World Academy of Science, Engineering and Technology
International Journal of Nutrition and Food Engineering
Vol:15, No:12, 2021

A Risk Assessment Tool for the Contamination of Aflatoxins on Dried Figs based on Machine Learning Algorithms

Kottaridi Klimentia, Demopoulos Vasilis, Sidiropoulos Anastasios, Ihara Diego, Nikolaidis Vasileios, Antonopoulos Dimitrios

***Abstract*—**Aflatoxins are highly poisonous and carcinogenic compounds produced by species of the genus Aspergillus spp. that can infect a variety of agricultural foods, including dried figs. Biological and environmental factors, such as population, pathogenicity and aflatoxinogenic capacity of the strains, topography, soil and climate parameters of the fig orchards are believed to have a strong effect on aflatoxin levels. Existing methods for aflatoxin detection and measurement, such as high-performance liquid chromatography (HPLC), and enzyme-linked immunosorbent assay (ELISA), can provide accurate results, but the procedures are usually time-consuming, sample-destructive and expensive. Predicting aflatoxin levels prior to crop harvest is useful for minimizing the health and financial impact of a contaminated crop. Consequently, there is interest in developing a tool that predicts aflatoxin levels based on topography and soil analysis data of fig orchards. This paper describes the development of a risk assessment tool for the contamination of aflatoxin on dried figs, based on the location and altitude of the fig orchards, the population of the fungus *Aspergillus* spp. in the soil, and soil parameters such as pH, saturation percentage (SP), electrical conductivity (EC), organic matter, particle size analysis (sand, silt, clay), concentration of the exchangeable cations (Ca, Mg, K, Na), extractable P and trace of elements (B, Fe, Mn, Zn and Cu), by employing machine learning methods. In particular, our proposed method integrates three machine learning techniques i.e., dimensionality reduction on the original dataset (Principal Component Analysis), metric learning (Mahalanobis Metric for Clustering) and K-nearest Neighbors learning algorithm (KNN), into an enhanced model, with mean performance equal to 85% by terms of the Pearson Correlation Coefficient (PCC) between observed and predicted values.

***Keywords*—**Aflatoxins, *Aspergillus* spp., dried figs, k-nearest neighbors, machine learning, prediction.

I. Introduction

AFLATOXINS are a group of mycotoxins produced mainly by species of *Aspergillus* spp. and are among the most potent mutagenic and carcinogenic substances formed in nature. The group includes four aflatoxins, B1, B2, G1 and G2 and although their degree of toxicity varies from organism to organism, the general order is B1 >G1 > B2 >G2 [1].

The aflatoxins detected in agricultural products are mainly derived from the fungi A. flavus and A. parasiticus which cause both field and post-harvest diseases. Both species produce large amounts of aflatoxins, especially under conditions of high humidity and temperature. Nuts (peanuts, almonds), dried figs, spices and cereals present a serious problem. Aflatoxins are even detected in the milk of animals fed on feed infected by the fungi [2], [3].

Human exposure to aflatoxins occurs mainly directly through the consumption of infected nuts, cereals and other agricultural products but also indirectly through the consumption of animal products from animals fed with feed, contaminated with aflatoxins. However, it has been found that both humans and animals can be exposed to aflatoxins in other ways, such as by inhaling spores or textures of *Aspergillus* spp., or powder from infected cereals, or even breast milk [4]. Prolonged human exposure to aflatoxins can lead to aflatoxinosis, a pathological condition that manifests itself in a wide variety of symptoms such as: confusion, chronic fatigue, difficulty swallowing, choking, high blood pressure, heavy sweating during sleep, headaches, diarrhea, heart arrhythmia, pain in the liver, spleen or kidneys, vision problems, asthma, thyroid problems, allergies, suppression of the immune system, dementia and various cancers especially in the kidneys and liver. Aflatoxin B1 is even considered to be the most potent known liver carcinogen. In animals, there is immunodeficiency, toxicity to the liver and kidneys, decreased weight gain, milk production, egg production and reproduction, increased susceptibility to disease, subcutaneous bleeding and eventual death. This wide variety of symptoms makes the diagnosis of aflatoxinosis very difficult [5].

The aflatoxin biosynthesis is affected by various biotic and abiotic factors. Several biological factors including cultivar, soil type, viable fungal species in the soil, and plant metabolites have been documented to influence the aflatoxin formation [6]. The survival of aflatoxins is greatly affected by the type of soil. Sandy loam soil led to rapid decomposition and shorter persistence of aflatoxins than silt loam and silty clay loam soils [6], [7]. Heavier soil with a high water-holding potential declined the level of aflatoxin contamination. While light and sandy soil promoted the growth of A. flavus and thereby increased the likelihood of aflatoxin contamination [6], [8].

Aflatoxin biosynthesis is also affected by certain physiological attributes including the culture pH, developmental stage of crop and oxidative stress. Fungal aflatoxin production increased by almost 5-10 times at the pH levels of 4 or 5 than pH 8 [6], [9]. Low pH led to activation of

Kottaridi Klimentia is with the University of the Peloponnese (corresponding author, e-mail: k.kottaridi@go.uop.gr).

Demopoulos Vasilis, Antonopoulos Dimitrios and Antonopoulos Dimitrios are with the University of the Peloponnese

Sidiropoulos Anastasios and Ihara Diego are with the University of Illinois at Chicago

aflatoxin-producing genes therefore acidic medium favored the aflatoxin biosynthesis by A. flavus [6], [10]. Aflatoxigenic fungi can grow during the growing, ripening and drying periods on fig fruits, but they thrive especially through the different ripening phases. Toxigenic fungi can be both in the inner cavity and on the surface of the dried figs. Ripening phase of fruits on the tree is the critical period for aflatoxin formation of dried figs [11], [12]. In addition to that, storage is also another phase for mycotoxin production [12]. Environmental conditions that occur during the processing of dried figs and storage before they reach the market, where temperatures are rarely controlled, also support A. flavus growth and development [13].

Nutritional sources including amino acids, carbon, nitrogen, lipids and trace elements have been documented to affect the aflatoxin biosynthesis [6], [14]. Moreover, environmental attributes like topography, climate, weather, temperature, drought, rainfall, and water activity have a great impact on aflatoxin production. In addition to the impact of individual weather elements, studies have also revealed the combined effect of various weather determinants on aflatoxin biosynthesis [6]. Finally, inappropriate agricultural systems, such as sowing time, tillage, crop rotation, irrigation, and application of fertilizers also contributed to A. flavus infestation followed by aflatoxin production [8].

The cultivation of figs for the production of dried figs is an important, usually complementary, income for the producers of the southern Peloponnese. Much of the total production is exported to EU countries, North America and the Middle East. The content of aflatoxins in dried figs is an important quality characteristic for the importing countries, especially in Europe and America. For that reason, the Producers Group cooperates with private laboratories and spends high amounts annually on quality controls of dried figs, both those intended for the domestic market and, above all, those intended for export, in order to check their aflatoxin content.

Given the scope and complexity of the problem, the development of predictive tools that can assist in managing aflatoxin as well as in early detection and appropriate handling of risk prone crops, becomes really important. This research aimed at predicting the risk factor for the presence of aflatoxins in dried figs, based on the location and altitude of the fig orchards, the population of *Aspergillus* spp. and the soil parameters (soil characteristics and nutrients). A machine learning model was developed based on a dataset collected from 45 fig orchards located in the Southern Peloponnese (namely 36 from Messenia and nine from Laconia).

II. Literature Review

We here present a review of studies that utilize machine learning regarding the prediction of aflatoxin contamination on various agricultural products.

Reference [15] proposed the development of a genetic algorithm/neural network hybrid in which a genetic algorithm was used to find weight assignments for a neural network that predicted aflatoxin contamination levels in peanuts based on environmental data. Environmental values included length of drought stress period (days), mean soil temperature (°C), crop age (days), and accumulated heat units (°C days). The results showed the genetic algorithm network to be as effective as simple back propagation for training networks to predict aflatoxin contamination levels in peanuts.

Reference [16] proposed a machine learning method for detecting aflatoxin-contaminated figs using multispectral pictures captured under UV light. The figs were classified using two distinct methodologies, with error rates of 9.38% and 11.98%, respectively, based on their surface mold and aflatoxin levels. For the evaluation of the classification performance, a 2 × 2-fold cross-validation, which equally distributed the samples into test and training sets, was used. The classification was performed by incrementally adding the ranked features on SVC (Support Vector Classifier) and ANN (Artificial Neural Network) and considering as labels either the aflatoxin level or the surface mold concentration.

Reference [17] conducted a study in order to analyze how ozonation (oxidation method for the detoxification of aflatoxins in foods) reduced the aflatoxin content of cotton seed. They designed an advanced neural network model for the prediction of aflatoxin in cotton seed. The performance of the prediction was assessed by measuring MSE (mean square error), NMSE (normalized mean square error), MAE (mean absolute error), and SMAPE (mean absolute percentage error). Comparison results between predicted and actual values showed that the proposed model could accurately predict the amount of aflatoxin. According to the results, increases in both the time of ozonation and storage of ozonated samples induced increases in the reduction rate of aflatoxin B1.

Reference [18] proposed a method for the optical detection of aflatoxins B (B1 and B2) in grained almonds using fluorescence spectroscopy and machine learning algorithms. Contamination of aflatoxins B, on the almonds used in the experiments, were in the range of 2.7-320.2 ng/g. Following pre-processing steps, a binary classification model based on SVM (Support Vector Machine) algorithm was used to apply spectral analysis. The classification findings were then subjected to a majority vote procedure. With a threshold set at 6.4 ng/g, the best result was a classification accuracy of 94% (with a false negative rate of 5%).

III. Theoretical Background

This section reviews basic concepts like k-nearest neighbors (KNN), Principal Component Analysis (PCA), Repeated k-fold Cross Validation and Mahalanobis Metric for Clustering (MMC).

A. K-Nearest Neighbors (KNN)

KNN is a lazy learning and non-parametric algorithm, which is very useful for nonlinear data because there is no assumption about the data in this algorithm. It is a simple, easy-to-implement supervised machine learning algorithm that may be wont to solve both classification and regression problems [19].

The observations are presented in a d- dimensional space, where d is the number of attributes or characteristics which the observation has [20]. Given a new point, it is classified according to its similarity to the rest of the data points as

determined by some similarity measure employed by the model. The k-nearest neighbor regressor explores the pattern space for the k training samples that are closest to the unknown sample and delivers a real-valued prediction when given an unknown sample. "*Closeness"* is a way to determine similarity and is measured by a distance function [21].

The distance function and the value of k are the only two parameters necessary to implement KNN. In practice, there is no best solution for choosing k; it depends on the problem in hand. If k is chosen too large, the algorithm may erroneously include known data that is too distant from the unknown sample as its nearest neighbors. If k is too small, the algorithm is prone to over fit the data because of the noise in the training data. This will affect the generalization ability [20]. The optimal size of k is the one that minimizes the classification or prediction error. The most common distance function that is used to measure similarity is the Euclidian distance and it is defined by:

$$d_{Euclidean}(x,y) = \sqrt{\sum_i (x_i - y_{i)}{}^2} \quad (1)$$

The shorter the distance between x and y the more similar x and y are.

B. Mahalanobis Metric for Clustering (MMC)

The performance of KNN algorithm depends crucially on the way that distances are computed between different examples. When no prior knowledge is available, most implementations of KNN compute simple Euclidean distances. While Euclidean distance is useful in low dimensions, it doesn't work well in high dimensions since it ignores the similarity between attributes and treats each attribute as totally different from all the others.

A number of researchers have demonstrated that the performance of KNN can be greatly enhanced by learning an appropriate distance from the data [22]. This is the so-called problem of distance metric learning. Distance metric learning aims at automatically constructing task-specific distance metrics from data. A key advantage of metric learning is that it can be applied beyond the standard supervised learning setting (data points associated with labels), in situations where only weaker forms of supervision are available (e.g., pairs of points that should be similar/dissimilar) [23].

Mahalanobis Metric for Clustering (MMC) is a weakly supervised metric learning algorithm that goals to minimize the distances between similar points while maximizing the distances between dissimilar ones. MMC learns a distance, by taking as input tuples of points and labels for theses tuples, indicating similarity or dissimilarity between them. The goal is to transform points in a new space, in which the tuple-wise constraints between points are respected [23]. Learning a Mahalanobis Metric is equivalent to learning a linear transformation function and computing the Euclidean distance over the transformed data domain [24].

C. Principal Component Analysis (PCA)

Principal Component Analysis is probably the oldest and certainly the most popular technique for computing lower-dimensional representations of multivariate data [25]. PCA consists of finding two linear transformations, one that compresses the data to a smaller space, and another that decompresses them in the original space, so that in the process of compression and decompression the minimum information is lost [26]. So, PCA is a tool to reduce feature vector to lower dimension while retaining most of the information.

Principal components are new variables that are created by combining the starting variables in such a way that the new variables are uncorrelated and the first components include the majority of the information from the beginning variables. By rejecting the components with little information, PCA aims to put the most possible information in the first component, then the most remaining information in the second, and so on, allowing dimensionality reduction without losing much information. From a geometrical point of view, principal components represent the directions of the data that explain a maximal amount of variance.

The steps required to perform PCA are: standardize the dataset, create a covariance matrix from the standardized data, calculate principal components (eigenvectors) and their corresponding eigenvalues, sort the components by their respective eigenvalues, plot the graph, and finally select top n features that explain most variance in the data. The eigenvectors are actually the directions of the axes where there is the most variance and that we call Principal Components while the eigenvalues are simply the coefficients associated with eigenvectors, which give the amount of variance transferred to each Principal Component.

Integrating PCA with KNN can not only reduce the data dimensionality to speed up the calculation of KNN, but also reduce redundancy information while remaining effective information, and improve the performance of KNN prediction [27]. Low noise sensitivity, reduced capacity and memory requirements, and enhanced efficiency due to processes taking place in smaller dimensions are all advantages of PCA [28].

D. Repeated k-fold Cross-Validation

The objective of k-fold cross validation is to estimate the performance of the machine learning model on a test set: data not used to train the model. The main idea behind cross – validation is that each sample in the dataset has the opportunity of being tested. This is achieved by iterating over the dataset k times, splitting the dataset into k parts every time. One part is used for validation and the remaining k-1 parts are merged into a training subset. Cross-validation performance is computed as the arithmetic mean over the k performance estimates from the validation tests [29].

The k-fold cross-validation approach may produce an unrepresentative estimate of model performance after just one run. Different data splits can produce quite different findings. Repeated k-fold cross-validation is a technique for improving a machine learning model's estimated performance, by simply repeating the cross-validation technique several times and returning the mean result across all folds from all runs.

IV. Materials and Methodology

A. Materials

For the study of the endemic population of *Aspergillus* spp., 45 fig orchards were selected in the fig growing areas of the Southern Peloponnese (namely 36 from Messenia and 9 from Laconia). The distribution was based on the quantity of dried figs produced during the two years 2010-12. Soil samples were taken from the fig orchards to determine the population of *Aspergillus* spp. in the summer (July - August) of 2013.

Six soil samples (replicates) were taken from each fig orchard, inoculated into the selective Modified Rose Bengal Chloramphenicol Agar (MRBCA) with dichloran and streptomycin, incubated at 28°C for 6 days and developed *Aspergillus* spp. colonies were counted [30]. Assuming that each colony originated from conidia or a part of the mycelium of the fungus, the population of *Aspergillus* spp. was computed and expressed as the "number of infectious units" (Colony Forming Unit, CFU) per g of soil. For each fig orchard, the average population of the 6 soil samples was calculated.

The results showed a strong spatial unequal distribution of the endemic population of *Aspergillus* spp. and therefore correspondingly different levels of risk of infection of figs by the fungus. The population of each fig orchard was classified based on $CFU \cdot g^{-1}$ as: very high (>250 $CFU \cdot g^{-1}$), high (101-250 $CFU \cdot g^{-1}$), moderate (51-100 $CFU \cdot g^{-1}$), low (11-50 $CFU \cdot g^{-1}$) or minimum (0-10 $CFU \cdot g^{-1}$).

High and very high population of *Aspergillus* spp. was found in 24% of the sycamores studied, all in the region of Messenia where they represent 31%. One third (1/3) of the fig orchards appeared to have a minimal population of *Aspergillus* spp. and in fact in three of them (two in Laconia and one in Messenia) the presence of the fungus in the soil was not detected. Minimum population of *Aspergillus* spp. was detected in 18% of the fig groves of Messenia and in 78% of those of Laconia.

For the in vitro study of the aflatoxinogenic capacity of the endemic population of *Aspergillus* spp., 209 isolates were inoculated on Yeast Extract with Supplements (YES) medium and incubated at 30°C for 8 days. This nutrient was rich in sugars in proportion to dried figs and promoted the production of aflatoxins. The composition of the nutrient was (concentration in $g^{-1} \cdot l^{-1}$): sucrose 150.00, yeast extract 20.00, agar 20.00. Five of 0.8 cm diameter disks were cut from the culture material and placed for 15 min in 5 ml of methanol, stirred for 30 sec and the suspension was infiltrated through a filter with a pore diameter of 0.45 μm. The determination of the total aflatoxins was done by the ELISA method and the use of a kit for the determination of the total aflatoxins (Agra Quant Aflatoxin 1-20 ppb Romer Labs) [31].

The aflatoxin values determined ranged from 0.0 to 113.1 $ng \cdot \mu l^{-1}$. Many countries have taken strict regulations and measures to control the level of aflatoxin contamination. In Turkey and USA, the generally accepted level of aflatoxin in food is 20 $\mu g \cdot kg^{-1}$. In the European countries, the maximum level of total aflatoxin is determined as 10 $\mu g \cdot kg^{-1}$ and the maximum level of aflatoxin B1 is determined as 5 $\mu g \cdot kg^{-1}$ [32]. Overall, the endemic strains of *Aspergillus* spp. produce mainly aflatoxin B1 (71.6%) which is the most toxic, to a large extent G1 (28%) which is the next in degree of toxicity and the other two less toxic aflatoxins B2 and G2 in negligible percentages.

Extracts from the culture medium of 44 strains of *Aspergillus* spp. were also analyzed for the ratio of the four aflatoxins they produce using chromatographic method (HPLC) [33]. Of the 44 strains studied, 15 produced one aflatoxin (9 B1 and 6 G1), 24 produced two (23 B1-G1 and 1 B1-B2), 5 produced three (all B1-B2-G1) and 4 produce all four aflatoxins (B1-G1-B2-G2). The strains of the *Aspergillus* spp. were categorized according to their aflatoxinogenic power as: very high (>100 $ng \cdot \mu l^{-1}$), high (20-100 $ng \cdot \mu l^{-1}$), medium (10-20 $ng \cdot \mu l^{-1}$), low (1-10 $ng \cdot \mu l^{-1}$), and not aflatoxinogenic (<1 $ng \cdot \mu l^{-1}$).

Based on the aflatoxinogenic capacity of the strains and the population of *Aspergillus* spp., the risk factor for the presence of aflatoxins in dried figs was calculated for each of the 45 fig orchards separately. The risk factor was calculated from the multiplication of the population of *Aspergillus* spp. per g of soil (CFU) with the average aflatoxinogenic capacity of the strains studied for each orchard separately. The fig orchards were categorized according to their risk factor for the presence of aflatoxin as: very high (>400 $CFU \cdot g^{-1} \cdot ng \cdot \mu l^{-1}$), high (100-400 $CFU \cdot g^{-1} \cdot ng \cdot \mu l^{-1}$), medium (25-100 $CFU \cdot g^{-1} \cdot ng \cdot \mu l^{-1}$), low (5-25 $CFU \cdot g^{-1} \cdot ng \cdot \mu l^{-1}$), and minimum (<5 $CFU \cdot g^{-1} \cdot ng \cdot \mu l^{-1}$).

High and very high risk factor for the presence of aflatoxins in dried figs was found in 60% of the fig orchards, while 27% of the fig orchards presented a small to medium risk. A percentage of 13% of fig orchards were found to have no risk for the presence of aflatoxins.

In the context of investigating the factors that affect the distribution of the endemic population of *Aspergillus* spp. in the orchards of the Southern Peloponnese, the effect of altitude and soil characteristics was studied. The soil samples collected from the orchards were dried using forced air at ambient temperatures <36°C to constant weight and then passed through a 2 mm sieve [34]. Saturation percentage (SP) was determined after the saturation of soil samples [35], [36]. Soil pH was measured in a soil/water slurry [37] and soil salinity (Electrical Conductivity, EC) was measured in the saturation paste extract [38], using a Consort C 835 multi-channel analyzer. Particle size analysis (proportions of sand, silt and clay) was determined according to the hydrometer method [39]. Organic matter (OM) content of soil samples was determined with colorimetric titration according to the Walkley-Black method [40]. Exchangeable cations (Ca, Mg, K and Na) were extracted with a 1 M NH4OAc (ammonium acetate) solution at pH: 7.0 [41]. Micronutrients (Fe, Mn, Zn and Cu) were extracted using DTPA extractant method [42]. Concentrations of exchangeable cations and micronutrients were determined by atomic absorption spectrophotometry (Shimadzu AA6200) in an air–acetylene flame. For the Ca and Mg determinations, La2O3 was added to both the standard and the diluted samples to achieve a concentration of 4,500 mg l-1La [43]. Extractable Phosphorus (P) was determined colorimetrically according the sodium bicarbonate soil test [44]. Boron (B) was extracted using DTPA solution and determined colorimetrically according to azomethine-H method [45], [46]. For both elements, P and B, a

Shimadzu UV-1700 spectrophotometer was used.

The purpose of the present research is to predict the risk factor for the presence of aflatoxins in dried figs, based on the location and altitude of the fig orchards, the population of the fungus *Aspergillus* spp. and the soil parameters (soil characteristics and nutrients). The target variable ("*Risk Factor*") was calculated from the multiplication of the population of *Aspergillus* spp. per g of soil (CFU) with the average aflatoxinogenic capacity of the strains studied for each orchard separately, and was normalized to 0-1 scale. The predictor variables from the available dataset that were taken into account for the prediction of the target variable were: location (1=Messenia, 2= Laconia), altitude, CFU, SP, pH, EC, Ca, Mg, K, Na, P, B, O.M, Sand, Silt and Clay, Fe, Mn, Zn, Cu. The collected dataset contained 45 instances and 21 numeric attributes.

B. Methodology

i. Proposed Model

Our proposed method aims to enhance the performance of KNN algorithm used for the prediction of the aflatoxin risk factor values, by constructing an integrated PCA-MMC-KNN model. Based on the K-Nearest Neighbor (KNN) regression model, a Principal Component Analysis (PCA) was initially applied to the dataset in order to reduce redundancy information and data dimensionality. Performance of KNN was further improved by learning a Mahalanobis distance metric from the data (MMC), i.e., learning a linear transformation of the input space that precedes KNN using Euclidean distances. Repeated 5-fold cross validation was used to estimate the performance of the training algorithm.

The detailed flow of our PCA-MMC-KNN algorithm is presented in Table I.

TABLE I
STEPS OF OUR MODEL BUILDING PROCESS

Step No.	Steps
1	Input Dataset
2	Apply PCA for dimensionality reduction
3	Split Dataset into 5 parts: 4 parts (80%) for training and 1 part (20%) for testing
4	Standardize Train and Test data
5	Learn a distance metric from Train data using MMC metric learning algorithm and transform Train and Test data points into the learned linear space
6	Fit KNN model to the transformed Train Data
7	Repeat Steps 3-6 for 5 times, until each part is used exactly once as the testing data
8	Compute average prediction performance of the 5 runs
9	Repeat Steps 3-8 for 10 times
10	Compute overall average prediction performance (across 5 folds from 10 runs)

In *Step No. 2* of Table I, Principal Component Analysis (PCA) method is used in order to reduce the dimensionality of the dataset, while preserving as much information as possible. The steps of the PCA process are described in Table II.

TABLE II
APPLY PCA FOR DIMENSIONALITY REDUCTION

Step No.	Steps
1	Compute the covariance d x d symmetric matrix (where d=21 the total number of feature variables) to summarize the correlations between all the possible pairs of variables
2	Compute the eigenvectors and eigenvalues of the covariance matrix
3	Order the eigenvectors by their eigenvalues in descending order
4	Discard eigenvectors with low eigenvalues and form with the remaining ones a symmetric matrix of vectors p x p (feature vector), where p is the number of eigenvectors (components) we decided to keep.
5	Reorient the data from the original axes to the ones represented by the principal components, by multiplying the transpose of the original data set by the transpose of the feature vector: $ReorientedDataset = FeatureVector^T * OriginalDataset^T$
6	Return $ReorientedDataset$

In *Step No. 5* of Table I, the metric learned by our model, puts points, whose absolute difference between target values is below a predefined lower bound (0.1), closer together in the transformed space, and points, whose absolute difference between target values is above a predefined upper bound (0.2), further away from each other.

The detailed flow for *Step No. 5* of Table I to "*learn a distance metric from Train data using MMC metric learning algorithm*", is presented in Table III.

TABLE III
LEARN A DISTANCE METRIC FROM TRAIN DATA USING MMC METRIC LEARNING ALGORITHM

Step No.	Steps
1	Input Train and Test Sets
2	Initiate a *lower* and *upper bound* (lower=0.1 and upper=0.2)
3	Generate a list from all $pairs$ of indices from the tuples in the Train Set
4	For each $pair$ of indices in the list do:
5	Compute the absolute difference between target values (risk factor values for aflatoxin) i.e., given that y_train is the vector containing the testing target values, compute the absolute difference: $y\_train[pair[0]] - y\_train[pair[1]]$
6	If the absolute difference between the target values is lower than the *lower bound* (0.1), mark the corresponding tuples in the Train Set as similar
7	Else if the absolute difference between the target values is higher than the *upper bound* (0.2), mark the corresponding tuples in the Train Set as dissimilar
8	End For
9	Give the original dataset of points to the estimator so that it knows the points the indices refer to.
10	Fit the metric learning algorithm MMC with this type of input
11	Transform Train and Test data points into the learned linear space
12	Return Transformed Train and Test data points

In *Step No. 6* of the proposed PCA-MMC-KNN algorithm, described in Table I, the KNN model is fitted to the transformed Train Data. The detailed flow for *Step No. 6* of Table I is presented in Table IV.

TABLE IV
FIT KNN MODEL TO THE TRANSFORMED TRAIN DATA

Step No.	Steps
1	Input Train and Test Sets
2	Calculate the Euclidean distance between the first point in Test Set and each point in Train Set
3	Select the k closest training points, based on the distances calculated in Step 2
4	Predict target value for the testing point by taking the average of the target values of the selected k training points
5	Repeat Steps 2-4 for all data points in the Test set
6	Calculate the prediction performance for the testing points

ii. Performance Metric

In order to assess the goodness of fit of our model, we have used the Pearson Correlation Coefficient (PCC) between predicted and observed values, which is a commonly used metric for regression problems.

PCC is a metric for determining the linear relationship between two random variables, X and Y. If σ_{XY} is the covariance between X and Y, and σ_X and σ_Y are the standard deviations of X and Y, respectively, the Pearson's Correlation Coefficient ρ_{XY} is calculated as follows [47]:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X * \sigma_Y} \quad (2)$$

The value of ρ ranges between -1 and +1 because the covariance is always smaller than the product of the individual standard deviations.

PCC quantifies the degree and direction to which two variables are related and tells how much one variable tends to change when the other one does. When it is positive, there is a trend that one variable goes up as the other one goes up, and when it is negative, there is a trend that one variable goes up as the other one goes down.

In our model, PCC is computed by *corrcoef()* function implemented by Python Numpy library and measures the degree of linear relationship between the predicted and the true values of the risk factor for aflatoxin. A high positive PCC indicates that the higher the predicted values are the higher are the true risk factor values for aflatoxin and the opposite.

iii. Hyperparameter Tuning

A variety of options must be investigated in order to develop an optimal ML model. Hyper-parameter tuning is the process of creating the ideal model architecture with the best hyper-parameter configuration. Because hyperparameters form the architecture of an ML model, they cannot be directly predicted from data learning and must be defined before training an ML model [48]. Hyperparameters can be set manually to optimize the performances of the different machine learning algorithms.

There are no fixed universal best settings of hyperparameters of a machine learning algorithm, so experiments over a set of possible values are usually needed in order to choose the best settings according to the dataset. Determining the optimal (hyperparameter) settings for a machine learning model is crucial for the bias-reduced assessment of the model's predictive power [49].

In our enhanced PCA –MMC – KNN model, five-fold (20% training, 80% testing) partitioning repeated 10 times was chosen for performance estimation, resulting in a total number of 50 different dataset partitions, with an 80/20 training/test split in each fold. Hyperparameter tuning was performed for PCA and KNN algorithms and the hyperparameter setting with the highest mean PCC across all tuning folds was used to train the model on the training set.

As already mentioned, PCA can be thought of as a projection method where data with m-columns (features) is projected into a subspace with m or fewer columns (principal components), whilst retaining the essence of the original data. The hyperparameter tuned for the PCA algorithm is the number of principal components used to represent our high dimensional data in lower dimension.

The most important hyper-parameter in KNN is the number of considered nearest neighbors, k. The model will be under-fitting if k is too little; if k is too large, the model will be over-fitting and require high computational time. Furthermore, the distance metric and the power parameter of the Minkowski metric can both be tuned for slight improvements [48].

For the purpose of determining the highest mean performance of our model, we have manually set the number of PCA components in a range from 2 to 8 and the number of nearest neighbors in KNN from 1 to 34, this way identifying the combination of these two parameters that optimized our model's performance across all tuning folds. Concerning the distance metric in KNN, we have used the Euclidean distance between new data and training data.

Two other parameters defined in our model, were the lower and upper bounds, used to define similarity and dissimilarity between tuples of points from the training dataset, during the metric learning process. After repeatedly setting the values for these bounds and estimating each time the highest mean model performance, we have concluded that the optimal values are 0.1 for lower bound and 0.2 for upper bound.

V. RESULTS AND DISCUSSION

We compared the performance of the proposed PCA-MMC-KNN model to the conventional PCA-KNN and KNN models on the original dataset, tuning the k parameter (number of neighbors in KNN) from 1 to 34 and p parameter (number of Principal Components in PCA) from 2 to 8. The performance of the three models (KNN, PCA-KNN, PCA-MMC-KNN) was evaluated for all possible combinations of tuning parameters, using 5-fold cross validation (80% training, 20% testing), repeated 10 times. The overall model performance was determined by the mean Pearson Correlation Coefficient (PCC) over 5 folds across 10 runs.

For the KNN model, the highest mean PCC between the predicted and the true risk factor values for the presence of aflatoxin, was 0.52 and was achieved for k values equal to 1 and 2. For PCA-KNN model, the highest mean PCC achieved, was 0.71, for k value equal to 5 and number of Principal Components equal to 2. Consequently, performing dimensionality reduction to the original dataset with the PCA method before applying the conventional KNN model has

improved mean model performance by 13.65% (from 0.52 to 0.71).

Our proposed PCA-MMC-KNN Model achieved the best performance in terms of the PCC, which was 0.85 for k value equal to 6 and number of Principal Components equal to 4. In particular, our proposed enhanced KNN model (PCA-MMC-KNN model), where dimensionality reduction with the PCA method and distance metric learning with the MMC algorithm precede the KNN model, improved mean model performance by 11.97% (from 0.71 to 0.85) compared to the PCA-KNN Model and by 16.35% compared to the conventional KNN model.

Comparison of PCA-KNN and PCA-MMC-KNN models for different k and p values, in terms of the PCC performance metric, is presented in Fig. 1.

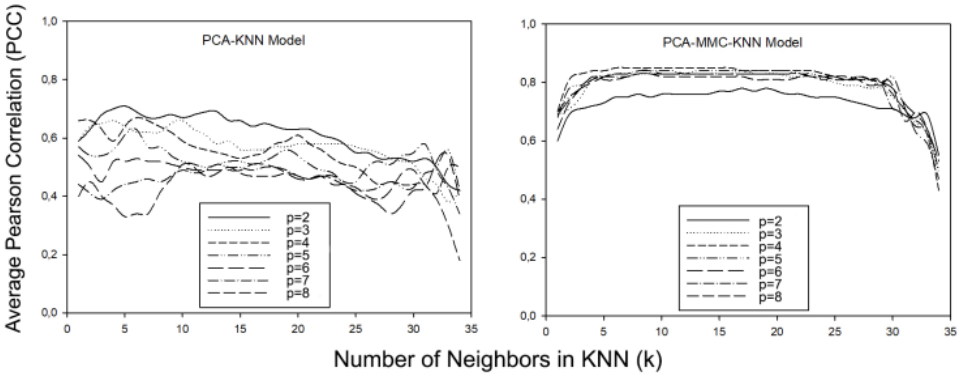

Fig. 1 Comparison of PCA-KNN and PCA-MMC-KNN models for different k and PCA values, in terms of the PCC performance metric

We have shown that performance of KNN model for regression can be greatly improved by applying two successive linear transformations to the data: a) reorienting the original data from the original axes to the ones represented by the principal components b) learning a linear transformation (distance metric learning) from the reoriented space that precedes KNN, which is finally applied to the transformed data using Euclidean distances.

The improvement of the mean prediction performance among the models is summarized in Table V and visualized in Fig. 2. The hyperparameters (k=number of nearest neighbors, p=number of Principal Components) which achieve the highest mean performance for each model are indicated inside the bars of Fig. 2.

TABLE V
PERFORMANCE COMPARATIVE RESULTS BY AVERAGING OVER 10 RUNS ON RANDOMLY GENERATED 80/20 SPLITS OF THE DATASET

Performance Metric	KNN	PCA- KNN	PCA-MMC-KNN
Pearson Correlation Coefficient (PCC)	0.52	0.71	0.85

Highest mean model performances were visually assessed by comparing scatter plots of estimated risk factor values for aflatoxin against the x-axis vs. observed values against the y-axis. Reference [50] observed that for graphical evaluation of model performance, a plot of observed values versus predicted is preferred to predicted versus observed.

Predictions of target values were produced across 5 folds (80% training, 20% testing) from 10 runs. Every time the data (sample size=45) was split, one part (validation sample size =9) was used for validation and the remaining 4 parts (training sample size=36) were merged into a training set. Over a total of 10 iterations of 5 data splits i.e. 50 train/test data splits, the total number of predicted values was 450 (50 iterations x 9 validation sample size).

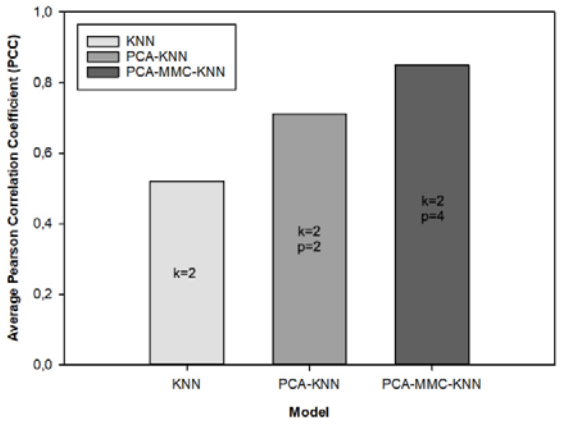

Fig. 2 Performance comparative results among the three models under review

Scatter plots in Fig. 3 depict a total of 450 predicted values against 450 observed values, in terms of KNN, PCA-KNN and PCA-MMC-KNN models. Since each sample in the dataset had the opportunity of being tested once in a 5-fold dataset split, over a total of 10 iterations of 5-fold splits, each target point was included in the testing set 10 times, and consequently appears in the scatter plot 10 times. For each model, hyperparameters (k nearest neighbors, number of Principal Components) having achieved the highest mean performance for each model were chosen.

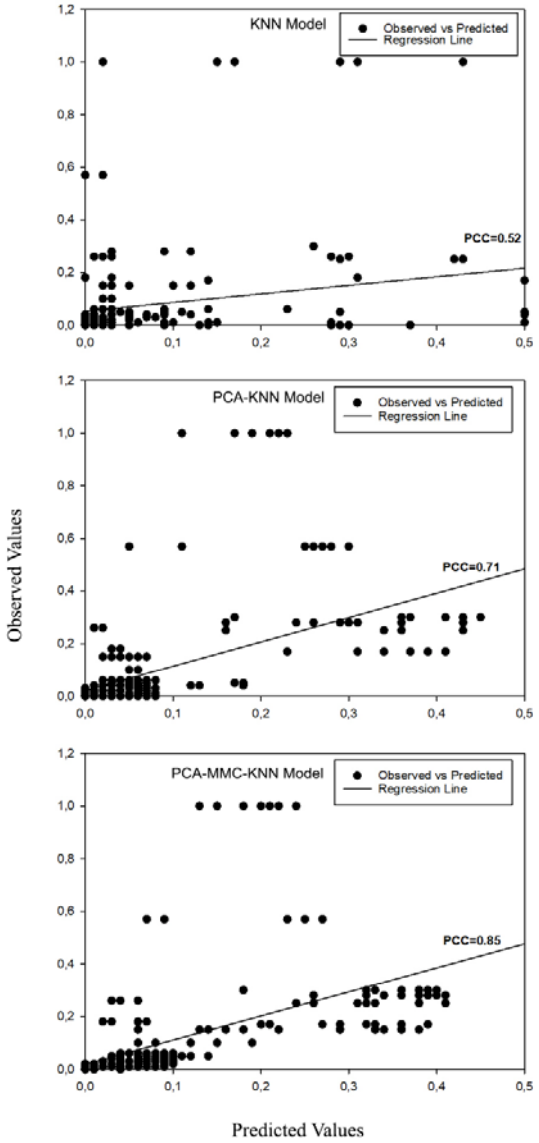

Fig. 3 Scatter plots and linear regression for *"observed against predicted"* risk coefficient values for the presence of aflatoxin in dried figs, in terms of the three models under review

As shown in Fig. 3, the correlation between the observed and predicted values is much stronger (PCC=0.85) for our enhanced PCA-MMC-KNN model, compared to the other two models under review, PCA-KNN (PCC=0.71) and conventional KNN (PCC=0.52).

VI. Conclusions

Aflatoxin contamination in dried figs and agricultural products in general has been a serious and long-standing problem around the world, since it is associated with severe health hazards in humans and animals, especially immunosuppression and cancer. Conventional optical methods for detection and quantification, such as high performance liquid chromatography (HPLC) are accurate and widely accepted, however they are toilsome, time consuming, require well-trained personnel and are costly because they need a significant investment in consumables, equipment and maintenance.

This study was aimed at developing a risk assessment tool for the prediction of aflatoxin contamination in dried figs, based on the location and altitude of the fig orchards, the population of the fungus *Aspergillus* spp. and the soil parameters, by employing machine learning methods. Our proposed model performed very satisfactorily, with respect to PCC (Pearson Correlation Coefficient) between predicted aflatoxin and actual values, by combining three machine learning techniques, i.e. dimensionality reduction with PCA (Principal Component Analysis), metric learning with MMC (Mahalanobis Metric for Clustering) and K-nearest neighbors learning algorithm (KNN). The correlation between predicted and observed value, expressed by PCC, proved to be strong and equal to 0.85.

Few highly inaccurate predictions of our proposed model are the result of the relatively small number of validation data points in specific classes. Future efforts will focus on obtaining additional observations as well as including as predictors even more features, such as environmental attributes (like topography, climate, weather, temperature, drought, rainfall, and water activity) and storage conditions.

VII. CRediT Authorship Contribution Statement

*Vasilis Demopoulos***:** Supervision, Project administration, Resources, Writing – review and editing. *Anastasios Sidiropoulos***:** Supervision, Conceptualization, Methodology, Validation. *Klimentia Kottaridi***:** Formal analysis, Validation, Writing – original draft, Writing – review and editing, Software, Data curation, Visualization. *Diego Ihara***:** Software, Formal analysis. *Vasileios Nikolaidis*: Conceptualization, Writing – review and editing. *Dimitrios Antonopoulos***:** Resources, Investigation.

VIII. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

We are thankful for the provision of the dataset, which was part of the research project 'Strategy of managing the aflatoxin problem in dried figs within the framework of the integrated management of fig tree cultivation', co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program 'Education and Lifelong Learning' of the National Strategic Reference Framework (NSRF) - Research Funding Program: ARCHIMEDES III. Investing in knowledge society through the European Social Fund (2013 – 2015).

References

[1] M. Moss, "Mycotoxicology: Introduction to the mycology, plant pathology, chemistry, toxicology, and pathology of naturally occurring mycotoxicoses in animals and man," in *British Veterinary Journal*, vol. 144, no. 104, 1988.

[2] P. J., Cotty, "Influence of field application of an atoxigenic strain of Aspergillus flavus on the populations of A. flavus infecting cotton bolls and on the aflatoxin content of cottonseed," in *Phytopathology*, vol. 84, no. 11, 1994, pp. 1270-1277.

[3] P. J., Cotty, "Comparison of four media for the isolation of Aspergillus flavus group fungi," in *Mycopathologia*, vol. 125, no. 3, 1994, pp. 157–162.

[4] J. E., Smith, *Handbook of Plant and Fungal Toxicants*, D'Mello,. J. P. F. (Ed.). CRC Press, Boca Raton, FL, 1997, pp. 269-285.

[5] J., Yu, D., Bhatnagar, and K. C., Ehrlich, "Aflatoxin biosynthesis," in *Revista iberoamericana de micologia*, vol. 19, no.4, 2002, pp. 191–200.

[6] M., Iqbal, M., Abbas, M., Adil, A., Nazir, and I., Ahmad, "Aflatoxins Biosynthesis, Toxicity and Intervention Strategies: A Review," in *Chemistry International*, vol. 5, no. 3, 2019, pp. 168-191. [Online]. Available: https://ssrn.com/abstract=3407341

[7] J. S., Angle, "Aflatoxin decomposition in various soils," *Journal of Environmental Science and Health,* Part B, vol. 21, no. 4, pp. 277-288, 1986.

[8] A. M., Torres, G. G., Barros, S. A., Palacios, S. N., Chulze, P., Battilani, "Review on pre- and post-harvest management of peanuts to minimize aflatoxin contamination," in *Food Research International* vol. 62, 2014, pp. 11-19.

[9] N. P., Keller, C., Nesbitt, B., Sarr, T. D., Phillips, and G. B., Burow, "pH regulation of sterigmatocystin and aflatoxin biosynthesis in A. Spp." in *Phytopathology,* vol. 87, no. 6, 1997, pp. 643-648.

[10] A. G., Marroquín-Cardona, N. M., Johnson, T. D., Phillips, and A. W., Hayes, "Mycotoxins in a changing global environment–a review," in *Food and Chemical Toxicology*, vol. 69, 2014, pp. 220-230.

[11] E. M., Embaby, L. F., Hagagg, and M. M., Abdel-Galil, "Decay of Some Fresh and Dry Fruit Quality Contaminated by Some Mold Fungi," in *Journal of Applied Sciences Research*, vol. 8, no. 6, 2012, pp. 3083-3091.

[12] D., Heperkan, A., Moretti, C. D., Dikmen, and A. F., Logrieco, "Toxigenic Fungi and Mycotoxin Associated with Figs in the Mediterranean Area," in *Phytopathologia Mediterranea*, vol. 51, no. 1, 2012, pp. 119-130.

[13] A. I., Galván, A., Rodríguez, A., Martín, M., Serradilla, A., Martínez-Dorado, M., Córdoba, "Effect of Temperature During Drying and Storage of Dried Figs on Growth, Gene Expression and Aflatoxin Production," in *Toxins*, vol. 13, 2021.

[14] R. H., Luchese, and W. F., Harrigan, "Biosynthesis of aflatoxin—the role of nutritional factors," *Journal of Applied Bacteriology*, vol. 74, no. 1, 1993, pp. 5-14.

[15] C., Henderson, W., Potter, R. W., McClendon, and G., Hoogenboom, "Aflatoxin Prediction Using a GA Trained Neural Network," in *FLAIRS Conference*, 1998.

[16] H., Kalkan, A., Güneş, E., Durmuş, and A., Kuscu, "Non-invasive detection of aflatoxin-contaminated figs using fluorescence and multispectral imaging," in *Food additives and contaminants, Part A, Chemistry, analysis, control, exposure and risk assessment*, vol. 31, no. 8, 2014, pp. 1414–1421, [Online]. Available: https://doi.org/10.1080/19440049.2014.926398

[17] V. Z., Romina, and A., Mohamadi Sani, "Use of artificial intelligence Algorithms to predict reduction of Aflatoxin in Cotton seed meal treated with ozone," in *IJALS*,vol. 9, no. 3, 2016, pp. 326-330.

[18] F. R., Bertani, L., Businaro, L., Gambacorta, A., Mencattin, D., Brenda, D., Giuseppe, A., De Ninno, M., Solfrizzo, E., Martinelli, and A., Gerardino, "Optical detection of Aflatoxins B in grained almonds using fluorescence spectroscopy and machine learning algorithms," in *arXiv*, 2020. [Online]. Available: https://arxiv.org/abs/2003.04096

[19] A., Godiya, and Dr. A., Kothari, "Study of Different Disease in Potato and their Detection Technique Using Leaf Image," in *International Journal of Innovative Research in Technology*, vol. 6, no. 12, 2020, pp. 246 - 254.

[20] A., Mohamed, "Comparative Study of Four Supervised Machine Learning Techniques for Classification," in *International Journal of Applied Science and Technology*, vol. 7, no. 2, Jun 2017.

[21] H. J., Oh, F., Ozkaya, and R., LaRose, "How does online social networking enhance life satisfaction? The relationships among online supportive interaction, affect, perceived social support, sense of community, and life satisfaction," in *Computers in Human Behavior*, vol. 30, 2014, pp. 69-78.

[22] K., Weinberger, J., Blitzer, and L., Saul, "Distance Metric Learning for Large Margin Nearest Neighbor Classification," in *Journal of Machine Learning Research*, vol. 10, 2009, pp. 207-244.

[23] W., De Vazelhes, C. J., Carey, Y., Tang, N., Vauquier, and A., Bellet, "metric-learn:Metric Learning Algorithms in Python", in *Journal of Machine Learning Research*, vol. 20, 2020, pp. 1-6.

[24] B., Shi, and J., Liu, "Nonlinear Metric Learning for kNN and SVMs through Geometric Transformations," in *Neurocomputing*, vol. 318, 2018, pp. 18-29.

[25] Y., Qu, G., Ostrouchov, N., Samatova, and Al., Geist, "Principal Component Analysis for Dimension Reduction in Massive Distributed Data Sets," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, 2002.

[26] J., Suárez, S., García, and F., Herrera, "A tutorial on distance metric learning: Mathematical foundations, algorithms, experimental analysis, prospects and challenges," in *Neurocomputing*, vol. 425, 2021, pp. 300-322.

[27] L., Tang, H., Pan, and Y., Yao, "K-Nearest Neighbor Regression with Principal Component Analysis for Financial Time Series Prediction," in *Proc. Int. Conf. on Computing and Artificial Intelligence (ICCAI)*, 2018, pp. 127-131. [Online]. Available: https://doi.org/10.1145/3194452.3194467

[28] S., Karamizadeh, S., Abdullah, A., Manaf, M., Zamani, and A., Hooman, "An Overview of Principal Component Analysis," in *Journal of Signal and Information Processing*, vol. 4, 2013, pp. 173-175.

[29] S. Raschka, "Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning," *arXiv,* 2018. [Online]. Available: https://arxiv.org/abs/1811.12808

[30] W. I., Baggerman, "A modified Rose Bengal medium for the enumeration of yeasts and moulds from foods," in *European Journal of Applied Microbiology and Biotechnology*, vol. 12, 1981, pp. 242-247.

[31] D., Nilufer, and D., Boyacioglu, "Comparative Study of Three Different Methods for the Determination of Aflatoxins in Tahini," in *Journal of agricultural and food chemistry*, vol. 50, 2002, pp. 3375-9. [Online]. Available: https://doi.org/10.1021/jf020005a

[32] Ö. B., Özlüoymak, and E., Güzel, 2 "Prediction of aflatoxin contamination on dried fig (ficus carica) samples by spectral image analysis in comparison with laboratory results," *Fresenius Environmental Bulletin*, vol. 27, no.2, 2018, pp. 681-689.

[33] M., Namjoo, F., Salamat, N., Rajabli, R., Haji-Hoseeini, F., Niknejad, F., Kohsar, and H., Joshaghani, "Quantitative Determination of Aflatoxin by High Performance Liquid Chromatography in Wheat Silos in Golestan Province, North of Iran," *Iranian journal of public health*, vol. 45, 2016, pp. 905-910.

[34] K., Elk, and R. H., Gelderman, "Soil sample preparation," in *Recommended chemical soil test procedures for the North Central Region*, no. 221, W.C. Dahnke, Ed. North Dakota: Agric. Exp. Stn. Bull., 1988, pp. 2-4.

[35] M. G., Klages, "Reproducibility of saturation percentage of soils" in *Proc. of the Montana Academy of Sciences (USA)*, vol. 44, 1984, pp. 67-69.

[36] J. D., Rhoades, "Salinity: Electrical Conductivity and Total Dissolved Solids," in *Methods of Soil Analysis: Part 3 Chemical Methods,* D.L.Sparks, Ed. USA Madison: SSSA Book Series, 1996.

[37] Y. P., Kalra, "Determination of pH of soils by different methods: collaborative study," in *Journal of the Association Off. Analytical Chemistry International*, vol. 78, no. 2, 1995, pp. 310-321.

[38] H. H., Janzen, "Soluble salts," in *Soil Sampling and Methods of Analysis,* M.R. Carter, Ed. Boca Raton, FL: Lewis Publishers, 1993, pp. 161–166.

[39] R. O., Miller, J., Kotuby-Amacher, J. B., Rodriguez, "Western States Laboratory Proficiency Testing Program-Soil Plant and Analytical Methods," Ver. 4.10, 1998.

[40] A., Walkley, and I. A., Black, "An Examination of the Degtjareff Method for Determining Soil Organic Matter and a Proposed Modification of the Chromic Acid Titration Method," in *Soil Science*, vol. 37, 1934, pp. 29-38.

[41] L., van Reeuwijk, "Procedures for Soil Analysis (6th Edition)," ISRIC, FAO, Wageningen, 2002.

[42] W. L., Lindsay, and W. A., Norvell, "Development of a DTPA soil test for zinc, iron, manganese and copper," in *Soil Science Society of America Journal*, vol. 42, 1978, pp. 421-428.

[43] D., Warncke, and J. R., Brown, "Potassium and Other Basic Cations" in *Recommended Chemical Soil Test Procedures for the North Central Region*, J.R., Brown, Ed. Columbia: NCR Publication No. 221, Missouri Agricultural Experiment Station, 1998, pp. 31-33.

[44] S. R., Olsen, C. V., Cole, F. S., Watanabe, and L.A. Dean, "Estimation of available phosphorus in soils by extraction with sodium bicarbonate," Circular, vol. 939, Ed. Washington, DC: US Department of Agriculture, pp.19, 1954.

[45] S. K., Gupta, J.W.B., Stewart, "The extraction and determination of plant available boron in soil" in *Schweizerische landwirtschaftliche Forschung* vol.14, pp. 153-169, 1975.
[46] B., Wolf, "Improvements in the azomethine-H method for the determination of boron," in *Communications in Soil Science and Plant Analysis*, vol. 5, 1974, pp. 39-44.
[47] A. G., Asuero, A., Sayago, and A., González, "The Correlation Coefficient: An Overview," in *Critical Reviews in Analytical Chemistry*, vol. 36, 2006, pp.41 - 59.
[48] L., Yang, and A., Shami, "On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice," in *arXiv*, 2020. [Online]. Available: https://arxiv.org/abs/2007.15745
[49] P., Schratz, J., Muenchow, E., Iturritxa, J., Richter, and A., Brenning, "Performance evaluation and hyperparameter tuning of statistical and machine-learning models using spatial data," in *arXiv*, 2018. [Online]. Available: https://arxiv.org/abs/1803.11266
[50] G., Piñeiro, S., Perelman, J., Guerschman, and J., Paruelo, "How to Evaluate Models: Observed vs. Predicted or Predicted vs. Observed?" in *Ecological Modelling*, vol. 216, issue 3, 2008, pp. 316-322.