BJES Basrah Journal for Engineering Sciences

Basrah Journal for Engineering Sciences, Vol. 25, No. 2, (2025), 114-119

Original Article

Journal homepage: https://bjes.edu.iq/bjes
ISSN (Online): 23118385, ISSN (Print): 18146120

Early Prediction of Tomato Diseases in Iraqi Agriculture using AI-Based Temperature and Humidity Data: A Random Forest Approach

Marwan Adnan Al-Ahbabi [1]*, Jamal Al-Tuwaijari [2], Awf A. Al-Jbory [3]

[1,2] *Department of Computer Science, College of Science, University of Diyala, Diyala, Iraq*
[3] *Department of Plant Protection College of Agriculture, University of Tikrit, Salah_Aldin, Iraq*
E-mail addresses: scicomphd222309@uodiyala.edu.iq , dr.altuwaijari@uodiyala.edu.iq , awfabd91@tu.edu.iq

Article Info

Article history:
Received: 22 February 2025
Revised: 3 May 2025
Accepted: 5 May 2025
Published: 31 December 2025

Keywords:
Iraqi agriculture,
Machine learning,
Artificial intelligent,
Tomato diseases.

https://doi.org/10.33971/bjes.25.2.13

Abstract

Plants and agriculture are important in Iraq and the world because they are among the essential basics of life; their importance lies in several fields, such as industry and food. Plant diseases are the first direct influence on plant production and the Iraqi economy. The primary contribution of this work is developing an efficient early warning system for tomato plant diseases based on readily available environmental data, demonstrating the usefulness of machine learning methods in real agricultural environments. This research investigates the use of artificial intelligence (AI) for the early prediction of tomato diseases in Iraqi agriculture, based on temperature and humidity data collected from Salah al-Din Governorate. Two major diseases were studied: Tomato Yellow Leaf Curl Virus (TYLCV) and late blight. The data were pre-processed and used to train predictive models, including linear regression and Random Forest Regressor (RFR). Results show that RFR outperformed linear regression, achieving a lower Root Mean Square Error (RMSE) of 0.053852 and a Mean Absolute Error (MAE) of 0.45000, indicating its superior accuracy in predicting disease occurrences.

1. Introduction

Iraqi agriculture is one of the essential sectors of the Iraqi economy. One of the most critical factors affecting agriculture is the climatic factors closely linked to it [1]. One of the most direct influences on plants is plant diseases. Temperature and humidity are essential in spreading various plant diseases that spread when suitable environmental conditions are available [2]. Tomato yellow leaf curl virus (TYLCV) is one of the most common plant diseases that affects plants, especially tomatoes. The whitefly spreads this virus, which transmits it in agricultural fields. Temperature and humidity significantly influence the spread of the whitefly and the virus [3]. Late blight is a dangerous disease that affects plants, especially tomatoes. A fungus called Phytophthora infestans causes this disease. It spreads in high humidity and moderate temperatures, leading to the growth and spread of these fungi [4]. Modern technologies such as artificial intelligence, machine learning, and the Internet of Things contribute to improving disease prediction capabilities, increasing accuracy, saving time, and increasing efficiency [5].

2. Related studies

This study showed that the integration between artificial intelligence and late blight disease that affects plants requires an integration process between machine learning algorithms and climate data and may rely on artificial neural networks to analyze the relationship between the emergence of diseases and climate data with high accuracy of prediction. This study also indicated accelerating and delaying the emergence and spread of diseases linked to environmental factors such as temperature and humidity [6]. The researchers used the importance of combining machine learning and automation in their study. These diverse models can be used to determine the relationship between environmental factors such as temperature and humidity and the spread of various agricultural diseases. They also explained the process of improving agricultural resource efficiency [7]. The authors used in their study the importance of integrating artificial intelligence with environmental conditions to control the prediction of various diseases. This study focused specifically on diseases affecting tomatoes. It proved that using machine-learning algorithms could contribute to analyzing climate data and identifying the conditions that precede the spread of diseases in tomatoes. It also pointed out the importance of building innovative systems linked to sensing technologies through which environmental data such as temperature and humidity are collected and linking this data to diseases affecting tomatoes. These models were considered effective in predicting diseases and making appropriate decisions before diseases occur and spread [8]. Babu et al. demonstrated the importance of combining IoT sensors and machine learning algorithms in smart precision agriculture. Through this, an integrated system was developed that integrates continuous monitoring of plant field soil using IoT technologies and some predictive models based on hybrid integrated machine learning algorithms. This enables high accuracy in predicting tomato diseases, especially those associated with environmental

factors such as late blight and TYLCV. In addition, the study demonstrated the role of ecological data analysis in the agricultural decision-making process, as the predictive system can send early recommendations and alerts to farmers to alert them to the possibility of diseases to prevent them [9]. In this study, a comparison is made between different machine learning techniques such as neural networks, machine support vectors, and other algorithms based on criteria such as accuracy, ease of application, and processing time, emphasizing that the integration between artificial intelligence and the Internet of Things provides an excellent opportunity to develop intelligent systems and make them more effective and integrated [10]. This study refers to using artificial intelligence to predict environmental factors that affect the agricultural production process. The environmental factors that were relied upon were relative humidity and evaporation. Deep learning algorithms such as convolutional neural networks and recurrent neural networks were used. Agricultural procedures were linked to the outputs of artificial intelligence, controlling irrigation, improving ventilation, and controlling tomato diseases such as late blight [11]. Joshi and Deepti et al. [12] present techniques based on artificial intelligence and linear regression models for evaluating the effect of conditions such as temperature and relative humidity on the crop's life cycle before the emergence of late blight disease. The study showed the great potential of using predictive models in the process of predicting the emergence of diseases. Machine learning and statistical models were used in the prediction process. Several algorithms were used, such as SVM and artificial neural networks, which proved more accurate than linear regression.

3. Materials and methods

This section explains the process of building a prediction system based on artificial intelligence techniques to predict plant diseases before they occur. First, the structure of building the system will be presented, from the first step, the data collection process, to the last step, the system evaluation step. Fig. 1 below shows the main structure of building the system.

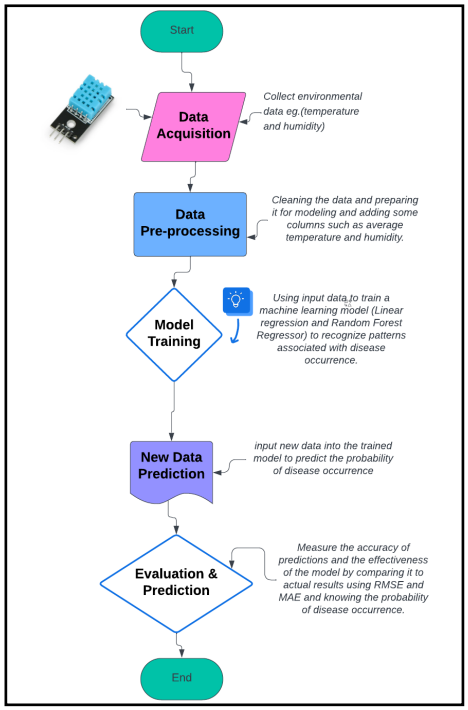

Fig. 1 The main flowchart of the prediction system.

The data pre-processing and model development were conducted using Google Colab, a cloud-based Python environment. Python libraries such as Pandas, NumPy, Scikit-learn, and Matplotlib were employed for data manipulation, analysis, modeling, and visualization. The data were pre-processed by normalizing temperature and humidity values and splitting them into training (80%) and testing (20%) sets. We evaluated model performance using RMSE, MAE, and R-squared metrics.

3.1. Data acquisition

Data were collected from Iraqi fields. Data include average temperature, humidity, and date and disease occurrence. The recorded temperature and humidity values for each documented day included morning and evening measurements. The highest and lowest values were obtained for each day, and a daily average was calculated. This approach ensured that the environmental conditions reflected the full range of temperature and humidity variations throughout the day.

3.2. Data pre-processing

At this stage, the data is prepared before entering the next stage. Some additional columns derived from the original data are added, such as calculating the disease incubation period, the average disease incubation temperature, the average humidity, the difference between each day and the average disease incubation temperature, and the difference between each day and the average humidity for the incubation period for each of the two diseases: the first disease is tomato yellow leaf curl (TYLCV) and the second disease is late blight.

3.3. Model training

At this stage, the data is divided into 80% training and 20% testing. After that, the models are trained, and two different models are used for two different algorithms: linear regression and the random forest algorithm regressor.

3.4. Linear regression

Linear regression was used to predict disease occurrence based on environmental information specific to each disease using independent and disease-related variables. The following equation was used to calculate linear regression [13]:

$$M = N_0 + N_1 S + E \quad (1)$$

M = dependent variable.
S = independent variable.
N_0 = the constant representing M expected value when S = 0.
N_1 = the regression coefficient.
E = random error.

3.5. Random forest regressor

The random forest algorithm is considered one of the most famous and influential algorithms in prediction in general. Several trees form this algorithm. Each tree makes a specific decision based on the data entered into it, and then all the trees are collected to make the final decision for the algorithm and make the prediction. In this algorithm, trees are calculated using the following equation [14]:

$$Y = \frac{1}{N}\sum_{i=1}^{N} H_i(x) \quad (2)$$

Where Y : is the final predicted value, N : is the total number of decision trees, $H_i(x)$ the prediction from the samples of the tree.

3.6. New data prediction

At this stage, data was obtained through a request submitted to the Ministry of Agriculture to obtain environmental data through a station affiliated with it close to the fields from which the data was collected in Salah al-Din Governorate/Balad District. Data included the average temperature, average relative humidity, and their date. This data makes it possible to verify the correctness of the program's operation and predict plant diseases before they occur.

3.7. Evaluation

In the final stage, the model is evaluated using several measures to measure the accuracy of the prediction, such as Root Mean Squared Error (RMSE) & Mean Absolute Error (MAE).

3. Result and discussion

In this section, the results obtained by the prediction system are displayed. Diseases are predicted 7 days before they occur, the incubation period for diseases affecting tomatoes. The results of the two diseases will be displayed. The first disease is TYLCV, and the second disease is late blight. Table 1 shows a sample of data collected in agricultural fields in Salah al-Din Governorate, Balad District, for the first disease TYLCV.

Table 1. Sample of data TYLCV disease.

Date	Average Temperature	Average Humidity	Disease 1 onset
14/3/2024	19.49	41.52	0
15/3/2024	20.08	50.185	0
16/3/2024	19.16	38.975	0
17/3/2024	18.56	44.115	0
18/3/2024	18.475	58.16	0
19/3/2024	17.14	75.05	0
20/3/2024	15.73	61.715	0
21/3/2024	17.89	52.69	1

Table 2 shows a sample of data collected in agricultural fields in Salah al-Din Governorate, Balad District, for the second disease, late blight.

The data was trained using temperature and humidity and the incubation average of 7 days. Two different algorithms were used: Linear Regression (LR) and Random Forest Regressor (RFR) algorithm. Various measures were used to measure the prediction accuracy and error rate of the two algorithms, RMSE and MAE. The results of the algorithms and the disease incidence will be shown in the following. Figure 2 shows the results of linear regression in predicting the TYLCV disease; the RMSE ratio was equal to 0.1370, the MAE ratio was equal to 0.095, and the probability of the disease occurrence was 72.58%.

Table 2. Sample of data Late blight disease.

Date	Average Temperature	Average Humidity	Disease 2 onset
17/3/2024	18.56	44.115	0
18/3/2024	18.475	58.16	0
19/3/2024	17.14	75.05	0
20/3/2024	15.73	61.715	0
21/3/2024	17.89	52.69	0
22/3/2024	15.06	70.85	0
23/3/2024	17.54	64.145	0
24/3/2024	15.97	73.88	0
25/3/2024	17.985	69.7125	1

```
The model for Disease1 (TYLCV):
Test RMSE: 0.137063
Test MAE: 0.095175
------------------------------

Sample predictions for Disease1 (TYLCV):
    Actual  Predicted
8        0   0.226948
4        0   0.000000
13       0   0.000000
3        0   0.153752

Prediction for the last day in the new data:
Date: 2025-03-21
Prediction for Disease1 (TYLCV): 72.58%
```

Fig. 2 Results of prediction of the TYLCV disease using LR.

Figure 3 shows the results of LR in predicting the late blight disease, the RMSE ratio was equal to 0.1370, the MAE ratio was equal to 0.095, and the probability of the disease occurrence was 72.58%.

```
The model for Disease2 (Late Blight):
Test RMSE: 0.204180
Test MAE: 0.131124
------------------------------

Sample predictions for Disease2 (Late Blight):
    Actual  Predicted
8        0   0.059294
4        0   0.000000
13       0   0.398486
3        0   0.066717

Prediction for the last day in the new data:
Date: 2025-03-25
Prediction for Disease2 (Late Blight): 58.19%
```

Fig. 3 Results of prediction of the Late blight disease using LR.

To further clarify the results, Fig. 4 shows the difference between the actual and predicted values for the disease TYLCV using linear regression.

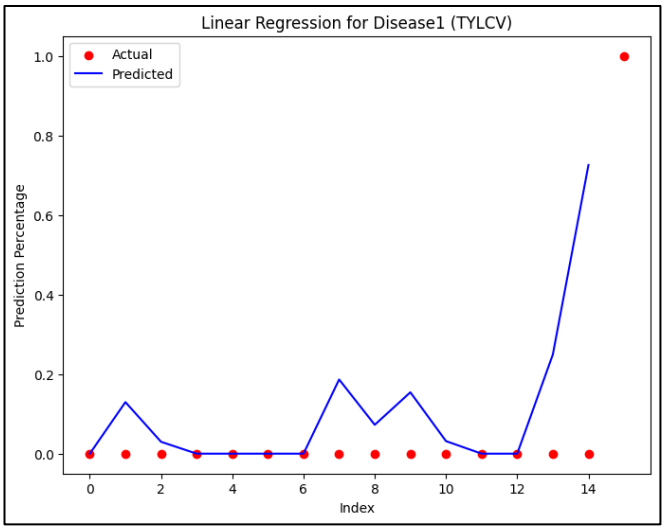

Fig. 4 The difference between actual and predicted values of the late blight disease.

In addition, Fig. 5 shows the difference between the actual and predicted values for late blight disease using LR.

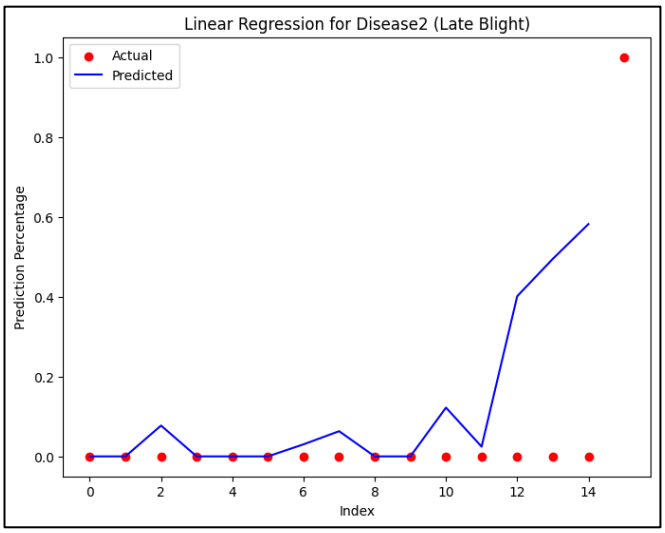

Fig. 5 The difference between actual and predicted values of the late blight disease.

Figure 6 shows the results of RFR in predicting the late blight disease, the RMSE ratio was equal to 0.080, the MAE ratio was equal to 0.045, and the probability of the disease occurrence was 82.00%.

```
The model for Disease1 (TYLCV):
Test RMSE: 0.080623
Test MAE: 0.045000
------------------------------

Sample predictions for Disease1 (TYLCV):
    Actual  Predicted
8        0       0.02
4        0       0.00
13       0       0.16
3        0       0.00

Prediction for the last day in the new data:
Date: 2025-03-21
Prediction for Disease1 (TYLCV): 82.00%
```

Fig. 6 Results of prediction of the TYLCV disease using RFR.

Figure 7 shows the results of RFR in predicting the late blight disease; the RMSE ratio was equal to 0.053, the MAE ratio was equal to 0.045, and the probability of the disease occurrence was 68.00 %.

```
The model for Disease2 (Late Blight):
Test RMSE: 0.053852
Test MAE: 0.045000
------------------------------

Sample predictions for Disease2 (Late Blight):
    Actual  Predicted
8        0       0.06
4        0       0.04
13       0       0.08
3        0       0.00

Prediction for the last day in the new data:
Date: 2025-03-25
Prediction for Disease2 (Late Blight): 68.00%
```

Fig. 7 Results of prediction of the Late blight disease using RFR.

Figure 8 shows the difference between the actual and predicted values for TYLCV disease using the RFR algorithm to clarify the results further. Fig. 9 also shows the difference between the actual and predicted values for late blight disease using the RFR algorithm.

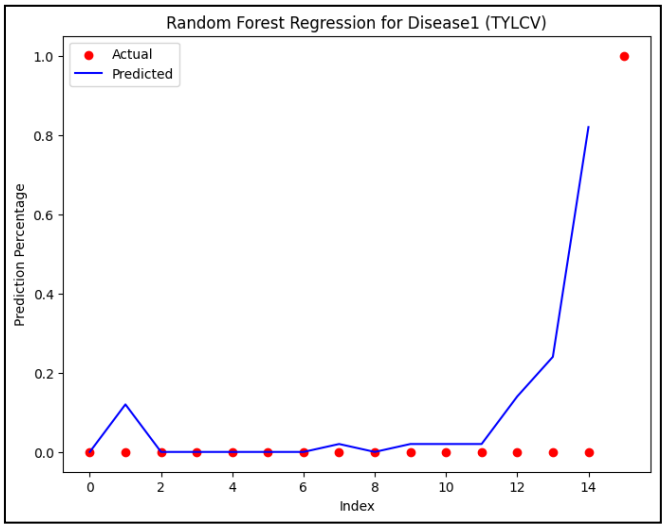

Fig. 8 The difference between actual and predicted values of the TYLCV disease.

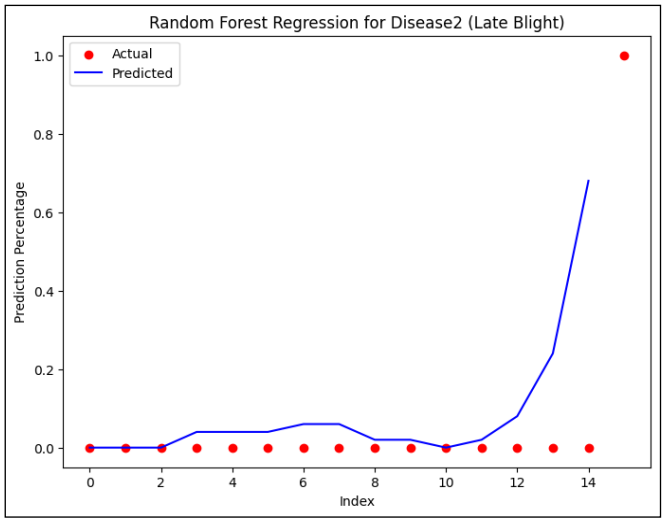

Fig. 9 The difference between actual and predicted values of the late blight disease.

After evaluating the two algorithms, the second algorithm, random forest regression, proved its effectiveness in predicting plant diseases. It is better than linear regression and has a much lower RMSE and MAE measure error rate.

To clarify the practical part, Fig. 10 shows the field in which the work was done and the DHT11 sensor that was used and connected to a Raspberry Pi device installed on a railway in the agricultural field to collect data and predict plant diseases.

Fig. 10 Demonstrates the use of equipment and data collection from Iraqi agricultural fields.

4. Conclusions

The ability to predict plant diseases proactively is a crucial step in improving agricultural production and reducing production losses for farmers and countries. The current research shows that predictive algorithms based on modern technologies such as artificial intelligence, such as the Random Forest Regressor (RFR) algorithm, achieve high accuracy and lower error compared to conventional algorithms such as linear regression, which has already been proven. RMSE 0.1370 for TYLCV disease using linear regression reached 0.080 for the RFR algorithm, as well as the MAE 0.095 for linear regression, while 0.045 for RFR, which has proven better performance in terms of reducing the error rate in prediction based on environmental data, such as average temperature and humidity. The Random Forest Regressor outperformed linear regression in all metrics, demonstrating its potential for real-time disease prediction. Future work will incorporate real-time data and additional AI models to enhance predictive accuracy. This study is limited by the regional data from Salah al-Din Governorate, and further studies are needed to assess the model's generalizability across different geographical locations.

Additionally, overfitting might occur due to the model's reliance on a relatively small dataset. This study highlights the potential of AI models, particularly Random Forest Regressor, for early predicting plant diseases in Iraq. The results demonstrate that the model could significantly improve agricultural practices by enabling proactive disease management. Future work will focus on integrating IoT-based real-time data, exploring other machine learning models, and expanding the dataset to enhance model accuracy. Future recommendations of this research also include increasing the scope of input data, establishing a broader early system for farmers, and cooperation between research institutions and farmers to enhance agricultural production in Iraq.

5. Acknowledgment

We would like to extend our thanks and appreciation to all those who contributed to the completion of this research, especially our esteemed university, the University of Diyala, and both the College of Science and the College of Agriculture. We would also like to extend our sincere thanks and appreciation to a group of farmers in Salah al-Din Governorate/Balad District for their cooperation in completing this research, and we would like to thank everyone who contributed to the completion directly or indirectly.

References

[1] M. Abdaki, A. Al-Iraqi, and R. M. Faisal, "Predicting long-term climate changes in Iraq," IOP Conference Series: Earth and Environmental Science, vol. 779, no. 1, p. 012053, 2021. https://doi.org/10.1088/1755-1315/779/1/012053

[2] K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine learning in agriculture: A review," Sensors, vol. 18, no. 8, p. 2674, 2018. https://doi.org/10.3390/s18082674

[3] M. Lapidot, O. Lachman, U. Zelinger, and A. D. Friedman, "Development of a scale for evaluation of Tomato yellow leaf curl virus resistance level in tomato plants," Phytopathology, vol. 96, no. 12, pp. 1404-1408, 2006. https://doi.org/10.1094/PHYTO-96-1404

[4] D. C. Erwin and O. K. Ribeiro, Phytophthora diseases worldwide. St. Paul, MN, USA: American Phytopathological Society, 1996, p. 562.

[5] P. Delfani, P. D. S. S. V, S. S. S. S, and R. S. S. S, "Integrative approaches in modern agriculture: IoT, ML and AI for disease forecasting amidst climate change," Precision Agriculture, pp. 1-25, 2024. https://doi.org/10.1007/s11119-024-10164-7

[6] G. Fenu and F. M. Malloci, "Artificial intelligence technique in crop disease forecasting: A case study on potato late blight prediction," in Proceedings of the 12th KES International Conference on Intelligent Decision Technologies (KES-IDT 2020), Singapore, 2020, pp. 79-89. https://doi.org/10.1007/978-981-15-5925-9_7

[7] D. Mishra and D. Deepa, "Automation and integration of growth monitoring in plants (with disease prediction) and crop prediction," Materials Today: Proceedings, vol. 43, pp. 3922-3927, 2021. https://doi.org/10.1016/j.matpr.2021.01.973

[8] S. A. Wagle and R. Harikrishnan, "Prediction of tomato plant disease with meteorological condition and artificial intelligence," ECS Transactions, vol. 107, no. 1, p. 20377, 2022. https://doi.org/10.1149/10701.20377ecst

[9] G. R. Babu, M. Gokuldhev, and P. S. Brahmanandam, "Integrating IoT for soil monitoring and hybrid machine learning in predicting tomato crop disease in a typical South India station," Sensors, vol. 24, no. 19, p. 6177, 2024. https://doi.org/10.3390/s24196177

[10] R. R. Patil, S. Kumar, and R. Rani, "Comparison of artificial intelligence algorithms in plant disease prediction," Revue d'Intelligence Artificielle, vol. 36, no. 2, pp. 307-314, 2022. https://doi.org/10.18280/ria.360202

[11] D. Joshi, T. P. S. S. S, and A. K. S. S. S, "AI-based machine learning and multiple linear regression approach to simulate the effect of weather on the crop age at first appearance of potato late blight (Phytophthora infestans (Mont.) de Bary) disease," Potato Research, pp. 1-24, 2024. https://doi.org/10.1007/s11540-024-09795-0

[12] D.-H. Jung, T. S. S. S, S. H. S. S, and J. W. S. S, "A deep learning model to predict evapotranspiration and relative humidity for moisture control in tomato greenhouses," Agronomy, vol. 12, no. 9, p. 2169, 2022. https://doi.org/10.3390/agronomy12092169

[13] D. C. Montgomery, E. A. Peck, and G. G. Vining, Introduction to linear regression analysis, 6th ed. Hoboken, NJ, USA: John Wiley & Sons, 2021.

[14] B. A. Goldstein, E. C. Polley, and F. B. S. Briggs, "Random forests for genetic association studies," Statistical Applications in Genetics and Molecular Biology, vol. 10, no. 1, Article 32, 2011. https://doi.org/10.2202/1544-6115.1691