Journal of Innovations in Pharmaceutical and Biological Sciences (JIPBS)
ISSN: 2349-2759
Available online at www.jipbs.com

Research article

Prediction of reclamation time of fresh coal mine overburden spoil through different soil quality indicators using artificial neural network

Jitesh Kumar Maharana, Amiya Kumar Patel*

Department of Biotechnology and Bioinformatics, Sambalpur University, At/Po- Jyoti Vihar, Burla- 768019, Odisha (India).

Keywords: Artificial neural network, Coal mine overburden spoil, Homology modeling, MOE, QSAR, Reclamation.

***Corresponding Author: Amiya Kumar Patel,** Department of Biotechnology and Bioinformatics, Sambalpur University At/po- Jyoti Vihar, Burla, Dist.-Sambalpur, (Pin- 768019), Odisha, India.

Abstract

Coal mines are unfavorable habitat for the growth of microorganisms and revegetation due nutrient deficiency, heavy metals toxicity and pyrite contamination. So, assessment of soil parameters in different age series coal mine overburden spoil provides better understanding for mine spoil reclamation and the implementation of appropriate strategies is needed for the pace and progress of restoration by using minimum datasets and valuable parameters. Hence, the artificial neural network is essential to validate the concept. About 9 mine spoil parameters were selected in order to develop the QSAR equation based on brute-force method and genetic function approximation for the prediction of mine spoil reclamation required for the fresh coal mine overburden spoil to reach the mean soil features of the nearby native forest soil. The training and the test sets with statistically best fitted with R^2= 0.994 and R^2_{LOO}= 0.881. The predictive ANN model with 9- 7- 1 structure was predicted as the best model which illustrated the time period required for the mine spoil genesis across the sites. The standard error for the proposed model was estimated to be 0.001, which can be used as an indicator of the robustness of the fit and suggested that the predicted years for the mine spoil reclamation across the sites based on the model is reliable. The validity of the developed model was confirmed with the highest calculated value of the squared correlation coefficient determination (R^2= 0.975) and lower root mean square error (RMSE= 0.28), which revealed good predictability. Hence, OB_0 shall take ~ 39.277 years to reach the mean soil features of the nearby native forest soil depending on the variability in physico-chemical properties, enzyme activities, microbial community structures and fungal PLFA biomarkers as the sensitive and reliable indicators influencing the mine spoil reclamation in different age series coal mine overburden spoil overtime.

Introduction

Coal is the most abundant fossil fuel that has several important uses worldwide. For the development of industries and other man-made requisites the trees are shredded and cut down for the excavation of the coal. These mining activities results in the loss of forest cover which was under the dense forest before the mining and causes tremendous effects to its nearby located peripheral zone. The impact of coal mining activities results in decrease of soil quality. Besides, being deficient in available nutrients due to the lack of biologically affluent topsoil, the mine overburden spoil represents dis-equilibriated geomorphic system and poses problems for pedogenesis, revegetation [1-3] and restoration of the coal mine overburden spoil. Hence, a comparative assessment of different age series coal mine overburden spoil in chronosequence is pre-requisite in order to implement suitable management strategies for recuperation of the legacy of mining sites in order to reach the soil features of nearby native forest soil.

The assessment of the microbial diversity in soil ecosystems is necessary to gain knowledge and information on the soil quality [4]. The successive amelioration of microbial diversity in mine overburden spoil over time play important roles in bringing about changes through the process of pedogenesis, by periodic monitoring and analyzing the soil variables influencing microbial community structure and their associated functioning which further improves root growth and reduce undesirable effects of the microclimatic conditions. Therefore, the assessment of various soil parameters of coal mine overburden spoil is not only important for better understanding of soil ecosystem functioning, in different age series coal mine overburden spoil over a geologic period of time but also gives detail information in implementing appropriate reclamation strategies which can lead to improve the soil quality, indicating the pace and progress of reclamation [5]. For the purpose, the use of statistical method like artificial neural network is essential to validate the concept.

The determination of the duration required for the coal mine spoil restoration through experimentation is broad and tedious [6]. Conversely, an alternative approach for such analysis is performed through the advanced computational technique for prediction of the expected time required for the fresh coal mine spoil reclamation to reach the soil features of nearby native forest soil. Such assessment can be performed by the artificial neural network (ANN), which is an empirical modeling tool, and is analogous to the behavior of biological neural structures [7]. It has two modes of operation such as training mode and operation/testing mode. In training mode, the neurons are trained by utilizing a particular input pattern to produce the desired output pattern. In operation/testing mode, when a trained input pattern is detected at the input, the ANN will produce its associated output. Artificial neural networks are potent techniques that have the ability to identify highly complex relationships from input output data only. It is efficaciously utilized for modeling, identification, prediction and control of complex process with nonlinearities, instabilities and uncertainties [8]. A neural-network model can determine the input-output relationship for a complicated system based on the strength of their interconnections presented in a set of sample data [9]. Such a model can provide data approximation and signal filtering functions beyond the optimal linear techniques [10].

Therefore, the neural-network models provide more robust outcomes for complicated system analysis as compared to the conventional mathematical models. It is also additionally valuable in modeling problems in which the connection between the dependent and independent variables is inadequately comprehended and has the probability to identify the highly complex relationships from the input-output data only. Further, the back-propagation algorithm is a non-linear augmentation of least mean square (LSM) algorithm for multi-layer perceptrons. It is effectively applied in model-free function, estimation for the pattern recognition, approximation/mapping of non-linear functions, classifications and time series prediction. The neural network is conventionally layered, where the layers are fully interconnected to each other. The first inputs layer receive external information datasets, which are standarized within the limit values generated from the activation functions and results in better numerical precision for the mathematical and scientific operations performed by the network. It was supported with the second hidden layer composed of neurons, which are responsible for extracting patterns associated with the internal processes being analyzed from the network. The quantity of neurons in each hidden layer can vary according to the complexity and unpredictability of the problem [11-12]. However, the final network output is produced representing the third output layer, which results from the processing performed by the neurons in the hidden layers.

Over a couple of decades this artificial neural networks (ANNs) and feed forward artificial neural networks (FANNs) have been widely studied in order to present the process models and their beneficiary use in industrial fields [13]. Besides, ANNs have been applied to various geotechnical engineering problems such as pile capacity prediction, modeling soil behavior, site characterization, earth retaining structures, design of tunnel and underground openings, liquefaction, soil permeability and hydraulic conductivity, soil compaction, soil swelling and classification of soils *etc* [14-18]. Likewise, it is additionally connected and applied for the prediction of organic matter content in soil [19], soil erosion [20], hydraulic conductivity of coarse grained soil samples [21], determination of volumetric soil moisture content [22] and modeling of soil electrical conductivity [23].

Unlike logical approaches, the ANNs require no explicit mathematical equation and no limiting presumptions of normality or linearity [3]. The benefits of ANN over traditional physiology-based predictive models includes (i) the involvement of intense parallel computations during the training process, (ii) the capability of quick speculations *i.e.* once the ANN is trained for a particular system, its operation is generally quicker and the unknown input patterns can be rapidly identified in a real-time environment, (iii) estimation of non-linear relationships between the input data and desired outputs, (iv) the data processing applications such as image recognition, (v) the classification based on land use changes, (vi) the utilities in land drainage engineering, (vii) the estimation of crop evapotranspiration as well as yield prediction for a new set of input conditions and thereby support the use of mechanistic simulation tools by providing the initial condition values or site-specific parameters and guide parameter estimation in agricultural machinery models [7, 14, 24-28].

Considering the tropical dry deciduous forest as natural vegetation in the study site, an attempt was made in the present study to predict the time period required for fresh coal mine overburden spoil (OB_0) to reach the mean soil features of the nearby native forest soil based on the variability in soil properties in six different age series coal mine overburden spoil ($OB_0 \rightarrow OB_{10}$) in chronosequence over time through mine spoil restoration using the multivariate predictive modeling technique *i.e.* artificial neuron network (ANN). This prediction model is considered to be superior in comparison to the non-parametric statistical benchmark methods, which provide valuable and significant information about mine spoil genesis influencing the pace and progress of mine spoil restoration. For each dataset, the ANN predictive models were designed and all the three datasets (image-scale, field-scale and lab-scale) revealed significant network performances for training, testing and validation indicating the good network generalization for predicting mine spoil restoration with the passage of time. The study reveals a clear discrimination among the different age series coal mine overburden spoil in chronosequence as well as the nearby NF soil through various soil quality indicators.

Experimental

Materials and methods

Study site

The present study was carried out in the Basundhara (west) open cast colliery in the Ib valley of Mahanadi Coalfields Limited (MCL), Odisha, India (Geographical location: 22º 03' 58" - 20º 04' 11" north latitude and 83º 42' 46" - 83º 44' 45" east longitude). The coal mine overburden spoil have been grouped into six different age series (fresh: OB_0, 2 yr: OB_2, 4 yr: OB_4, 6 yr: OB_6, 8 yr: OB_8 and 10 yr: OB_{10}) on the basis of their formation since inception located within a peripheral distance of 10 km from the core mining area. Tropical dry deciduous forest was considered to be the natural vegetation of the study site, which experiences a semi-arid climate (1300 mm rainfall y^{-1}, annual average temperature 26°C and relative humidity 15%) with three distinct seasons *i.e.* summer, rainy and winter.

Quantitative analysis of soil parameters

Textural composition of six different age series coal mine overburden spoil and nearby NF soil includes the estimation of clay (< 0.002 mm), as per the method prescribed in TSBF handbook. The water holding capacity (WHC) was estimated [29]. Soil organic C was estimated through titration method suggested by Walkley and Black [29]. In addition, the amylase activity in different age series coal mine overburden spoil was determined in adaptation to the procedures described by Somogyi and Roberge [30] determined by spectrophotometric method using starch as substrate. The Phosphatase activity was determined by spectrophotometric method as per the methodology prescribed by Tabatabai and Bremner using *p*-nitrophenol as the substrate. The dehydrogenase activity was estimated spectrophotometrically through the reduction of 2,3,5-triphenylotetrazolium chloride (TTC) as electron acceptor to triphenyl formazon (TPF). Besides, the microbial enumeration such as the heterotrophic aerobic bacterial population (HAB) was performed by serial dilution technique using nutrient agar.

Moreover, the Phospholipid fatty acids (PLFA) analysis of six different age series coal mine overburden spoil and nearby forest soil was performed through lipid extraction based on fractionation and quantification. Further, the fundamental differences in the bacterial and fungal physiology and ecology would suggest that the biogeography of each group would be controlled by separate edaphic factors, which may vary among different mine spoil profiles. As bacteria and fungi are likely to have distinct functional roles in different soil profiles, the more robust understanding of the specific effects of land-use and edaphic factors on microbial groups will improve our ability to predict specific effects of land-use changes in microbial community structure and function [31].

Neural network data mapping model development

In this study a back-propagation neural-network model was created using Stuttgart Neural Network Simulator package [SNNS version 4.2; Institute for Parallel and Distributed High Performance Systems (IPVR) at the University of Stuttgart, Germany] and trained using the physico-chemical soil variables (textural composition, WHC, organic C), enzyme activities (amylase, phosphatase, and dehydrogenase), microbial enumeration (HAB), PLFA markers (16:1ω5c) and the ratio of F:B as inputs and predicted the reclamation time required for fresh coal mine overburden spoil to reach the soil features of nearby NF soil in years as the output.

The topological structure of this neural-network model consisted of 9 input neurons in the input layer and one output neuron in the output layer to match 9:1 input-output pattern of the training datasets. One hidden layer with 7 neurons was the optimal topology for the neural-network model determined by trial and error method (Figure 1). The evaluation criterion for determining the optimal topology was the best correlation value of the training set. The neural network model was trained in an iterative training process using the obtained training datasets. The first nine numbers referred to different soil parameters that are most important towards the reclamation process, and the last number is the predicted year of the fresh coal mine overburden spoil to attain the soil features of the nearby native forest soil (Figure 1). To avoid possible bias, the order of input-output data pair in a training dataset was randomized before the training process.

During the training process, the back propagation training algorithm compares the estimated output value with the target value (namely the measured value), then tunes weighted values followed by connecting all the neurons to minimize the difference between the estimated and the target values until the error is smaller than the predefined level or until the number of the iteration reached a preset maximum number. The constructed model was trained with the input data for an epoch of 10,000 with 0.1 learning rate. After completion of the training process, all the weighting indices describing the interconnection strengths between neighboring neurons are fixed and the neural network model will then be capable of mapping input variables to an estimated output promptly and accurately.

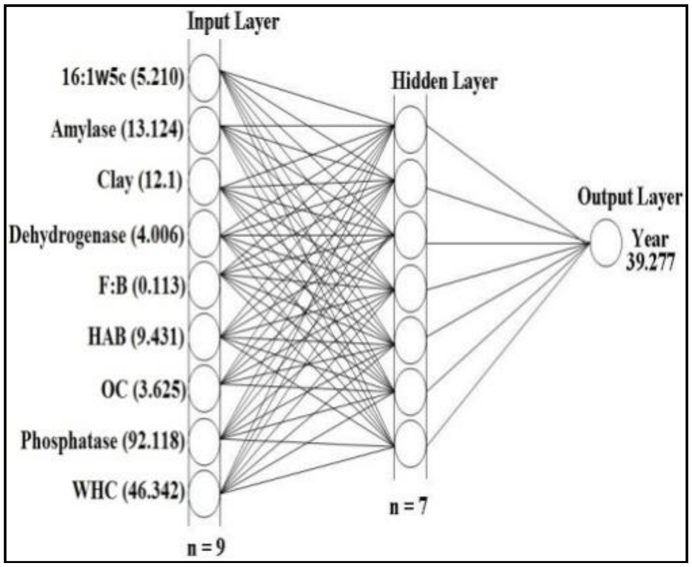

Figure 1. Layers and connection of a feed-forward back-propagating ANN. The neural network model developed here applies the sigmoid transfer function to compute the strength of interconnection between each pair of neurons.

Data processing and development of prediction model

A total of 34 soil parameters were used including physico-chemical parameters (sand, silt and clay percentage, bulk density, moisture content, water holding capacity, pH, organic C, total N and extractable P), microbial biomass pool (Microbial biomass-C, Microbial biomass-N, Microbial biomass-P and basal soil respiration), enzyme activity (amylase, invertase, protease, urease, phosphatase and dehydrogenase), microbial CFUs (azotobacter, arthrobacter, rhizobia, heterotrophic aerobic bacteria, sulfate reducing bacteria, actinomycetes, yeast and fungi), PLFAs (16:1ω5c, 18:1ω9c and 18:2ω6c), fungal: bacterial biomass ratio, gram-positive/gram-negative ratio and anaerobes for the purpose. The calculated soil parameters were collected in a data matrix (D), where the rows represents different mine spoil samples from six different age series coal mine overburden spoil ($OB_0 \rightarrow OB_{10}$) and the columns represent different soil parameters. In order to minimize the effect of colinearity and to avoid redundancy, the correlation among different soil parameters with each other was investigated and those pairs with higher relationships were determined. Among the collinear parameters, those with the lowest correlation with soil properties were removed from the data matrix. Among the remaining parameters, the set of parameters that provide statistically best prediction model was selected using the genetic function approximation (GFA) [32] within the evolution module (ga.svl) of the MOE program.

The evolutionary genetic tool enables automated prediction modeling on the fly and is available through the SVL exchange. The GFA algorithm starts with the creation of a population of randomly generated parameter sets. The algorithm was set up to discover the soil parameters relevant for mine spoil restoration by linear polynomial terms. One hundred random initial equations with four variables were used (adding constants wherever necessary) to search for the equations of unlimited length but with the acceptable lack-of-fit (LOF) scores [33], new 'child equations' were generated using the multiple linear regression method. Child equations were mutated (*i.e.* changed at "birth") 50% of the time after their generation by addition of randomly selected new terms. The number of generations of equation evolution required in the dataset was gauged by the attainment of adjusted R^2 values and minimal LOF scores. Creation of a consecutive generation involves crossovers between set contents as well as mutations. Total number of crossovers was set to 50 msp 14000 with the auto-termination factor of 1000 (meaning that the calculation was stopped when the fitness function value does not change during 1000 crossovers). The equations were evaluated for statistical soundness by the Friedman LOF score, R^2adjusted, R^2, least-squares error and correlation coefficient after cross-validation statistics.

The Friedman LOF score is expressed by the following equation:

$$LOF = LSE / \{1 - (c + dp) / m\} \qquad (1)$$

Where LSE is the least-square error, c is the number of basic functions in the model, d is smoothing soil parameters, p is the number of soil parameters, and m is the number of spoil samples in the training dataset.

The smoothing parameter, which controls the scoring bias between equations of different sizes, was set at default value of 1.0. The set of 9 soil parameters (Clay, Water holding capacity, Organic carbon, Amylase, Phosphatase, Dehydrogenase, Heterotrophic aerobic bacteria, 16:1ω5c, Fungal: Bacterial) were found to be the best relevant influencing the reclamation process of coal mine overburden spoil with the passage of time, which were used in the ANN designing for the development of prediction model.

Validation of the prediction model

The predictive capability of the developed prediction model was further validated based on several statistical tests such as leave-one-out cross-validation and a Y randomization test using a svl script (QSARwizard.svl). The cross-validation regression coefficient (R^2_{LOO}) was calculated based on the prediction error sum of squares (PRESS) and sum of squares of deviation of the experimental values 'Y' from their mean (SSY) using the following equation:

$$R^2_{LOO} = 1 - \frac{PRESS}{SSY} = 1 - \frac{\sum_{i=1}^{n}(Y_{\exp} - Y_{pred})^2}{\sum_{i=1}^{n}(Y_{\exp} - \bar{Y})^2}$$

Where Y_{exp}, Y_{pred}, and ȳ are the observed, predicted, and mean values of the observed activity which belongs to the training datasets of the soil parameters respectively. The Y-randomization test was done y repeatedly shuffling the data set and developing new prediction models and then comparing the predicted years with the years of the original QSAR model generated from nonrandomized data. This process was repeated 100 times. If the original prediction model is statistically significant, its predictive value should be significantly better than those from the permuted data. We have used a parameter, R_p^2, which penalizes the model for the difference between the squared mean correlation coefficient of randomized models (R_r^2) and the squared correlation coefficient of the nonrandomized model (R^2). The R_p^2 parameter was calculated by the following equation:

$$R_p{}^2 = R^2 . \sqrt{R^2 - R_r{}^2}$$

The parameter R_p^2 ensures that our prediction model thus developed is not obtained by chance. We have assumed the value of R_p^2 should be greater than 0.5 for an acceptable model. The determination coefficient in prediction using the test set (R^2 test) was calculated using the following equation [34,35].

$$R^2_{test} = 1 - \frac{\Sigma(Y_{pred_{test}} - Y_{exp_{test}})^2}{\Sigma(Y_{exp_{test}} - \bar{Y}_{exp_{train}})}$$

where R_{test}^2 is the squared pearson correlation coefficient for regression calculated using Y= a + bx; a is referred to as the y-intercept, b is the slope value of regression line, and R_{test0}^2 is the squared correlation coefficient for regression without using the y-intercept, and the regression equation was Y = bx. To further check the intercorrelation between soil parameters used in the final prediction model, we performed variance inflation factor (VIF) analysis. The VIF value was calculated from 1 / (1 – R^2), where R^2 is the multiple correlation coefficient of one parameter's effect regressed onto the remaining soil parameters. If the VIF value is larger than 10 for a parameter, its information could be hidden by other parameters [34].

Results and discussion

Comparative assessment of 9 selected soil parameters including physico-chemical properties, enzyme activities, Microbial CFU counts, PLFAs [35] in six different age series coal mine overburden spoil ($OB_0 \rightarrow OB_{10}$) and nearby native forest soil (NF) have been represented (Table 1). The textural composition revealed an increasing trend with respect to clay fraction was observed *i.e.* maximum in OB_{10} (11.3%) and minimum in OB_0 (5.4%). However, the clay content in the nearby native forest soil was found to be 12.1%. The gradual establishment of vegetation cover on mine overburden may be the reason for increase in clay formation [36] [2]. Besides, the root of vegetational component specifically root exudates in form of organic acids promotes the disintegration of coarse particles to finer clay particles [37]. Further, the absence of vegetation cover makes clay prone to loss. Additionally, the vegetational development on degraded barren land was reported to check the loss of clay and promotes its conservation [2]. Clay being an important primary particle contributes to soil structural stability [37]. Progressive increase in clay fraction in mine spoil indicated the progressive development of soil structural stability, aggregation and developed resistance to erosion with the increase in age of mine overburden [2].

The water holding capacity is the total amount of water a soil can hold. The water holding capacity of a soil is of great value to practical agriculture, because it provides a simple means of determining moisture contents required for soils for good plant growth. Besides, the water holding capacity (WHC) showed a reverse trend *i.e.* minimum in OB_0 (27.5%) and progressively increased showing maximum in OB_{10} (43.8%) over time (Table 1). The OB_8 (41.2%) exhibited relatively higher WHC as compared to OB_6 (38.3%). Similarly, relatively higher WHC was observed in OB_4 (36.1%) as compared to OB_2 (31.3%). The comparative assessment revealed that the nearby NF soil exhibited relatively higher WHC (46.34%) in comparison to the six different age series coal mine overburden spoil across the sites. The progressive improvement in moisture over time may be due to the positive influence of canopy cover on OB_{10} that prevents loss of water through evaporation by not allowing the direct exposure of soil surface to incoming radiation [38].

Organic Carbon (OC) is the main source of energy for soil microbes. The ease and speed of available soil organic C is based on the size and breakdown of soil organic matter, which acts as the source of energy and triggers nutrient availability through mineralization. Organic C enters into soil through the decomposition of plant/animal residues, root exudates, living/dead microbes and soil biota that vary with the decomposition rate and turnover time. The organic C in association with primary soil particles is reported to promote macroaggregation. A wide variability with respect to soil organic carbon content was exhibited by the different age series coal mine spoil, which showed a range from 0.151 mgC/g spoil (OB_2) to 2.004 mgC/g spoil (OB_{10}) across the sites (Table 1). However, the analysis indicated that the organic C in OB_0 was found to be beyond detectable limit. The study revealed relatively higher level of OC in OB_{10} as compared to OB_8 (1.553 mg C/g spoil) and OB_6 (1.057 mg C/g spoil). Besides, it was evident from the data that comparatively higher level of OC were recorded in the nearby NF soil (3.625 mg C/g soil) as compared to the different age series coal mine overburden spoil across the sites. The study suggested that there was gradual increase in organic carbon content from the nutrient deficient mine overburden spoil (OB_0) to an enriched nearby NF soil over time across the sites. Increase in OC was found to be correlated with the increase in clay fraction in the ecologically disturbed lands. The clay acts as an absorption sink for organic material. Increase in organic fraction with the increase in clay can also be due to the absorption of organic complexes onto the clay surface are being physically protected against decomposition, which lead to accumulation of organic C level with respect to the increase in age of mine overburden spoil.

Table 1. Comparative assessment of 9 soil parameters in different age series coal mine overburden spoil ($OB_0 \rightarrow OB_{10}$) and nearby native forest soil (NF) selected for the ANN study.

Soil parameters	Different age series coal mine overburden spoil						
	OB_0	**OB_2**	**OB_4**	**OB_6**	**OB_8**	**OB_{10}**	**NF soil**
Clay (%)	5.4	6.9	8.7	9.9	10.7	11.3	12.1
WHC (%)	27.5	31.3	36.1	38.3	41.2	43.8	46.34
OC (mg C/g spoil)	nd*	0.151	0.770	1.057	1.533	2.004	3.625
Amylase (μg glucose/g spoil/hr)	nd*	1.253	2.034	2.263	3.655	4.571	13.124
Phosphatase (μg PNP/g spoil /hr)	nd*	nd*	10.108	26.495	35.407	49.617	92.118
Dehydrogenase (μg TPF/g /hr)	0.056	0.144	0.291	0.458	0.948	1.275	4.006
HAB (Log CFU)	3.462	3.491	3.544	5.342	5.69	7.792	9.431
16:1ω5c (%)	0	0	0	0.34	1.14	1.44	5.21
Fungal: Bacterial	0.0216	0.0277	0.0307	0.0649	0.0757	0.0990	0.1128

nd*: beyond detectable limit.

Amylases are complex enzymes belong to glycoside hydrolase group of enzymes [α-amylase (α-1,4–glucan-4 glucanohydrolase; E.C. No. 3.2.1.1), β-amylase (β-1,4-glucanmaltohydrolase; E.C. 3.2.1.2), glucoamylase (α-1,4–glucanglucohydrolase; E.C. 3.2.1.3)]. The amylase activity showed a range from (1.253 μg glucose/g spoil/hr) to (4.571 μg glucose/g spoil/hr) with minimum in OB_2 and maximum in OB_{10}. Relatively higher amylase activity was observed in OB_{10} as compared to OB_8 (3.655 μg glucose/g spoil/hr) and OB_6 (2.263 μg glucose/g spoil/hr). Similarly, OB_4 (2.034 μg glucose/g soil/hr) exhibited relatively higher amylase activity as compared to OB_2. The amylase activity in OB_0 was found beyond the detectable limit. However, the amylase activity estimated in the NF soil (13.124 μg glucose/g soil/hr) was comparatively higher in comparison to different age series coal mine overburden spoil across the sites. Wide variation in amylase activity across the sites may be attributed to the variation in available soil nutrients and diversity of microbiota. Microbes are the major source of amylases, which hydrolyze starch mainly to form glucose or dextrins or oligosaccharides and small quantities of maltose. Amylase can also be used as biomarker to assess the soil quality basing on their sensitivity to soil management practices, importance in nutrient cycling, organic matter decomposition and bioremediation activities.

A phosphatase is an enzyme (Orthophosphoric monoester phosphohydrolase E.C. 3.1.3.2) that removes a phosphate group from its substrate by hydrolyzing the orthophosphoric monoester to alcohol and orthophosphate ion and a molecule with a free hydroxyl group, which acts as intermediary enzyme involved in the transformation of organic phosphates into inorganic form [39]. Besides, phosphatase activity appeared to be more dependent on the metabolic state of soil, biological activity of soil microbial population and hence their activity level can be used as an index of soil microbial activity. Besides, a wide variation in phosphatase activity was exhibited by the different age series coal mine overburden spoil, which ranged from OB_4 (10.108 μg PNP/g spoil/hr) to OB_{10} (49.617 μg PNP/g spoil/hr). The phosphatase activity in OB_0 and OB_2 were beyond detectable limits (Table 1). Relatively higher phosphatase activity was observed in OB_8 (35.407 μg PNP/g spoil/hr) as compared to OB_6 (26.495 μg PNP/g spoil/hr). However, the phosphatase activity in the nearby NF soil was estimated to be (92.118 μg PNP/g soil/hr). The data indicated an increasing trend in phosphatase activity with the increase in the age of mine overburden spoil ($OB_0 \rightarrow OB_{10}$) across the sites.

The dehydrogenase activity in the different age series coal mine overburden spoil in chronosequence showed consistent increase from OB_0 (0.056 TPF/g spoil/hr) to OB_{10} (1.275 TPF/g spoil/hr) across the sites (Table 1). Relatively higher dehydrogenase activity was observed in OB_8 (0.948 μg TPF/g spoil/hr) as compared to OB_6 (0.458 μg TPF/g spoil/hr) and OB_4 (0.291 μg TPF/g spoil/hr). Besides, minimal difference in dehydrogenase activity was evident from the comparison between OB_0 (0.056 TPF/g spoil/hr) and OB_2 (0.144 μg TPF/g spoil/hr). Relatively higher dehydrogenase activity was observed in the nearby NF soil (4.006 μg TPF/g soil/hr) as compared to the different age series coal mine overburden spoil. Dehydrogenases are linked to oxidation-reduction reactions involved in microbial respiration [39]. Being intracellular, the dehydrogenase activity is considered as the index of endogenous microbial activity and metabolic status that exists only in viable microbial cells [40]. The estimation of dehydrogenase activity is attractive, because they are an integral part of soil microbes involved in electron transport system and require an intracellular environment (viable cells) to express its activity. Comparative assessment of enzyme activities represent direct expression of soil microbial communities to metabolic requirements and hence provide the valuable information about the linkages between resource availability, microbial community structure and ecosystem processes. The change in microbial indices in terms of the soil enzyme activities correlated well with the extent of land degradation, which also exhibited a rapid response to both the natural and anthropogenic disturbances and therefore it serve as sensitive biomarkers for reclamation studies.

Heterotrophic aerobic bacteria (HAB) are the aerobic consumers of simple carbon compounds that take an active part in the natural recycling of substances. They decompose the soil organic matter and use the residual organic carbon compounds as their source of energy. Some HAB breaks down the pesticides and pollutants present in the soil thereby providing nutrition to the soil and prevent the loss of nutrients from plant root zone which colonizes other groups of microbes in soil leading to the vegetational development. Many heterotrophic aerobic bacteria utilize sugar, alcohol, and organic acids. However, there are specialized heterotrophic bacteria capable of decomposing cellulose, lignin, chitin, keratin, hydrocarbons, phenol, and other substances.

The unsaturated fungal biomarker 16:1ω5c typically represent arbuscular mycorrhizal fungi in different soil profiles. PLFA 16:1ω5c derived from arbuscular mycorrhizal fungi are known to contribute substantially to fungal biomass in the soil, which responds to the changes in the available C. The percentage composition of reliable fungal PLFA biomarker showed a wide variation across the sites from 0.34% (OB_6), 1.14% (OB_8) and 1.44% (OB_{10}). Higher relative distribution of PLFA 16:1ω5c reflected the dominance of arbuscular mycorrhizal fungi in OB_{10} (0.52%) than OB_8 (0.25%). However, the fungal marker in OB_0, OB_2 and OB_4 were not found due to the pyrite contamination and presence of poor soil nutrients as the land becomes less fertile (Table 1). In addition, higher relative abundance and distribution of arbuscular mycorrhizal fungal PLFA 16:1ω5c (*cis*-11-palmitoleic acid) was observed in the nearby NF soil (5.21%) as compared to the different age series coal mine overburden spoil across the sites. PLFA 16:1ω5c derived from arbuscular mycorrhizal fungi are known to contribute substantially to fungal biomass in NF soil, which responds to the changes in the available C.

Further, the F:B ratio was reported to be the potential tool for discrimination of the disturbed from undisturbed soil system [35]. The F:B ratio exhibited an increasing trend from OB_0 (0.0216) to OB_{10} (0.0990) across the sites. Comparatively higher F:B ratio was estimated in OB_4 (0.0307) as compared to OB_2 (0.0277). In addition, OB_8 (0.0757) exhibited relatively a higher F:B ratio in comparison to OB_6 (0.0649). However, the difference in F:B ratio in chronosequence coal mine overburden spoil was less pronounced due to extreme environmental conditions as well as heavy metal contamination mainly the pyrite and other mineral ores [29].

The Fungal to bacterial ratio (F:B) is the potential tool in order to discriminate the disturbed from undisturbed soil system [34]. Higher F:B ratio was observed in nearby NF soil (0.1128) as compared to the different mine overburden spoil, which may be due to the higher prevalence of fungal diversity exhibiting higher C:N ratio and low bulk density.

The increase in F:B ratio also improves the soil pH towards neutral due to the gradual accumulation of available nutrients leading to the shift in microbial community composition [34,40] across the chronosequence coal mine overburden spoil over time. The capacity of fungi for translocation N to C sources is thought to be important in NF soil with high C:N ratio. In addition, the NF soil was supported with distinct microbial communities, which are correlated well with the factors that define the land-use history and associated soil quality influencing microbial community composition [35]. The study substantiates the findings, which indicated that the disturbed ecosystems have lower F:B ratio whereas the organically managed soil systems have increased F:B ratio than the conventional system.

Further, the values of the soil vulnerability potentials to degradation and hence soil quality of different age series mine spoil can be determined based on the variations in soil properties, which influence mine spoil restoration over time. The time period required for mine spoil restoration can be estimated through the development of prediction model based on the feed forwarded back propagation ANN. Mine spoil samples from 18 mining sites with their efficiency towards mine spoil restoration were randomly divided into a training dataset of 12 mine sites and a test dataset of 6 mine sites. Out of the total 11 parameters, 9 soil parameters *i.e.* Clay, Water holding capacity, Organic carbon, Amylase, Phosphatase, Dehydrogenase, Heterotrophic aerobic bacteria, 16:1ω5c, Fungal: Bacterial were screened out using GFA and used for development of the QSAR equation. Taking a brute-force approach, the number of variables were increases in the QSAR equation one by one and the effect of addition of a new terms on the statistical quality of the proposed model was evaluated. The prediction model (equation 1) with robust prediction of the time period (in years) required for fresh coal mine overburden spoil (OB_0) to reach the soil feature of NF soil has been deduced as per following equation.

$$\text{Year} = 16.6906 + 1.10304\ 16{:}1\omega5c + 0.056\ \text{amylase} - 0.030\ \text{clay} + 0.058\ \text{dehydrogenase} + 0.161\ \text{F:B} - 0.029\ \text{HAB} + 0.164\ \text{OC} - 0.000996\ \text{phosphatase} + 0.345\ \text{WHC}).$$

(n = 12; $R^2 = 0.994$; LOF = 0.0001; $F = 74928646.0$; p = 0.0001; $q^2 = 0.984$).

Where, n is the number of spoil samples in the training set, R^2 is the squared correlation coefficient between observed and predicted years of the mine spoils, *F*-test is the measure of variance that compares two models differing by one or more variables to determine if the complexity of the model correlates positively with its reliability (the model is supposed to be good if the *F*-test is above threshold value), and q^2 is the cross validated R^2. R^2_{LOO} is the square of the correlation coefficient of the cross validation using the leave-one-out (loo) cross-validation technique. The prediction model developed in this study is statistically best fitted (R^2= 0.994, R^2_{LOO} = 0.881) and consequently used for the prediction of years of spoil of training and test sets (Tables 2 and 3). Similarly, the quality of the prediction models for the training set is shown (Figure 2). The R^2 and R^2_{LOO} values of the model corroborate the criteria for a highly predictive model.

Table 2. Statistical assessment of QSAR models for estimation of the predicted year for mine spoil restoration with varying numbers of soil parameters in training set.

Sites	Observed Year	Predicted Year
OB_0_S1	0.00	0.03
OB_0_S2	0.00	0.19
OB_2_S1	2.00	1.91
OB_2_S2	2.00	1.80
OB_4_S1	4.00	4.02
OB_4_S2	4.00	3.81
OB_6_S1	6.00	6.19
OB_6_S2	6.00	6.00
OB_8_S1	8.00	8.08
OB_8_S2	8.00	6.27
OB_{10}_S1	10.00	10.62
OB_{10}_S2	10.00	9.98

The standard error of estimate for the model was 0.001, which is an indicator of the robustness of the fit and suggests that the predicted years of the mine spoils based on the model is reliable. It is necessary that the parameters involved in the equation should not be intercorrelated with each other.

Table 3. Statistical assessment of QSAR models for the estimation of predicted year for mine spoil restoration with varying numbers of soil parameters in the test set.

Sites	Observed Year	Predicted Year
OB_0_S3	0.00	0.11
OB_2_S3	2.00	1.855
OB_4_S3	4.00	3.915
OB_6_S3	6.00	6.095
OB_8_S3	8.00	7.175
OB_{10}_S3	10.00	10.3

To further check the intercorrelation of parameters, VIF analysis was performed. VIF values of these parameters are Clay (5.208), Water holding capacity (12.345), Organic carbon (4.629), Amylase (1.358), Phosphatase (4.310), Dehydrogenase (1.386), Heterotrophic aerobic bacteria (2.932), 16:1ω5c (4.424), Fungal: Bacterial (1.322) respectively. The predicted and the observed years of the mine spoils of the test set are presented (Table 3). Based on the VIF analysis, it was found that the parameters used in the final model have very low intercorrelation. Satisfied with the robustness of the prediction model developed using the training set, we next applied the model to an external data set of spoils comprising the test set. Similarly, the quality of prediction models for the test set is shown (Figure 3).

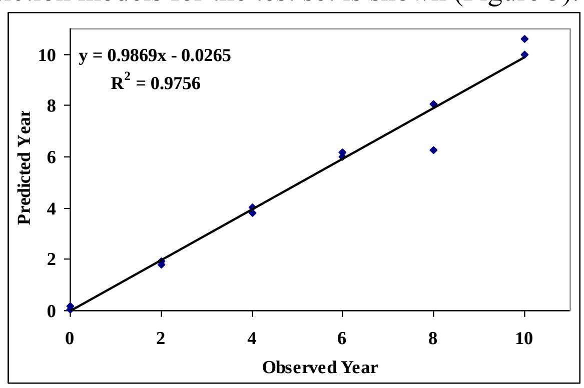

Figure 2. Quantitative structure-activity relationship (QSAR) model revealed the relationship between the predicted and observed year for the training set soil parameters.

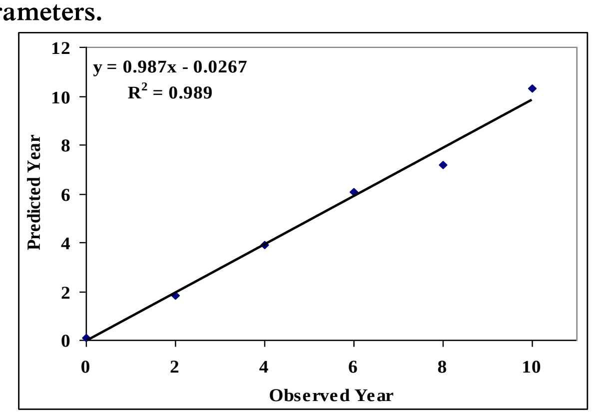

Figure 3. Quantitative structure-activity relationship (QSAR) model revealed the relationship between the predicted and observed year for the test set soil parameters.

The overall root mean square error between the observed and the predicted years was found out to be 0.28, which revealed good predictability. The overall root mean square error is a frequently used measure of the differences between values (sample and population values) predicted by a model or an estimator and the values actually observed. The RMSE represents the sample standard deviation of the differences between predicted values and observed values. These individual differences are called residuals when the calculations are performed over the data sample that was used for estimation, and are called prediction errors when computed out-of-sample. The RMSE serves to aggregate the magnitudes of the errors in predictions for various times into a single measure of predictive power. RMSE is a measure of accuracy, to compare forecasting errors of different models for a particular data and not between datasets, as it is scale dependent. Similarly the data for the nearby native forest soil representing different soil parameters is shown (Table 4)

Table 4. Data representing different soil parameters of nearby native forest soil.

	Clay	WHC	OC	Amylase	PHase	DHase	HAB	SRB	ACT	16:1ω5c	F:B
NF_S1	12.6	47.175	3.875	14.277	95.225	4.121	9.644	2.297	4.849	5.230	0.125
NF_S2	12.1	46.342	3.625	13.124	92.118	4.006	9.431	2.176	4.707	5.210	0.112
NF_S3	11.6	45.509	3.375	11.971	89.011	3.891	9.218	2.055	4.565	5.190	0.092

Conclusion

The squared correlation coefficient between the observed and the predicted years for the test set is also significant (R^2=0.975) that shows the quality of the fit (Figure 3). The estimated correlation coefficient between observed and predicted years with intercept (R^2) and without intercept (R^2_0) were found to be 0.9756 and 0.9755, respectively. The value of $[(R^2 - R^2_0)/R^2]$ = (0.9756 – 0.9755)/0.9756 = 0.0001 is also less than the stipulated value of 0.1. Values of R^2_{test} = 0. 817 and rm^2=0.90 were in the acceptable range, thereby indicating the good external predictability of the prediction model. Hence, the prediction model was used to predict the approximate years required for the reclamation of fresh coal mine overburden spoil as par the characteristic features of the soil of nearby native forest soil. The approximant 39.277 years that will be needed is predicted by the ANN prediction model by taking the input values of the 9 parameters of the coal mine overburden spoil.

Acknowledgements

The authors are thankful to the Head, P.G. Department of Biotechnology and Bioinformatics), Sambalpur University, Odisha for providing the necessary laboratory facilities during the study. Further, the help rendered by several persons during sampling, data analysis and interpretation of data in several counts were duly acknowledged.

References

1. Bell, T.J. and Ungar, I.A. Factors affecting the establishment of natural vegetation on a coal strip mine spoil bank in southeastern Ohio, American Midland Naturalist Journal 1981; 105:19-31.
2. Jha, A.K. and Singh, J.S. Spoil characteristics and vegetation development of an age series of mine spoils in a dry tropical environment, Vegetation 1991; 97: 63-76.
3. MathWorks, Neural network toolbox. Natick, Mass, The MathWorks, Inc. 2005.
4. Zhao, Z.Q., Shahrour, I., Bai, Z.K. Soils development in opencast coal mine spoils reclaimed for 1-13 years in the west-northern Loess Plateau of China, European Journal of Soil Biology 2013; 55: 40-46.
5. Bartuska, M. and Frouz, J. Carbon accumulation and changes in soil chemistry in reclaimed opencast coal mining heaps near Sokolov using repeated measurement of chronosequence sites, European Journal of Soil Science 2015; 66: 104-111.
6. Arshi, A. Reclamation of coal mine overburden dump through environmental friendly method, Saudi Journal of Biological Sciences 2017; 24: 371-378.
7. Yang, C.C., Prasher, S.O., Lacroix, R., Sreekanth, S., Patni, N.K. and Masse, L. Artificial neural network model for subsurface-drained farmlands, Journal of Irrigation and Drainage Engineering 1997; 123(4): 285-292.
8. Howard, D. and Mark, B. Neural network toolbox for computation, visualization and programming-user's guide, The MathWorks, 1st (Eds), 2000; 323.
9. Clifford, G. and Lau, Y. Neural Networks. Theoretical Foundations and Analysis. New York, NY: IEEE 1992.
10. Erzin, Y., Rao, B.H., Patel, A., Gumaste, S.D. and Singh, D.N. Artificial neural network models for predicting electrical resistivity of soils from thermal resistivity, International Journal of Thermal Sciences 2010; 49(1): 118-130.
11. Keskin, T. and Makineci, E. Some soil properties on coal mine spoils reclaimed with black locust (Robinia pceudoacacia L.) and umbrella pine (Pinus pinea L.) in AgacliIstanbul, Environmental Monitoring and Assessment 2009; 159: 407- 414.
12. Banu- Ikizer, S., Aytekin, M., Vekli, M. and Kocabas, F. Prediction of swelling pressure of expensive soils using artificial neural networks, Advances in Engineering Software 2010; 41: 647-655.
13. Chik, Z. Aljanabi, Q.A., Kasa, A. and Taha, M.R. Tenfold cross validation artificial neural network modeling of the settlement behaviour of a stone column under a highway embankment, Arab Journal of Geosciences 2014; 7(11): 4877-4887.
14. Kalinli, A., Acar, M.C. and Gunduz, Z. New approaches to determine the ultimate bearing capacity of shallow foundations based on artificial neural networks and ant colony optimization, Engineering Geology 2011; 117(1): 29-38.
15. Kim, M. and Gilley, J.E. Artificial neural network estimation of soil erosion and nutrient concentration in runoff from land application areas, Computers and Electronics in Agriculture 2008; 64(2): 268-275.
16. Sulewska, M.J. Applying artificial neural networks for analysis of geotechnical problems, Computer Assist Mechanical Engineering Sciences 2011; 18: 231-241.
17. Ingleby, H.R. and Crowe, T.G. Neural network models for predicting organic matter content in Saskatchewan soils, Canadian Biosystems Engineering 2001; 43: 7-12.
18. Licznar, P. and Nearing, M.A. Artificial neural networks of soil erosion and runoff prediction at the plot scale, Catena 2003; 51: 89-114.
19. Akbulut, S. Artificial neural networks for predicting the hydraulic conductivity of coarse grained soils. Eurasian Soil Science 2005; 38: 392-398.

20. Chai, S.S. Veenendaal, B. West, G. and Walker, J.P. Backpropagation neural network for soil moisture retrieval using Nafe'05 data: a comparison of different training algorithms, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Beijing 2008 37:1345-1350.
21. Davood, N.K., Shorafa, M., Omid, M. and Mahmou, F.S. Application of artificial neural networks in modeling soil solution electrical conductivity, Soil Science 2010; 175(9): 432-437.
22. Behrens, T., Forster, H., Scholten, T., Steinrucken, U., Spies, E.D. and Goldschmitt, M. Digital soil mapping using artificial neural networks, Journal of Plant Nutritional and Soil Sciences 2005; 168:1-13.
23. Carpenter, G.A. Gopal, S. Macomber, S. Martens, S. and Woodcock, C.E. A neural network method for mixture estimation for vegetation mapping, Remote Sensing of Environment 1999a; 70(2):138-152.
24. Das, S.K. and Basudhar, P.K. Prediction of residual friction angle of clays using artificial neural network, Engineering Geology 2008; 100(3/4):142-145.
25. Gopal, S., Woodcock, C.E. and Strahler, A.H. Fuzzy neural network classification of global land covers from a 1° AVHRR data set, Remote sensing of Environment 1999; 67(2): 230-243.
26. Odhiambo, L.O., Yoder, R.E., Yoder, D.C. and Hines, J.W. Optimization of fuzzy evapotranspiration model through neural training with input-output examples, Transactions of American Society of Agricultural Engineers 2001; 44(6):1625-1633.
27. Zelles, L. Fatty acid patterns of phospholipids and lipopolysaccharides in the characterization of microbial communities in soil: A review, Biology and Fertility of Soils 1999; 29: 111-129.
28. Zhao, Z., Chow, T.L., Rees, H.W., Yang, Q., Xing, Z. and Meng, F.R. Predict soil texture distributions using an ANN model, Computers and electronics in agriculture. 2009; 65: 36-48.
29. Mishra, R.R. Ecology work book. Oxford and I.B.H. Publication Co., New Delhi, 1968.
30. Roberge, M.R. Methodology of soil enzyme measurement and extraction, In: Soil Enzymes, (Eds) Burns, R.G. London, Academic Press, 1978; 341-369.
31. Claassens, S., Riedel, K.J. Van. Rensburg, L., Bezuidenhout, J.J. and Rensburg Jansen van, P.J. Microbial community function and structure on coal mine discard under rehabilitation, South African Journal of Plant Soil 2006; 23(2):105-112.
32. Friedman, J. Multivariate adaptive regression splines technical report no. 102. Laboratory for computational statistics, department of statistics. stanford university: Stanford, 1990.
33. Naik, P.K., Alam, A., Malhotra, A. and Rizvi, O. Molecular modeling and structure activity relationship of the podophyllotoxin and its congeners, Journal of Biomolecular Screening 2010; 15: 528-540.
34. Naik, P.K., Sindhura, S.T. and Singh, H. Quantitative structure activity relationship (QSAR) of the insecticides: the development of predictive invivo insecticide activity models, SAR and QSAR in Environmental Research 2009; 20: 551-556.
35. Maharana, J.K. and Patel, A.K. Microbial community PLFA responses to ecosystem restoration in a chronosequeces coal mine overburden spoil and implications of soil quality, International Journal of Current Microbiology and Applied Sciences 2014; 3(6): 45-71.
36. Banerjee, S.K., Das, P.K. and Mishra, T.K. Microbial and nutritional characteristics of coal mine overburden spoils in relation to vegetation development, Journal of Indian Society of Soil Science 2000; 48: 63-68.
37. Dutta, R.K. and Agarwal, M. Effect of tree plantations on the soil characteristics and microbial activity of coal mine spoil land, Tropical Ecology 2002; 43(2): 315-324.
38. Vimmerstedt, J.P., House, M.C., Larson, M.M., Kasile, J.D. and Bishop, B.L. Nitrogen and carbon accretion on Ohio coal mine soils: Influence of soil forming factors, Land Urban Planning 1989; 17: 99-111.
39. Harris, J.A. and Birch, P. Soil microbial activity in opencast coal mine restoration, Soil Use and Management 1989; 5: 155-160.
40. Urbanova, M., Kopecky, J., Valaskova, V., Mareckova, M.S., Elhottova, D., Kyselkova, M., Loccoz, Y.M. and Baldrian, P. Development of bacterial community during spontaneous succession on spoil heaps after brown coal mining, FEMS Microbiology Ecology 2011; 78: 59-69.