Information and Control, vol. 25, pp 357-370,1974.

[25] Fan Z, Herrick JE, Saltzman R, Matteis C, Yudina A, Nocella N,Crawford E, Parker R,Van Zee J (2017) Measurement of soil color:a comparison between smartphone camera and the munsell color charts. Soil Sci Soc Am J 81(5):1139–1146

[26] Viscarra Rossel, Raphael & Fouad, Youssef & Walter, Christian. (2008). Using a digital camera to

measure soil organic carbon and iron contents. Biosystems Engineering. 100. 149-159.

[27] Vibhute, Amol & Kale, Karbhari & Dhumal, Rajesh & Mehrotra, Suresh. (2015). Soil type classification and mapping using hyperspectral remote sensing data, MAMI 2015. 1-4.

[28] Qiu Z, Chen J, Zhao Y, Zhu S, He Y, Zhang C. Variety Identification of Single Rice Seed Using Hyperspectral Imaging Combined with Convolutional Neural Network. Applied Sciences. 2018; 8(2):212.

[29] Chatnuntawech I, Tantisantisom K, Khanchaitit P, Boonkoom T,Bilgic B, Chuangsuwanich E (2018) Rice classification using hyperspectral imaging and deep convolutional neural network. arXiv preprint arXiv:1805.11491.

[30] Stiglitz R, Mikhailova E, Post C, Schlautman M, Sharp J (2016) Evaluation of an inexpensive sensor to measure soil color. Comput Electron Agric 121:141–148

[31] Stiglitz R, Mikhailova E, Post C, Schlautman M, Sharp J (2017) Using an inexpensive color sensor for rapid assessment of soil organic carbon. Geoderma 286:98–103

[32] Gómez-Robledo L, López-Ruiz N, MelgosaM, Palma AJ,Capitán- Vallvey LF, Sánchez-Marañón M (2013) Using the mobile phone as munsell soil-color sensor: an experiment under controlled illumination conditions. Comput Electron Agric 99:200–208

[33] Stiglitz R,Mikhailova E, Post C, Schlautman M, Sharp J, Pargas R, Glover B, Mooney J (2017) Soil color sensor data collection using a gps-enabled smartphone application. Geoderma 296:108–114.

[34] Breul P, Gourves R. Field Soil Characterization: Approach Based on Texture Image Analysis. J Geotech Geoenviron Eng 2006;132(1).

[35] Aziz, M.M, Ahmed, D.R., Abraham, B.F, 2016. “Determine the pH of Soil by using Neural Network Based on Soil's Colour”. International Journal of Advanced Research in Computer science and Software Engineering, Vol.: 6, Issue: 11, pp: 51-54, 2018.

[36] Aditya, A., Chatterjee, N., Pradhan, C., “Computation and Storage of Possible Pouvoir Hydrogen Level of Soil using Digital Image processing”, International Conference on Communication and Signal Processing, India. pp: 205-209, 2017.

[37] Babu, C.S.M. and Pandian, M.A, “Determination of Chemical and Physical Characteristics of Soil using Digital Image processing”, International Journal of Emerging Technology in Computer Science & Electronics, Vol.: 20, Issue: 2, pp: 331-335, 2016.

[38] Gurubasava, Mahantesh S.D., “Analysis of Agricultural soil pH using Digital Image Processing”, International Journal of Research in Advent Technology, Vol.: 6, Issue: 8, pp: 1812-1816, 2018.

[39] Kumar, V., Vimal, B., Kumar, R., Kumar, R., & Kumar, M, “Determination of soil pH by using digital image processing technique”. Journal of Applied and Natural Science, Vol.: 6, Issue: 1, pp: 14-18, 2014.

[40] Mohan, R.R., Mridula S., Mohanan P., “Artificial Neural Network Model for Soil Moisture Estimation At Microwave Frequency”, Progress In Electromagnetics Research M, Vol.: 43, pp: 175–181, 2015.

[41] Abu, M.A., Nasir, E.M.M. and Bala, C.R, “Simulation of Soil PH Control system using Fuzzy Logic Method”, International Journal of Emerging Trends in Computer Image & Processing. Vol.: 3, Issue: 1, pp: 15-19, 2014.

[42] ROBERT, (2019). pH-recognition, Version 1, Retrieved November 23, 2022, from https://www.kaggle.com/datasets/robjan/ph-recognition.

[43] Mrs. Monika Soni. (2015). Design and Analysis of Single Ended Low Noise Amplifier. International Journal of New Practices in Management and Engineering, 4(01), 01 - 06. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/33

[44] Kuna, S. L. ., & Prasad, A. K. . (2023). Deep Learning Empowered Diabetic Retinopathy Detection and Classification using Retinal Fundus Images. International Journal on Recent and Innovation Trends in Computing and Communication, 11(1), 117–127. https://doi.org/10.17762/ijritcc.v11i1.6058

[45] Gupta, S. K., Lanke, G. R., Pareek, M., Mittal, M., Dhabliya, D., Venkatesh, T., & Chakraborty, S. (2022). Anamoly detection in very large scale system using big data. Paper presented at the IEEE International Conference on Knowledge Engineering and Communication Systems, ICKES 2022, doi:10.1109/ICKECS56523.2022.10059870 Retrieved from www.scopus.com