Machine Learning for Bioelectromagnetics: Prediction Model using Data of Weak Radiofrequency Radiation Effect on Plants

Malka N. Halgamuge
Department of Electrical and Electronic Engineering
The University of Melbourne, Parkville, VIC 3010, Australia

***Abstract*—Plant sensitivity and its bio-effects on non-thermal weak radio-frequency electromagnetic fields (RF-EMF) identifying key parameters that affect plant sensitivity that can change/unchange by using big data analytics and machine learning concepts are quite significant. Despite its benefits, there is no single study that adequately covers machine learning concept in Bioelectromagnetics domain yet. This study aims to demonstrate the usefulness of Machine Learning algorithms for predicting the possible damages of electromagnetic radiations from mobile phones and base station on plants and consequently, develops a prediction model of plant sensitivity to RF-EMF. We used raw-data of plant exposure from our previous review study (extracted data from 45 peer-reviewed scientific publications published between 1996-2016 with 169 experimental case studies carried out in the scientific literature) that predicts the potential effects of RF-EMF on plants. We also used values of six different attributes or parameters for this study: frequency, specific absorption rate (SAR), power flux density, electric field strength, exposure time and plant type (species). The results demonstrated that the adaptation of machine learning algorithms (classification and clustering) to predict 1) what conditions will RF-EMF exposure to a plant of a given species may not produce an effect; 2) what frequency and electric field strength values are safer; and 3) which plant species are affected by RF-EMF. Moreover, this paper also illustrates the development of optimal attribute selection protocol to identify key parameters that are highly significant when designing the in-vitro practical standardized experimental protocols. Our analysis also illustrates that Random Forest classification algorithm outperforms with highest classification accuracy by 95.26% (0.084 error) with only 4% of fluctuation among algorithm measured. The results clearly show that using K-Means clustering algorithm, demonstrated that the Pea, Mungbean and Duckweeds plants are more sensitive to RF-EMF ($p \leq 0.0001$). The sample size of reported 169 experimental case studies, perhaps low significant in a statistical sense, nonetheless, this analysis still provides useful insight of exploiting Machine Learning in Bioelectromagnetics domain. As a direct outcome of this research, more efficient RF-EMF exposure prediction tools can be developed to improve the quality of epidemiological studies and the long-term experiments using whole organisms.**

***Keywords*—Machine learning; plants; prediction; mobile phones; base station; radiofrequency electromagnetic fields; RF-EMF; plant sensitivity; classification; clustering**

I. Introduction

Mobile phone technology has exhibited remarkable growth in recent years, heightening the debates on the changes in plant growth due to non-thermal weak radio-frequency electromagnetic fields (RF-EMF). In order to preserve green living and biodiversity, one of the major ground-level concerns is environmental damage and its effects on plants. Modeling plant sensitivity due to RF-EMF is an important task for both agriculture sector and for epidemiologist, on the other hand, it is a useful tool to assist a better understanding of this phenomenon and eventually advance it. Reported studies showed significant effects on plants that exposed to the radiofrequency radiation or plant sensitivity to the RF-EMF [1].

The fields of machine learning and big data analytics helps to extract high-levels of knowledge from raw data and improve automated tools that can aid the health domain. Machine learning is a key tool in analytics, where algorithms iteratively learn from data to discover hidden insights [2]. It is quite challenging for experts to overlook the important details of billions of data, hence, alternatively, use of automated tools to analyze raw data and extract stimulating high-level information is exceptionally important for the decision-makers [3].

Machine learning techniques have been used in big data analysis; nonetheless, the challenge is to build a prediction model for the data with multiple variables. The raw-data grasps crucial information, such as patterns and trends, which can be used to advance decision-making and optimize achievements. This paper uses machine learning in bioelectromagnetics; that consequently, develops a prediction model of plant sensitivity to RF-EMF.

The controversy or the contention exists about the physiological and morphological changes that affect sensitivity in plants due to the non-thermal weak radio-frequency electromagnetic fields (RF-EMF) effects from mobile phones and base station radiation. On the other hand, the world has been challenged with recent environmental concerns and the loss of green living that has caused dilemma and re-evaluation of implications, especially in agriculture. While developing the country economically, citizens expect political measures to be taken for a greener environment. Nonetheless, one of the major ground-level concerns is external environmental effects on plants. There is a need to understand the trends and patterns that occur in the non-thermal weak radio-frequency electromagnetic field (RF-EMF) and its effects caused by mobile phones and base station radiation activities on plants and trees. Also, it is important to understand the significance of environmental attributes which have impacted the classification algorithm for better prediction. There is no single study that sufficiently covers machine learning concept in bioelectromagnetics domain yet.

This study tries to demonstrate the usefulness of Machine Learning algorithms for predicting the possible damages of electromagnetic radiations on plants and consequently, develops a prediction model of plant sensitivity to RF-EMF. hence, this proposes a novel solution to apply machine learning concepts and techniques by using raw data from our previous review study. Similarly, this study will replicate the former study to validate former study and to perform predictions extracting high-levels of knowledge from raw data using different classifications and clustering algorithms. This study will also presents and outline the following: 1) development of optimal attribute selection protocol to identify key parameters that should be used in in-vitro laboratory experiments; 2) K-mean clustering algorithms to analyze and predict what conditions will RF-EMF exposure impacts plant of a given species may not produce an effect; 3) which frequency and electric field strength values are safer; 4) classification algorithms for prediction of RF-EMF effect on plants species; and 5) the verification of the performance of the classification and clustering algorithms.

II. Classification Algorithms, Clustering Algorithms and Performance Evaluation Methods

This section discusses 1) classification framework; 2) classification algorithms (Bayesian Network Classifiers, Naive Bayesian Model Classifier, Decision Table, JRip, OneR, J48, Random Forest, Random Tree); 3) test modes (k-fold Cross-validation, Data Percentage Split Criteria); 4) performance evaluation of classification algorithms (Percentage of Correct Classifications, Root-mean-square error, Confusion Matrix, Time Performance); 5) clustering algorithms (K-Means Clustering, Cannopy Clustering, Expectation Maximization (EM) Clustering, Filtered Clustering, Hierarchical Clustering); 6) performance evaluation of clustering algorithms (Cluster Sum of Squared Error, Silhouette coefficient); 7) data collection; and 8) data analysis, that we used for our analysis.

A. Classification Algorithms

A classification algorithm is used to train a data sets to build a model that can be used to assign unclassified records into one of the defined classes. Classification algorithms are most appropriate for predicting or labeling new data sets (test data) with numeric, binary or nominal categories (nominal data types that represent the text data and ordinal data types that represent the data with pre-defined options). The classification algorithms or techniques are used in this study to predict the expected outcomes are Bayes Net, Nave Bayes, Decision Table, JRip, OneR, J48, Random Forest and Random Tree.

The list of symbols are defined in Table 1. Consider n-dimensional attribute vector $\vec{X} = (X_1, X_2, \ldots, X_n)$. Let there be m classes variables $C = \{c_1, c_2, \ldots, c_j, \ldots, c_m\}$.

*1) **Bayesian Network Classifiers (Bayes Net)**:* The learning task consists of finding an appropriate Bayesian network [4] given a data set D over $\vec{X}$.

Using the Bayes theorem $(P(c_j|\vec{X}) = P(c_j)p(\vec{X}|c_j|)/\left[\sum_j P(c_j)p(\vec{X}|c_j|)\right]$, Bayesian classification, C_{BN} or (h_b) is given by

$$C_{BN} = h_b(\vec{X}) = \arg \max_{j=1,\ldots,m} P(c_j)p(\vec{X}|C_j) \qquad (1)$$

TABLE 1. List of Symbols and Descriptions

Symbol	Description
D	Dataset
N	Number of labels
C	Number of class variables, i.e., $C =$ (changed, unchanged)
$\vec{X}$	Attribute vector $\{F, SAR, P, E, T, p\}$
$\vec{X_1}$	$\{F_1, SAR_1, P_1, E_1, T_1, p_1\}$, First instance
$\vec{X_2}$	$\{F_2, SAR_2, P_2, E_2, T_2, p_2\}$, Second instance
Ω_x	Map x^{th} instance (data point)
Ω_c	Value of c^{th} classifier (class label)
h_c	Hypothesis that correctly predict classification
c_j	j^{th} classifier
D_i^t	Training data set
s	Cross-validation number (e.g. $S \in 5, 10, 20$)
D_i	i^{th} data partition for cross validation
a_i	Actual value of i^{th} instance
p_i	Predicted value of i^{th} instance
F	F-measure or harmonic mean
r	Recall or sensitivity
p	Precision
K	The number of clusters
E_{ss}	Cluster sum of squared error
cl_K	Centroid off the K^{th} cluster
s_i	Set of objects in the i^{th} cluster
x_i	An object or i^{th} attribute vector
μ_i	Center point of the i^{th} cluster
M	Maximization step
E	Expectation step

where $P(c_j)$ is *"a priori"* or prior probability distribution and $p(\vec{X}|C_j)$ is the conditional probability density. For example, for 2 class problem (c_1, c_2) Bayes rule is given by:

$$C_{BN} = \begin{cases} c_1 & P(c_1|\vec{X}) > P(c_2|\vec{X}) \\ c_2 & \text{otherwise.} \end{cases}$$

*2) **Naive Bayesian Model Classifier (Naive Bayes)**:* This is based on the Bayesian theorem and uses the method of maximum likelihood for attribute estimation. Naive Bayesian Model classifier (Naive Bayes) [5] requires a small amount of training data to predict the data attributes. The Naive Bayes classifier predicts whether $\vec{X}$ belongs to class c_i, if $p(c_i|\vec{X}) > p(c_j|\vec{X})$ for $1 \leq j \leq m$. Using Bayes' theorem, the maximum posteriori hypothesis is given by $p(c_i|\vec{X} = p(c_i)p(\vec{X}|c_i)/p(\vec{X})$. This maximize $p(c_i)p(\vec{X}|c_i)$ and $p(\vec{X})$ is a constant. If we have many attributes, it is computationally costly to evaluate $p(\vec{X}|c_i)$. Hence, Naive assumption of "class conditional independence" is given by $p(\vec{X}|c_i) = \prod_{k=1^n} p(X_k|c_i)$.

*3) **Decision Table**:* Decision table classification algorithm can be efficiently used to decide the most important attributes

in a given dataset [6]. This evaluates feature of subsets by using best-first search and cross-validation mode that can be used for evaluation. In this method, attributes are not considered as an independent that is differentiated, from a verified model.

*4) **JRip**:* This classification algorithm implements a propositional rule, "Incremental Pruning is to Produce Error Reduction" (RIPPER), which uses sequential covering algorithms for creating ordered rule lists [7]. The algorithm goes through a few stages: building (growing, pruning), optimization and selection [7].

*5) **OneR**:* A simple classification algorithm that produces one rule for each predictor in the data and uses the minimum-error attribute for prediction [8].

*6) **J48**:* The J48 is a classification algorithm generates decision tree which generates a pruned or unpruned C4.5 decision tree [9] and is used for the classification of the data.

*7) **Random Forest**:* Random forests classification algorithm considers amalgamation of tree predictors (each tree depends on the independent values of a random vector sampled) and uses similar distribution for all trees in the forest [10]. When a number of trees in the forest become large, the generalization error for forests converges to a limit. This error of the forest tree classifiers depends on the vigour of the individual trees as well as the correlation between them [11].

*8) **Random Tree**:* Random Tree classification algorithm [12] uses a class for building a tree, which considers x randomly chosen attributes at each node and it does not perform pruning. Furthermore, it has an option to estimate the class probabilities established on a hold-out set or back-fitting.

B. Test Modes

Cross-validation is a technique used for estimating the error (accuracy) of the algorithm. This works by splitting the data into k subsets of approximately equal size. The performance evaluation method of eight classifiers or classification algorithms (described above) were obtained by using two different test modes: k-fold cross-validation and percentage split. Hence, for this paper, 10-fold cross validation and data percentage split criteria are used for model assessment.

*1) **K-fold cross-validation ($k-foldcv$)**:* The $k-foldcv$ splits the data set D in s equal parts $D_1, \ldots, D_s$, where typical values for s are 5, 10 and 20. Here training data set D_i^t is given by removing i^{th} data portion, D_i from D with $s = N$, $k-fold$ cross-validation. Each data point used once for testing and $s-1$ times for training. When $s = N$, $k-fold$ cross-validation becomes $loo-cv$. For an example, in this work, we use k-fold cross-validation (k =10) method. Hence, these splits the data into 10 equal parts then uses first 9 parts for training and the final fold is for testing purposes.

*2) **Data percentage split criteria**:* The data percentage split mode is a mode that splits the dataset into training data and testing it with different percentage ratios. In this test mode, the identified percentage of the train data: test data split ratio, e.g., 90%:10%, 80%:20%, etc.

C. Performance Evaluation Methods of Classification Algorithms

Outputs are then compared to understand the classifier performances using: 1) percentages of correctly classified instances (PCC); 2) mean absolute error (MAE); 3) root-mean-squared error (RMSE); 4) confusion matrix; and 5) computational time or CPU time (sec). For the confusion matrix we considered True Positive (TP) Rate, False Positive (FP) Rate, Precision (p), Recall (r) and F-measure (F).

*1) **Percentage of correct classifications (PCC)**:* The classification algorithms are frequently evaluated using the percentage of correct classifications (PCC) and this is denoted as:

$$\Psi(i) = \begin{cases} 1 & \text{if } a_i = p_i \\ 0 & \text{else} \end{cases}$$

where a_i denotes the actual value and p_i denotes the predicted value for the i^{th} instance. Using $\Psi(i)$, PCC is defined,

$$PCC = \sum_{i=1}^{n} \Psi(i)/N \times 100\%. \quad (2)$$

*2) **Mean absolute error (MAE)**:* In this analysis, we calculated the average of the absolute errors: MAE, where is the prediction forecasts and the true value. If the prediction instances are $p_1, p_2, \ldots, p_n$ and actual values are $a_1, a_2, \ldots, a_n$. Mean absolute error of testing data value is given by $(p_i - a_i)+, \ldots, +(p_n - a_n)/n$ for n different predictions.

*3) **Root-mean-square error (RMSE)**:* The root mean squared error (RMSE) is a popular metric [13]. A large PCC (i.e. near 100%) suggests a suitable classifier, while a regressor should exist a low global error (i.e. RMSE close to zero). Root-mean-square error (RMSE) = $\sqrt{\sum_{i=1}^{n}(p_i - a_i)^2/n}$ for n different predictions.

*4) **Confusion matrix**:* We also calculated the rate of each classifier that we used to predict the actual plant sensitivity and see if it changes using test data. Moreover, the weighted average of precision (p), recall (r) and F-Measure (harmonic mean) are obtained by using the 10-fold cross-validation approach.

Prediction model is defined using Confusion Matrix [14] as, 1) true positive, TP or correct hit (actual plant sensitivity changes in instances that were correctly classified), 2) true negative, TN or correct rejection (non-sensitivity changes in instances that were not classified as changes), 3) false positive, FP or the false alarm (non-sensitivity changes instances that were classified as changes, Type II error), and 4) false negative, FN or a miss (actual plant sensitivity changes in instances that were not classified as changes, Type II error). For prediction, in this paper, we also calculated other measures: 1) precision, p is the percentage of predictive items that are correct where $p = TP/(TP+FP)$; and 2) recall or sensitivity, r is the percentage of correct items that are predicted where $r = TP/(TP + FN)$. The F-measure, the harmonic mean of p and r, can be calculated as $F = 2pr/(p + r)$.

*5) **Time performance (CPU time)**:* CPU time, in this case, the time taken to build a model [14], was calculated for every algorithm (both clustering and classification). Time taken to build model was observed to identify their characteristics in

the tool. Classification algorithms are used to train a data set to build a model, and then the model can be used to allocate unclassified records into one of the well-defined classes. A test set is used to decide the accuracy of the model. Usually, the given data set is divided into the training data and tests data sets. The training set is used to build the model and test sets are then used to validate it.

D. Clustering Algorithms

Clustering is the task of grouping a set of data instances in a way that it assigns the data instances to the same group (intra-cluster) that is more alike to each other than to those in other groups (inter-clusters) [15]. Clustering is a common form of unsupervised learning, where no human expert who has assigned data to clusters ever before. Moreover, unsupervised learning methods do not construct a hypothesis prior to this analysis [16]. The clustering algorithms or techniques used in this study are: Simple K-Means, Cannopy, EM, FathestFirst, Filtered Clusterer and Hierarchical Clutterer. These are to create clusters with a set of similar behavioral points that are consistent internally.

1) ***K-Means Clustering****:* Here we use K-Means clustering algorithm to discover sensitive plants to RF-EMF. Consider a set of instances (attribute vector) $x_1, x_2, \ldots, x_n$ and K clusters where $\vec{K} = (K_1, K_2, \ldots, K_r)$. Denote the centre of the clusters (centroids) as $cl_1, cl_2, \ldots, cl_K$, where cl_K is centroid of K^{th} cluster. Initially, each centroid will be randomly placed. The Euclidean distance between i^{th} data point, x_i, in cluster and cluster centre cl_j is $\arg\min_j ||x_i - cl_j||$ to calculate nearest centroid where x_i is the i^{th} data instance (attribute vector) in K^{th} cluster. For each cluster $j = 1, \ldots, K$ we calculate cluster centre

$$cl_j(a) = \frac{1}{n_j} \sum_{x_i \to c_j} x_i(a)$$

for $a = 1, 2, \ldots, d$. Hence, each instance (e.g. attribute vector X_1) will be assigned to a cluster with the nearest centroid. Each centroid will be moved to the mean of the instances assigned to it. The K-Means clustering algorithm [17] continues until no data point changed the cluster membership, and then we stop the algorithm as it has converged. After the clustering is completed, we calculate the distance between a data point and a cluster centroid.

2) ***Cannopy clustering****:* The canopy clustering algorithm is an unsupervised pre-clustering algorithm that is usually used as the pre-processing step for the Hierarchical clustering algorithm or K-Means algorithm. This algorithm speeds up clustering procedures on large data sets [18].

3) ***Expectation Maximization (EM) clustering****:* The EM iteration swaps between performing an expectation (E) step and a maximization (M) step. During the step E it constructs a function for the expectation of the log-likelihood using the current estimate for the attributes. During the step M it calculates attributes maximizing the expected log-likelihood discovered during the step E [19]. This expectation maximization is used to establish the distribution of variables and continues until it gets the optimal value.

4) ***Filtered clustering****:* This clustering method uses an arbitrary clusterer on data that has been approved through an arbitrary filter. The structure of the filter is totally based on the training data points and test data points that will be managed by the filter without changing their structures [14].

5) ***Hierarchical clustering****:* This method [20] follows a collection of closely related clustering algorithms that produce a hierarchical clustering by merging two closest clusters until it becomes a single set. This divides a data set into the sequence of nested partitions.

E. Performance Evaluation Methods of Clustering Algorithms

We then evaluate the performance of how well each data point places within its cluster using each clustering algorithm by using three methods: 1) Cluster Sum of Squared Error (E_{ss}); and 2) Silhouette coefficient; and 3) Time performance (CPU time). The evaluation is based on log-likelihood, if clustering scheme creates a probability distribution [21].

1) ***Cluster Sum of Squared Error (E_{ss})****:* Given a set of data points, cluster sum of squared error (E_{ss}) is given by

$$E_{ss} = \sum_{i=1}^{K} \sum_{\beta \in s_i} ||\beta - \mu_i||^2$$

where s_i set of objects in the i^{th} cluster ($i = 1, 2, \cdots, K$) and μ_i is the center point of the i^{th} cluster, β is a data instances in cluster s_i.

2) ***Calculate Optimal Number of Clustering using Silhouette Coefficient****:* Silhouette signifies to a method of explanation and justification of consistency within clusters of data and it is a quantitative method to evaluate the quality of a clustering. The algorithm provides a concise graphical representation of how well each data point places within its cluster [22]. Data points to a high silhouette value are considered satisfactory clustered, oppose to, the data points with a low value could be outliers. This method works well with K-Means clustering, and it is also used to define the optimal number of clusters, hence, we use this method for cluster evaluation.

III. Materials and Methods

This section discusses materials and methods used for this study: 1) classification framework; 2) data collection; 3) data analysis; and 4) statistical analysis. Fig. 1 shows the steps by step procedure to analyze the dataset (Algorithms 2 and 3), and how machine learning could implement in Bioelectromagnetics that are used. This is further explained in Algorithm 2 (an adaptation of classification algorithm) and Algorithm 3 (an adaptation of clustering Algorithm) for plant sensitivity to RF-EMF. The list of symbols is defined in Table 1 and attributes that we used for this analysis are shown in Table 2.

As explained in our previous review study (Halgamuge, 2016), in this study, physiological or morphological effects of plants (bio-effects) or plant response (changed/unchanged or effect/no effect) due to exposure to weak radiofrequency radiation from mobile phones and base station is defined as the changes in 1) plant growth rate; 2) seed germination rate (primary shoot and root length); 3) thermographic imaging;

4) carbohydrate metabolism; 5) oxidative damage/stress; 6) gene expression; 7) DNA damage; 8) reactive oxygen species (ROS); 9) cell function, enzyme activities; 10) mitotic index and mitotic abnormalities; 11) mutation rates and genomic stability; 12) pigmentation (chlorophyll concentration); and 13) chromosomal aberrations and micronuclei.

TABLE 2. ATTRIBUTE DESCRIPTION USED FOR ANALYSIS

Attribute	Symbol	Type	Description (Domain)
Plant	p	Nominal	29 different plant types
Frequency	F	Numeric	00 - 8000 (MHz)
SAR	SAR	Numeric	0 - 50 (W/kg)
Power Flux Density	P	Numeric	0 - 50 (W/m2)
Electric Field Strength	E	Numeric	0 - 100 (V/m)
Exposure Time	T	Numeric	0 - 6 years
Response	R	Binary	Changed or Unchanged

The list of symbols is defined in Table 1 and attributes that we used for this analysis are shown in Table 2.

A. Classification Framework

One of the key machine learning tasks is classification. The main task of classification is learning a target function f which maps each attribute sets and mapping an input attribute set Ω_x into its appropriate class label Ω_c. Although the classification is made by generating a predictive model of data, interpreting the model normally offers information for distinguishing labeled classes in data [13]. In this paper, we used 2 class variables: plants growth responses that are changed or unchanged due to non-thermal weak RF-EMFs.

Consider a data set D with N labeled and C classifiers. Then the data split into two parts: training data (used to train the classifier) and test data (used to estimate the error rate of the trained classifier). Train data is used to learn the algorithm and to test data set that will only be accessible during the classifier prediction. Classifier is a mapping method from unlabeled instances (new data points) to classes (in our case, 2 class variables: changed or unchanged). Hence, define a classifier as a function (f) assigns a class variables $C \in \Omega_c = \{c_1, c_2, \ldots, c_m\}$ to objects described by a set of attribute variables such that $\vec{X} \in \Omega_x = \{X_1, X_2, \ldots, X_n\}$ (n dimensional attribute vector), then map $f : \Omega_x \rightarrow \Omega_c$, ($x^{th}$ instance to c^{th} classifier). The classification can be divided into two phases: learning phase to train data and classification phase for test data. A classifier $h : \Omega_x \rightarrow \Omega_c$ is a function that maps an instance of Ω_x to a value of Ω_c. Now consider the classifier or the hypothesis (h_c) that can correctly predict the classification of the new scenario and its a function that maps an instance $h_c : \Omega_x \rightarrow \Omega_c$ ($\vec{X} \rightarrow y$).

The classifier is learned or becomes proficient from a data set D consisting of samples, (Ω_x, Ω_c). Given the probability $P(c_j|\vec{X})$ where x belongs to a certain class rather than a simple classification. Here $\vec{X}$ is a n-dimensional attribute vector. Then we map $\vec{X} \rightarrow P(C|\vec{X})$, $j = 1, \ldots, m$

$$\lim_{j=1,\ldots,m} P(c_j|\vec{X})$$

where c_j is the j^{th} classifier. Finally, classification is defined as

$$C = h_c(\vec{X}) = \arg \max_{j=1,\ldots,m} P(c_j|\vec{X}). \quad (3)$$

Example: Consider attributes: frequency (f_1), specific absorption rate (SAR_1), power flux density (p_1), electric field strength (E_1), exposure time (T_1), biological material (m_1). So, consider two new data attributes vectors: $\vec{X_1} = \{f_1, SAR_1, p_1, E_1, T_1, m_1\}$ and is the first instance and $\vec{X_2} = \{f_2, SAR_2, p_2, E_2, T_2, m_2\}$ in the second instance. The binary type of class variables, i.e., $C =$ changed, unchanged will be used. Now, considering the classifier that can correctly predict the classification of the new scenario then the classification could be selected as one of the two class variables (changed, unchanged) to allocate to each instances $\vec{X_1}$ and $\vec{X_2}$, based on how classification algorithm calculates the probabilities of predicting that option.

B. Data Collection

The raw-data holds crucial information, such as patterns and trends, that can be used to improve decision-making and optimize the achievements. This paper used raw-data of plant exposure from our previous review study [1] (extracted data set from 45 peer-reviewed scientific publications (1996-2016) with 169 experimental observations carried out in the scientific literature, e.g. [23] and performed prediction extracting high levels of knowledge from raw data using different classification algorithms and performance evaluation methods. Moreover, we used these data sets for clustering algorithms. The collected dataset comprises of 8 attributes and 169 experimental case studies or instances.

C. Data Analysis

In our analysis, we considered the class variables, attributes (characteristics), classification algorithms, performance evaluation methods of classification algorithms, clustering algorithms, performance evaluation methods of clustering algorithms, as shown in Table 3.

D. Statistical Analysis

The statistical significance is a technique that does not vary in outcome when applying it to the same dataset. All studies require a statistically significant method to analyze their data to come up with the final analysis of whether the hypothesis of the radio frequency radiation affects the plants or not. In order to detect whether or not a frequency may have an effect on plant sensitivity, we performed clustering algorithms, as outlined in Section II. We perform cluster analysis tests to observe whether *intra-cluster-variance* (V_{intra}) of some data points are smaller compared to *inter-cluster-variance* (V_{inter}). We consider variability among mean of the sum of squared distances within groups which are smaller than distances between the groups. Hence, the null hypothesis (H_0) in this analysis is that there are no subsets of observed data that are more alike to each other than the rest of the data, in other words, cluster analysis tests whether *intra-cluster-variance* (V_{intra}) of some data points are small compared to *inter-cluster-variance* (V_{inter}). The alternative hypothesis (H_A) is that the probabilities are statistically different. In

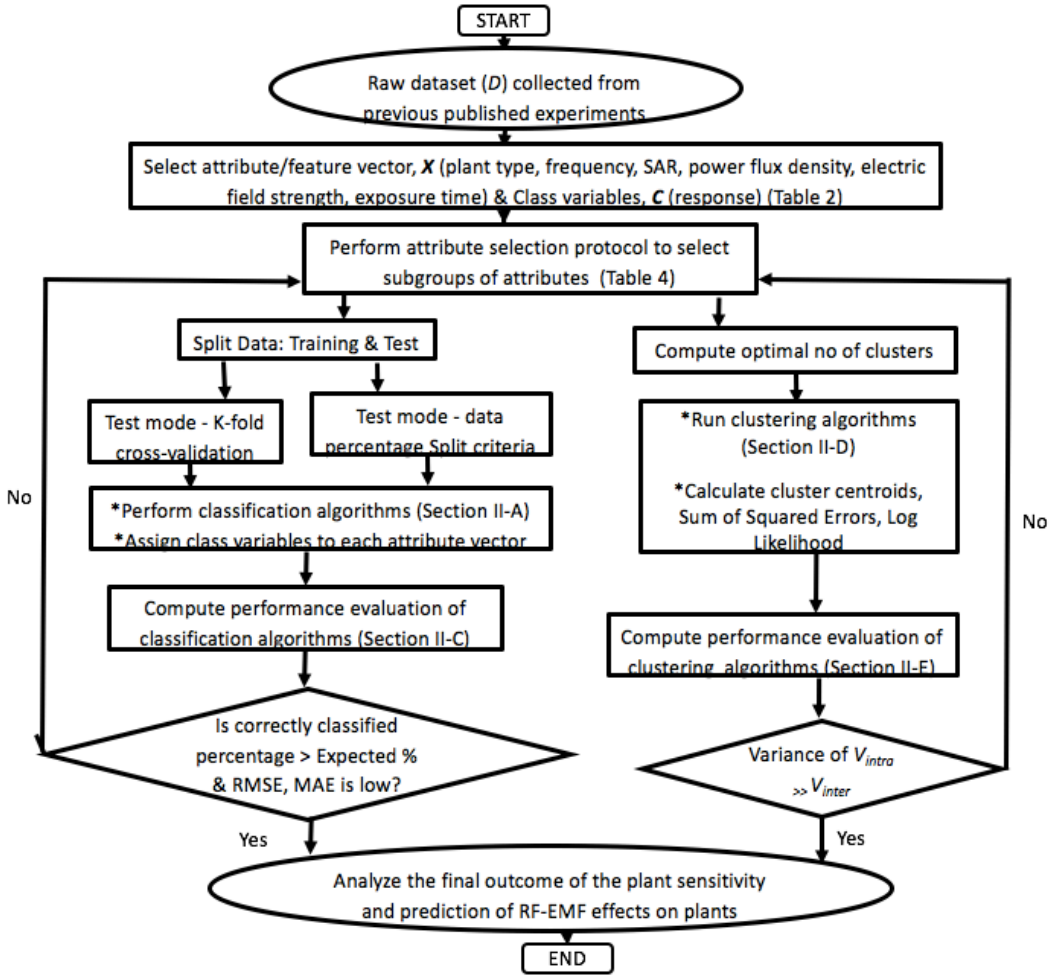

Fig. 1. Plant sensitivity to RF-EMF analysis and prediction tool (Tables 2, 3, and 4)

TABLE 3. LIST OF PARAMETERS USED IN THE PROPOSED RF-EMF DATA ANALYSIS

Type	Number	Description
Data Instances	169	Data from 169 published studies gathered in our previous work [1]
Class variables	2	Changed, Unchanged
Attributes	6	Plant, frequency, SAR, power flux density, electric field strength, exposure time
Classification algorithms	8	Bayes net, NaiveBayes, Decision Table, JRip, OneR, J48, Random Forest and Random Tree
Performance evaluation methods of classification algorithms	5	Percentage of Correct Classifications (PCC), Mean absolute error (MAE), Root-mean-square error (RMSE), Confusion Matrix, Time performance (CPU time)
Clustering algorithms	6	Simple K Mean, Cannopy, EM, FathestFirst, Filtered Clusterer, Hierarchical Clutterer
Performance evaluation methods of clustering algorithms	2	Cluster Sum of Squared Error (E_{ss}), Time performance (CPU time)

Algorithm 1 : Optimal Attribute Selection

1: **Load** raw Dataset D
2: **Split** Data into $\Rightarrow$ *Training : Test*
3: **Load** $\vec{X} \in \Omega_x = \{f, SAR, P, E, T, p\}$ (complete attribute vector)
4: **Find** all compulsory attributes $\Rightarrow$ in-vitro experiments
5: **Select** sub-set of attribute vector $\vec{x_j} < \vec{X}$ (e.g. Case *A*, Case *B*)
6: **Select** classification algorithm
7: **Perform** attribute selection protocol to select subgroups of attributes
8: **for** $\forall\ \vec{x_j}$ **do**
9: **Run** classification algorithm
10: **Evaluate** both test modes **do**
11: Select Test mode: *K-fold cross-validation*
12: Select Test mode: *Data Percentage Split Criteria*
13: **end** for
14: **Select** attribute set $\Rightarrow$ Training : Test score is minimized
15: **Allocate** model type (Case *A*, Case *B*) for each attribute vector, $\vec{X}$
16: **End**

TABLE 4. EXPERIMENTAL PROTOCOL FOR SELECTION OF SUBGROUPS OF APPROPRIATE ATTRIBUTES (ALGORITHM 1): THE VARIOUS SCENARIOS OF ATTRIBUTES SELECTION (PARAMETER) FOR CLASSIFICATION AND CLUSTERING ANALYSIS

Model Type	Plant	Frequency	SAR	Power Flux Density	Electric Field Strength	Exposure Time
Case *A*	Yes	Yes	Yes	Yes	Yes	Yes
Case *B*	Yes	Yes	Yes	Yes	Yes	Yes
Case *C*	-	Yes	Yes	Yes	Yes	Yes
Case *D*	Yes	Yes	-	Yes	Yes	Yes
Case *E*	Yes	Yes	Yes	Yes	-	Yes
Case *F*	-	Yes	-	Yes	-	Yes
Case *G*	-	Yes	Yes	Yes	Yes	Yes
Case *E*	Yes	Yes	-	Yes	Yes	Yes
Case *I*	Yes	Yes	Yes	Yes	-	Yes
Case *J*	Yes	Yes	-	Yes	Yes	Yes

this analysis, 95% of confidence level ($p < 0.05$) to estimate statistical significance. The null hypothesis is rejected if $y < 0.05$ i.e. at confidence level ($p < 0.05$). This study we use cluster sum of squared error (E_{ss}), hence, the hypothesis is given as

$$H = \begin{cases} H_o & \text{if } V_{intra} \text{ of } E_{ss} >> V_{inter} \text{ of } E_{ss} \\ H_A & \text{otherwise.} \end{cases}$$

The MATLAB (MathWorks Inc., Natick, MA, USA) R2015b, one-way ANOVA procedure in SPSS Statistics (Version 23) and Weka tool (Waikato Environment for Knowledge Analysis, Version 3.9) have been used to carry out analysis on a computer with an Intel Core Intel Core i7 CPU.

IV. RESULTS

This section briefly explains the results and the aim of this study to develop a tool using machine learning to analyze data in bioelectromagnetics domains. In order to measure plant sensitivity to non-thermal weak RF-EMF, the different classification and clustering algorithms are used. We have used extracted data from the 45 peer-reviewed scientific publications published between 1996-2016 with 169 experimental case studies carried out in the scientific literature with 6 attributes and 2 class variables to analyze the prediction performance of algorithms. For our evaluation, we used 8 classification algorithms specifically using 2 test modes, 5 performance evaluation methods of classification algorithms, 6 clustering algorithms, 2 performance evaluation methods of clustering algorithms.

A. Attribute Selection

Our proposed attribute selection protocol (ten different cases, as shown in Table 4) and performed it under 10 different scenarios to observe the highest important attribute that demonstrates certain aspects of the proposed method. Tables 4 and 5 demonstrates $Case\ C$ (frequency, SAR, power flux density or electric field strength and exposure time) attribute group is the more appropriate parameter group to predict most correctly classified instances. The optimal attribute selection protocol is beneficial to identify key parameters that should be used in in-vitro laboratory experiments.

Algorithm 2 : Adaptation of Classification Algorithm for Plant Sensitivity to RF-EMF

1: Collect raw dataset D with C classifiers
2: Select attribute vector, $\vec{X} \in \Omega_x = \{f, SAR, P, E, T, p\}$
3: Select class variables $C \in \Omega_c = \{c_1, c_2, \ldots, c_m\}$
4: Select classification algorithms, a_1
5: Perform attribute selection protocol to select subgroups of attributes
6: **Repeat**
7: **for** all attribute selection protocols Class A to Class J **do**
8: **for** classification algorithms $a_1 = 1, 2, 3, \ldots, p$ **do**
9: **for** both test modes **do**
10: Select Test mode: *K-fold cross-validation*
11: Select Test mode: *Data Percentage Split Criteria*
12: **for** all dataset D **do**
13: **Split** Data into $\Rightarrow$ *Training : Test*
14: Perform classification algorithm
15: Assign class variable using classifier to each attribute vector, $h : \Omega_x \rightarrow \Omega_c$
16: **end for**
17: **end for**
18: Compute Percentage of Correct Classifications (PCC)
19: Compute Mean absolute error (MAE)
20: Compute Confusion Matrix
21: Compute Time Performance (CPU time)
22: **if** PCC $< 80\%$
23: **else if** MAE > 1
24: **else if** CPU Time > 1 sec **then**
25: **stop**
26: **end for**
27: Select next attribute set
28: **end for**
29: **until** there is attribute selection protocol to test

B. Classification

In this subsection, we further analyze RF-EMF sensitivity it caused on the plants using classification algorithms. Ten test cases (as in Table 4) were designed to demonstrate certain aspects of the proposed method. After carrying out the Multivariate Analysis of plants, six classification algorithms (Bayes

Algorithm 3 : Adaptation of Clustering Algorithm for Plant Sensitivity to RF-EMF

Collect raw dataset D
Select attribute vector, $\vec{X} \in \Omega_x = \{f, SAR, P, E, T, p\}$
Select clustering algorithms, a_2
Perform attribute selection protocol to select subgroups of attributes
Repeat
for all attribute selection protocols Class A to Class J **do**
  for classification algorithms $a_1 = 1, 2, 3, \ldots, q$ **do**
  Compute optimal number of clusters using silhouette coefficient
  Calculate cluster centroid, $cl_1, cl_2, \ldots, cl_K$ where $\arg\min_j ||x_i - cl_j||$
  Calculate distances between data points and a cluster centroid
for No of Clusters $K = 1, 2, \ldots, K$ **do**
  Compute Cluster Sum of Squared Error (E_{ss})
  Compute Time Performance (CPU time)
    if V_{intra} of $E_{ss} >> V_{inter}$ of E_{ss} **then**
    H_0
    else H_A
    end if
  end for
 end for
Select next attribute set
end for
until there is attribute selection protocol to test

Network, J48, JRIP, Naive Bayes, OneR and PART) were used to make the best predictions for the given dataset. In order to test each algorithm, mainly three different testing techniques were used: using 1) full training set; 2) cross-validation with 10 folds; and 3) percentage split. Table 5 shows the correctly classified percentages of each classification algorithm.

This study has found that the Random Forest algorithm shows a high percentage of accuracy by 95.26% (0.084 error) with only 4% of fluctuation among algorithm measured. We also used Nave Bayes algorithm and found the least classification percentage. Hence, we removed it from tables. We developed an optimal attribute selection protocol and performed it under 10 different scenarios to observe the most important attribute (parameter) for classification and clustering. This is vital to identify key parameters that are highly significant in the in-vitro laboratory experiments. The protocol of various scenarios is described in Table 4. The optimal attribute selection protocol is vital to identify key parameters that are highly significant when designing the in-vitro practical standardized experimental protocols.

1) ***Changed or unchanged prediction: k-fold cross-validation of raw data method****:* This work has used k-fold cross-validation (k =10) method. This method splits the data into 10 equal parts and then uses the first 9 parts for training, and final fold is for testing purposes. The classification model performance uses a confusion matrix-10-folds cross-validation method (Table 6) shows a comparative study between the classifiers to obtain which classifier is the best for the given dataset. Computational time seems to be low due to the smaller sample size. The obtained results reveal that (Table 5) the Random Forest algorithm is the most accurate and most suitable classification algorithm to be used in effect of the plant for their further data analysis and predictions. Random Forest classification algorithm outperforms with highest classification accuracy by 95.26% (0.084 error) followed by JRip with 94.08% (0.235 error) and Bayes Net with 94.08% (0.2349 error) (Table 5). Table 6 shows a comparative study between the classifiers. The weighted average values of changed or unchanged prediction were considered by using "$Case\ C$" parameter selection, as shown in Table 4.

2) ***Changed or unchanged prediction: percentage split of raw data method****:* The dataset was verified by splitting the data into different percentages whereas Train%: Test%. In this technique, the model will be trained and constructed with a certain percentage of data and then tested with the rest of the percentage. Table 7 shows the correctly classified percentage of each classification algorithm. The bold values are marked as the best within the classification type. According to this analysis, Bayes net and Random Forest algorithms show the high percentage of accuracy. Our results suggest to disregard differentiating plant type (i.e. tomato, soybean) then the classification prediction accuracy is the highest (Table 5). The classification results (PCC values (%) and RMSE values are in the bracket, underline is the best model, bold values are the best within the classification type (Table 7). The "$Case\ C$" data set has been used for this analysis (Test mode: Percentage Split test method (Train Data: Test Data)).

Considering the classification of algorithms, Random Forest gives the best results with a strong connection among attributes. Nevertheless, the overall of all eight algorithms demonstrates good results. For instance, results show that the fluctuation among the correctly classified percentages of algorithms is less than 4%.

C. Clustering

In this sub-section, we try to find data points from our datasets with similar behaviors in groups. Six clustering algorithms were used to cluster the data sets from 169 experimental records. Evaluation of different clustering algorithms is shown in Table 8. It is visually clear that there are three distinct clusters. Moreover, we visualized the potential clusters using Simple K-Means clustering algorithm. The K-Means is the simplest clustering algorithm among all the clustering methods. Hence, we used it for visualizing the clusters. Table 8 shows 1) the cluster instances and percentages of 2 clusters; 2) CPU time, a number of iterations, log-likelihood value, and cluster sum of squared error (E_{ss}) for the different clustering algorithms. The optimal number of clusters were obtained using Silhouette value. Our analysis gives optimal results when $K = 3$ (Fig. 2 and 3). Cluster sum of squared error (E_{ss}) for K-Means clustering was 148.08 and Filtered cluster method also shows the same error. Log-likelihood value for Expectation Maximization clustering (EM) method was -37.42.

Table 9 shows Duckweed is the most common plant species that is very sensitive to RF-EMFs in any given number of clusters. We also observed that when $K = 4$, Duckweed species repeatedly shows the sensitivity to RF-EMF in more than one cluster. Using optimal clustering ($K = 3$, silhouette plots (Fig. 2 and 3), our data showed Duckweeds, Mungbean,

TABLE 5. CLASSIFICATION RESULTS (PCC VALUES (%) AND RMSE VALUES ARE IN BRACKET, UNDERLINE IS THE BEST MODEL, BOLD VALUES ARE THE BEST WITHIN THE INPUT SETUP. TEST MODE: 10-FOLD CROSS VALIDATION METHOD)

Classification Type	Case *A*	Case *B*	Case *C*	Case *D*	Case *E*	Case *F*	Case *G*	Case *H*	Case *I*	Case *J*
Bayes net	93.49% (0.2370)	92.89% (0.2447)	94.08% (0.2349)	93.49% (0.237)	94.08% (0.2298)	93.49 % (0.2295)	87.57% (0.2587)	92.89% (0.2447)	92.89% (0.2385)	92.89% (0.2447)
NaiveBayes	73.96% (0.4204)	55.62% (0.5216)	59.17% (0.5201)	77.51% (0.408)	92.89% (0.2654)	88.16% (0.3300)	44.37% (0.6212)	54.43% (0.5188)	90.53% (0.2947)	54.43% (0.5188)
Decision Table	92.31% (0.2522)	92.89% (0.2431)	93.49% (0.2465)	92.30% (0.2522)	92.30% (0.2524)	94.08% (0.2377)	94.08% (0.2375)	92.899% (0.2431)	92.89% (0.2433)	92.89% (0.2431)
JRip	94.08% (0.2345)	94.08% (0.2347)	94.08% (0.235)	92.89% (0.2525)	**94.67% (0.2224)**	94.67 % (0.2234)	94.08% (0.2355)	94.08% (0.2347)	94.08% (0.2351)	94.08% (0.2347)
OneR	88.16% (0.344)	88.16% (0.344)	93.49% (0.2551)	88.16% (0.3440)	88.16% (0.3440)	94.67 % (0.2308)	94.67% (0.2308)	88.16% (0.344)	88.16% (0.3440)	88.16% (0.344)
J48	93.49% (0.2457)	94.67% (0.2233)	92.30% (0.2686)	93.49% (0.2457)	93.49% (0.2469)	**94.67 % (0.2224)**	**94.67% (0.2224)**	**94.67% (0.2233)**	93.49% (0.2461)	**94.67% (0.2233)**
Random Forest	**94.08% (0.2222)**	**94.08% (0.2243)**	**95.26% (0.084)**	**94.08% (0.2251)**	94.08 % (0.2232)	94.67 % (0.2291)	94.67% (0.2272)	93.49% (0.2269)	**94.08% (0.2263)**	93.49% (0.2269)
Random Tree	93.49% (0.2478)	92.89% (0.2595)	92.89% (0.2556)	94.08% (0.2382)	94.67 % (0.2249)	91.12 % (0.2858)	91.12% (0.2908)	93.49% (0.2475)	94.08% (0.2374)	93.49% (0.2475)

TABLE 6. CLASSIFICATION MODEL PERFORMANCE USING CONFUSION MATRIX (WEIGHTED AVERAGE). TEST MODE: 10-FOLD CROSS VALIDATION METHOD USING CASE *C* DATA SET

Classifier	PCC (%)	MAE	RMSE	TP Rate	FP Rate	Precision (p)	Recall (r)	F-Measure (F)	CPU Time (sec)
Bayes net	93.49%	0.0725	0.2345	94.1%	37.3%	93.6%	94.1%	93.7%	0.02
NaiveBayes	76.33%	0.2681	0.4068	59.2%	**30.7%**	86.7%	59.2%	67.2%	0.00
Decision Table	92.30%	0.1263	0.2521	93.5%	47.7%	92.9%	93.5%	92.7%	0.05
JRip	94.08%	0.0980	0.2345	94.1%	42.5%	93.6%	94.1%	93.5%	0.01
OneR	89.94%	0.1006	0.3172	93.5%	47.7%	92.9%	93.5%	92.7%	0.00
J48	93.49%	0.1092	0.2458	92.3%	47.9%	91.5%	92.3%	91.7%	0.02
Random Forest	**94.08%**	0.0824	0.2242	**95.3%**	31.9%	**95.0%**	**95.3%**	**95.0%**	0.20
Random Tree	91.71%	0.0841	0.2786	92.9%	37.4%	92.6%	92.9%	92.7%	0.00

TABLE 7. CLASSIFICATION RESULTS (PCC VALUES (%) AND RMSE VALUES ARE IN BRACKET, UNDERLINE IS THE BEST MODEL, BOLD VALUES ARE THE BEST WITHIN THE CLASSIFICATION TYPE. TEST MODE: PERCENTAGE SPLIT TEST METHOD (TRAIN DATA: TEST DATA) USING CASE *C* DATASET)

Classification Type	Train 90%: Test 10%	Train 80%: Test 20%	Train 70%: Test 30%	Train 60%: Test 40%	Train 50%: Test 50%	Train 40%: Test 60%	Train 30%: Test 70%	Train 20%: Test 80%	Train 10%: Test 90%
Bayes net	88.23% (0.2971)	94.12% (0.2801)	94.12% (0.2631)	94.12% (0.2503)	94.04% (0.2454)	94.05% (0.2422)	94.91% (0.2249)	**95.55% (0.2260)**	94.07% (0.2341)
NaiveBayes	64.70% (0.4487)	73.52% (0.4641)	72.54% (0.5082)	70.58% (0.4849)	72.61% (0.4568)	76.23% (0.4351)	94.06% (0.2489)	93.33% (0.237)	**94.07% (0.2428)**
Decision Table	**94.11% (0.2473)**	94.11% (0.2480)	**94.11% (0.2473)**	94.11% (0.2379)	92.85% (0.2570)	93.06% (0.2543)	94.06% (0.2351)	94.07% (0.2326)	90.13% (0.3059)
JRip	**94.11% (0.2359)**	94.11% (0.2407)	94.11% (0.2363)	92.64% (0.2690)	92.85% (0.2584)	93.06% (0.2565)	94.06% (0.2369)	94.07% (0.2395)	94.07% (0.2433)
OneR	**94.11% (0.2425)**	**94.11% (0.2425)**	**94.11% (0.2425)**	94.11% (0.2425)	92.85% (0.2673)	93.06% (0.2633)	94.06% (0.2436)	94.07% (0.2434)	94.07% (0.2433)
J48	94.11% (0.2353)	94.11% (0.2358)	**94.11% (0.2351)**	92.64% (0.2466)	94.04% (0.2390)	94.05% (0.2417)	92.37% (0.2762)	94.07% (0.2395)	86.18% (0.3717)
Random Forest	94.11% (0.2469)	94.11% (0.2471)	94.11% (0.2390)	92.64% (0.2349)	92.85% (0.2375)	93.06% (0.2448)	93.22% (0.2322)	94.07% (0.2209)	**95.39% (0.2245)**
Random Tree	88.23% (0.2953)	**94.11% (0.2425)**	88.23% (0.3430)	92.64% (0.2716)	90.47% (0.3124)	92.07% (0.2743)	94.06% (0.2473)	93.33% (0.2553)	93.42% (0.2484)

TABLE 8. CLUSTERING RESULTS (PERCENTAGE OF INSTANCES IN EACH CLUSTER, CPU TIME, LOG LIKELIHOOD, CLUSTER SUM OF SQUARED ERROR (E_{ss}) USING CASE *C* DATA SET)

Clustering Algorithm	**Cluster 1**	**Cluster 2**	**Cluster 3**	**CPU Time**	**Log Likelihood**	E_{ss}
Simple K Mean	55 (33%)	65 (38%)	49 (29%)	0.05	-	148.08
Cannopy	66 (39%)	70 (41%)	33 (20%)	0.01	-	-
EM	17 (10%)	62 (40%)	85 (50%)	0.06	-37.42	-
FathestFirst	154 (91%)	13 (8%)	2 (2%)	0.01	-	-
Filtered Clusterer	55 (33%)	65 (38%)	49 (29%)	0.01	-	148.08
Hierarchical Clusterer	165 (98%)	3 (2%)	1 (1%)	0.10	-	-

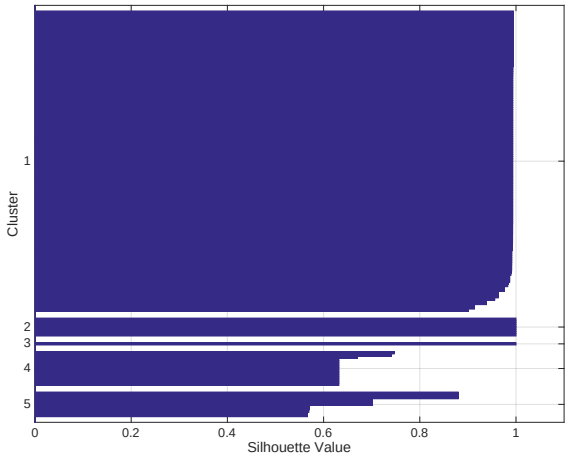

Fig. 2. Silhouette coefficient to determine optimal number of clusters - Calculating the silhouette plots.

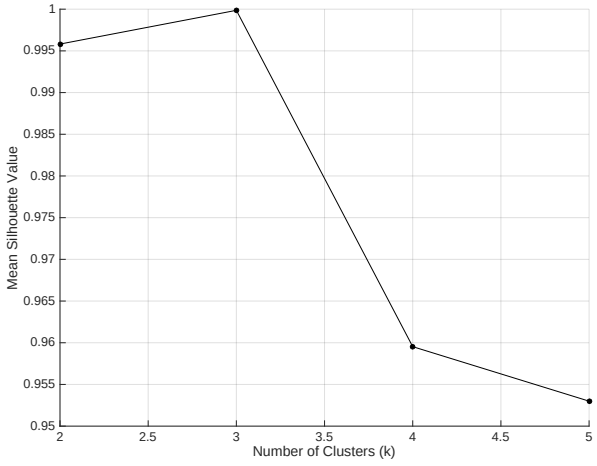

Fig. 3. Silhouette coefficient to determine optimal number of clusters - The optimal Silhouette value is obtained when $K = 3$.

TABLE 9. COMPARISON OF DATA ANALYSIS FROM OUR PREVIOUS METHOD AND SIMPLE K-MEAN CLUSTERING USING SAME DATASET [1] USED BY OUR PREVIOUS STUDY

Study/ No of Clusters	**Sensitive Plants for RF-EMF**
Analysis from our previous Study [1] where $p < 0.05$	Duckweeds, Mungbean, Pea, Broadbean, Maize, Fenugreek, Roselle, Tomato, Onions
K-Mean clustering $K = 1$	Duckweeds
K-Mean clustering $K = 2$	Duckweeds, Mungbean
K-Mean clustering $K = 3$, Optimal clustering using silhouette plots (Fig. 2 and Fig. 3), $p < 0.009$	Duckweeds, Mungbean, Pea
K-Mean clustering $K = 4$	Duckweeds, Maize, Pea, Mungbean
K-Mean clustering $K = 5$	Duckweeds*, Mungbean, Roselle, Onions
K-Mean clustering $K = 6$	Duckweeds*, Mungbean, Roselle, Onions, Fenugreek
K-Mean clustering $K = 7$	Duckweeds*, Mungbean, Roselle, Broadbean, Maize, Fenugreek
K-Mean clustering $K = 8$	Duckweeds*, Mungbean, Pea, Broadbean, Maize, Fenugreek, Roselle

Pea species are more sensitive to RF-EMFs ($p < 0.0001$). These values were then compared with the results from our previous review study [1] and observed similar behaviors. In our previous review, we found Maize, Roselle, Pea, Fenugreek, Duckweeds, Tomato, Onions and Mungbean plants are more sensitive to RF-EMF ($p < 0.0001$). In this paper, we used simple K-Means clustering algorithm and observed Pea, Mungbean, and duckweeds plants are more sensitive to RF-EMF ($p < 0.0001$).

To interpret the clusters, we compared our previous analysis of electric field strength values (effect or no effect) for different frequencies (raw data from 45 case studies) (Fig. 4(a)) [1] and clustering using K-mean algorithm (Fig. 4(b)). As clearly shown in the figure, we observed the robust connection using the K-mean clustering and it is clearly grouped no-effect data instances. This proves that K-mean clustering algorithms can be successively used in Bioelectromagnetics to observe which frequency and which electric fields strengths are more sensitivity (bio-effects) or more effective on plants (Fig. 4). Hence, this paper provides the useful insights about under

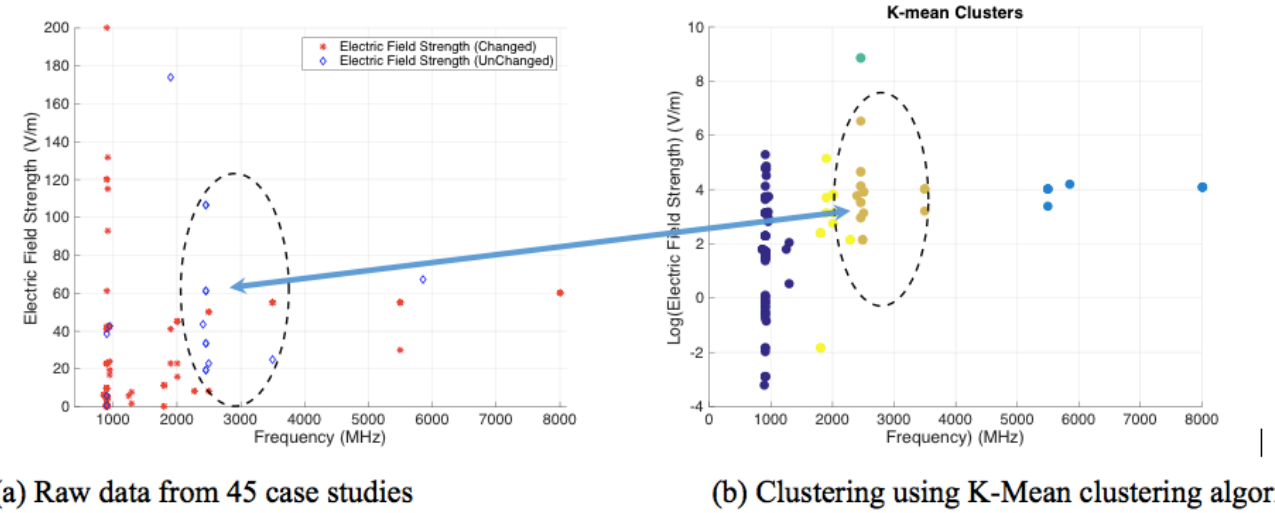

Fig. 4. K-Mean clustering algorithm to data from plants exposed to RF radiation in experiments that reported results (plant sensitivity response (changed/unchanged or effect/no effect)) from 45 publications describing 169 experimental observations to detect the plant sensitivity changes in plants due to the non-thermal RF-EMF effects from mobile phones and base station radiation. Please note that due to identical exposure conditions there were overlaps of data points. This figure demonstrates the robust connection using the K-mean clustering and it is clearly grouped no-effect data instances. This shows that K-mean clustering algorithms can be successively used to observe which frequency and which electric fields strengths are more sensitivity (bio-effects) or more effective on plants.

what conditions will RF-EMF exposure of given plant species may not produce an effect. Ultimately, the observational data for this study agrees with our earlier study, and suggest that Machine learning is an important tool, as it verifies some unexplained correlations in bioelectromagnetics domain.

V. Discussion

In order to preserve green living and biodiversity, one of the foremost ground-level concerns is environmental damage and its effects on plants. Modelling plant sensitivity due to RF-EMF is an important task for both agriculture sector and for the epidemiologist. It is also a beneficial tool to assist a better understanding of this phenomenon and ultimately advance it. On the other hand, mobile phone technology has exhibited remarkable growth in recent years, heightening the debates on the impact and changes it causes in plant growth due to non-thermal weak radio-frequency electromagnetic fields (RF-EMF). Nonetheless, mobile phone technology is updated and upgraded every day. Consequently, the importance of combining the importance of conserving plants, and technology, guarantees sustainability by identifying the effects of RF-EMF on plant species. As the diversity changes and the requirement of its understanding increase, at the same time of technology, it assists people to find more precise responses quicker than ever. Hence, using the technology, machine learning algorithms gives a better understanding of diversity. This study has developed a prediction tool to investigate the effect of RF-EMF to plant species in order to identify key variables that affect plant sensitivity (bio-effect). This approach shows changed/unchanged levels by using big data analytics and machine learning concept in bioelectromagnetics domain to reveal hidden patterns and unknown correlations. We used raw-data of plant exposure from our previous work [1] (extracted data from 45 peer-reviewed scientific publications published between 1996-2016 with 169 experimental case studies carried out in the scientific literature) and performed predictions, obtaining high-level of knowledge from raw data.

The number of mobile phones usage boosted in a drastic way due to the 1) decreasing communication cost; 2) excessive usability of web services, send and receive emails; and 3) using services from entertainment, education, banking, and medicine. With the remarkable advancement in the use of this technology, the controversy remains to exists about the physiological and morphological or bio-effect in the plants due to non-thermal weak RF-EMF effects from mobile phones and base station radiation. Our results suggest that a good predictive accuracy can be succeeded, if the information is provided about the frequency, SAR, power flux density, electric field strength, and exposure time. Hence, optimal attribute selection protocol to identify key parameters that are highly significant when designing the in-vitro practical standardized experimental protocols. Nevertheless, for the field of bioelectromagnetics and medical science accuracy is the key objective as they deal with sensitive data and a single error that can lead to the wrong conclusion. The advancement of Information Technology, and interest in big data analytics, machine learning has led to exponential growth of business organizational databases. This data holds beneficial information, such as trends and patterns, consequently, can be utilized to improve decision making that inadvertently optimizes success. Experts overlooked important details from billions of data which are quite challenging, thus, alternatively, using automated tools to analyze raw data and obtain stimulating high-level information for the decision-maker is quite significant [3].

Machine learning concepts have also been used in many research communities, including medicine [24], [25], crime prediction [26] and education [3]. However, no single study

exists which adequately covers machine learning concept in bioelectromagnetics domain. Due to attributes that influence (in our case, attributes are: frequency, SAR, power flux density, electric field strength, exposure time and plant type) to RF-EMF effects on plant sensitivity, it is very challenging to predict the growth of changes with high accuracy. On the other hand, machine learning concepts have not been generally accepted due to their inherent stochastic behavior [24]. Consequently, the results may not provide a sufficient reproducibility to adequately facilitate thoughtful scientific studies, as machine learning techniques use the probability approach. Therefore, it allows small fluctuation of incorrectly classified instances in different classifiers. However, with the advancement of technology, the reproducibility became sufficient to permit serious scientific studies [24]. On the other hand, advancement of the modern technology, intelligent data analysis will show a vital role due to the vast amount of information produced and stored [24]. To accommodate that, current machine learning algorithms provide sophisticated tools that can considerably help the science community to uncover new relationships in the data and its behavior.

Results revealed attributes set selected using the developed algorithm is consistent with in-vitro experiments. Once the raw data is fed, using K-Means clustering algorithm, demonstrated that the Pea, Mungbean, and Duckweeds plants are more sensitive to RF-EMF and statistical analysis revealed the same results evidencing precision ($p < 0.0001$). The cluster sum of squared error (E_{ss}) has been used to evaluate how well all the data points are clustered. To support these results, our previous research [1] found Maize, Roselle, Pea, Fenugreek, Duckweeds, Tomato, Onions, and Mungbean plants are more sensitive to RF-EMF ($p < 0.0001$). Additionally, this study shows that K-mean clustering algorithm can be successively used to predict what conditions will RF-EMF exposure given to plant species produce has an effect. Another possibility to obtain statistical significance (p-value) is using the Silhouette coefficient. We use the Silhouette coefficient to estimate the optimal number of clusters. Then, the ratio between intra-cluster-distances: inter-cluster-distances should be in-between -1 to $+1$. Clustering algorithms have been extensively used by research in areas for energy minimization [27], [28] that could also have been trained in this area as well. Similar to our results, the findings of previous research [29], [30] show that extensive thoughtful and computational attributes that can be used with K-Means clustering approach using medical data could be ideal. Their results have also suggested that K-Means have the potential to classify medical data.

Our results show that in bioelectromagnetics domain, the various classifiers are accomplished the same way, and the similar outcomes were obtained by another group of physicians in medical data obtained similar outcomes [24]. However, we cannot generalize this as we had a small sample size. In different classifiers who have different explanation capability [24], suitable for each classifier which could depend on the explanation that fits our own data. We used 7 different classification algorithms to select the best classifier for our data. This idea was supported by a previous research. Selecting a single best classifier that could be an option, nonetheless, the best solution could also be to use all of them and combine their judgment when solving a new problem [24].

In bioelectromagnetics domain, obtaining of SAR data is generally difficult and time-consuming. Therefore, it is appropriate to have a classifier that is able to consistently identify with a less amount of data about some attributes. Our results show that getting the appropriate subgroup of attributes could play a significant role in obtaining the high percentage of correctly classified instances. This observation was also supported by [2], whereas, selecting an appropriate subgroup of attributes (parameters or characteristics) is a key thing when using machine learning algorithms [2], as the selection is completed during the learning.

Despite its benefits, there is no single study that adequately covers machine learning concept in bioelectromagnetics domain yet, nonetheless, in the future, this technique might play a vital role to predict the potential effects of RF-EMF in order to study the possible interaction mechanism between RF-EMFs and living beings. Though this research was conducted only for in-vitro studies, it can be applied to in-vivo and epidemiology studies as well. Hence, as a direct outcome of this research, more efficient RF-EMF exposure prediction tools can be developed, in order to improve the quality of epidemiological studies and the long-term laboratory experiments using whole organisms (in-vivo). As a direct outcome of this research, more efficient prediction tools can be developed, reducing the environmental exposure and enhancing the quality of life using more raw data. More research is essential in order to understand whether and how some attributes (e.g. frequency, SAR, exposure time, power flux density) affect the prediction of effects/no-effects in plants. The difference between classification and clustering may not seem pronounced. Nevertheless, these two algorithms are fundamentally different, as the classification is a form of supervised learning while clustering is a form of unsupervised learning. In general, classification and clustering display to be a promising tool for weak radio-frequency radiation effect prediction on plants.

Machine learning technique also could be used to incorporate data from field observations in which appropriate variables are taken with an identical methodology (e.g. field strength, SAR, radiation frequency, damage types found, species affected, distance to radiation source etc.), however, more experiment records are needed for that analysis. Even without a thorough knowledge of plants or RF-EMF, it is promising to use machine learning algorithm in bioelectromagnetics domain. Nonetheless, its limitation is that it demands a large number of data to provide adequate results [2] and the quality of the predictions depends on the dataset. However, the results obtained by our study shows only 4% of fluctuation among correctly classified percentage, proving that the results are significant. Besides, the sample size of reported 169 experimental case studies, perhaps low significant in a statistical sense, nonetheless, this analysis still provides useful insight of using Machine Learning in Bioelectromagnetics domain. This investigation should be further analyzed with a bigger sample size (more data) in the future.

VI. Conclusion

Using mobile phone has triumphed as it has become a crucial part of our society, as it serves as a social and informative tool. Big data analytics and machine learning techniques allow a high-level extraction of awareness from

raw data which offers remarkable opportunities to predict the future trends and outcomes of the impact of handheld devices and its impacts on living beings. There is no single study that adequately covers machine learning concept in bioelectromagnetics domain. However, this paper has analyzed prediction models and their accuracies in order to identify the best classification algorithm to be used in analyzing data that shows environmental effects from mobile phones and base station radiation on plants. This analysis has helped us understand different types of attributes that have shown effects and impact on plants. Random Forest algorithm stands out producing better prediction among all the classification algorithms. Using K-Means clustering algorithm we found, Pea, Mungbean, and Duckweeds plants are more sensitive to RF-EMF. Moreover, this study shows that K-mean clustering algorithms can be successively used to predict conditions will RF-EMF exposure of given plant species are affected by RF-EMF (bio-effects). Moreover, this paper also illustrates the development of optimal attribute selection protocol to identifies key parameters that should be used when designing the in-vitro practical standardized experimental protocols. Our results show that clustering and classification are, in general, a promising prediction tool which can be practically used to predict plant effect changes due to non-thermal weak RF-EMF. Although this research was conducted only data from in-vitro studies, it can be applied to in-vivo and epidemiology studies. Hence, as a direct outcome of this research, more efficient RF-EMF exposure prediction tools can be developed, in order to improve the quality of epidemiological studies and the long-term laboratory experiments using whole organisms (in-vivo). Machine learning is an important tool, to validate some mysterious occurrences in bioelectromagnetics domains, which is not used by the community, so far, however, in the future, this might play a fundamental role to predict the potential effects of environment on plants and to study the possible interaction mechanism between RF-EMF and living being.

VII. Declaration of Interest

The authors declare no conflict of interest.

References

[1] M. N. Halgamuge, "Review: Weak radiofrequency radiation exposure from mobile phone radiation on plants," *Electromagnetic Biology and Medicine*, vol. 26, no. 2, pp. 213–235, September 2016.

[2] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," *Artif Intell Med.*, vol. 23, no. 1, pp. 89–109, August 2001.

[3] P. Cortez and A. Silva, "Using data mining to predict secondary school student performance," in *Proceedings of 5th Annual Future Business Technology Conference, EUROSIS,*, 2008.

[4] N. R. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine Learning*, vol. 29, pp. 131–163, 1997.

[5] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, ser. ISBN 978-0137903955. Prentice Hall, 1995, vol. 2.

[6] R. Kohavi, "The power of decision tables," in *8th European Conference on Machine Learning*. Springer, 1995, pp. 174–189.

[7] W. W. Cohen, "Fast effective rule induction," in *Twelfth International Conference on Machine Learning*. Morgan Kaufmann, 1995, pp. 115–123.

[8] R. Holte, "Very simple classification rules perform well on most commonly used datasets," *Machine Learning*, vol. 11, pp. 63–91, 1993.

[9] R. Quinlan, "C4.5: Programs for machine learning." Morgan Kaufmann Publishers, 1993.

[10] T. K. Ho, "Random decision forests," in *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal, August 1995, pp. 278–282.

[11] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[12] B. Reed, "The height of a random binary search tree," *Journal of the ACM*, vol. 50, no. 3, pp. 306–332, 2003.

[13] I. Witten and E. Frank, "Data mining: Practical machine learning tools and techniquescal machine learning tools and techniques with java implementations," Morgan Kaufmann, San Francisco, CA., 2005.

[14] I. Witten, E. Frank, and M. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, January 2011.

[15] K. Bailey, "Numerical taxonomy and cluster analysis," *Typologies and Taxonomies*, p. 34.

[16] A. V. Solanki, "Data mining techniques using WEKA classification for sickle cell disease," *International Journal of Computer Science and Information Technology*, vol. 5, no. 4, pp. 5857–5860, 2014.

[17] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, 1967, pp. 281–297.

[18] A. McCallum, K. Nigam, and L. H. Ungar, "Efficient clustering of high dimensional data sets with application to reference matching," in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2000, pp. 169–178.

[19] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society, Series B.*, vol. 39, no. 1, pp. 1–38, 1977.

[20] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York: Springer, 2009.

[21] E. Frank, "Machine learning with weka," University of Waikato, New Zealand, Tech. Rep., 1999.

[22] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[23] M. N. Halgamuge, S. K. Yak, and J. L. Eberhardt, "Reduced growth of soybean seedlings after exposure to weak microwave radiation from GSM 900 mobile phone and base station," *Bioelectromagnetics*, vol. 36, no. 2, pp. 87–95, February 2015.

[24] M. Kukar, "Estimating the reliability of classifications and cost sensitive combining of different machine learning methods," Ph.D. dissertation, Faculty of Computer and Information Science, University of Ljubljana, , Ljubljana, Slovenia, 2001.

[25] C. Wanigasooriya, M. N. Halgamuge, and A. Mohamad, "The analyzes of anticancer drug sensitivity of lung cancer cell lines by using machine learning clustering techniques," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 8, no. 9, September 2017.

[26] A. Gupta, A. Mohammad, A. Syed, and M. N. Halgamuge, "A comparative study of classification algorithms using data mining: Crime and accidents in denver city the usa," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 7, no. 7, pp. 374 – 381, August 2016.

[27] M. N. Halgamuge, S. M. Guru, and A. Jennings, "Energy efficient cluster formation in wireless sensor networks," in *Proceedings of IEEE International Conference on Telecommunication (ICT'03)*, vol. 2. Tahity, French Polynesia: IEEE, March 2003, pp. 1571–1576.

[28] C. Brewster, P. Farmer, J. Manners, and M. N. Halgamuge, "An incremental approach to model based clustering and segmentation," in *Proceedings of International Conference on Fuzzy Systems and Knowledge Discovery (FSKD'02)*, Singapore, November 2002, pp. 586–590.

[29] A. K. Dubey, U. Gupta, and S. Jain, "Analysis of k-means clustering approach on the breast cancer wisconsin dataset," *Int J Comput Assist Radiol Surg*, June 2016.

[30] A. Dharmarajan and T. Velmurugan, "Lung cancer data analysis by k-means and farthest first clustering algorithms," *Indian Journal of Science and Technology*, vol. 8, no. 15, July 2015.