Guava Trees Disease Monitoring Using the Integration of Machine Learning and Predictive Analytics

Mahmoud Elsayed*, Nourhan Hassan†, Marina Maher‡,
Nouran Waleed§, Rehab Reda¶, Haitham Sharaf Eldin ‖, and Hassan Mostafa **

***Abstract*—The increase in population, food demand, and the pollution levels of the environment are considered major problems of this era. For these reasons, the traditional ways of farming are no longer suitable for early and accurate detection of biotic stress. Recently, precision agriculture has been extensively used as a potential solution for the aforementioned problems using high resolution optical sensors and data analysis methods that are able to cope with the resolution, size and complexity of the signals from these sensors. In this paper, several methods of machine learning have been utilized in order to study pests, their types, population, and agricultural conditions in terms of soil and climate for some crops such as potatoes, guava, and cotton, which are among the main Egyptian crops. In the process of obtaining a suitable estimate of insects population affecting each of the aforementioned crops, a hardware model control, based on the results provided by the predictive analysis, an estimate of the electromagnetic force is applied to the cultivated areas to get rid of the pests as well as giving a background to farmers about the possibility of infecting a crop such as Potato with Late Blight, according to climatic conditions.**

***Index Terms*—Precision Agriculture(PA), Machine Learning(ML), Guava trees, Model Predictive Control(MPC),Model Predictive Analysis(MPA)**

I. Introduction

Agriculture occupies an important sector in Egypt. About a quarter of the Egyptian workers are farmers. These farmers help in providing about 17% of the local Egyptian income [10]. It is one of the most important sectors that need to pay great attention from the technological point of view due to its problems related to relying heavily on the human factor. Here we present precision cultivation as a solution to the problem at hand, which is the difficulty of controlling some pests that have a destructive effect on the main seasonal crops in Egypt. In addition to accurate calculations of the quantity of crops planted, proper planning of land areas and calculation of the amount of nutrients needed by the soil, the fight against unwanted pests is done in ways that are safer on the health of crops than the use of insecticides and that through the electromagnetic force that targets the eggs of unwanted insects and eliminates them.

Sensor systems and data interpretation sensor systems can provide high resolution data concerning agricultural crop stands. By anticipating the importance of precision agriculture and its promising effects, the focus has been on choosing the best machine learning models to give an estimate of the number of insects and parasites likely to be present in the guava tree crop, and predicting Late Blight affects potatoes crops according to the information offered by sensor systems. Recent research has been done in Egypt by Ahmed Tageldin [1] about the usage of precision agriculture in reducing the potato blight and the cotton leafworm effects on potato, and cotton crops respectively. The work has been conducted in a similar environment in Egypt, therefore, his research was the starting point for this work.

II. Dataset and Setup

The datasets used in this work for guava trees, cotton leafworm are a research study by Dr. Haitham Sharaf a professor at Cairo University, faculty of Agriculture, in order to study the relationship between fertilizers and the number and types of parasites and insects affecting crops, which contains data of weather conditions inside a controlled greenhouse system where Guava trees are planted. The weather data has been collected manually for the past two years.

Seasonal abundance of mealybug species and their associated predators and parasitoid on guava trees in Egypt have been surveyed. The survey has been conducted in Giza, Egypt spanning two years (Jan. 2014 to Dec. 2015). Fifteen plants have been randomly chosen and five leaves have been biweekly collected, Each leave has been picked either from the middle of the inspected trees or the four cardinal directions.The dataset includes four pest mealybug species, four predator species, two parasitoids attack predators, four primary parasitoid species, and one hyper parasitoid as shown in Table I.

TABLE I. Insects species and names

Species Classification	Insect name	Index
Pest mealybug	Ferrisia virgata	1
	Icerya seychellarum	2
	Icerya purchasi	3
	Planococcus citri	4
Predator	Scymnus syriacus	5
	Cydonia vicina	6
	Chrysoperla carnea	7
	Rodalia	8
Parasitoids attack predators	Homalotylus vicinus	9
	Homalotyloidea	10
Primary parasitoid species	Leptomastix	11
	Leptomastidae	12
	Gyranusoidea indica	13
	Aenasius	14
Hyper parasitoid	Chartocerus subaeneus	15

The most dominant insect species are: the mealybug Ferrisia virgata, the predator Scymnus syriacus, the parasitoid

978-1-6654-2157-7/21/$31.00 ©2021 IEEE

that attacks Homalotylus vicinus, the primary parasitoid of mealybugs, and the hyperparasitoid, Chartocerus subaeneus . F.virgata is the first recorded insect on guava trees in Egypt and the primary parasitoid Aenasius spieces is recorded for the first time in the Egyptian guava harvest. Based on the present study, it is clear that the surveyed natural enemies have played a weak role in controlling the mealybugs attacking guava trees, due to the effect of the parasitoids attacking predators and the hyperparasitoids. It is found that the role played by these bio-agents could have been of great importance in the process of predictive analysis and studying the relationship between these insects to accurately help reduce the harmful effect of insects or parasites that cannot be overcome by other insects.Consequently, in finding the correlation and statistical distribution between insects, some of which were among the most important factors for conducting a good predictive analysis.

The dataset used for cotton Late Blight analysis was collected by Dr. Mohamed Fahim from the department of Plant Pathology at Cairo University. The set contains data of weather conditions within potato areas that were collected during four consecutive seasons, i.e. 2002 to 2006. The weather data were recorded manually in the Badasshin region.

III. METHODS

Several methods of machine learning have been utilized for these analyses such as support vector machines ,Decision Tree, Random Forest, Gradient Boosting algorithms for classification (supervised learning); k-means and self-organizing maps for clustering (unsupervised learning). These methods are able to calculate both linear and non-linear models, require few statistical assumptions and adapt flexibly to a wide range of data characteristics. The collected dataset includes the insects popularity on a biweekly basis, meanwhile the weather data is available on a daily basis. Thus some preprocessing on the data has to be done to make use of the available daily weather data.Here, unsupervised learning techniques played the solution to predict the relationship between climatic and soil features and population of insects. Mainly two methods have been used: predicting the numbers of each insect separately on a daily basis, and predicting the numbers of all insects simultaneously together on a daily basis as well. In each of these methods one or a combination of the following techniques have been applied: linear interpolation, data imputation using mean/median value, data imputation using KNN, k-fold cross-validation, standard scaler, and min-max scaler as shown in fig. 2.

Fig. 1 shows a diagram of the algorithm followed in this paper. It starts with data preparation which includes preprocessing, features inclusion/exclusion, and merging of data of different nature, the two main methods used for predicting the numbers of insects come next. Fig. 2 is considered to be a subclass of fig. 1 as it illustrates the data preparation block in detail. Numbers in fig. 2 are being referred to in results tables II, IV, and V as methods IDs.

Linear interpolation means predicting intermediate values between two endpoints assuming they follow a linear relationship [2]. Data imputation using the mean/median method replaces missing values between two endpoints using their mean/median value [3]. Imputation Using KNN uses 'feature similarity' to predict the values of any new data points. The new point is assigned a value based on how closely it resembles the points in the training set [4]. K-fold cross-validation splits the input data to user-defined k-folds, then the model uses (k-1)folds for the sake of training and one fold for testing [5]. Standard scaler scales data to have a mean value of 0 and standard deviation value of 1 [6]. Minmax scaler scales the data to be in the range of [0,1] [6].

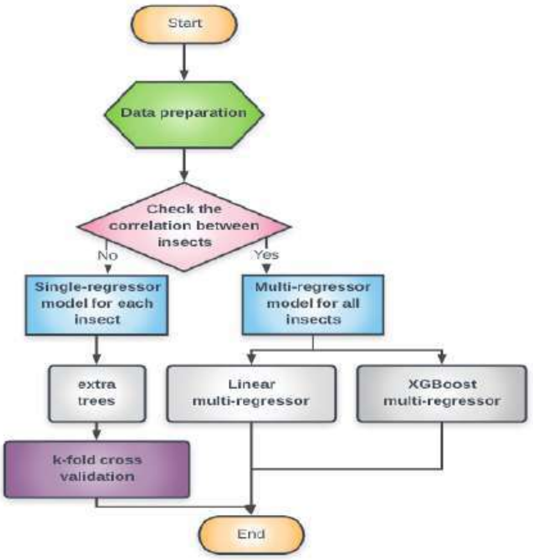
Start
Data preparation
Check the correlation between insects
No
Yes
Single-regressor model for each insect
Multi-regressor model for all insects
extra trees
Linear multi-regressor
XGBoost multi-regressor
k-fold cross validation
End

Fig. 1: Algorithm diagram

Traditional linear regression and linear discriminant analysis, are based on predefined distributions and model assumptions. These methods are applicable without loss in accuracy only for data that complies with these demands. Therefore, the field of application is limited for these kinds of analysis methods. Advanced methods of machine learning, such as k-means, Decision Tree and Gradient Boosting algorithms require less prior information and are applicable to a wider range of tasks as they derive the underlying distributions and model assumptions implicitly from training data. This capacity to adapt to almost any kind of data makes them well suited for tasks with limited prior knowledge about a suitable interpretation model or complex data characteristics like non-linearity, non-Gaussian noise and outliers. These methods are particularly suitable for the interpretation of data from sensors because the included noise factors may be compensated by a sufficient amount of representative training data. The resulting models are able to reduce the influence of the unknown variability significantly and provide more reliable decisions.

Choosing one of the two main methods is based on the degree of correlation between insects. fig. 3 shows that, there is much variance in correlation values between insects, and there is no general trend either strong correlation or weak correlation between all insects. Thus the performance of the two methods has to be compared.

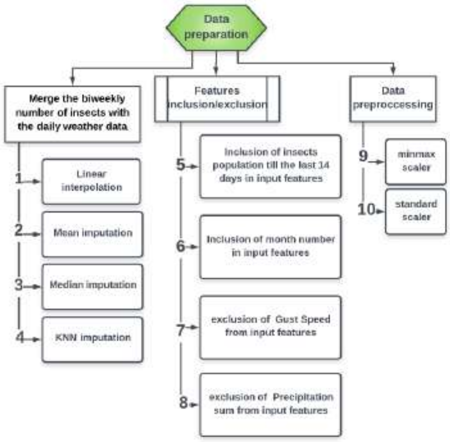
Data preparation
Merge the biweekly number of insects with the daily weather data
Features inclusion/exclusion
Data preproccessing
1
Linear interpolation
2
Mean imputation
3
Median imputation
4
KNN imputation
5
Inclusion of insects population till the last 14 days in input features
6
Inclusion of month number in input features
7
exclusion of Gust Speed from input features
8
exclusion of Precipitation sum from input features
9
minmax scaler
10
standard scaler

Fig. 2: methods used for data preparation

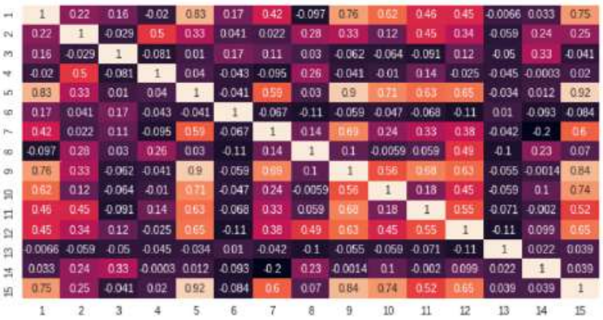

Fig. 3: Correlation between insect species

A. *Predicting numbers of each insect separately*

In this method, a separate machine learning model is trained for each insect individually. First, the biweekly number of insects is merged with the daily weather data. To do so, linear interpolation is used per every two weeks to estimate the number of insects daily as shown in fig. 4.

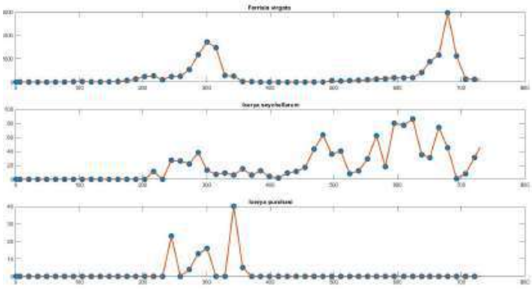

Fig. 4: Linear interpolation of biweekly number of insects

In fig. 4, the blue circles show the actual bi-weekly numbers of three random insects, and the orange lines show the estimated numbers of these insects on every day. Second, the following machine learning models: extra trees, and random forest have been tested without tuning to pick the most suitable one out of them. Finally, some tuning is made to the selected machine learning algorithm. It turns out that the extra trees regressor suits all insects the best. Extra trees is a well-known ensemble method, and it is considered to be substantial improvement on simple decision trees. Basically, an extra trees regressor combines multiple decision trees, each of them is trained over the whole sample, and the final prediction value is the arithmetic mean of the predictions of all the decision trees. It is worth noting that, unlike random forest, extra trees pass the whole sample of input features to each decision tree, but random forest sub samples the input features with replacement. Another difference between random forest and extra trees lies in the selection of cut points. Extra trees chooses a random split, while random forest uses the optimum split [7]. Using extra trees regressor only, the error percentage value has been large for all insects, so some tuning has been made to improve the performance of the model.

Linear interpolation and imputation using mean value were used interchangeably because their performance has been comparable. Also, the result of using extra trees without tuning is not mentioned because error percentages exceeded 100%. The next part compares the results of the methods in table II. In one of the techniques used, the number of insects available till a certain day in the past is added as input feature to analyze the effect of including recent output in the input features on the prediction accuracy of the number of insects in the future. Table. II shows the results of the methods mentioned in fig 2. Each column in table II represents an insect, each cell contains the root mean square error percentage for each insect, and each row represents one method of the available methods.

It is obvious that the performance is comparable for all methods except for the last one. Moreover, no single method out of the first three methods is the best for all insects, instead, one method can be good for one insect, and another method can be better for another insect. Though this is not the case with the last method in the results, it works better than other methods with all insects.

To investigate this difference between the first three methods and the last one, the features importance scores have been analyzed using both the standard scaler (method 1 is used as a candidate for it) and the minmax scaler (method 4 is the only candidate for it). The analysis is done on a sample of 10 random insects to compare their scores. Table III shows only scores for three insects for readability purpose, besides the scores are approximately the same for all insects, thus no need to include features scores of all insects.

B. *Predicting numbers of all insects together*

In this method, correlation between insects population is considered. The strong correlation between some insects as shown in fig. 3 is considered the motivation behind predicting all insects simultaneously together using one model.

TABLE II. results of single regressors

	Insect index																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		
RMSE percentage	3.40	5.80	4.90	4.20	1.90	5.40	5.70	10.3	3.40	6.20	4.30	4.80	7.00	8.60	2.20	1,10	method
	1.73	5.24	6.73	2.50	2.03	6.90	3.80	7.80	3.00	4.96	5.06	8.07	7.34	9.76	1.73	1,5,10	
	1.90	3.49	3.04	1.80	1.24	3.91	4.01	6.61	2.07	5.28	3.19	3.79	4.90	6.30	1.25	1,2,10	
	0.31	0.89	0.46	0.20	0.19	0.48	0.87	0.92	0.37	0.54	0.43	0.75	0.52	0.75	0.19	1,2,9	

TABLE III. Features importance

Insect 1				Insect 2			
Minmax scaler		Standard scaler		Minmax scaler		Standard scaler	
Temp.(°C),low	0.123414	Temp.(°C),low	0.120835	Wind(km/h),avg	0.175701	Wind(km/h),avg	0.182498
Temp.(°C),avg	0.116195	Temp.(°C),avg	0.098230	Temp.(°C),low	0.122033	Temp.(°C),low	0.115217
Temp.(°C),high	0.098598	Temp.(°C),high	0.096773	Temp.(°C),avg	0.073381	Temp.(°C),avg	0.082685
Wind(km/h),avg	0.093936	Wind(km/h),avg	0.085529	DewPoint (°C),avg	0.062944	DewPoint (°C),high	0.060008
DewPoint (°C),low	0.068870	DewPoint (°C),low	0.071940	Temp.(°C),high	0.062361	DewPoint (°C),avg	0.059153
DewPoint (°C),avg	0.061190	SeaLevelPressure(hPa),avg	0.058799	Wind(km/h),high	0.057184	Temp.(°C),high	0.056060
SeaLevelPressure(hPa),low	0.061079	SeaLevelPressure(hPa),low	0.057235	DewPoint (°C),high	0.054073	Wind(km/h),high	0.053150
Humidity (%),avg	0.049439	Humidity (%),low	0.056906	SeaLevelPressure(hPa),high	0.052058	SeaLevelPressure(hPa),high	0.051526
Humidity (%),low	0.047817	DewPoint (°C),avg	0.056667	SeaLevelPressure(hPa),avg	0.044710	SeaLevelPressure(hPa),avg	0.043547
SeaLevelPressure(hPa),high	0.046545	Humidity (%),avg	0.055286	DewPoint (°C),low	0.042314	Visibility (km),avg	0.041279
Wind(km/h),low	0.044108	SeaLevelPressure(hPa),high	0.046607	Humidity (%),avg	0.042089	SeaLevelPressure(hPa),low	0.040460
SeaLevelPressure(hPa),avg	0.043947	Wind(km/h),low	0.044842	Visibility (km),avg	0.039240	Humidity (%),avg	0.039934
Humidity (%),high	0.032638	DewPoint (°C),high	0.042361	SeaLevelPressure(hPa),low	0.035601	DewPoint (°C),low	0.038730
DewPoint (°C),high	0.031729	Wind(km/h),high	0.030739	Humidity (%),high	0.035124	Humidity (%),low	0.034587
Wind(km/h),high	0.027796	Humidity (%),high	0.027231	Humidity (%),low	0.034743	Wind(km/h),low	0.034097
Visibility (km),avg	0.022703	Visibility (km),low	0.024036	Wind(km/h),low	0.033208	Humidity (%),high	0.033782
Visibility (km),low	0.022331	Visibility (km),avg	0.019371	Visibility (km),low	0.028999	Visibility (km),low	0.029636
Precip.(mm),Precip.(mm)	0.005284	Precip.(mm),Precip.(mm)	0.004367	Visibility (km),high	0.002690	Visibility (km),high	0.002258
Visibility (km),high	0.002384	Visibility (km),high	0.002246	Precip.(mm),Precip.(mm)	0.001547	Precip.(mm),Precip.(mm)	0.001393

First, the biweekly number of insects is merged with the daily weather data as shown in fig. 2 using the following two methods:
Median imputation per every two weeks to estimate the number of insects daily, and imputation using kNN per every two weeks to estimate the number of insects daily. fig. 5 shows the Scymnus syriacus population before and after KNN imputation is applied.

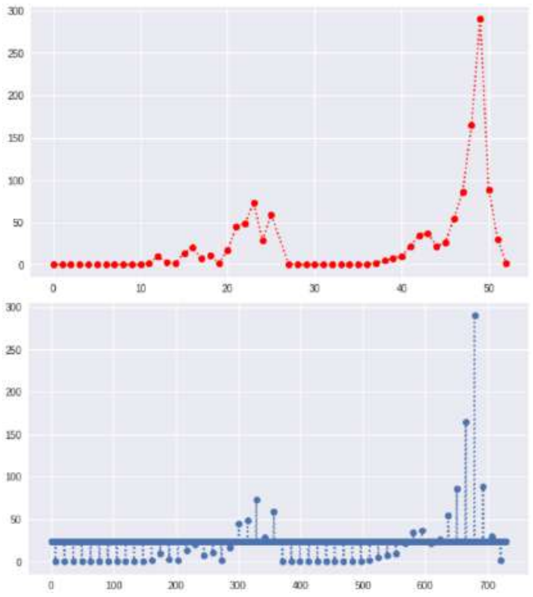

Fig. 5: distribution of Scymnus syriacus population before and after KNN imputation

Second, two types of multi machine learning regression models have been tested: linear multi-regressor, Xgboost multi-regresssor [8]. Both of the two strategies consist of fitting only one regressor for all targets. This is a simple strategy for extending regressors that do not natively support multi-target regression. Multiple regression analysis [9] is a powerful technique used for predicting the unknown value of multiple variables from the known value of two or more variables. More precisely, multiple regression analysis helps in predicting the value of multiple output values. The Linear and XGBoost models were implemented using two different scalers: standard scaler, and minmax scaler.
The following part compares the results of the methods used in multi XGB multi-regressor in table IV and the same methods used in multi linear multi-regressor in table V. Each column represent an insect, and each cell contains the root mean square error percentage for each insect, and each row represent one method of four different methods.

Again the performance of the multi regressor models using standarad scaler is totally different than the performance using minmax scaler as shown in result tables.
This difference has been analyzed in the first subsection of the methods. Here, the same investigation process is repeated to confirm that the same conclusion is reached. Besides studying the features importance scores of the XGB multi regressor, weights of the linear XGB multi regressor model are studied as well. Fig 6 shows similar scores as ones from the single regressor results, and fig 7 shows that expected important features have higher weights. This investigation supports the conclusion reached in the previous subsection as discussed in

TABLE IV. results of multi regressors

	Insect index																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		
RMSE percentage	20.7	151	19.8	451	16.9	5.43	44.2	163	15.6	2.19	27.9	44.5	34.1	25.	11.2	4,10,6,7,8	method
	6.83	145	6.14	463	17.6	9.77	3.76	155	14.5	0.74	32.0	41.7	55.6	36.9	20.6	3,10,6,7,8	
	0.19	0.70	0.17	0.82	0.18	0.18	0.28	0.64	0.18	0.21	0.19	0.38	0.26	0.21	0.18	4,9,6,7,8	
	0.12	0.98	0.12	0.86	0.14	0.11	0.15	0.55	0.15	0.11	0.15	0.29	0.10	0.12	0.13	3,9,6,7,8	

TABLE V. results of linear multiregressor

	Insect index																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		
RMSE percentage	15.24	124.8	8.64	451.3	14.68	9.36	38.16	162	14.23	2.16	26.35	37.5	7.96	8.97	10.79	4,10, 6,7,8	method
	31.14	124.3	18.5	215.1	39.05	20.15	30.51	127	37.6	14,73	52.61	61.3	19.48	29.9	31.6	3,10, 6,7,8	
	0.03	0.68	0.02	0.74	0.03	0.01	0.20	0.52	0.04	0.003	0.04	0.18	0.008	0.02	0.02	4,9,6,7,8	
	1.53	9.76	0.75	9.09	1.82	0.61	1.99	7.22	2.07	0.58	2.23	4.28	0.54	1.25	1.35	3,9,6,7,8	

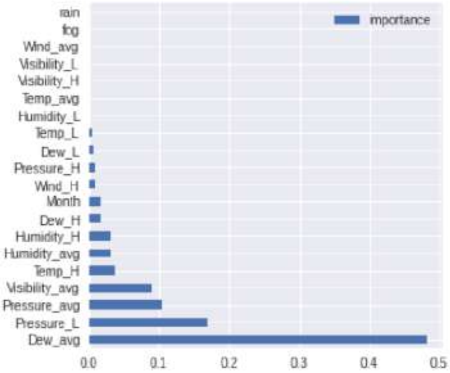

Fig. 6: XGBoost multi-regressor feature importance With standard scaler.

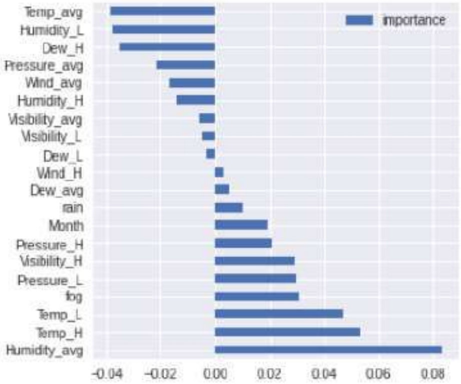

Fig. 7: linear multi regressor coefficients with minmax scaler

Conclusion below.

IV. Conclusion

The choice of an optimal data analysis method strongly depends on the problem. Therefore, it is not possible to provide general recommendations, but some criteria could help to identify applicable algorithms. There are some basic properties that are important for a method selection: the number of features, the number of training samples, information about the data distributions and quality of data itself. Based on this research, the following conclusions have been drew out; the most important features using either the minmax or the standard scaler have very similar scores, the most important features scores are very similar for all insects regardless of the scaler used, and the maximum and minimum features scores for a single insect are approximately the same using either the minmax scaler or the standard scaler. Moreover, results of both of single regressors and multi regressors models are comparable with high accuracy. Besides, both of XGB and linear multi regressors have similar prediction accuracy; however, the linear multiregressor model is slightly better. To sum up, It is clear that both scalers give comparable importance to the same features. Thus the the high variance in results of the minmax scaler and the standard scaler is due only to the way that each scaler of them handles the data.

References

[1] Ahmed Tageldin, ”Integration of machine learning and predictive analytics in agriculture to optimize plant disease detection and treatment in Egypt,” M.S. thesis, Department of Electrical and Computer Engineering, TUM, 2020.

[2] ScienceDirect, ”Linear Interpolation,” [online]. Available: https://www.sciencedirect.com/topics/engineering/linear-interpolation

[3] Brownlee, J. ”Data preparation for machine learning: Data cleaning, feature selection, and data transforms in Python,” 2020.

[4] García-Laencina, P. J., Sancho-Gómez, J., Figueiras-Vidal, A. R., & Verleysen, M. ”K nearest neighbours with mutual information for simultaneous classification and missing data imputation.” Neurocomputing, 72(7-9), 1483-1493. 2009.

[5] Carnege Mellon University, ”Cross Validation,” [online]. Available: https://www.cs.cmu.edu/~schneide/tut5/node42.html

[6] kanoki, ”Sklearn data Pre-Processing using Standard and Minmax scaler,”[online]. Available: https://kanoki.org/2020/06/01/sklearn-data-pre-processing-using-standard-and-minmax-scaler/

[7] QuantDare, ”What is the difference between extra trees and random forest,”[online]. Available: https://quantdare.com/what-is-the-difference-between-extra-trees-and-random-forest/

[8] Chen, T., & Guestrin, C. ”XGBoost: A Scalable Tree Boosting System.” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016.

[9] Makgahlela, M., Mäntysaari, E., Strandén, I., Koivula, M., Nielsen, U., Sillanpää, M., & Juga, J. ”Across breed multi-trait random regression genomic predictions in the nordic red dairy cattle.” Journal of Animal Breeding and Genetics, 130(1), 10-19. 2012.

[10] Middle East Institute, ”Greening the Egyptian Economy with Agriculture ,” [online]. Available: https://www.mei.edu/publications/greening-egyptian-economy-agriculture