# Prediction of Ecological Restoration Following Mine Spoil Genesis in Chronosequence Coal Mine Overburden Spoil Using Machine Learning Techniques

**Payal Agrawal[1], Ankita Agrawal[2] Sibaram Panigrahi[3] and Amiya Kumar Patel***

*[1,2] *Department of Biotechnology and Bioinformatics, Sambalpur University, Jyoti Vihar, Burla- 768019, Odisha, India,*
*Department of Computer Science and Engineering. Sambalpur University Institute of Information Technology (SUIIT), Jyoti Vihar, Burla- 768019,Odisha, India.*

## Abstract

Ecological restoration should be dogmatic and involves holistic approach emphasizing the role of microbial community composition that varies in accordance with the physiological and nutritional status of mine spoil profiles. Ecological restoration of coal mine overburden spoil in response to multiple stresses has attracted considerable research attention, which requires periodic assessment using physico-chemical, microbiological, biochemical indices, and analysis *via* suitable models using machine learning techniques that determine the nonlinear relationships among them to predict the time required for mine spoil restoration. In the study, 40 mine spoil variables were selected to develop MOE model based on brute-force approach and genetic function approximation for prediction of mine spoil restoration required for fresh coal mine overburden spoil to reach the soil features of nearby NF soil. The training and test sets with statistically best fitted with $r^2$ = 1.0 and $r^2_{LOO}$ = 0.996. The predicted MOE model with 11 mine spoil variables was recognized as best model illustrating the time period required for mine spoil genesis. The validity of predicted model was confirmed with higher calculated value of squared correlation coefficient determination ($r^2$ = 0.999) and lower root mean square error, which revealed good predictability. Thus, IB0 shall take □□29.257 years to reach soil features of NF soil based on 11 variables as reliable minimal datasets influencing ecological restoration in chronosequence coal mine overburden spoil.

**Key words:** coal mine spoil, ecological restoration, mine spoil genesis, machine learning.

## 1. Introduction

Ecological restoration of coal mine overburden spoil involves initiation, acceleration and recovery of the diverse impacts of environmental degradation due to extensive mining activities with respect to soil quality, integrity and sustainability with an intension to assist the recovery of ecosystem structure and function. Ecological restoration involves progressive development through the state of recovery and revealed potential trajectory expressions in chronosequence coal mine overburden spoil with heterogeneity over time. Being the dynamic entities, the progress of ecosystem recovery should be assessed periodically and explicitly with respect to its integrity, which can be defined in terms of microbial diversity (species composition, microbial community structure and ecological function) and health (microbial organization, adaptation, resilience). Success of ecological restoration of coal mine spoil depends on heavy metal toxicity, establishment of vegetation, nutrients turnover, biogeochemical cycling, carbon sequestration, management practices, biodiversity as well as microbial mediated metabolic processes including microbial diversity, microbial community structure in chronosequence coal mine overburden spoil [1].

Restoration ecology has gained strong scientific concerns addressing problems related to land degradation, bringing new focus to existing ecological theory and fostering foothold novel ecological ideas. The community ecology is strongly linked with ecological restoration useful for designing models of succession including assembly, state-transition and diversity-function relationship. There

exists relationship between mine spoil restoration and r- or k- strategic microbes, where the r-strategists favour new sites or sites under restoration and the k-strategists favour stable environment [2]. Thus, the progress of mine spoil genesis can be evaluated based on microbial community dynamics and functional soil quality indices. The study substantiates the need of 'minimum data sets' (MDS) pre-requisite to elucidate ecological restoration of coal mine overburden spoil. There is no specific MDS recommended for the assessment for ecological restoration in India though microbial biomass pool, enzyme activities, basal soil respiration are being widely used [3,4]. Keeping the extensive coal mining activities, the study facilitates the development of MDS, which would provide the structured approach for the assessment of ecological restoration. MDS involved in the restoration ecology should satisfy following criteria such as (i) compatible with ecosystem processes, (ii) sensitive to management practices in acceptable time frames, (iii) easy to assess during long-term monitoring, (iv) robust methodology, (v) relevant to productivity and ecosystem sustainability and (vi) cost-effective with economic efficiency.

Assessment of time period required for ecological restoration in chronosequence coal mine overburden spoil through experimentation is extensive and time consuming. In contrast, the alternative approach is performed by empirical mathematical modeling to predict the time required for mine spoil restoration. Machine learning (MI) provides significant advantages over conventional statistical methods for analyzing larger ecological datasets including (i) the selection of relevant data and its preprocessing, (ii) selection of adequate algorithms, and (iii) its quality assessment [5]. Machine learning is used to execute algorithms allowing computer to mechanize data-driven model programming and build models based on artificial intelligence through methodical detection of patterns using statistically significant training data, which makes prediction without being explicitly programmed. The objectives of machine learning include (a) classification of datasets based on data-driven models, and (b) data-driven prediction of outcomes based on model.

Machine learning algorithms have been used to predict the time period required for ecological restoration of coal mine spoil. Linear regression is a statistical based predictive algorithm that shows linear relationship between a dependent and one or more independent variables, which revealed the pattern of shifting dependent variables according to the value of independent variables [6]. Assumptions of linear regression algorithm includes (i) linear relationship between dependent and independent variables, (ii) assumes minimal multicollinearity between independent variables, (iii) homoscedasticity (error term is same for all independent variables), (iv) no autocorrelation in error terms and thereby increase the accuracy of predictive model. Polynomial regression algorithm is used to model relationship between a dependent and independent soil variable as nth degree polynomial using training datasets of non-linear nature [7]. Polynomial regression depends on coefficients that are arranged in linear fashion instead of depending on the variables. Lasso (least absolute shrinkage and selection operator) regression performs variable selection and L1 regularization to enhance the prediction accuracy and interpretability of resulting statistical model. Lasso regression is referred as L1 regularization algorithm used for datasets with high multicollinearity, dimensionality and involves shrinkage that allows shrinking coefficients towards zero to avoid over fitting through variable elimination and feature selection [8]. Ridge regression algorithm is a tuning method based on L2 regularization approach, which is used to analyse the datasets with relatively high multicollinearity, which is mostly used to reduce over fitting thereby making predictive model more robust including all features by reducing complexity of the model through shrinking of coefficients [9].

Elastic net algorithm is referred as embedded method that performs variable selection, and L1 and L2 regularization simultaneously, which is mostly used when the dataset is greater than number of samples. Such approach combines feature elimination derived by Lasso and feature coefficient reduction from ridge regression to improve prediction accuracy using penalties to shrink coefficients of independent undesired variables [10].

Random Forest belongs to supervised learning technique used for both classification and regression problems based on ensemble learning**,** which combines the multiple classifiers to resolve complexity and improve model prediction accuracy [11]. Random Forest includes a number of decision trees based on variable datasets and takes the average of the prediction of each tree without relying on one decision tree to predict final output with higher predictive accuracy. Greater is the number of decision trees leads to higher accuracy and prevents the problem of overfitting. Random Forest has several appealing features such as (i) minimal training time compared to other algorithms,

(ii) predicts output with high accuracy even for larger datasets with high dimensionality, (iii) enhances model accuracy and prevents overfitting, and (iv) maintains accuracy even if large proportion of missing data.

Gradient boosting is supervised learning algorithm, built by combining decision trees with boosting to develop prediction model [12]. Gradient boosting is used in regression and classification problems, which predicts model combat with bias and variance based on decision trees. Gradient boosting performs well associated with unbalanced data and exhibits more trees than Random Forest. Decision trees are aggregated at the end in Random Forest, while result of each decision tree is used to calculate final result in Gradient boosting. XGBoost (Extreme gradient boosting) algorithm is decision tree-based ensemble algorithm using gradient boosting framework with scalable system for parallel tree boosting for supervised learning problems [13]. XGBoost provides the leading ML library dominated with structured datasets for regression, classification and regression predictive problems. XGBoost emphasizes functional space by reducing the cost of predictive model while Random Forest gives preferences to hyperparameters to optimize the model.

MLP Regressor (Multilayer perceptron) is a feed-forward artificial neural network, which is used ambiguously either loosely to mean any feed-forward ANN or strictly to refer to networks composed of the multiple layers of perceptron. MLP repressor is the basic deep neural network supplemented with a series of fully connected layers, which can be used to overcome the requirement of high computing power required by modern deep learning architectures and is used for predicting model of soil by optimization of the log-loss function using LBFGS or stochastic gradient descent [8].

SVR (Support vector regression) is a supervised learning algorithm, which is used as classification algorithm to predict discrete values using the concepts of SVM (Support vector machine). SVR acknowledges the presence of non-linearity in datasets and supplements the proficient prediction model by providing the flexibility in acceptable error in order to find out hyperplane accommodating maximum number of points [13]. Advantages of SVR include (i) its computational complexity does not depend on the dimensionality of input space, (ii) exhibit excellent generalization capability with high prediction accuracy.

Stacking regression is ensemble learning technique involving stacked generalization consisting of the output of individual estimator and uses predictions for multiple nodes to compute final prediction of the test datasets through multiple regression model [14]. The principle of ensembling is to combine predictions of various models built with learning algorithm to improve robustness over single model. Stacking involves heterogeneous weak learners, learns them in parallel, and combines them by training a meta-learner to output a prediction based on the different weak learner's predictions.

Genetic algorithm is a stochastic global optimization algorithm, which is most popular and widely known biologically inspired algorithm. Genetic algorithm is an approximation algorithm and used for solving both constrained and unconstrained optimization procedure with binary representation based on natural selection that drives biological evolution. The method is robust and highly efficient but does not guarantee an exact identification of the optimum solution. However, it does permit the localization of an optimum set of solutions close to this optimum [15]. Genetic algorithm allows to choose the suitable data sets according to the Pareto frontier and to guarantee the coherence between the tests, which is part of a predictive benchmark. Several studies revealed the use of genetic algorithm as an efficient tool of the prediction of ecological restoration of mine overburden spoil [16, 17, 18].

Considering tropical dry deciduous forest as the natural vegetation in the study site, the present study was designed to predict the time period required for fresh coal mine overburden spoil to reach the soil features of nearby forest soil based on the variations in soil quality indices in chronosequence coal mine overburden spoil by multivariate predictive machine learning techniques. Such approach is considered to be superior compared to the non-parametric statistical benchmark methods, which provide the accurate prediction of ecological restoration. The predictive models were developed using datasets to evaluate the performance for training, testing and validation indicating the good network generalization for predicting mine spoil restoration in chronosequence coal mine overburden spoil.

## 2. Materials and Methods

### 2.1 Study site

Basundhara (west) open cast colliery under Ib valley of Mahanadi Coalfields Limited located in Sundargarh district of Odisha (22°03'58"-20°04'11" north latitude and 83°42'46"-83°44'45" east longitude) was selected as the study site, which was topologically hilly sloppy (244m above sea level) to plateau. The study area exhibited top soil with thickness that varies from 0.15-0.30 m with an average of 0.22 m. Climatic condition of study site is considered to be Aw according to the Köppen-Geiger climate classification, which is broadly hot, dry and semi-arid with an average of 1483 mm annual rain fall, 26°C temperature and 58.58% humidity per year. Natural vegetation of the area is considered to be tropical dry deciduous forest. Coal mining activities (open cast) lead to generation of huge mine overburden spoil, which were categorized based on their age since inception namely fresh mine spoil (OB0), 3 yr (OB3), 6 yr (OB6), 9 yr (OB9), 12 yr (OB12) and 15 yr (OB15).

### 2.2 Mine spoil sampling

Each coal mine overburden was randomly demarcated into 5 blocks for the sampling of mine spoil. Pits of (15×15×15) $cm^3$ size were dugged up and five samples were collected from each block from (0-15) cm soil depth that were referred as "sub-samples', which were mixed thoroughly to form one 'composite sample' for each mine overburden. Similar strategy of mine spoil sampling was performed from chronosequence overburden (OB0 → OB15) and nearby forest soil (NF), which was taken as reference. Aseptically collected samples were homogenized, sieved with 0.2 mm mesh and kept at 4°C for further analysis.

### 2.3 Quantitative analysis of mine spoil variables

#### 2.3.1 Physico-chemical characterization

Textural composition of chronosequence mine overburden spoil and NF soil were estimated following TSBF handbook [19]. Moisture content (MC), bilk density (BD) and water holding capacity (WHC) were estimated [20]. Soil pH (1:2.5 ratio of soil: water) was measured. Soil organic C by titration method [20], total N by Kjeldahl method [21] and extractable P by chlorostannous reduced molybdophosphoric blue colour method [22] were estimated.

#### 2.3.2 Microbiological characterization

Microbial biomass C by fumigation extraction method [23], microbial biomass N by $CHCl_3$ fumigation method were estimated and expressed on oven dry weight basis [24]. Microbial biomass P was estimated [25], where inorganic P was extracted by 0.5M $NaHCO_3$ (pH 8.5) adjusted with NaOH. Extracted P was determined by $CHCl_3$ fumigation [26]. Microbial basal respiration was estimated by alkali absorption method [27].

Enumeration of microbial population in chronosequence coal mine overburden spoil and NF soil was performed through serial dilution upto $10^{-10}$ fold following spread plate technique. Azotobacter population (AZB) using azotobacter mannitol agar (ATCC 1992), arthrobacter population (ARB) using arthrobacter medium supplemented with 0.01% cycloheximide and 2% NaCl, rhizobial count (RZB) using yeast extract mannitol agar with congo red dye, heterotrophic aerobic bacterial population (HAB) using nutrient agar, sulfate reducing bacterial population (SRB) count using sulfate reducing medium, actinomycetes population (ACT) using starch-casein agar supplemented with streptomycin (40 µl/ml) and griseofulvin (50µl/ml) to inhibit bacterial and fungal growth, yeast count (YES) using potato sucrose agar and fungal count (FUN) using rose bengal agar supplemented with streptomycin (50µl/ml) to inhibit bacterial contaminants were estimated.

#### 2.3.3 Enzyme activities

Different enzyme activities exhibited by the chronosequence coal mine spoil and NF soil was determined such as amylase activity [28], invertase [29], protease activity [30] using sodium caseinate as substrate, urease activity [31], phosphatase activity [32] and dehydrogenase activity [33].

### 2.3.4 Phospholipid fatty acid profiling (PLFA)

PLFAs in chronosequence coal mine overburden spoil and NF soil was performed through lipid extraction based on fractionation and quantification [34]. Extracts were cleaned up by SPE chromatography using $NH_2$ SPE column, the samples were dissolved in equal volume of hexane: methyl tert- butyl ether (1:1 v/v) and quantified by GC-MS. Fungal to bacterial (F/B) ratio of microbial biomass was used to study the state of microbial community structure in chronosequence coal mine overburden spoil. The fungal biomass was calculated based on PLFAs 18:1 w9c and18:2ω6c. Total bacterial biomass was obtained by the summation of PLFAs 14:0, 15:0, a15:0, i15:0, i16:0, 16:1ω7c, 16:1ω11c, 10Me 16:0, 17:0, a17:0, cy17:0, i17:0, 17:1ω8c, 10Me 17:0, 18:0 2OH, 18:1ω5c, 18:1ω7c, 10Me 18:0, 19:1ω6c and cy19:0ω8c [35].

### 2.3.5 Community level physiological profiling (CLPP)

CLPP profiling was performed using BIOLOG™ Ecoplates supplemented with 31 carbon sources. Five groups of carbon substrates were used such as carbohydrates, carboxylic and ketonic acids, amino acids, polymers, and amines and amides [36] and absorbance was taken at 590 nm using BIOLOG™ microstation at 24, 48, 72 and 96 hrs of incubation. Microbial response expressed in average well colour development (AWCD) was derived from mean difference among absorbance values of 31 response wells.

$$\text{AWCD} = \frac{1}{31}\sum_{i=1}^{31}(A_i - A_0)$$

Where, $A_i$ represents absorbance reading of the well $i$, which is corrected by subtracting the absorbance value of blank well $A_0$ (without carbon source) from the value of each well. Shortest incubation time that allows better resolution is used to calculate AWCD, richness (R) and Shannon-Weaver index (H) [36]. Richness value was calculated as the number of oxidized C substrates using absorbance value of 0.25 as threshold for positive response [37]. Shannon-Weaver index was calculated as: $H = -\sum pi\ (\ln pi)$; where, $pi$ is the ratio of activity on each substrate to sum of activities including all substrates.

### 2.4 Data sets used for prediction model

In the study, 40 mine spoil attributes were used to discriminate the chronosequence coal mine overburden spoil and NF soil. Soil attributes were classified into the following categories: (i) Subclass-I includes physico-chemical attributes such as clay percentage, bulk density, water holding capacity, moisture content, pH, organic C, total N and extractable P; (ii) Subclass-II includes microbiological attributes such as Microbial biomass C, N and P, basal soil respiration, microbial counts (azotobacter, arthrobacter, rhizobium, heterotrophic aerobic bacteria, actinomycetes, yeast, fungi and sulfate reducing bacteria); (iii) Subclass-III includes enzyme activities (amylase, invertase, protease, urease, phosphates, dehydrogenase); (iv) Subclass-IV includes the relative distribution of microbial community structure based on PLFAs distribution (18:1 w9c, 18:2 w6,9c, Anaerobes, 16:1w5c, 10-Methyl, Gram positive bacteria, Gram negative bacteria, and fungal to bacterial ratio); (v) Subclass-V contains the attributes of CLPP including AWCD and substrate utilization (carbohydrates, carboxylic and ketonic acids, amino acids, polymers, amine and amides). The following data sets were used to determine the progress of mine spoil genesis following ecological restoration, which can be ultimately used for the prediction of time period required by fresh mine overburden spoil in chronosequence to reach the soil features of nearby forest soil.

### 2.5 Machine learning algorithms used for prediction

In the study, 12 different machine learning algorithms such as Linear regressor (LR), Polynomial regressor (PR); Lasso regressor (LSR), Ridge regressor (RR), Elastic net Regressor (ENR), Random Forest (RF), Gradient boosting (GB), Extreme gradient boosting (XGB), Multilayer perceptron regressor (MLPR), Support vector regressor (SVR), Stacking regressor (SR) and Genetic algorithm (GA) were used to fit the train data. Since some of the machine learning algorithms are stochastic in nature, the simulations were repeated 10 times and RMSE value is measured. The model providing the lowest mean RMSE is used to predict the time period required for ecological restoration

of coal mine overburden spoil. Excluding genetic algorithm, all other models were implemented using Python 3.6. The number of input features to ML algorithms was set same to the number of features *i.e.* 40 and the remaining parameters of machine learning models were set to default values in Python 3.6. The genetic algorithm is implemented in SVL script for the development of MOE equation.

### 2.6 Screening of descriptors and development of the model

A set of 40 parameters that discriminate the six different age series of coal mine overburden spoil in chronosequence was used for developing the prediction model using genetic algorithm. The set of parameters that provide the statistically best prediction model out of 40 parameters were selected through 10,000 prediction model using genetic function approximation implemented in SVL script of MOE (Molecular Operating Environment). The genetic algorithm starts with the creation of a population of randomly generated parameters sets. The algorithm was set to determine the soil parameters relevant for mine spoil genesis by linear polynomial terms. The usage probability of a given parameter from the active set was 0.5 in any of the initial population sets. The sets were then compared according to their objective functions. Parameters set used for genetic algorithm includes mutation 0.1, crossover 0.9, population 1000, number of generations 10,000, $r^2$ floor limit 50% and objective function $r^2/N\_par$. The form of the objective function favors sets, which have $r^2$ value as high as possible while minimizing the number of parameters used as descriptors. Higher the score, the higher is the probability that a given set will be used for creation of the next generation of sets. Creation of a consecutive generation involves crossovers between set contents as well as mutations. The algorithm runs until the desired number of generations is reached. Equations were developed between the observed activity and descriptors. The best equation was taken based on the statistical parameters such as squared regression coefficient ($r^2$) and leave-one out cross-validated regression coefficient ($R^2_{LOO}$).

### 2.7 Validation of the prediction model

The predictive capability of the developed prediction model was validated using the leave-one-out cross-validation method. 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 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 Ypred, Yexp and ȳ represents the predicted, observed and mean values of observed activity belonging to the training datasets of soil variables respectively. The determination coefficient in prediction using the test set ($R^2$ test) was calculated [38] using the following equation:

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

Where, $R^2$ test 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 $R^2$ test is the squared correlation coefficient for regression without using the y-intercept, and the regression equation was Y = bx. Further, the intercorrelation between different variables used in final prediction model was checked through variance inflation factor (VIF) analysis. 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 variables. If VIF value is larger than 10 for a given variable, its information could be hidden by other variables [38].

## 3. Results and Discussion

Wide variation in different physico-chemical properties (clay %, bilk density, water holding capacity, moisture, soil pH, organic C, total N and extractable P), microbiological properties (MB-C, MB-N, MB-P, BSR), microbial enumeration (AZB, ARB, RZB, HAB, ACT, YES, FUN, SRB),

enzyme activity (amylase, invertase, protease, urease, phsophatase and dehydrogenase), PLFAs (18:1 w9c, 18:2 w6,9c, anaerobes, 16:1w5c, 10-Methyl, Gram positive, Gram negative and F:B ratio), patterns of substrate utilization based on community level physiological profiling (Average well colour development, carbohydrates, carboxylic and ketonic acid, amino acids, polymers, amines/amides) revealed the shift in microbial community composition in chronosequence mine overburden spoil and NF soil (Table 1).

**Table 1. Comparative Assessment of Mine Spoil Attributes in Chronosequence Coal Mine Overburden Spoil (OB0 → OB15) and Nearby NF Soil.**

| Parameters | OB0 | OB3 | OB6 | OB9 | OB12 | OB15 | NF Soil |
|---|---|---|---|---|---|---|---|
| Clay | 5.3 | 7.5 | 9.1 | 10.3 | 11.2 | 11.8 | 12.9 |
| BD | 1.712 | 1.584 | 1.389 | 1.321 | 1.293 | 1.268 | 1.236 |
| WHC | 26.73 | 33.29 | 39.13 | 42.45 | 44.67 | 45.36 | 47.13 |
| MC | 6.913 | 7.328 | 7.967 | 8.672 | 9.547 | 10.238 | 11.319 |
| pH | 6.12 | 6.21 | 6.35 | 6.42 | 6.59 | 6.68 | 6.92 |
| OC | 0 | 0.358 | 1.118 | 1.634 | 2.118 | 2.684 | 3.705 |
| TN | 0 | 32.963 | 83.562 | 335.523 | 915.658 | 1267.25 | 1733.12 |
| EP | 0 | 6.359 | 14.137 | 54.522 | 108.452 | 171.152 | 272.531 |
| MB-C | 0 | 14.357 | 65.436 | 173.554 | 421.982 | 596.358 | 947.564 |
| MB-N | 0 | 1.983 | 5.364 | 23.365 | 65.613 | 93.453 | 141.288 |
| MB-P | 0 | 0 | 1.935 | 8.618 | 17.549 | 29.667 | 48.543 |
| BSR | 0.129 | 0.213 | 0.258 | 0.297 | 0.356 | 0.394 | 0.463 |
| AZB | 1.3617 | 2.0607 | 3.3118 | 4.1847 | 4.4914 | 4.7404 | 5.7993 |
| ARB | 3.0414 | 3.3617 | 4.0792 | 4.5563 | 4.7404 | 4.9191 | 5.5682 |
| RZB | 1.8129 | 1.9685 | 3.2304 | 4.0969 | 4.3222 | 4.5441 | 5.3802 |
| HAB | 3.3424 | 3.5441 | 4.1614 | 5.5563 | 7.6128 | 7.8751 | 9.3617 |
| ACT | 2.4472 | 3.1761 | 3.7993 | 4.0414 | 4.2041 | 4.3424 | 4.7076 |
| YES | 1.4472 | 2.0792 | 2.3617 | 2.959 | 3.1761 | 3.3802 | 3.8921 |
| FUN | 1.4771 | 1.9031 | 2.9542 | 3.1139 | 3.3979 | 3.7853 | 5.2788 |
| SRB | 7.5051 | 6.2304 | 5.1139 | 4.9542 | 4.0414 | 3.8451 | 1.9542 |
| Amylase | 0 | 1.564 | 2.259 | 3.963 | 5.894 | 8.671 | 13.124 |
| Invertase | 0 | 7.139 | 35.361 | 126.106 | 361.549 | 623.472 | 849.335 |
| Protease | 0 | 4.137 | 18.654 | 31.364 | 46.357 | 88.674 | 215.813 |
| Urease | 0 | 4.532 | 8.667 | 15.862 | 22.539 | 36.784 | 57.913 |
| Phosphatase | 0 | 5.325 | 21.329 | 32.467 | 50.264 | 62.338 | 89.175 |
| Dehydrogenase | 0.048 | 0.198 | 0.635 | 0.959 | 2.115 | 2.684 | 4.138 |
| 18:1 w9c | 0.65 | 1.23 | 1.67 | 1.75 | 1.98 | 3.12 | 4.15 |
| 18:2 w6,9c | 0.28 | 0.54 | 0.91 | 1.15 | 1.63 | 1.82 | 2.18 |
| Anaerobes | 5.12 | 4.56 | 4.83 | 3.85 | 3.61 | 3.55 | 3.47 |
| 16:1w5c | 0 | 0.13 | 0.28 | 0.39 | 1.24 | 1.37 | 4.19 |
| 10-Methyl | 2.63 | 2.31 | 1.84 | 1.32 | 1.14 | 0.57 | 0.32 |
| Gram positive | 15.32 | 13.42 | 11.14 | 13.81 | 13.12 | 11.98 | 10.48 |
| Gram negative | 29.13 | 21.23 | 18.23 | 15.85 | 15.17 | 14.56 | 11.43 |
| F:B ratio | 0.055 | 0.094 | 0.107 | 0.17 | 0.208 | 0.288 | 0.348 |
| AWCD | 0.064 | 0.1063 | 0.2148 | 0.3132 | 0.4236 | 0.506 | 0.6994 |
| Carbohydrates | 0.0033 | 0.0126 | 0.0869 | 0.2139 | 0.4035 | 0.4758 | 0.6131 |
| Carboxylic & ketonic acid | 0.0004 | 0.0036 | 0.0139 | 0.1356 | 0.2954 | 0.3267 | 0.5228 |
| Amino acids | 0.0236 | 0.0368 | 0.0653 | 0.2049 | 0.3427 | 0.4126 | 0.5826 |
| Polymers | 0.0004 | 0.0137 | 0.0986 | 0.1257 | 0.1963 | 0.2316 | 0.2764 |
| Amines/amides | 0.0077 | 0.0085 | 0.0098 | 0.0106 | 0.0151 | 0.0348 | 0.0536 |

With relevance to mine spoil genesis, soil quality assessment through periodic monitoring, soil fertility status, available soil nutrients and biological monitoring, different models predicting ecological restoration is of prime importance [39,40]. Several models predicting ecological restoration were developed [4,7,41]. RMSE (Root mean square error) value for 10 independent simulations in predicting the time period for the ecological restoration of coal mine overburden spoil by the machine learning algorithms are presented (Table 2). The study suggested that the genetic algorithm provides the lowest mean RMSE value compared to other models used in the study. In contrast, the other ML algorithms have revealed poor performance with relatively higher RMSE value. Being the minimal RMSE value exhibited in 10 instances, genetic algorithm was found to be the most suitable predictive machine learning algorithm for the prediction of time period required by fresh coal mine overburden spoil (OB0) in chronosequence coal mine overburden spoil to reach the soil features of NF soil.

**Table 2. RMSE (Root Mean Square Error) Value of Different Machine Learning Predictive Models Used for Prediction of Time Period Required for Ecological Restoration in Chronosequence Coal Mine Overburden Spoil.**

| # | LR | PR | LSR | RR | ENR | RF | GB | XGB | MLPR | SVR | SR | GA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. | 0.08 | 0.08 | 0.467 | 0.045 | 0.285 | 0.412 | 0.117 | 0.008 | 0.047 | 3.547 | 0.037 | 0 |
| 2. | 0.115 | 0.115 | 0.793 | 0.475 | 0.475 | 0.398 | 0.719 | 0.001 | 18.62 | 2.144 | 0.241 | 0 |
| 3. | 0.139 | 0.139 | 0.737 | 0.073 | 0.442 | 0.444 | 0.315 | 1.497 | 26.487 | 2.137 | 0.379 | 0 |
| 4. | 0.154 | 0.154 | 0.261 | 0.061 | 0.167 | 0.464 | 0.258 | 1.503 | 30.287 | 2.079 | 0.158 | 0 |
| 5. | 0.115 | 0.115 | 0.381 | 0.123 | 0.247 | 0.579 | 0.424 | 0.002 | 29.575 | 2.958 | 0.187 | 0 |
| 6. | 0.114 | 0.114 | 0.787 | 0.165 | 0.476 | 0.39 | 0.205 | 0.002 | 0.283 | 3.837 | 0.247 | 0 |
| 7. | 0.174 | 0.174 | 0.27 | 0.106 | 0.21 | 0.532 | 0.337 | 0.008 | 40.262 | 1.584 | 0.13 | 0 |
| 8. | 0.052 | 0.052 | 0.334 | 0.073 | 0.209 | 0.526 | 0.225 | 0.008 | 0.223 | 1.584 | 0.802 | 0 |
| 9. | 0.258 | 0.258 | 0.459 | 0.098 | 0.281 | 0.586 | 0.46 | 1.503 | 7.211 | 3.236 | 0.82 | 0 |
| 10. | 0.256 | 0.256 | 0.536 | 0.059 | 0.31 | 0.355 | 0 | 0.008 | 0.804 | 3.553 | 0 | 0 |
| Mean | 0.145 | 0.145 | 0.502 | 0.127 | 0.310 | 0.468 | 0.306 | 0.454 | 15.379 | 2.665 | 0.30 | 0 |

**NB: LR: Linear Regressor; PR: Polynomial Regressor; LSR: Lasso Regressor; RR: Ridge Regressor; ENR: Elastic Net Regressor; RF: Random Forest; GB: Gradient Boosting; XGB: Extreme Gradient Boosting; MLPR: Multilayer Perceptron Regressor; SVR: Support Vector Regressor; SR: Stacking Regressor; GA: Genetic Algorithm.**

Out of 40 parameters, 11 parameters such as clay, OC, TN, MB-C, MB-N, MB-P, BSR, dehydrogenase, 18:1ω9c, F:B ratio and AWCD were screened out using Genetic algorithm implemented in SVL script for development of MOE equation. The number of input variables was reduced from 40 to 11 without affecting the predictive power of decision trees, which was substantiated by earlier studies [16,40]. Furthermore, the reduction in number of input variables allowed easing the evaluation of time period for ecological restoration [40,41].

**Table 3. Comparative Distribution of 11 Parameters Selected for Developing Prediction Model in Chronosequence Coal Mine Overburden Spoil (OB0 → OB15) and nearby NF Soil.**

| Parameters | OB0 | OB3 | OB6 | OB9 | OB12 | OB15 | NF Soil |
|---|---|---|---|---|---|---|---|
| Clay | 5.3 | 7.5 | 9.1 | 10.3 | 11.2 | 11.8 | 12.9 |
| OC | 0 | 0.358 | 1.118 | 1.634 | 2.118 | 2.684 | 3.705 |
| TN | 0 | 32.963 | 83.562 | 335.523 | 915.658 | 1267.25 | 1733.12 |
| MB-C | 0 | 14.357 | 65.436 | 173.554 | 421.982 | 596.358 | 947.564 |
| MB-N | 0 | 1.983 | 5.364 | 23.365 | 65.613 | 93.453 | 141.288 |
| MB-P | 0 | 0 | 1.935 | 8.618 | 17.549 | 29.667 | 48.543 |
| BSR | 0.129 | 0.213 | 0.258 | 0.297 | 0.356 | 0.394 | 0.463 |
| Dehydrogenase | 0.048 | 0.198 | 0.635 | 0.959 | 2.115 | 2.684 | 4.138 |
| 18:1 w9c | 0.65 | 1.23 | 1.67 | 1.75 | 1.98 | 3.12 | 4.15 |
| F:B ratio | 0.055 | 0.094 | 0.107 | 0.17 | 0.208 | 0.288 | 0.348 |
| AWCD | 0.064 | 0.1063 | 0.2148 | 0.3132 | 0.4236 | 0.506 | 0.6994 |

The prediction model with robust prediction of the time period (in year) required for fresh coal mine overburden spoil to reach the soil feature of nearby NF soil has been deduced as per the following equation.

**Year =** -7.53288 + (0.0869394*18:1ω9c) + (7.41153*AWCD) + (4.02414*BSR) + (1.35865*Clay) – (5.32018*Dehydrogenase) – (2.43722*F:B ratio) + (0.0716277*MB-C) – (0.00580868*MB-N) – (0.174259*MB-P) + (0.0185814*OC) – (0.0138142*TN)

(n = 11; $r^2$ = 1.0; LOF = 0.0001; F = 1615.4; $p$ = 0.0001; $r^2_{LOO}$ = 0.996).

Where, 'n' is number of mine spoil samples, $r^2$ is the squared correlation coefficient between observed and predicted years of 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 $r^2_{LOO}$ is the square of correlation coefficient of cross validation using the leave-one-out (loo) cross-validation technique [16,46,47]. Prediction model developed in this study is statistically best fitted ($r^2$= 1.0; $r^2_{LOO}$ = 0.996) and used for the prediction of time period (in years) for ecological restoration based on the training and test sets (Table 3 and 4).

**Table 3. Statistical Assessment of GFA Model for the Estimation of Predicted Year for Coal Mine Spoil Restoration with Varying Numbers of Soil Variables.**

| Sites | Observed Year | Predicted Year |
|---|---|---|
| OB0_S1 | 0.00 | 0.1226966 |
| OB0_S2 | 0.00 | 0.2808207 |
| OB3_S1 | 3.00 | 3.2285315 |
| OB3_S2 | 3.00 | 3.4965585 |
| OB6_S1 | 6.00 | 6.3784361 |

| | | |
|---|---|---|
| OB6_S2 | 6.00 | 6.5172648 |
| OB9_S1 | 9.00 | 9.5377215 |
| OB9_S2 | 9.00 | 9.8436099 |
| OB12_S1 | 12.00 | 12.981682 |
| OB12_S2 | 12.00 | 12.730812 |
| OB15_S1 | 15.00 | 15.879806 |
| OB15_S2 | 15.00 | 15.988135 |

**Table 4. Statistical Assessment of GFA Model for the Estimation of Predicted Year for Mine Spoil Restoration with Varying Numbers of Soil Variable in the Test Set.**

| Sites | Observed Year | Predicted Year |
|---|---|---|
| OB0_S3 | 0.00 | 0.4168766 |
| OB3_S3 | 3.00 | 3.7329298 |
| OB6_S3 | 6.00 | 6.762202 |
| OB9_S3 | 9.00 | 9.825289 |
| OB12_S3 | 12.00 | 13.356653 |
| OB15_S3 | 15.00 | 16.234989 |
| OB15_S3 | 15.00 | 16.234989 |

The quality of prediction models for the training set is represented (Figure 1a). The $r^2$ and $r^2_{LOO}$ values of the model corroborate the criteria for a highly predictive model. The standard error for the proposed model was found to be 0.276, which can be used as an indicator of robustness of the fit and suggests that the predicted year of mine spoil based on the model is reliable. Similarly, the quality of prediction models for the test set is shown (Figure 1b). The overall root mean square error (RMSE) between the observed and predicted years was found to be 0.194 that revealed good predictability. Squared correlation coefficient between the observed and predicted years for test set was significant ($R^2$ = 0.999) (Figure 1b). Estimated correlation coefficients between observed and predicted years with intercept ($R^2$) and without intercept ($R^2_0$) were found to be 0.9992 and 0.9991 respectively. The value of $[(R^2 - R^2_0)/R^2] = (0.9992 - 0.9991)/0.9992 = 0.0001$ is less than the stipulated value of 0.1. Thus, GFA prediction model can be used to determine the time required for fresh coal mine overburden spoil as par with the soil features of nearby NF soil taking into account input values of 11 variables was estimated to be ~ 29.257 years.

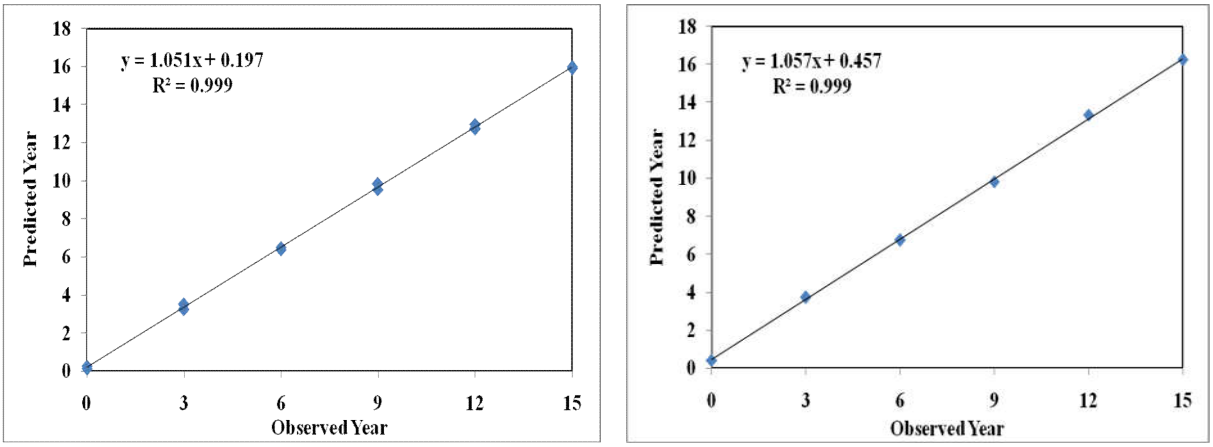


**Figure 1. GFA Model Revealed the Relationship Between the Predicted and Observed Year based on Datasets on Coal Mine Spoil Parameters using (A) Training Set; (B) Test Set.**

## 4. Conclusion

The study clearly indicated that the holistic approach based on the combination of quantitative biomarkers established connecting links between the fluxes driving nutrient pool, which can be used as efficient strategy for ecological restoration. The quality assessment will assist to determine soil variables for development of 'minimum datasets' (MDS) influencing the quality thresholds set for soil quality indicators depending on the impact of anthropogenic activities including mine activities over time. The multivariate predictive modelling based on genetic algorithm was designed using different mine spoil variables to predict the time period required for mine spoil restoration. The study revealed the selection of 11 parameters out of 40 mine spoil attributes as input variables, which is considered as powerful tool in predicting best model for mine spoil restoration in chronosequence coal mine spoil based on genetic function approximation. The validity of the developed GFA model was confirmed by squared correlation coefficient ($r^2$ = 0.999) and lower root mean square error (RMSE = 0.194 kPa). The study based on GFA predictive model determine the time period required for fresh mine overburden spoil to reach the soil features of NF soil shall take ~ 29.257 years. The study suggested that the 11 mine spoil variables can be used as the 'minimum datasets' to determine the progress of mine spoil genesis that influence ecological restoration. Therefore, there is need for screening minimum datasets with higher discriminating potential, which not only used to evaluate the role of microbial community structure influencing ecological restoration but also provide the information lacunae regarding the dynamic interplay between site-specific and landscape-scale assessment for critical assessment of microbial processes to implement effective reclamation strategies.

## 6. Acknowledgement

Authors were indebted to every individual involved in sampling of mine spoil and statistical analysis during the present investigation. In the same breath, due acknowledgement has been given to the HOD, Department of Biotechnology and Bioinformatics, Sambalpur University, Odisha for providing the necessary laboratory facilities and infrastructure to perform the experimental work to fulfill the objectives of the present study.

## 7. References


1. J. K. Maharana and A. K. Patel, "Assessment of microbial diversity associated with chronosequence coal mine overburden spoil using random amplified polymorphic DNA markers", International Journal of Recent Scientific Research, vol. 6, **(2015)**, pp. 4291-4301.
2. M. Kujur and A. K. Patel, "Quantifying the contribution of different soil properties on microbial biomass carbon, nitrogen and phosphorous in dry tropical ecosystem", International Journal of Environmental Sciences, vol. 2, no. 3, **(2012)**, pp. 2272-2284.
3. C. Hawes, J. N. Morris, C. D. Phillips, V. Mor, B. E. Fries and S. Nonemaker, "Reliability estimates for the Minimum Data Set for nursing home resident assessment and care screening (MDS)". The Gerontologist, vol. 35, no. 2, **(1995)**, pp. 172-178.
4. G. Li, J. Chen, Z. Sun, and M. Tan, "Establishing a minimum dataset for soil quality assessment based on soil properties and land-use changes", Acta ecologicasinica, vol. 27, no. 7, **(2007)**, pp. 2715-2724.
5. R. Muhamedyev, "Machine learning methods: An overview. Computer modelling & new technologies", vol. 19, no. 6, **(2015)**, pp. 14-29.
6. D. Maulud and A. M. Abdulazeez, "A review on linear regression comprehensive in machine learning", Journal of Applied Science and Technology Trends, vol. 1, no. 4, **(2020)**, pp. 140-147.
7. E. E. Golia and V. Diakoloukas, "Soil parameters affecting the levels of potentially harmful metals in Thessaly area, Greece: a robust quadratic regression approach of soil pollution prediction", Environmental Science and Pollution Research, vol. 29, no. 20, **(2022)**, pp. 29544-29561.
8. F. Wang, Z. Shi, A. Biswas, S. Yang and J. Ding, "Multi-algorithm comparison for predicting soil salinity", Geoderma, vol. 365, **(2020)**, pp. 114211.
9. C. J. Ransom, N. R. Kitchen, J. J. Camberato, P. R. Carter, R. B. Ferguson, F. G. Fernández, ... and J.F. Shanahan, "Statistical and machine learning methods evaluated for incorporating soil and weather into corn nitrogen recommendations", Computers and Electronics in Agriculture, vol. 164, **(2019)**, pp. 104872.
10. X. N. Bui, H. Nguyen, Q. H. Tran, H. B. Bui, Q. L. Nguyen, D. A. Nguyen, ... and V. V. Pham, "A lasso and elastic-net regularized generalized linear model for predicting blast-induced air over-pressure in open-pit mines". Inżynieria Mineralna, **(2019)**, pp. 21.
11. K. Tan, W. Ma, F. Wu and Q. Du, "Random Forest–based estimation of heavy metal concentration in agricultural soils with hyperspectral sensor data", Environmental monitoring and assessment, vol. 191, no. 7, **(2019)**, pp. 1-14.
12. J. Cai, K. Xu, Y. Zhu, F. Hu, and L. Li, "Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest", Applied Energy, vol. 262, **(2020)**, pp. 114566.
13. H. Nguyen, X. N. Bui, H. B. Bui, and D. T. Cuong, "Developing an XGBoost model to predict blast-induced peak particle velocity in an open-pit mine: a case study", Acta Geophysica, vol. 67, no. 2, **(2019)**, pp. 477-490.
14. S. Gruszczyński and W. Gruszczyński, "Supporting soil and land assessment with machine learning models using the Vis-NIR spectral response", Geoderma, vol. 405, **(2022)**, pp. 115451.
15. K. Gallagher and M. Sambridge, "Genetic algorithms: a powerful tool for large-scale nonlinear optimization problems", Computers & Geosciences, vol. 20, no. 7-8, **(1994)**, pp. 1229-1236.
16. S. Levasseur, Y. Malécot, M. Boulon, and E. Flavigny, "Soil parameter identification using a genetic algorithm", International journal for numerical and analytical methods in geomechanics, vol. 32, no. 2, **(2008)**, pp. 189-213.
17. A. Papon, Y. Riou, C. Dano and P. Y. Hicher, "Single-and multi-objective genetic algorithm optimization for identifying soil parameters", International Journal for Numerical and Analytical Methods in Geomechanics, vol. 36, no. 5, **(2012)**, pp. 597-618.
18. M. Pasayat and A. K. Patel, "Assessment of mine spoil genesis influencing restoration in chronosequence iron mine spoil using artificial neural network", International journal of recent scientific research, vol. 8, no. 7, **(2017)**, pp. 18120-18128.
19. J.M. Anderson and J.S.I. Ingram, "Tropical soil biology and fertility". A Handbook of methods 2nd (Eds) C.A.B. International, **(1992)**, pp. 221.
20. R. R. Mishra, "Ecology work book". Oxford and I.B.H. Publication Co., New Delhi, **(1968).**
21. M.L. Jackson, "Soil chemical analysis", Prentice Hall of India Ltd. New Delhi, **(1958)**, pp. 498-503.
22. S.R. Olsen and L.E. Sommers, "Phosphorous. In: Methods of Soil Analysis, Miller, A.L., and Keeney, D.R., 2nd (Eds)", American Society of Agronomy Inc., Madison, **(1982)**, pp. 403-430.
23. E.D. Vance, P.C. Brookes and D.S. Jenkinson "An extraction method for measuring soil biomass carbon C", Soil Biology and Biochemistry, vol. 19, **(1987)**, pp. 703-707.
24. P.C. Brookes, J.F. Kragt, D.S. Powlson and D.S. Jenkinson, "Chloroform fumigation and release of soil N: A rapid direct extraction method to measure biomass N in soil", Soil Biochemistry, vol. 17, **(1985)**, pp. 837-842.
25. S.R. Olsen, C.V. Cole, F.S. Watanabe, and L.A. Dean, "Estimation of available phosphorous in soils by extraction with sodium bicarbonate", United States Department of Agriculture Cric No. 939, **(1954)**, pp. 746.
26. P.C. Brookes, D.S. Powlson, and D.S. Jenkinson, "Measurement of microbial biomass phosphorous in soils", Soil Biology and Biochemistry, vol. 14, **(1982)**, pp. 319-321.
27. H. Ohya, S. Fujiwara, Y. Komai and M. Yamaguchi, "Microbial biomass and activity in urban soils contaminated with Zn and Pb", Biology and Fertility of Soils, vol. 6, **(1988)**, pp. 9-13.
28. M. R. Roberge, "Methodology of soil enzyme measurement and extraction", In: Soil Enzymes, (Eds) Burns, R.G., London, Academic Press, **(1978)**, pp. 341-369.
29. D. J. Ross, "Invertase and amylase activities as influenced by clay minerals, soil clay fractions and topsoil under grassland", Soil Biology and Biochemistry, vol. 15, **(1983)**, pp. 287-293.

30. J. N. Ladd and J. H. A. Butler, “Short term assay of soil proteolytic enzymes activities using proteins and dipeptide derivatives as substrates”, Soil Biology and Biochemistry, vol. 4, **(1972)**, pp. 19-30.
31. G. G. Hoffmann and K. Teicher, “Ein kolorimetrisches Verfahren zur Bestimmung der Urease-Aktivitat in Boden. Zeitschrift Fur Pflanzenernahrung Dungung Bodenkunde, vol. 91, **(1961)**, pp. 55-63.
32. M. A. Tabatabai and J. M. Bremner, “Use of p-nitrophenyl phosphate for assay of soil phosphatase activity”, Soil Biology and Biochemistry, vol. 1, **(1969)**, pp. 301-307.
33. P. Nannipieri, S. Greco, and B. Ceccanti, “Ecological significance of the biological activity in soil. In: Soil Biochemistry”, (Eds) Bollag, J.M., and Stotzky, G. Marcel Dekker Inc. NY., vol. 6, **(1990)**, pp. 293-355.
34. J. S. Buyer, J. R. Teasdale, D. P. Roberts, I. A. Zasada and J. E. Maul, “Factors affecting soil microbial community structure in tomato cropping systems”, Soil Biology and Biochemistry, vol. 42, **(2010)**, pp. 831-841.
35. J. M. Fraterrigo, T. C. Balser, and M. G. Turner, “Microbial community variation and its relationship with nitrogen mineralization in historically altered forests”, Ecology, vol. 87: **(2006)**, pp. 570-579.
36. M. Frąc, K. Oszust, and J. Lipiec, “Community level physiological profiles (CLPP), characterization and microbial activity of soil amended with dairy sewage sludge”, Sensors, vol. 12, no. 3, **(2012)**, pp. 3253-3268.
37. J. L. Garland, “Analysis and interpretation of community-level physiological profiles in microbial ecology”, FEMS microbiology ecology, vol. 24, no. 4, **(1997)**, pp. 289-300.
38. P. K. Naik, Sindhura, T. Singh, and H. Singh, “Quantitative structure - activity relationship (QSAR) for insecticides: development of predictive in vivo insecticide activity models” SAR and QSAR in Environmental Research, vol. 20, no. 5-6, **(2009)**, pp. 551-566.
39. S. Džeroski, J. Grbović, W. J. Walley and B. Kompare, “Using machine learning techniques in the construction of models. II. Data analysis with rule induction”, Ecological Modelling, vol. 95, no. 1, **(1997)**, pp. 95-111.
40. T. D'heygere, P. L. Goethals, and N. De Pauw, “Use of genetic algorithms to select input variables in decision tree models for the prediction of benthic macroinvertebrates”, Ecological Modelling, vol. 160, no. 3, **(2003)**, pp. 291-300.
41. T. M. Crow, C. A. Buerkle, D. E. Runcie, and K. M. Hufford, “Implications of genetic heterogeneity for plant translocation during ecological restoration”, Ecology and evolution, vol. 11, no. 3, **(2021)**, pp. 1100-1110.