



Machine Learning Modelling of Effective Cumin (*Cuminum cyminum* L.) Genotype for Antipyretic Property Based on Secondary Metabolites

Shubhi Saxena

Department of Biomedical Engineering,
School of Engineering and Technology,
Mody University, Lakshmangarh, Sikar,
Rajasthan- 332311, India

Manish Kumar

Department of Biomedical Engineering,
School of Engineering and Technology,
Mody University, Lakshmangarh, Sikar,
Rajasthan- 332311, India

S N Saxena

ICAR-National Research Centre on Seed Spices,
Tabiji, Ajmer-305206, India

P Shakti Prakash

Department of Biomedical Engineering,
School of Engineering and Technology,
Mody University, Lakshmangarh, Sikar,
Rajasthan- 332311, India

Dolly Agarwal

ICAR-National Research Centre on Seed Spices,
Tabiji, Ajmer-305206, India

Dawa Doma Bhutia

Department of Biomedical Engineering,
School of Engineering and Technology,
Mody University, Lakshmangarh, Sikar,
Rajasthan- 332311, India

Mukesh Kumar Vishal

ICAR-National Research Centre on Seed Spices,
Tabiji, Ajmer-305206, India and Indian Institute
of Technology Bombay, Powai, Mumbai-400076

ABSTRACT

Cumin is a flowering plant that has been used as a spice in multicuisine. It has various health benefits such as weight loss, cholesterol control, anti-diabetes, and more. It also consists of dietary fibers, vitamin B, vitamin E, and minerals, especially iron and magnesium. The present study was conducted for the effective implementation of two tasks, the prediction of cumin genotype in which the inputs were total phenolic, flavonoid, and antioxidant content and the classification of anti-pyretic activity. MLP and other ML algorithm simulations were committed to executing the tasks of prediction and classification. The effectiveness of the proposed model was compared with the various classification and regression techniques like SVM, Naïve Bayes, KNN, Logistic Regression, and Decision Tree. Along with the mentioned task, this paper also exhibits the implementation of feature selection techniques like PCA in ML-based prediction and classification. It was found that MLP with PCA has outperformed other algorithms.

Keywords: Cumin (*Cuminum cyminum* L.), Forecasting, Genotypes, Machine Learning, Secondary Metabolites

INTRODUCTION

Cumin is an angiosperm that belongs to the family Apiaceae, i.e., native to Egypt but cultivated widely in India and middle Asia for its aromatic seeds and its oil used for medicinal purposes. It is widely used in Indian, German, Middle Eastern, Italian, and Mexican cooking as spice [1]. The economic part of the cumin plant is its seeds are used dry both as a whole and as ground form, Figure 1-a. It is produced in areas with arid and semi-arid climatic conditions, which are found in most Asian countries like India, China, Iran, and Indonesia. However, India is the largest producer of cumin which is 70% of the entire world's production. The main cumin growing belt in India includes the state of Rajasthan and Gujrat^{[2]-[4]}, these areas are depicted in the Figure 1-b. It is very important cash crop for the farmers in the arid and semi-arid regions of the India, and requires low relative humidity specially at the time of maturity. Cumin has various advantages in health care applications as it helps in improving digestion, boosting immunity, and preventing the growth of cancerous cells. Due to its good anti-inflammatory as well as antipyretic properties, it helps in reducing pain and fever. Also, it is a good source of antioxidants [5].

In our earlier study, widespread cumin genotypes were collected from different cultivating areas of India and evaluated for the presence of phenolics, flavonoids, antioxidants [6],[7] and fatty acids, as well as their essential oil content and constituents like cumin aldehyde, gamma-terpinene, alpha-pinene, etc. [8], [9]. It also evaluated the effectiveness of different solvent extracts of cumin seeds and leaves on the extraction of phenol, flavonoids, and antioxidant contents. Maximum phenolic contents were found in the methanol extracts irrespective of cumin genotype, and a minimum of total phenolic contents was observed in genotype RZ-19. The experimental evaluation of the genotypic variation for the therapeutic use of cumin been carried out by Agrawal et al. [2]. In their work, the cumin seeds of different genotypes i.e., GC-1, GC-4, RZ-19, RZ-209, RZ-341 were extracted using hexane and methanol as solvent. Further, these crude seed extracts were used as a drug to study various medicinal activities such as antipyretic properties, anti-diabetic, anti-fungal and glucose tolerance test [10]-[13][12], [13]. The *in vivo* observation revealed that the pure genotype of cumin was found to have more additional therapeutic properties.

Nowadays, Machine Learning algorithms are increasingly important in biology due to their precision, accuracy, and robustness as they capture the complexity and pattern of small to large datasets [20]. Therefore, in the present study, a model was designed using Artificial Neural Network (ANN), taking the data set based on the results obtained from the above-mentioned work. Using this model, a simulation environment was created to classify and predict the effective genotypes of cumin based on secondary metabolites. The ANN belongs to the family of Machine Learning (ML), which mimics the human brain and how it performs every other function [21]. It uses a learning model that can perform adjustments independently. In return, ML is a part of Artificial Intelligence (AI) [27]-[30], and it functions to allow software applications to become more accurate at predicting the outcomes without being formally programmed to do so. ANN uses a learning model that can perform adjustments independently. This tool is very useful for non-linear statistical data modelling [14]. The different experiments and modelling have been performed using ANN, including prediction and classification. Despite initial research, the genotypic and phenotypic characteristic prediction still requires more attention [15]-[17]. The contribution of current work is mentioned as, in this work, two major tasks were performed with the help of the ANN model. The first one is the prediction of genotype and the second one is the classification of the antipyretic activity.

1. The developed technique will provide a parallel simulation environment that allows the researchers to predict not only genotype but also classify any therapeutic property of cumin.
2. With the use of such technique in performing any type of experiment, the process will become easy and less time-consuming and any unknown data can be tested using the existing data set as it will provide some level of probability and accuracy in the results.

This work, will strive to get a result that is satisfying and may be able to reduce the load of performing the experiments such that researchers who would want to further experiment on cumin can replace this part of the experimental phase with the help of our model.

MATERIAL AND METHODS

Data Set Formation and Algorithm Implementation

The data set was acquired from the experiment performed by Aggarwal, et. al. [2]. The author used the crude seed extract of cumin as a drug after extracting by hexane and methanol solvents. The experiment was performed on the albino mice by injecting the cumin seed crude extracts. The mice were first injected with carrageenan, which led to an increase in their body temperature (fever). To control their body temperature, the tone of the mice was given with allopathic medicine as control and others were given methanol and hexane crude seed extracts of different cumin genotypes as experimental drugs. The observations were taken at the time interval of 2 hours and were checked which among the extracts were able to bring the body temperature of the mice equal to or similar to the body temperature of the control mice. Most of the genotype extracts were able to control the body temperature of the mice but a little less than the control mice, it was found that genotype GC-1 of methanol extract was able to reach the level of control mice and worked effectively on it. Using this experiment and with the help of ML, this model was functioned for the classification of antipyretic properties based on the effectiveness of each genotype extract on defined time interval and prediction of genotype based on the different essential oil contents of all given genotype. For the prediction of genotype, total phenolic flavonoid, and anti-oxidant data present in the crude seed extract of cumin genotype were used and for the classification purpose, the data on how does the different

genotype of the cumin extracts has affected the mice on paw edema at a varied time interval were used. The availability of data was limited; hence to attain a certain level of accuracy, data pre-processing was performed.

For the step of data pre-processing, the raw data was converted into an efficient one and then performed the labelling step as presented in Table 1. Total phenolic, flavonoid and antioxidant content were labelled as 1, 2 and 3 (total phenolic content = 1, total flavonoid content = 2 and total antioxidant content = 3). The labelled genotypes were kept as RZ-209 = 1, RZ-19 = 2, RZ-341 = 3, GC-1 = 4 and GC-4 = 5. For validation purposes, antipyretic activity was classified into 0 = fever not controlled and 1 = fever controlled. The above-mentioned data set provided the total phenolic, flavonoid, and antioxidant content present in the hexane and methanol crude seed extract of cumin and classification of antipyretic activity also with methanol and hexane crude seed extract of cumin.

Proposed Multilayer Perceptron (MLP) Based Model for Estimation & Classification of Genotype and Antipyretic Activities

The multilayer perceptron neural network (MLPNN) has one or more hidden layers and performs better than a single-layer perceptron neural network in terms of computing efficiency [22]. Owing to this advantage this network is employed in current work. This model takes the data, *i.e.*, total labelled flavonoid, phenolic, antioxidant content, hexane and methanol extracts, are considered as input data for MLP network. Similarly, the labelled genotype and antipyretic activities are considered target data for the MLP network. Both input and target data are transformed into the space where it becomes linearly separable. Let's represent the input feature with X , as there are three inputs, so let the three inputs be named X_1 , X_2 , and X_3 .

$$X_1 = [X_{1_1}, X_{1_2}, X_{1_3} \dots X_{1_N}] \quad (1)$$

$$X_2 = [X_{2_1}, X_{2_2}, X_{2_3} \dots X_{2_N}] \quad (2)$$

$$X_3 = [X_{3_1}, X_{3_2}, X_{3_3} \dots X_{3_N}] \quad (3)$$

Hence input feature can be represented as:

$$X = [X_1, X_2, X_3] \quad (4)$$

where X_1 represents the total phenolic, flavonoid, and antioxidant content. X_2 represents the hexane extract's total phenolic, flavonoid, and antioxidant content. X_3 represents the total phenolic, flavonoid and antioxidant content in methanol extract. Similarly, let's define the target feature (output) with T , so let the two outputs be named T_1 and T_2 . Hence, the target features in the term of matrices as,

$$T_1 = [T_{1_1}, T_{1_2}, T_{1_3} \dots T_{1_N}] \quad (5)$$

$$T_2 = [T_{2_1}, T_{2_2}, T_{2_3} \dots T_{2_N}] \quad (6)$$

Hence the target feature can be written as:

$$T = [T_1, T_2] \quad (7)$$

The input and output layers are connected with hidden layers of MLP and each hidden layer is referred to as an intermediate layer with some weights and bias. The next step is the initialization of the weight, and the bias, where W stands for any random weight and B stands for the bias with iteration. It makes MLP network adaptive and it makes it as the universal approximator [18], [22]. The net network input can be calculated using;

$$Y_{in} = [X] * [W^T] + B \quad (8)$$

The output of the MLP network can be controlled by using a threshold or activation function and weight is adjusted significantly by applying Gradient descent algorithms like Least Means Squares, Recursive Least Means Squares, or Back-propagation techniques [19]. Learning Rate controls decides the speed of training and it is usually ranged between [0-1]. Training stops when the Mean Square Error reaches zero/minimum or when it reaches a predefined maximum number of epochs. Figure 3 demonstrates the basic structure of the proposed MLP network. The network output can be calculated using;

$$Y_{out} = \varphi[Y_{in}] \quad (9)$$

where φ represents the activation function, network error can be calculated as the difference between the target input and the network output.

$$\text{Network Error : } E = T - Y_{out} \quad (10)$$

With the help of backpropagation, It will perform weight and bias updating. Representation of backpropagation equation:

$$J \propto (W, B) \quad (11)$$

$$\frac{DE}{DB} = 0 \quad (12)$$

$$\frac{DE}{DW} = 0$$

where E represents error which is objective function and weight as well as bias will be updated on the basis of minimum error.

$$\text{weight update: } W_{new} = W_{old} - \alpha \frac{DE}{DW} \quad (13)$$

$$\text{bias update: } B_{new} = B_{old} - \alpha \frac{DE}{DB} \quad (14)$$

where alpha ' α ' stands for Learning Rate and for the testing purpose, the mathematical approach is,

$$\text{Testing} = \psi [\text{test_data} * (\text{Updated weights} + \text{updated bias})] \quad (15)$$

The efficiency of implemented algorithms is monitored on the various performance measures and they are,

$$MAE = \frac{1}{N} \sum_{j=1}^N y_j - \hat{y}_j \quad (16)$$

$$MSE = \frac{1}{N} \sum_{j=1}^N (y_i - \hat{y}_j)^2 \quad (17)$$

where, y_i is actual target value and $\hat{y}_j$ is output of the proposed MLP network. With the help of different classification tools various results were acquired. Two classes in the data-set “positive” and the “negative” class for the classification task. The terms mentioned above could be defined as follows. Accuracy was calculated using the formula;

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (18)$$

where True Positive (TP) refers to a sample belonging to the positive class, being correctly classified by a model. False Positive (FP) refers to a sample belonging to the negative class being incorrectly classified as belonging to the positive class. True Negative (TN) refers to a sample belonging to the negative class being correctly classified by the model. False Negative (FN) refers to a sample belonging to the positive class being incorrectly classified as belonging to the negative class.

RESULTS AND DISCUSSION

The Matlab software version 2021a has been used to perform the simulation task with an Intel(R) Core (TM) i5-9300HF CPU @ 2.40GHz, RAM 8.00 GB, graphic card 4GB. In the present study, initially, Neural Network Toolbox i.e., nntool (MATLAB) was used for the purpose of estimation and classification. In the proposed technique, as a target labelled genotype and classification of antipyretic activity data were considered with respect of inputs are total phenolic, flavonoid, and antioxidant content present in the hexane and methanol crude extract of cumin genotypes. It is a parallel operation that not only provide information of genotypes but same network will be used for the purpose of classification of anti-pyratic activity. Figure 4 represents the architecture of the proposed multi-layer perceptron network. It consists of three inputs, ten hidden layers and two output layers along with two major outputs that are types of genotypes and another one is antipyretic classification. The proposed model was then compared using different classification learner and regression learner tools of Matlab so that it can be compared with other state-of-the-art techniques such as other variants of Neural Network, Support Vector Machine [23], K' Nearest Neighbour [24], Naïve Bayes [25], Linear Regression and Decision Tree [26]. Principal Component Analysis (PCA) was applied with other ML techniques for better classification and regression. The use of these tools provided a better picture of the confusion matrix and comparison graphs of ROC, response plot, validation prediction v/s actual plot, and validation residual plot.

Figure 5 shows different regression graphs. The plot provides the information about how close the output from the model is to the actual target values. The straight line shows the actual output values, and the lines overlapping them in training, validation, and test graphs show the predicted outputs values. The accuracy of regression obtained in training $R^2 = 0.998$, for validation accuracy obtained is $R^2 = 0.97$ and for testing, accuracy appeared to be $R^2 = 0.999$. Hence it can be concluded that the overall accuracy of regression is $R^2 = 0.993$. As all values appears to be equal ~ 1 hence, this shows that the accuracy obtained is precise and high.

Figure 6 showing the performance plot and best validation performance appears to be 0.21574 at the epoch 24, where in graph blue represents the train, green is representing the validation and red represents the test. The graph also provides with the information about the plot of Mean Squared Error (MSE) vs Epoch. The learning and validation of the proposed MLP algorithm will stop at the 30th iteration, whereas by default, it was 1000. Figure 7 exhibiting the similarity between the actual and the output values, which is obtained after the completion of the training of the proposed MLP algorithm.

With the help of the classification tools of Matlab, the confusion matrix and ROC curve were acquired as a result for better assessment. Among all the classification techniques, Artificial Neural Network training showed the best result with an accuracy of 100%, and an AUC curve value of 1.00 and the least favourable results were shown by KNN tool with the accuracy of 46% and AUC curve of 0.46. Figure 8, Figure 9, and Figure 10 show the confusion matrix and ROC curve of other Neural networks with and without PCA and Support Vector Machine. The rest of the techniques used in ML classification are listed below in the Table 2, which provides information about the accuracy, prediction speed, training time, and Roc curve values. Among all the techniques used for classification purposes, the decision Tree of 2000 obs/sec showed the highest prediction speed, and the lowest prediction speed was offered by KNN and Naïve Bayes of 1300 obs/sec. Similarly, the fastest training time was shown by the MLP of 0.499 seconds, and the slowest training time recorded was by Naïve Bayes at 1.466 seconds.

With the help of the regression learner tool, different results of prediction were acquired, for this process results, was obtained by enabling and disabling PCA. The given below Figures gives information about the response plot, validation predicted v/s actual plot and validation residual plot. For the response plot, the blue dots represent the true values, the yellow dots represent the predicted values, and the red dots represent the errors. Similarly, in the validation prediction v/s actual plot, the straight black line represents the perfectly predicted values and the blue dots represents the observation taken by the model after training.

Other techniques that were used for regression purposes are mentioned in table 3, which provides the information about the RMS validation, R-square validation, MSE validation, MAE validation, prediction speed, and training time of each technique used. Among all the techniques used highest RMSE validation is shown by the SVM technique, and for the same, the lowest is shown by the Decision Tree. Similarly, the minimum prediction time is taken by linear regression with PCA of 1400 obs/sec, and the maximum time is taken by the Decision Tree of 8500 obs/sec. least training time was taken by Decision Tree of 0.55 secs and maximum training time was taken by MLP with PCA of 2.26 secs.

CONCLUSION

The present study highlights the successful implementation of two parallel tasks, among which the first task includes a prediction of the desired genotype where the input is taken was total phenolic, flavonoid, and anti-oxidant content and the second task was the classification of antipyretic property. The proposed work provides a simulated environment in which these two parameters can be estimated with the help of MLP or other ML algorithms. The efficiency of the proposed model was compared with the different classification and regression techniques like SVM, Naïve Bayes, KNN, Logistic Regression, and Decision Tree. Also, this paper showcases the application of feature selection techniques like PCA in ML-based forecasting as well as

classification. Among all the ML-based techniques MLP with PCA is outplayed the other supervised ML techniques.

References

- [1] Britannica, "Cumin," *Encyclopedia*, 2021. <https://www.britannica.com/plant/cumin> (accessed Apr. 29, 2022).
- [2] D. Agarwal, S. N. Saxena, P. . Dubey, B. . Mishra, K. Kant, and G. Lal, "Genotypic Variation in Pharmacological Potential of Seed Extracts of Cumin (*Cuminum cyminum* L.) Genotypes," *Indian J. Pharm. Sci.*, vol. 81, no. 5, pp. 946–954, 2019, [Online]. Available: www.ijpsonline.com.
- [3] Y. K. Sharma, K. Kant, R. K. Solanki, and R. P. Saxena, "Prevalence of cumin diseases on farmer ' s field : A survey of Rajasthan and Gujarat states Season Gujarat Gujarat," *Int. J. Seed Spices*, vol. 3(2), no. July, pp. 46–49, 2013.
- [4] H. Aslam, A. Zaheer, A. Imra, and M. Sajad, "Plant growth-promoting rhizobacteria (PGPR) for sustainable agriculture editors in chief," no. January 2015, p. 597, 2011, [Online]. Available: <https://www.researchgate.net/publication/271201449%0AMicrobial>.
- [5] S. S. Rathore, S. N. Saxena, and B. Singh, "Potential health benefits of major seed spices."
- [6] S. Mnif and S. Aifa, "Cumin (*Cuminum cyminum* L.) from traditional uses to Potential Biomedical Applications," *Chem. Biodivers.*, vol. 12, no. 5, pp. 733–742, 2015, doi: 10.1002/cbdv.201400305.
- [7] I. Bettaieb, S. Bourgou, W. A. Wannes, I. Hamrouni, F. Limam, and B. Marzouk, "Essential oils, phenolics, and antioxidant activities of different parts of cumin (*Cuminum cyminum* L.)," *J. Agric. Food Chem.*, vol. 58, no. 19, pp. 10410–10418, 2010, doi: 10.1021/jf102248j.
- [8] L. K. Sharma, D. Agarwal, S. S. Rathore, S. K. Malhotra, and S. N. Saxena, "Effect of cryogenic grinding on volatile and fatty oil constituents of cumin (*Cuminum cyminum* L.) genotypes," *J. Food Sci. Technol.*, vol. 53, no. 6, pp. 2827–2834, 2016, doi: 10.1007/s13197-016-2258-0.
- [9] K. Srinivasan, "Cumin (*Cuminum cyminum*) and black cumin (*Nigella sativa*) seeds: Traditional uses, chemical constituents, and nutraceutical effects," *Food Qual. Saf.*, vol. 2, no. 1, pp. 1–16, 2018, doi: 10.1093/fqsaf/fyx031.
- [10] R. S. Shetty, R. S. Singhal, and P. R. Kulkarni, "Antimicrobial properties of cumin," *World J. Microbiol. Biotechnol.*, vol. 10, no. 2, pp. 232–233, 1994, doi: 10.1007/BF00360896.
- [11] H. A. Hassan, M. M. Genaidy, M. S. Kamel, and S. F. Abdelwahab, "Synergistic antifungal activity of mixtures of clove, cumin and caraway essential oils and their major active components," *J. Herb. Med.*, vol. 24, no. August, p. 100399, 2020, doi: 10.1016/j.hermed.2020.100399.
- [12] A. Mansouri, A. S. Vahed, H. Shahdadi, F. Dashtban, and A. Arbabisarjou, "The effect of garlic and cumin on blood pressure and glycosylated hemoglobin in patients with type 2 diabetes," *Bali Med. J.*, vol. 7, no. 1, p. 156, 2018, doi: 10.15562/bmj.v7i1.849.
- [13] A. M. Al-saaedi, "Antipyretic Activity of The Aqueous Extract of Cumin (*Cuminum cyminum* L.) with Yeast Induced Pyrexia in Female Rats.," *Univ. Thi-Qar J. Sci.*, no. 1, pp. 33–35, 2021, doi: 10.32792/utq/utjsci/vol8/1/5.
- [14] E. T. Mohamad, R. S. Faradonbeh, D. J. Armaghani, M. Monjezi, and M. Z. A. Majid, "An optimized ANN model based on genetic algorithm for predicting ripping production," *Neural Comput. Appl.*, vol. 28, pp. 393–406, 2017, doi: 10.1007/s00521-016-2359-8.

- [15] S. Ghafouri-Fard, M. Taheri, M. D. Omrani, A. Daaee, and H. Mohammad-Rahimi, "Application of Artificial Neural Network for Prediction of Risk of Multiple Sclerosis Based on Single Nucleotide Polymorphism Genotypes," *J. Mol. Neurosci.*, vol. 70, no. 7, pp. 1081–1087, 2020, doi: 10.1007/s12031-020-01514-x.
- [16] S. Chesang, T. K. Muasya, and K. Ngeno, "Application of artificial neural network (ann) to evaluate extend of non-linearity among explanatory variables within and between genotypes and phenotypes," *Int. J. Vet. Sci. Anim. Husb.*, vol. 5, no. 3, pp. 21–26, 2020.
- [17] G. Rong, A. Mendez, E. Bou Assi, B. Zhao, and M. Sawan, "Artificial Intelligence in Healthcare: Review and Prediction Case Studies," *Engineering*, vol. 6, no. 3. Elsevier Ltd, pp. 291–301, Mar. 01, 2020, doi: 10.1016/j.eng.2019.08.015.
- [18] A. A. Heidari, H. Faris, I. Aljarah, and S. Mirjalili, "An efficient hybrid multilayer perceptron neural network with grasshopper optimization," *Soft Comput.*, vol. 23, no. 17, pp. 7941–7958, 2019, doi: 10.1007/s00500-018-3424-2.
- [19] M. Kumar and S. K. Mishra, "Jaya based functional link multilayer perceptron adaptive filter for Poisson noise suppression from X-ray images," *Multimed. Tools Appl.*, vol. 77, no. 18, pp. 24405–24425, 2018, doi: 10.1007/s11042-017-5592-y.
- [20] Greener, Joe G., Shaun M. Kandathil, Lewis Moffat, and David T. Jones. "A guide to machine learning for biologists." *Nature Reviews Molecular Cell Biology* 23, no. 1 (2022): 40-55.
- [21] Grossi, Enzo, and Massimo Buscema. "Introduction to artificial neural networks." *European journal of gastroenterology & hepatology* 19, no. 12 (2007): 1046-1054.
- [22] Madhiarasan, Manogaran, and Mohamed Louzazni. "Analysis of Artificial Neural Network: Architecture, Types, and Forecasting Applications." *Journal of Electrical and Computer Engineering* 2022 (2022).
- [23] Schölkopf, B., Williamson, R.C., Smola, A., Shawe-Taylor, J. and Platt, J., 1999. Support vector method for novelty detection. *Advances in neural information processing systems*, 12.
- [24] Uddin, Shahadat, Ibtisham Haque, Haohui Lu, Mohammad Ali Moni, and Ergun Gide. "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction." *Scientific Reports* 12, no. 1 (2022): 1-11.
- [25] Webb, Geoffrey I., Eamonn Keogh, and Risto Miikkulainen. "Naïve Bayes." *Encyclopedia of machine learning* 15 (2010): 713-714.
- [26] Charbuty, Bahzad, and Adnan Abdulazeez. "Classification based on decision tree algorithm for machine learning." *Journal of Applied Science and Technology Trends* 2, no. 01 (2021): 20-28.
- [27] Jakhar, Deepack, and Ishmeet Kaur. "Artificial intelligence, machine learning and deep learning: definitions and differences." *Clinical and experimental dermatology* 45, no. 1 (2020): 131-132.
- [28] Castiglioni, Isabella, Leonardo Rundo, Marina Codari, Giovanni Di Leo, Christian Salvatore, Matteo Interlenghi, Francesca Gallivanone, Andrea Cozzi, Natascha Claudia D'Amico, and Francesco Sardanelli. "AI applications to medical images: From machine learning to deep learning." *Physica Medica* 83 (2021): 9-24.
- [29] Miller, D. Douglas. "The medical AI insurgency: what physicians must know about data to practice with intelligent machines." *NPJ digital medicine* 2, no. 1 (2019): 1-5.
- [30] Kamruzzaman, M. M., Ibrahim Alrashdi, and Ali Alqazzaz. "New opportunities, challenges, and applications of edge-AI for connected healthcare in internet of medical things for smart cities." *Journal of Healthcare Engineering* 2022 (2022).

Table 1. Labelled data set [2]

Sl.no.	Total flavonoid, phenolic and antioxidant content	Hexane extract	Methanol extract	Genotypes	Antipyretic property	
					Hexane	Methanol
1.	Total Phenolic Content (TPC)	74.794	95.102	RZ-209	0	1
2.		47.546	70.464	RZ-19	0	1
3.		59.375	85.871	RZ 341	0	0
4.		79.956	84.366	GC-1	1	1
5.		56.298	92.435	GC-4	0	0
6.	Total Flavonoid Content (TFC)	30.666	53.278	RZ-209	0	1
7.		22.7	38.059	RZ-19	0	1
8.		22.859	55.073	RZ341	0	0
9.		25.825	46.289	GC-1	1	1
10.		24.276	41.158	GC-4	0	0
11.	Total Antioxidant Content (TAC)	19.93	19.93	RZ-209	0	1
12.		10.588	10.588	RZ-19	0	1
13.		12.566	12.566	RZ-341	0	0
14.		11.828	11.828	GC-1	1	1
15.		14.504	14.504	GC-4	0	0

Table 2. Different ML classification techniques

S.no.	ML Algorithms	Accuracy	Prediction speed	Training time	Area Under curve
1.	MLP	100%	~1600 obs/sec	0.499 sec	1.00
2.	MLP with PCA	100%	~1800 obs/sec	0.919 sec	1.00
3.	SVM	80%	~1600 obs/sec	1.209 sec	0.71
4.	KNN	80%	~1300 obs/sec	1.14 sec	0.89
5.	Naïve Bayes	82.7%	~1300 obs/sec	1.466 sec	0.96
6.	Logistic regression	80%	~1700 obs/sec	1.293 sec	0.74
7.	Decision tree	100%	~2000 obs / sec	1.1 sec	1.00

Table 3. Performance of regression tools

S.NO.	ML based Regression Techniques	RMSE validation	R-square validation	MSE validation	MAE validation	Prediction Speed	Training Time
1.	MLP without PCA	0.101	0.88	0.010	0.029	~8400 Obs/sec	1.256 sec
2	MLP with PCA	0.097	0.88	0.009	0.046	~1700 Obs/sec	2.269 sec
3	SVM without PCA	0.308	-0.16	0.095	0.252	~6000 Obs/sec	1 sec
4	SVM with PCA	0.293	-0.05	0.085	0.253	~ 2700 Obs/sec	0.777 sec
5	Linear Regression	0.293	-0.05	0.086	0.256	~6900 Obs/sec	0.654 sec
6	Linear Regression With PCA	0.294	-0.05	0.086	0.255	~1400 Obs/sec	1.913 sec
7	Decision Tree	0.081	0.92	0.006	0.048	~8500 Obs/sec	0.55 sec
8	Decision Tree with PCA	0.286	0.00	0.082	0.245	~2400 Obs/sec	0.647 sec

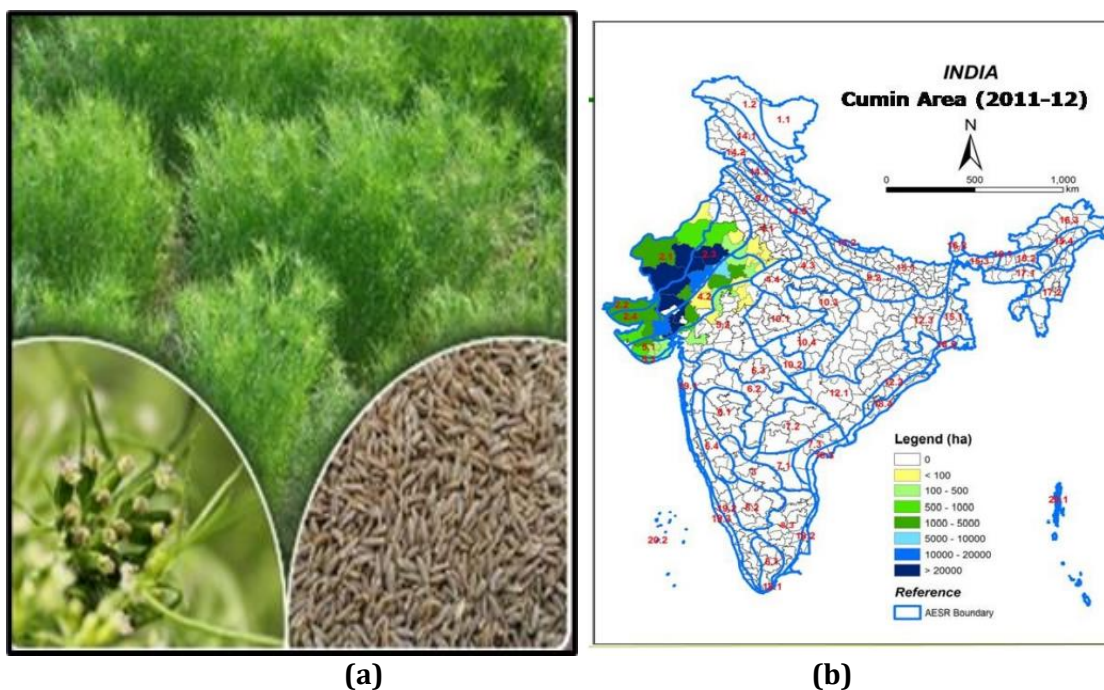


Figure 1. (a) Cumin-plant, flower and seed, (b) major cumin-growing regions of India



Figure 2. Cumin at the farmer's field, a low volume, and high-value seed spice, is a crucial cash crop for the livelihood of the farmers in arid and semi-arid regions in the western part of India.

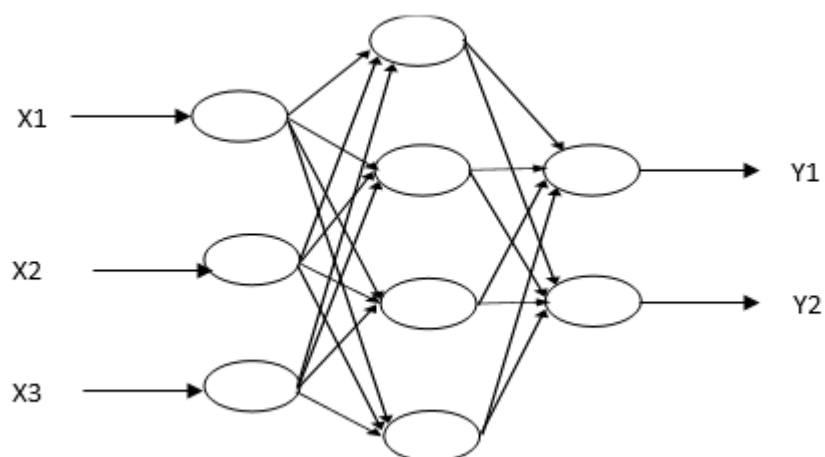


Figure 3. The architecture of Multi-Layer Perceptron

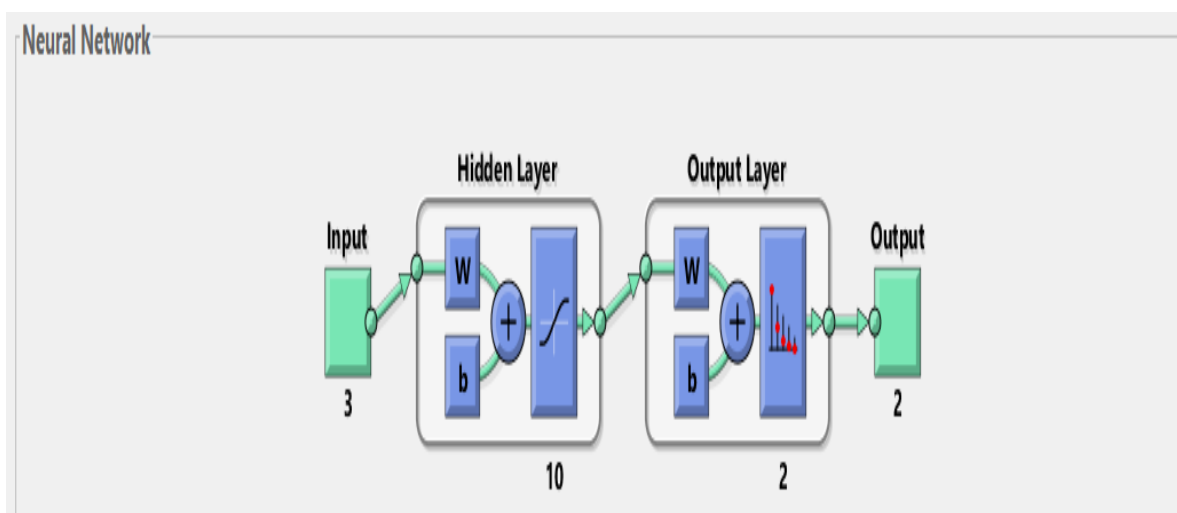


Figure 4. The architecture of the proposed multi-layer perceptron network

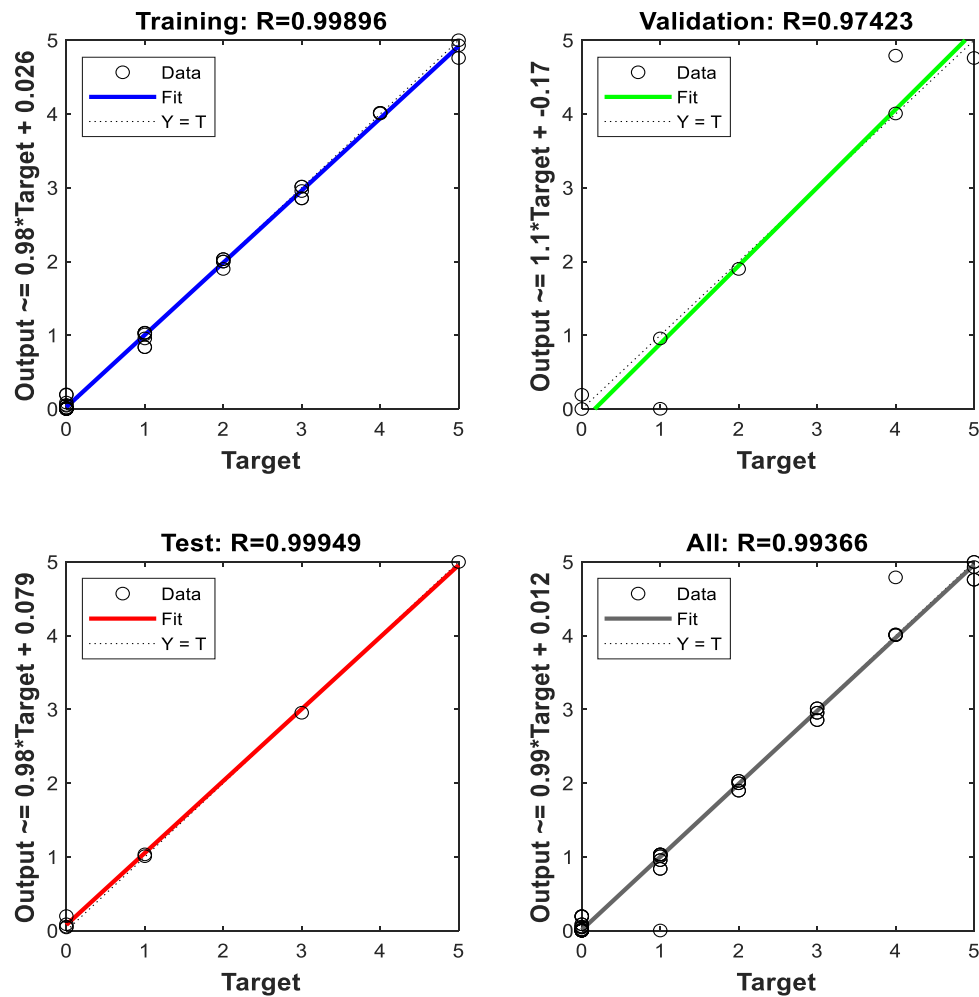


Figure 5. Regression plot

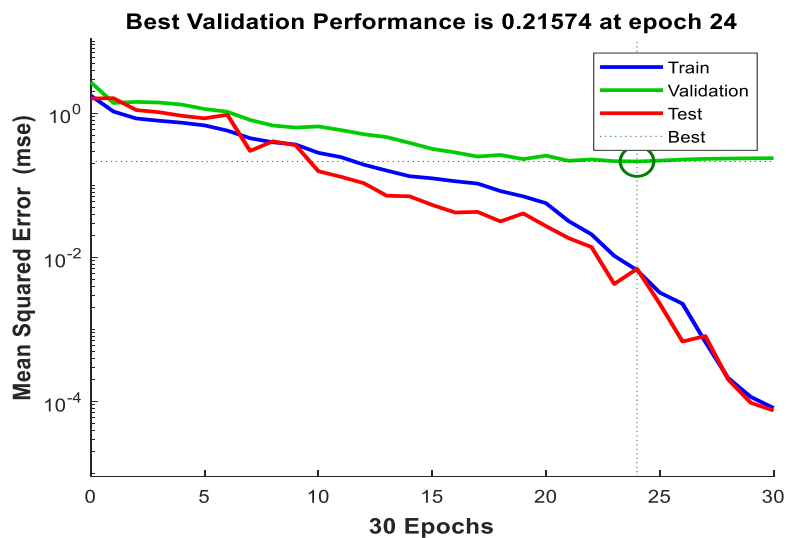


Figure 6. The performance plot

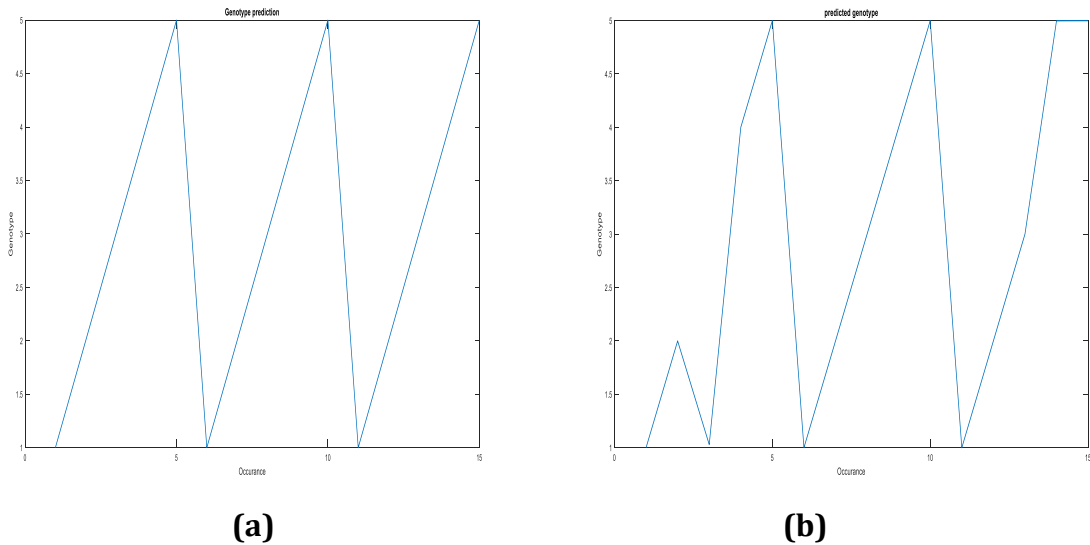


Figure 7. Graphs of (a) Actual and (b) Predicted Genotype

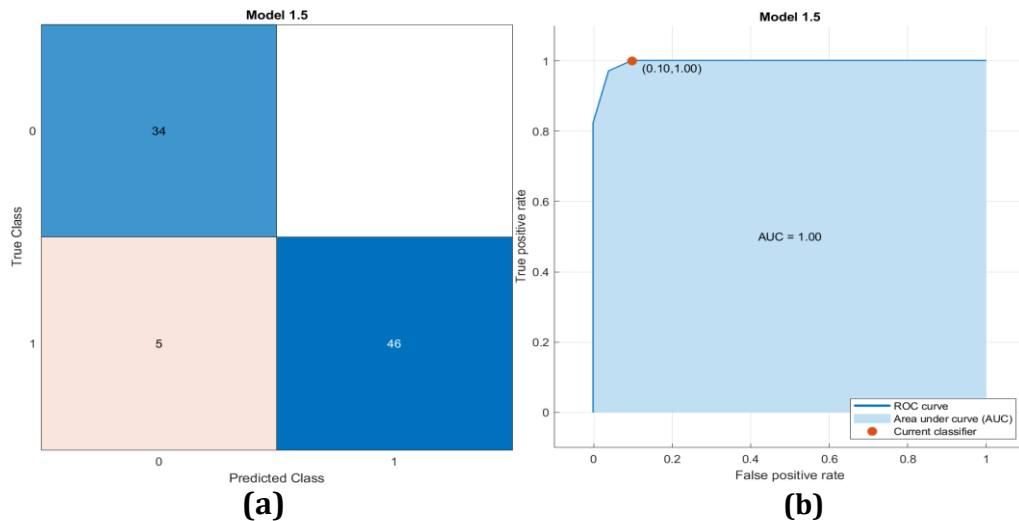


Figure 8. Plot of (a) confusion matrix of MLP (b) ROC curve of MLP with PCA

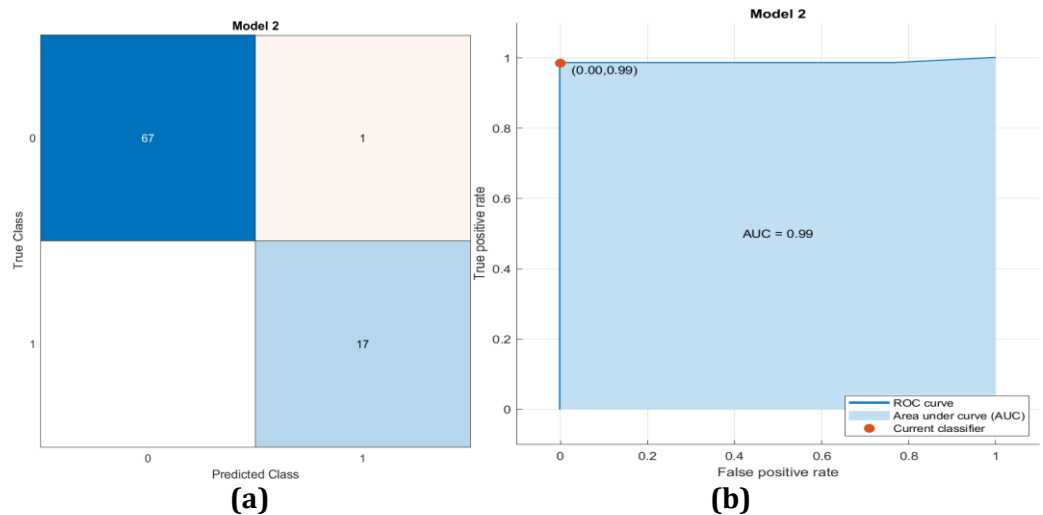


Figure 9. Plot of (a) Confusion matrix of MLP (b) ROC curve of MLP without PCA

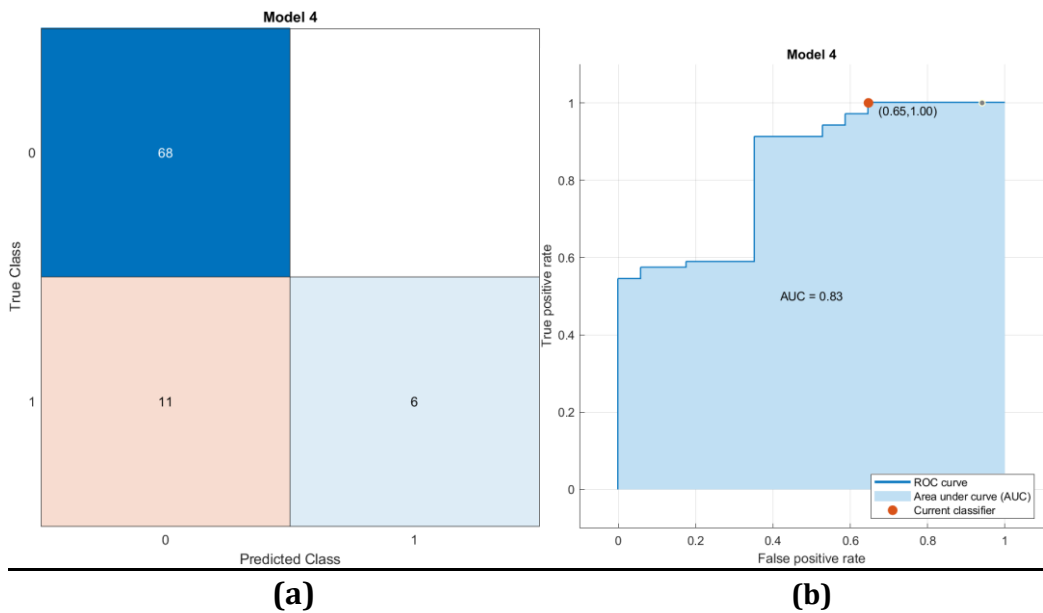


Figure 10. Plot of (a) confusion matrix of SVM (b) ROC curve of SVM with PCA

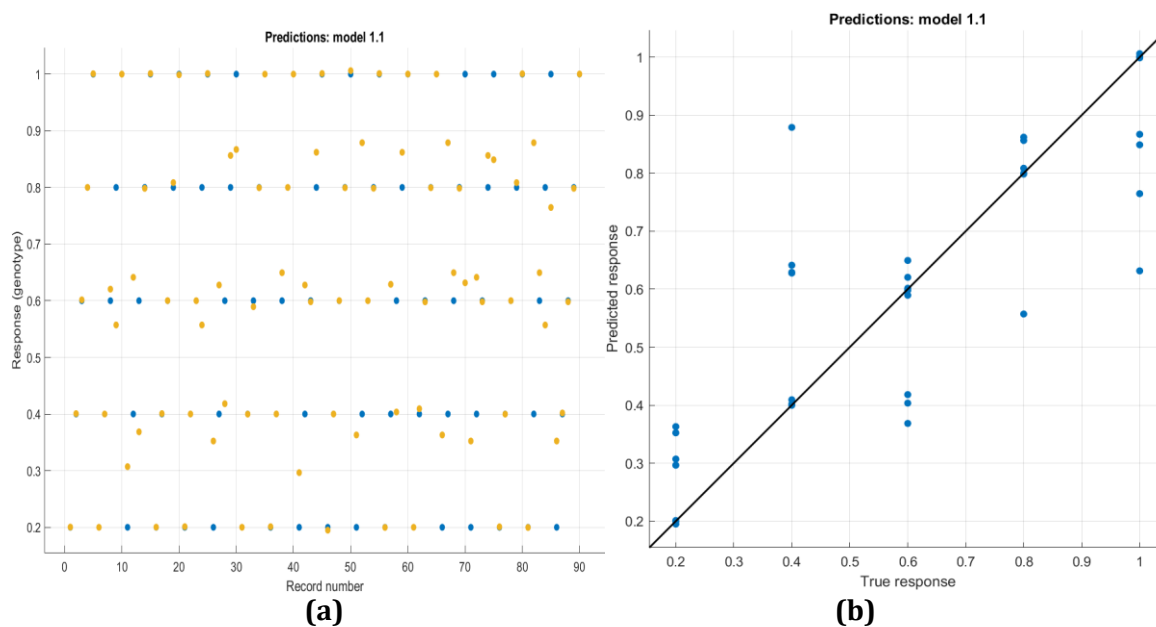


Figure 11. (a) Response plot (b) Predicted v/s Actual plot

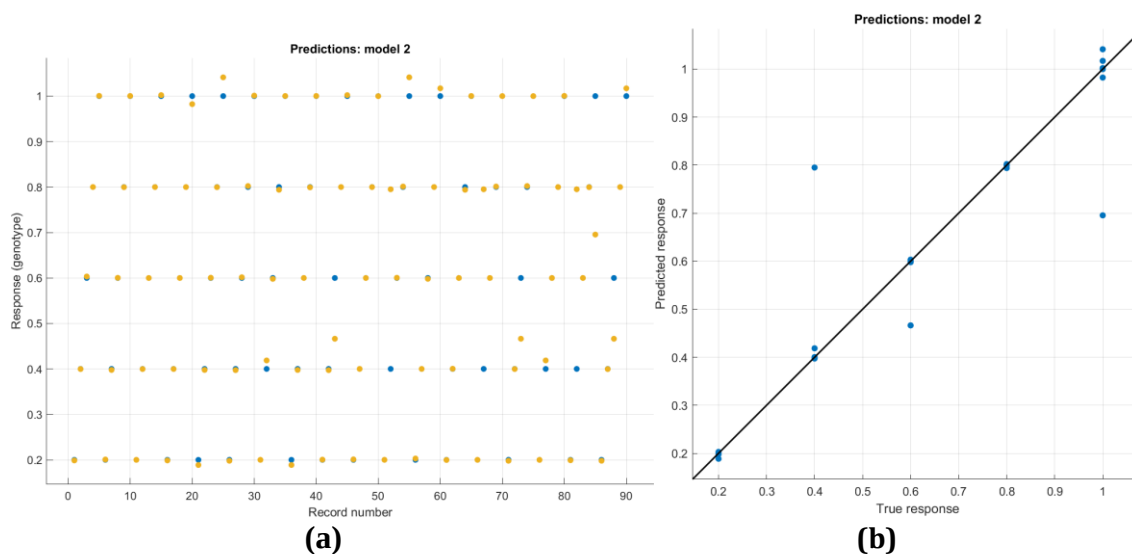


Figure 12. (a) Response plot (b) Predicted v/s Actual plot