Ann. For. Res. 65(1): 746-764, 2022
DOI: 10.5281/zenodo.7238175

ANNALS OF FOREST RESEARCH
https://www.e-afr.org/

EXPLORATORY DATA ANALYSIS AND OPTIMAL FEATURE SELECTION ON POMEGRANATE DATASET FOR ENHANCING THE PERFORMANCE OF POMEGRANATE DISEASE PREDICTION

Vaishali Nirgude
Research Scholar, Department of Computer Engineering, Thakur College of Engineering and Technology, Mumbai, India. vaishali.nirgude@thakureducation.org

Sheetal Rathi
Professor, Department of Computer Engineering, Thakur College of Engineering and Technology, Mumbai, India. sheetal.rathi@thakureducation.org

ABSTRACT
Exploratory Data Analysis (EDA) and Feature Selection (FS) are crucial in the fields of machine learning and data mining. FS is a dimensionality reduction technique to select optimal features from the original features by eliminating noisy, unimportant, and redundant features. FS simplifies the models by reducing the number of features, lower computational cost by decreasing the training time, reduces overfitting, solves the curse of dimensionality, improves the machine learning models' accuracy, and enhances the performance of the classification or prediction models.
In this paper, a data collection framework using agriculture drone and sensors have been designed to collect real field weather, soil, and water parameters. FS techniques are applied to select important features from the original features. Statistical methods are used to analyze the correlation between all micro-level parameters with pomegranate diseases. The Machine Learning (ML) approach is used to develop a prediction model. Optimal features are provided to the pomegranate disease prediction model to predict the accurate disease and recommend disease preventive measures. Also, studied and analyzed the impact of sudden changes in climatic conditions on pomegranate diseases.
Further, evaluated and compared the accuracy and loss of the various binary and multi-classification ML models. Experimental results prove that Random Forest (RF) has achieved excellent performance with an accuracy of 96.53%. The proposed method will help the agro-industry to detect and classify the most prominent diseases on pomegranate and to improve the growth and quality of fruits.
Keywords: Feature Selection, Exploratory Data Analysis, Machine Learning, Pomegranate, Disease Detection, Agriculture.

1. INTRODUCTION
Exploratory Data Analysis (EDA) is a vital step in research to study the dataset [1]. Different statistical methods and data visualization techniques for graphical representations are available to extract important features, discover different patterns, and relations, test hypotheses, and detect

outliers and anomalies within the data [2]. A feature is defined as a quantifiable property of a process also referred to as a parameter, attribute, or variable. Nowadays, due to multidisciplinary domains, the number of features has increased from a few to a huge. A large number of parameters increases cost, processing time, and computing resources. Therefore, a rise in data dimensions and computational cost leads to the overfitting and 'the curse dimensionality' problem [3]. The presence of noisy, redundant, and irrelevant parameters makes the ML models slow and reduces the model learning performance. Feature Selection (FS) is one of the important preprocessing steps that reduce the feature set by picking the relevant features (O) and eliminating the redundant features (R) from the original feature set (D) based on an evaluation criterion. In this paper, different FS techniques have been implemented to analyze the high-dimension dataset to select the most important features. Let D be the dataset with high dimensional parameter {P1, P2, P3, …., P_n}, n is a number of parameters in the data set. By applying various FS techniques select optimal features O from D. Optimal features help to improve the accuracy of ML detection and prediction tasks [4].

Based on the search policy, Feature selection is categorized into three types, i.e., filter methods, wrapper methods, and embedded methods [5]. The filter method analyzes each variable separately and then selects the top-ranking features. Feature selection should be performed before classification, prediction, or clustering tasks. Some of the univariant metrics are Information Gain/ Mutual information, Pearson correlation coefficient, Variance threshold, and Chi-square test. Information gain verifies the capability of independent variables to predict the dependent variable. The correlation coefficient removes the identical variables. Variance removes constant features. Filter methods are computationally very fast but determine individual features for selection. In the wrapper method, a subset of features gets selected to train the ML models. Based on the interpretation, we can add or remove features from the subset. Therefore, this method is computationally very expensive. Trendy wrapper methods are Forward feature selection, Backward feature selection, and Recursive feature selection. The embedded method blends the important features of filter and wrapper methods. It uses the ML algorithm to evaluate the features [6]. During experimentation, we have explored various ML algorithms such as Random Forest, Decision tree, Logistic Regression, Gradient Boosting, Ada Boosting, and KNN wrapped by Recursive Feature Elimination to select the most relevant features. **Fig. 1**, shows the detailed classification of filter, wrapper, and embedded methods.

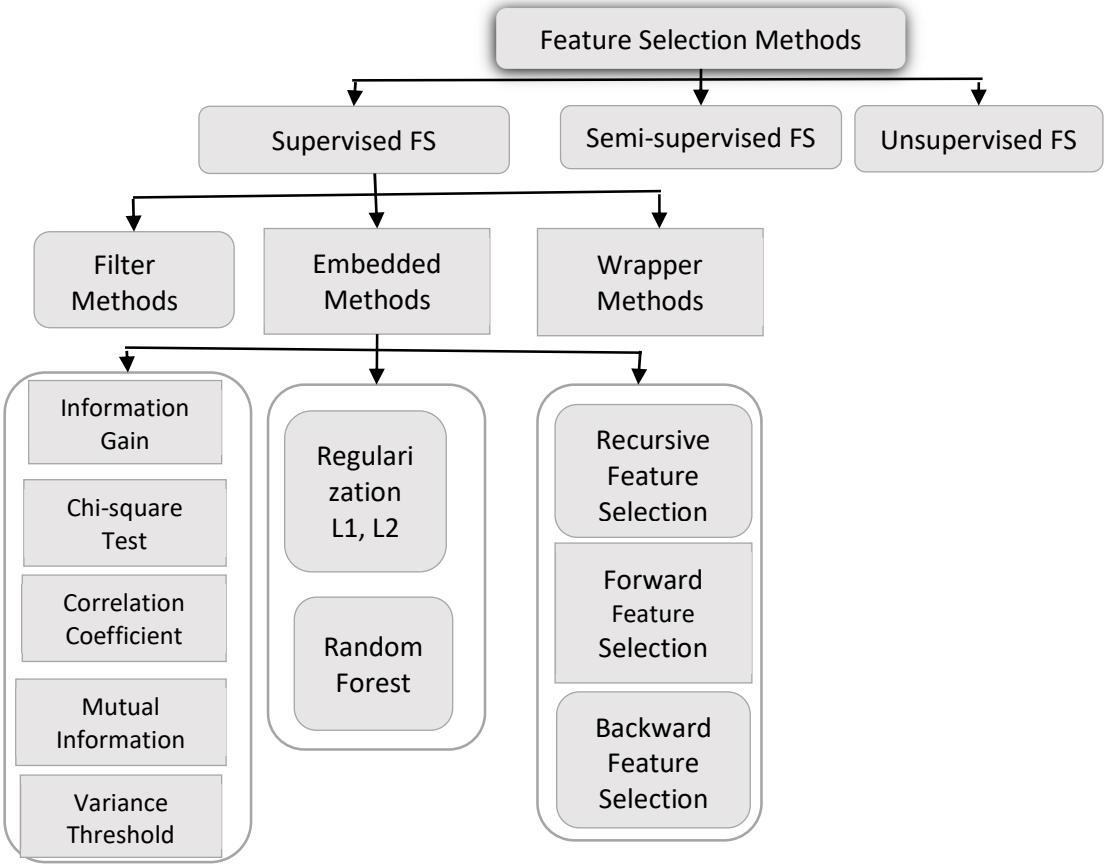


Fig.1 Feature selection techniques

Pathan et al. 2020, performed EDA and selected optimal features by applying FS methods to analyze the performance of ML models for heart disease prediction. Experimental results proved that even after reducing the number of features, the performance of the classification models improved considerably w.r.t. models training time as compared with models trained on the original feature set [7]. Wang et al. 2022, introduced a frequency-based measure to assess the strength of feature selection and evaluate the feature selection for the diagnosis of Alzheimer's Disease [8]. Senan et al. 2021, used feature score and Pearson correlation coefficient techniques to enhance the accuracy of ML algorithms for heart disease diagnosis [9]. Rajab and Wang, 2020, analyzed the challenges of filter-based methods and provided solutions to overcome them [10]. Pudjihartono et al. 2022, explored different FS methods with their strengths and weakness. No single FS method gives optimal features therefore multiple FS methods can be combined for selecting important features [11].

In this paper, detailed EDA and FS techniques are discussed to enhance the accuracy of pomegranate disease prediction and classification. Because of the huge impact of sudden changes

in climatic conditions on pomegranate growth and quality, several studies are being conducted to understand the causes of the spread and transmission of diseases on pomegranates [12]. Experimentation is carried out on the collected real-field weather, soil, and water parameters. Detailed Exploratory Data Analysis was performed to understand and analyze the dataset. Different Feature Selection methods have been implemented to select important features to improve pomegranate disease prediction and classification accuracy. A machine learning model was used to predict pomegranate diseases and recommended preventive measures if required.
The paper is organized as follows. **Section 2** describes the materials and methods used to correctly classify pomegranate diseases. **Section 3** shows the results and discussion. Finally, we presented the conclusions of the research work.

2. MATERIALS and METHODS

2.1 Research Methodology

The research methodology has been divided into two phases: (a) Pomegranate disease detection based on an Image-based approach and (b) Pomegranate disease prediction based on a data-driven approach. A real-time pomegranate disease management system is designed and developed using statistical, machine learning, and deep learning approaches. The overall system architecture has shown in **fig. 2**. First part of the research methodology is the study and performance analysis of the pomegranate disease detection system using DL techniques based on an image dataset. The models' performance has been measured using the Accuracy and Cross entropy loss function [13]. The second part of the research methodology is a study and performance analysis of the pomegranate disease prediction system. Statistical methods and Machine learning techniques are used to enhance the performance of disease prediction systems.

Ann. For. Res. 65(1): 746-764, 2022
DOI: 10.5281/zenodo.7238175

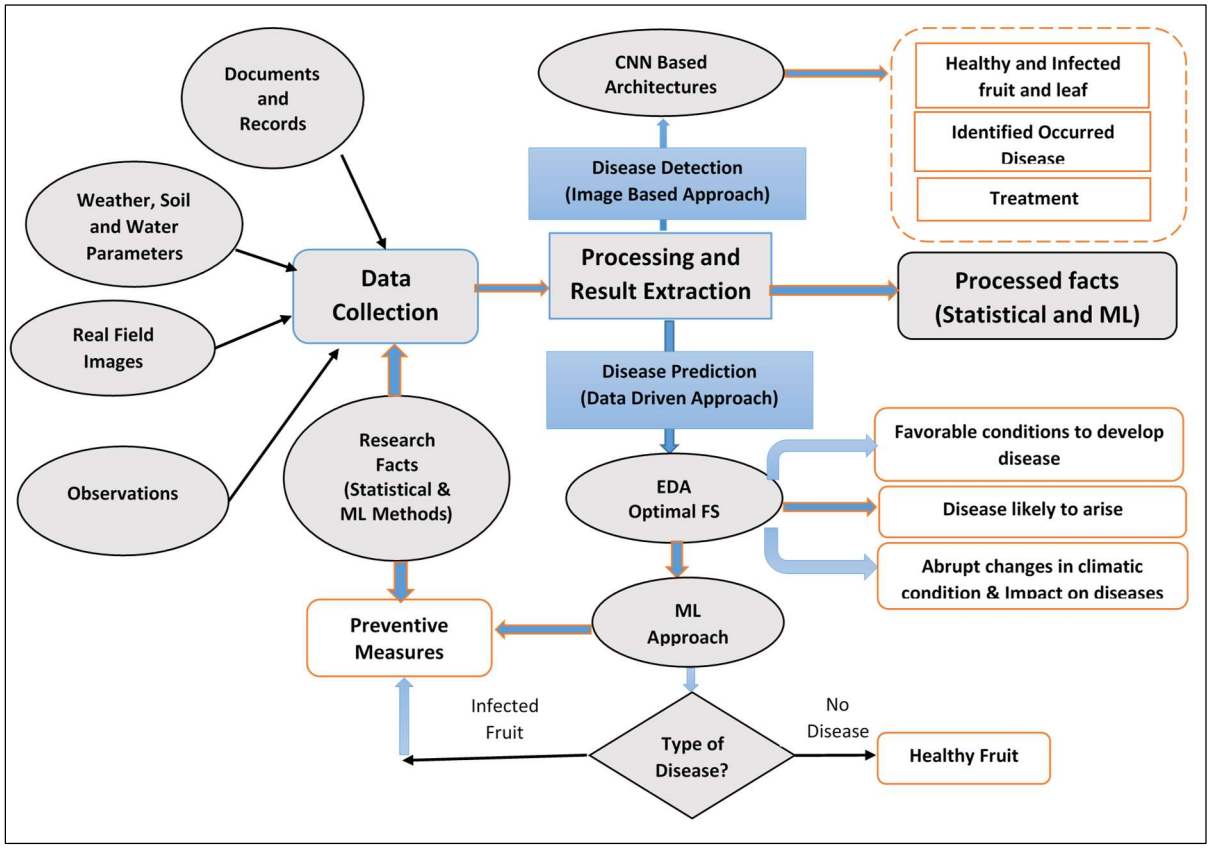


Fig. 2 System Architecture of pomegranate disease detection and prediction using DL and ML approach

2.2 Data Acquisition

The data collection framework has been designed and developed to collect weather, soil, and water parameters. An agriculture drone with sensors is used to collect the weather parameters (Temperature T, Relative Humidity RH). A TDS- meter is used to collect water TDS and a Soil pH-moisture combo tester kit is used to collect soil pH, soil moisture, and water pH. The remaining weather parameters namely wind speed WS, the number of sunshine hours SH, weather conditions, precipitation PP, and pressure P are collected through weather websites. The last 11 years (1st Jan 2010 to 1st Dec 2021: 4227 records) of historical data is collected for training the models and 6 months of real field data (1st Oct 2021-28th Feb 2022: 151 records) for testing models. As our dataset is collected via sensors and APIs, it is already in clean form. Still few values are null and missing therefore replaced by taking an average of values. By using one hot encoding technique, categorical variables are converted into numbers with grading 1-5 e.g., Weather descriptions. Detailed EDA and FS have been performed to analyze the data and to improve disease prediction and classification accuracy. The dataset contains 32 parameters namely Day DD, Month MM, Year YY, Temperature T (Morning, Afternoon, Evening, Average), Relative Humidity RH (Morning, Afternoon, Evening, Average), Wind Speed WS (Morning, Afternoon, Evening, Average),

Pressure P (Morning, Afternoon, Evening, Average), Pressure P (Morning, Afternoon, Evening, Average), Precipitation PP (Morning, Afternoon, Evening, Average) WeatherDesc (Morning, Afternoon, Evening, Average), Sunshine Hours SH, Soil Moisture SM, Soil pH, Water pH, Water TDS.

2.3 Exploratory Data Analysis

2.3.1 Average data normal distribution with a density of features

The normal distribution, also called the Gaussian distribution, has the probability density function (PDF). In a normal distribution, values of a variable are distributed symmetrically, and a bell shape curve is formed.

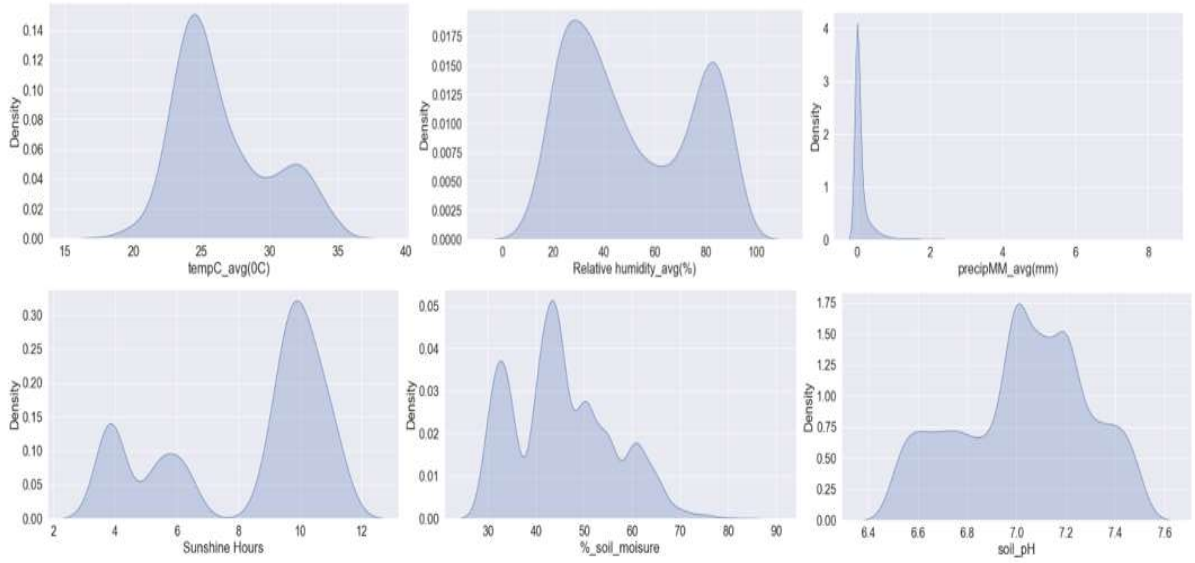


Fig. 3 Normal Distribution

2.3.2 Occurrence of Diseases year and month wise

It is found that across all 12 months, most pomegranate diseases are occurring in June, July, August, and September months due to high humidity, less number of sunshine hours, more soil moisture, cloudy weather conditions, and irregular rainfall.

Ann. For. Res. 65(1): 746-764, 2022
DOI: 10.5281/zenodo.7238175

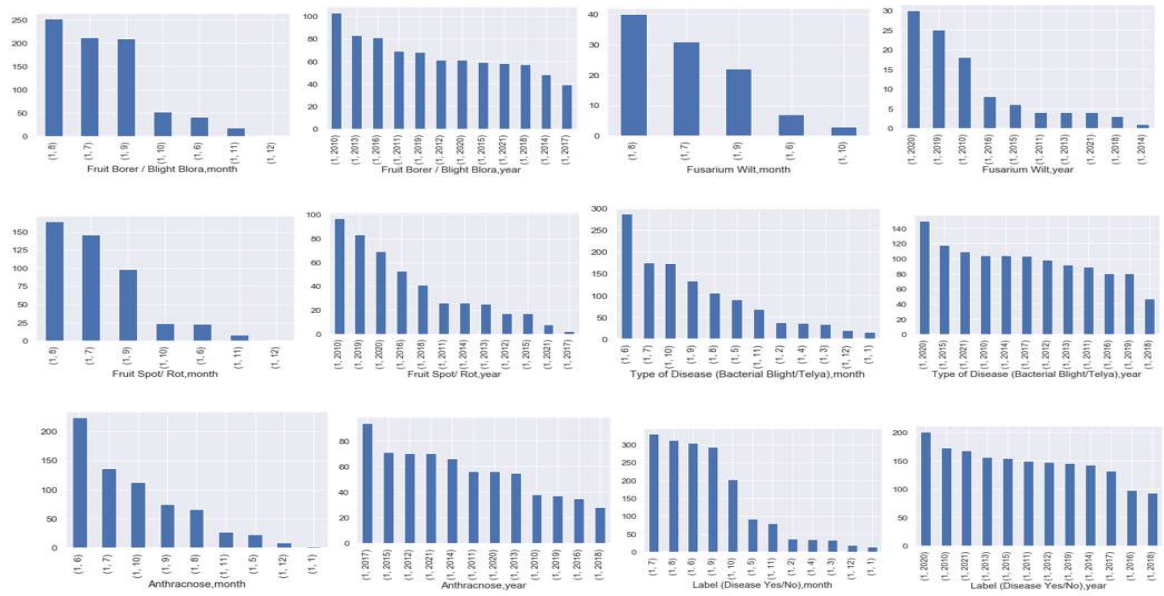


Fig. 4 Occurrence of Diseases year and month wise

2.3.3 Correlation between two variables with respect to the disease column

The correlation between all variables is observed for disease occurrences [14]. In fig. 5 on a sample basis correlation between relative humidity RH and temperature average T_{avg} has been shown. If the RH is > 31% and T_{avg} is in the range of 22^0 C >and <35^0 C, then there are more chances of disease occurrences.

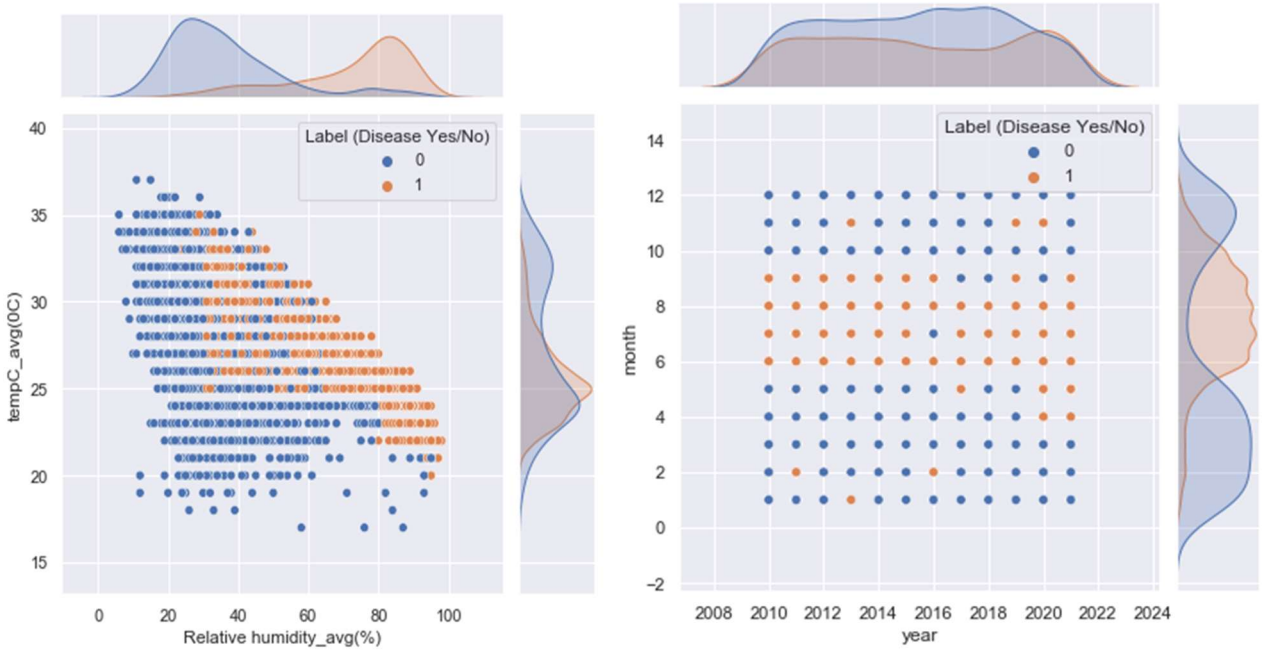


Fig. 5 Correlation between two variables with respect to the disease column

2.4 Feature Selection techniques

Optimal feature selection has been done by applying various statistical techniques. However, different feature selection methods return different subsets of features and multicollinear features affect model performance due to uncertainty in coefficient evaluations. Therefore, domain expert advice has been taken to retain the important features for improving the accuracy of classification models. The below section described the obtained results of different feature selection methods.

2.4.1 Pearson Correlation coefficient

2.4.1.1 **Micro-level Parameters Correlation Model (MPCM)**

We have designed and developed an MPCM to find (A) Correlation between all micro-level parameters (B) Correlation between parameters with diseases (C) Correlation between diseases (D) Correlation between parameters with disease occurrence. MCPM model analyses which parameters directly impact pomegranate diseases by showing positive or negative or no correlation. Correlated features do not provide any useful information to the model.

(A) Correlation between all Micro-Level Weather/Soil/Water Parameters

Correlation between all 32 micro-level parameters was found using the Pearson Correlation coefficient. Fig. 6 shows the negative correlation between relative humidity RH and temperature average T_{avg} (-0.52), Positive correlation between relative humidity RH and soil moisture SM (0.83). No correlation between water TDS and water pH (0).

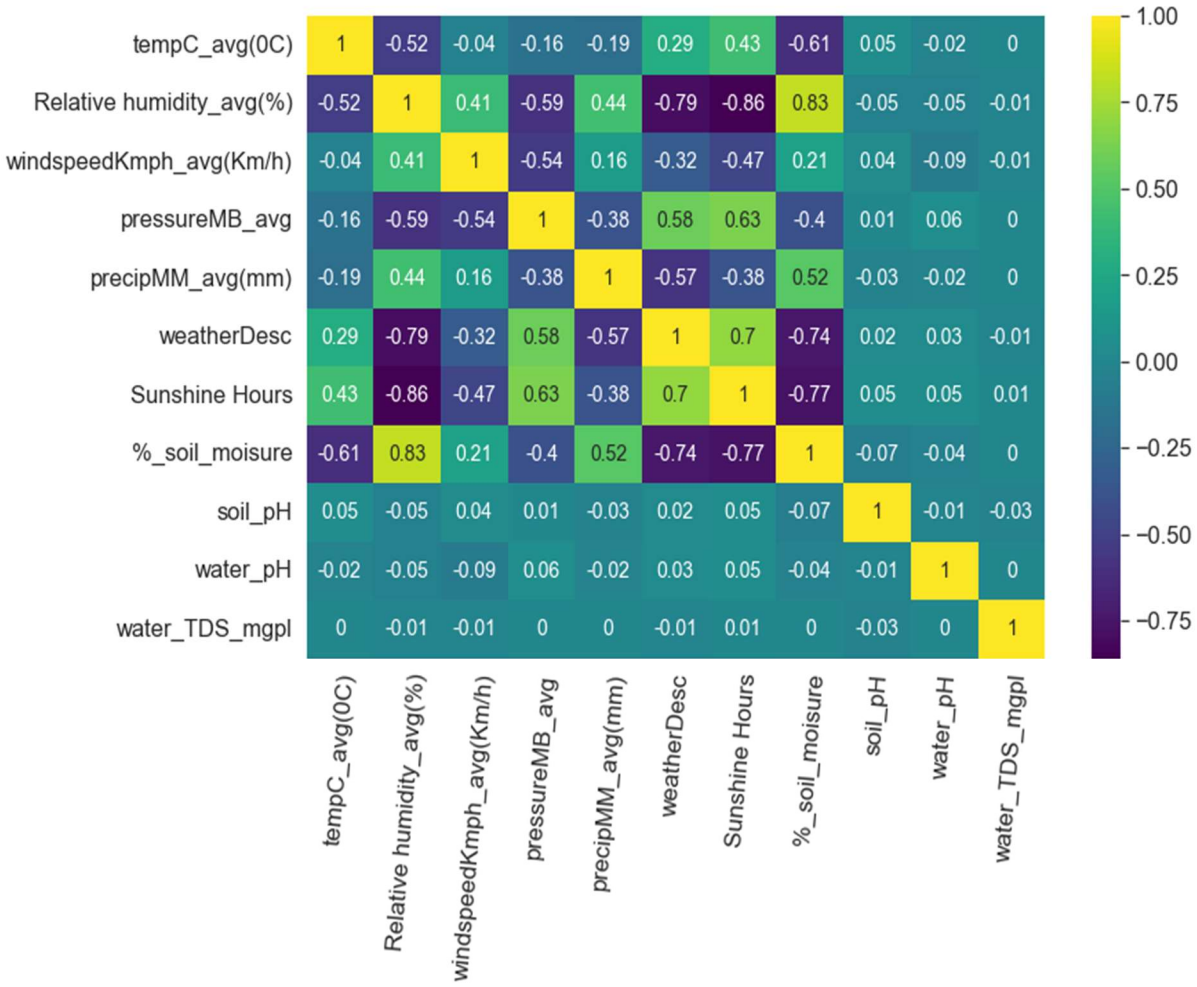


Fig. 6 Correlation between all Micro-Level Parameters

B) Correlation between parameters with diseases

The correlation between micro-level parameters versus target disease (Fruit Borer, Blight Blora, Fusarium Wilt, Bacterial Blight, and Anthracnose) was found in the collected pomegranate dataset. It was found that RH and SM have a strong positive correlation and SH has a strong negative correlation with Fruit Borer. Similarly, PP and SM have a strong positive correlation with Fusarium Wilt. In the case of Bacterial blight and Anthracnose RH and SM showed a positive correlation and SH & PP showed a negative correlation.

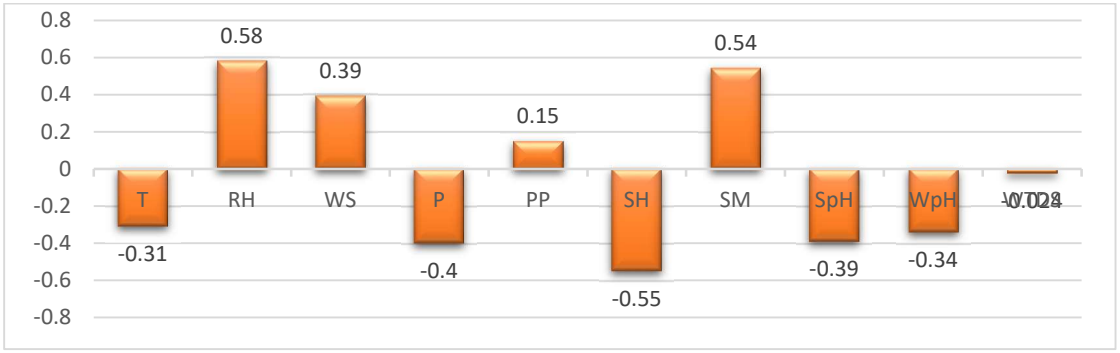


Fig. 7 Correlation of Micro-level Parameters with Fruit Borer / Blight Blora

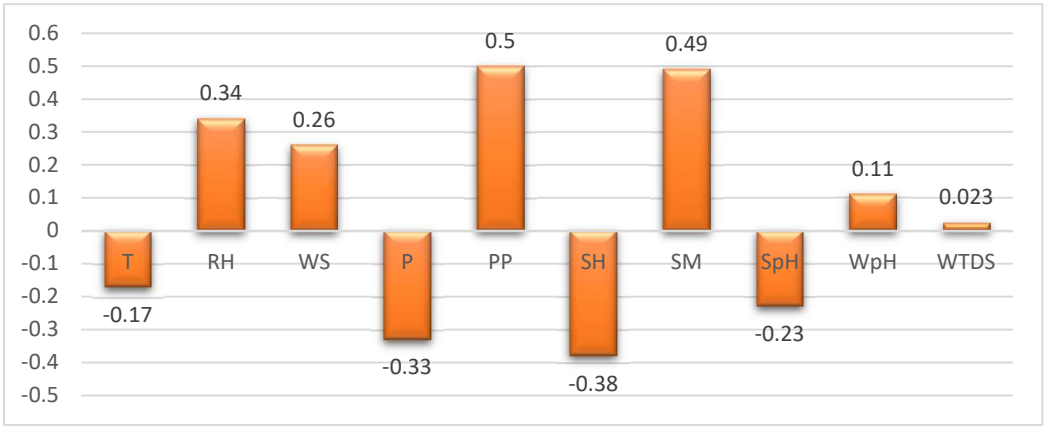


Fig. 8 Correlation of Micro-level Parameters s with Fusarium Wilt

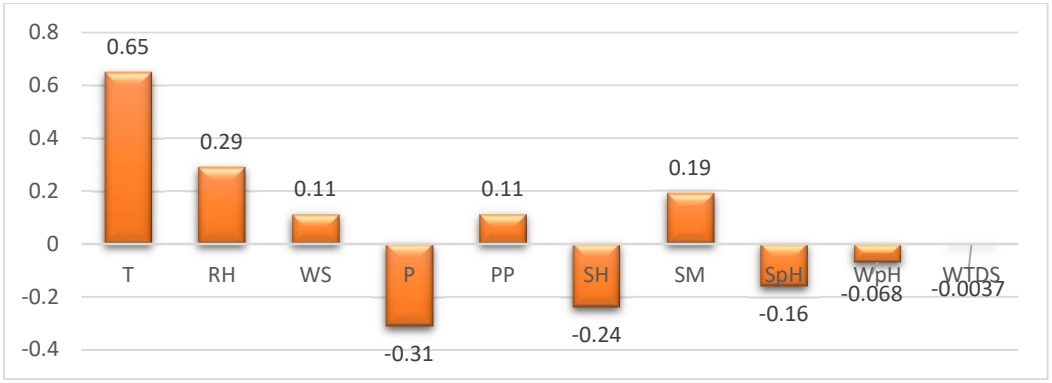


Fig. 9 Correlation of Micro-level Parameters with Bacterial Blight

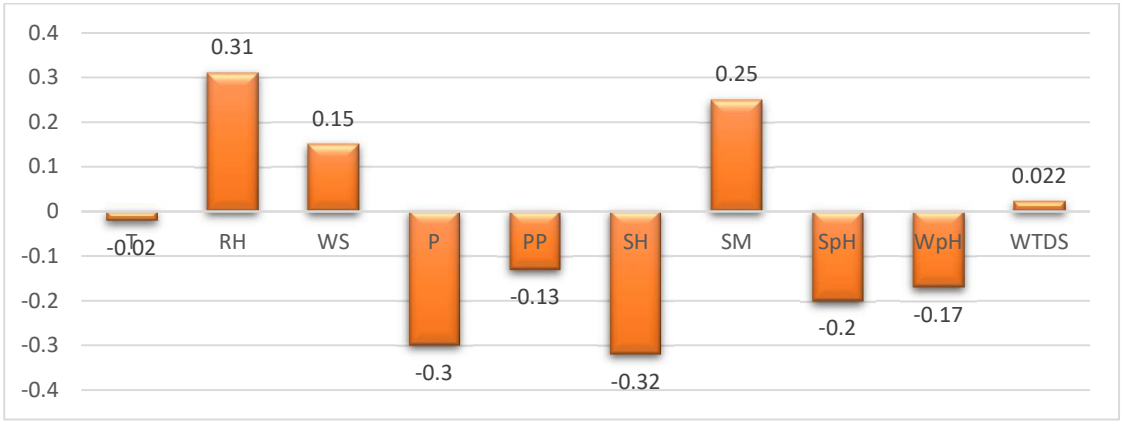


Fig. 10 Correlation of Micro-level Parameters with Anthracnose

C) Correlation between diseases

The correlation between diseases is examined and it has been observed that Bacterial Blight and Anthracnose were found to be strongly correlated (0.7). Apart from this, there was a positive correlation found among Fruit Spot, Fruit borer, and Fusarium Wilt (0.58).

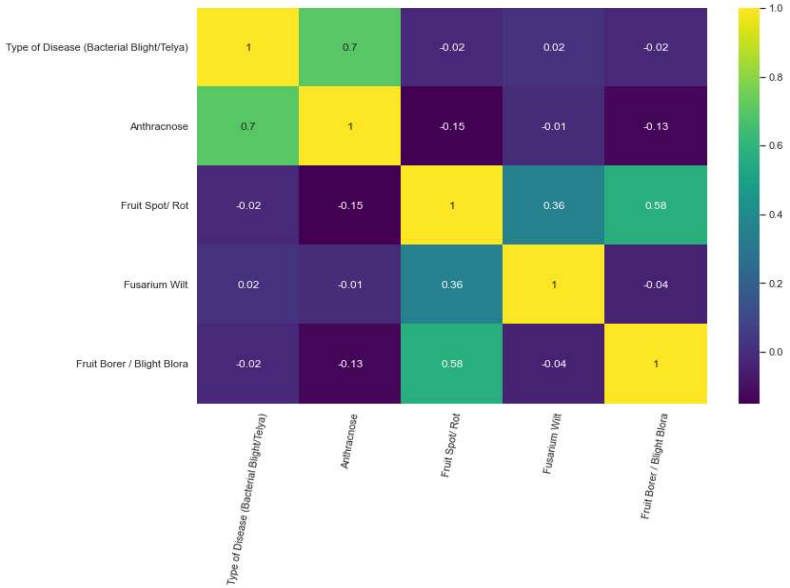


Fig. 11 Correlation between diseases

D) Correlation between parameters with disease occurrence

The correlation between selected optimal features (T_{avg}, RHavg, WS_{avg}, P_{avg}, PP_{avg}, WD, SH, SM) with the target (Disease Occurs Y/N) was investigated. It was found strong positive (RH_{avg}, WS_{avg}, PP_{avg}, SM) or negative (T_{avg}, P_{avg}, WD, SH) correlation with disease occurrence.

Ann. For. Res. 65(1): 746-764, 2022
DOI: 10.5281/zenodo.7238175

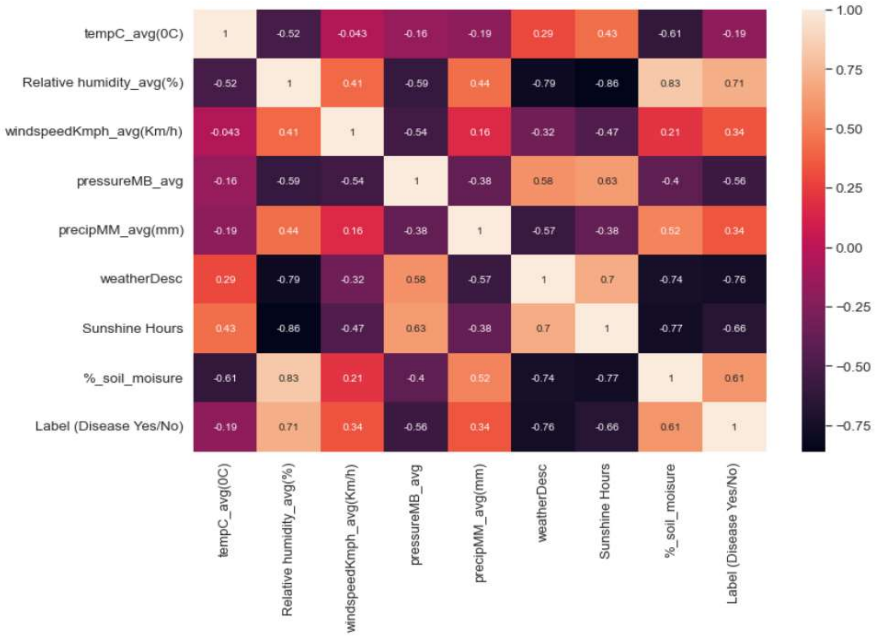


Fig. 12 Correlation between Micro-level parameters with disease occurrence (Y/N)

2.4.2 Information Gain

Information gain (IG) is calculated using entropy. It is used for FS by evaluating the IG of all independent variables i.e., IG(X1), IG(X2), …., w.r.t. dependent variable. Then rank the features in the descending order of their respective IG. Select the features which having high information score. As per the IG score and threshold value, we have selected the parameters namely RHavg, SH, T_{avg}, PP_{avg}, SM, and P_{avg} as shown in fig 13.

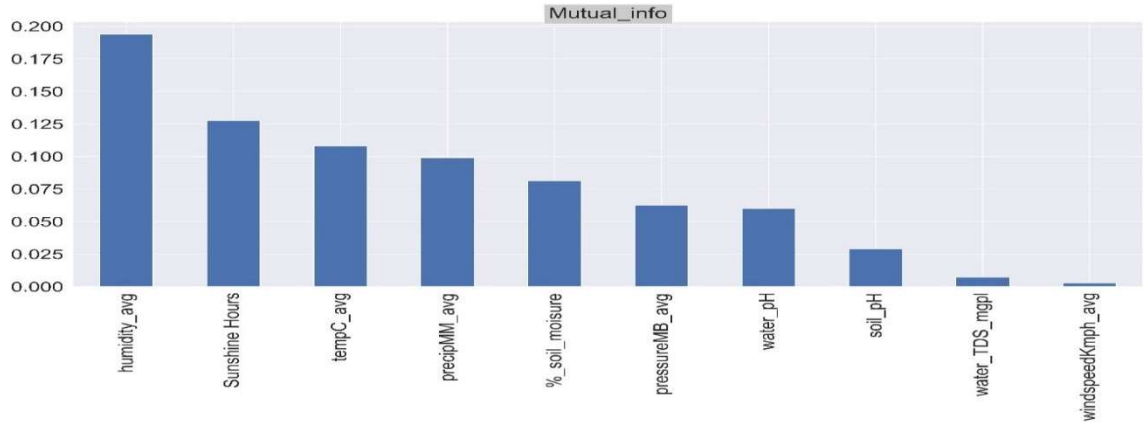


Fig. 13 Feature selection using Information Gain

2.4.3 Filter features by Variance Threshold

Ann. For. Res. 65(1): 746-764, 2022
DOI: 10.5281/zenodo.7238175

A variance threshold is a standard approach to FS. It eliminates all features whose variance is below the threshold value. The features with a higher variance contain more useful information than a lower variance.

Table 1: Filter features by Variance Threshold

	Sunshine Hours	%_soil_moisure	soil_pH	water_pH	water_TDS_mgpl	Label (Disease Yes/No)
Sunshine Hours	1.000000	-0.769105	0.047175	0.053492	0.010106	-0.662051
%_soil_moisure	-0.769105	1.000000	-0.073358	-0.035720	0.001171	0.606175
soil_pH	0.047175	-0.073358	1.000000	-0.013978	-0.033262	-0.032147
water_pH	0.053492	-0.035720	-0.013978	1.000000	0.003156	-0.041087
water_TDS_mgpl	0.010106	0.001171	-0.033262	0.003156	1.000000	-0.006777
Label (Disease Yes/No)	-0.662051	0.606175	-0.032147	-0.041087	-0.006777	1.000000

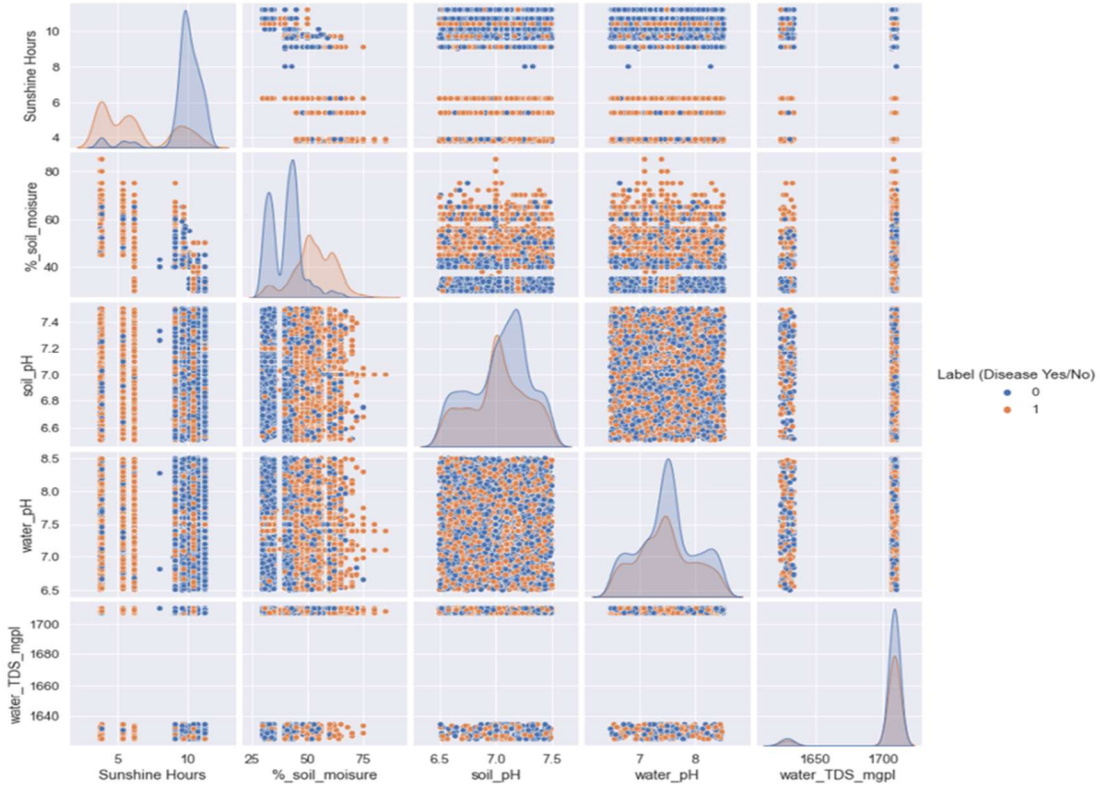


Fig. 14 Filter features by Variance Threshold

3. RESULTS and DISCUSSION

3.1 Binary Classifiers for Evaluating the Performance of Disease Prediction

Binary classification refers to predicting one of two classes i.e., Disease occurs Yes (1) or No (0). ML classifiers such as Logistic Regression (LR), K-Nearest Neighbour (KNN), Naïve Base (NB), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), and Ada

Boosting (AB) algorithms have been implemented for accurately predicting diseases. Out of these, RF gave the highest accuracy (96.53%) and minimum loss (3.47).

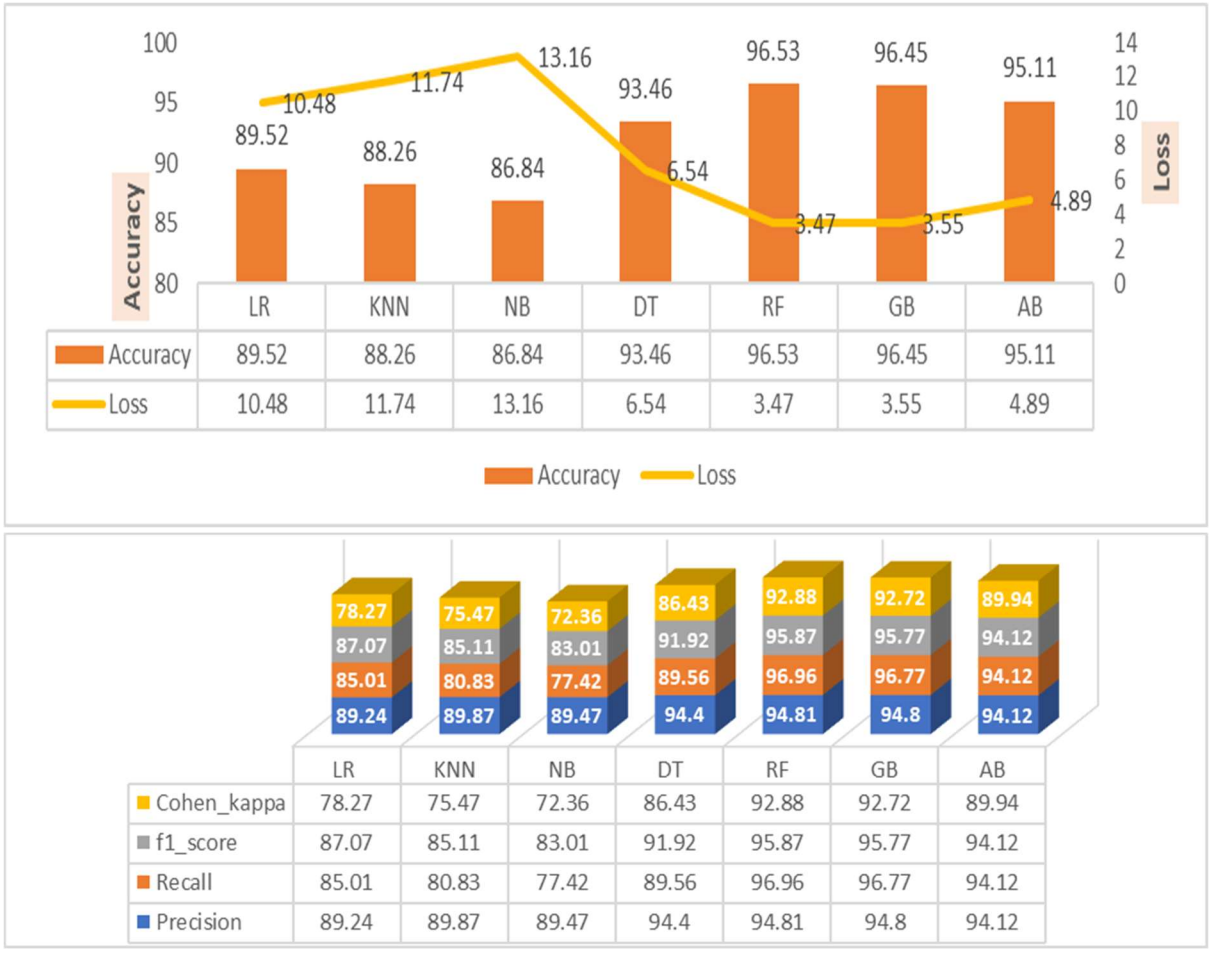
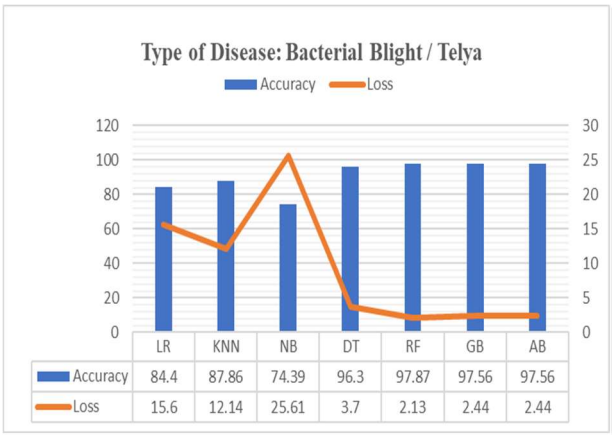
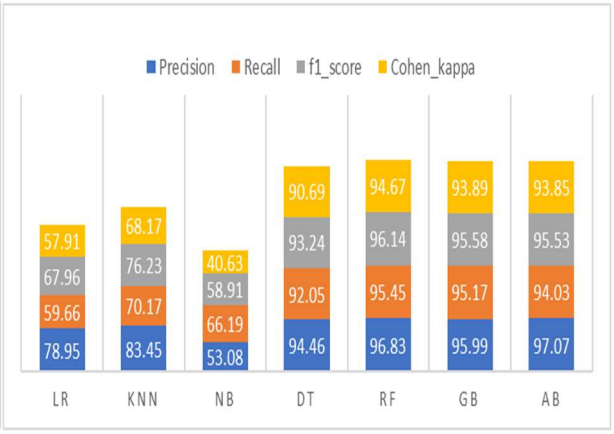


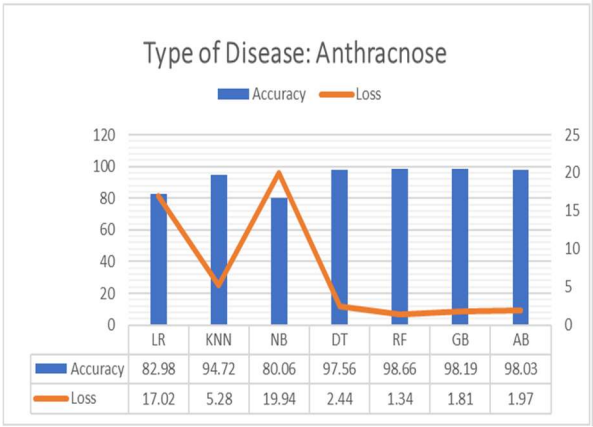
Fig. 15 **Binary Classifiers for Evaluating the Performance of Disease Prediction**

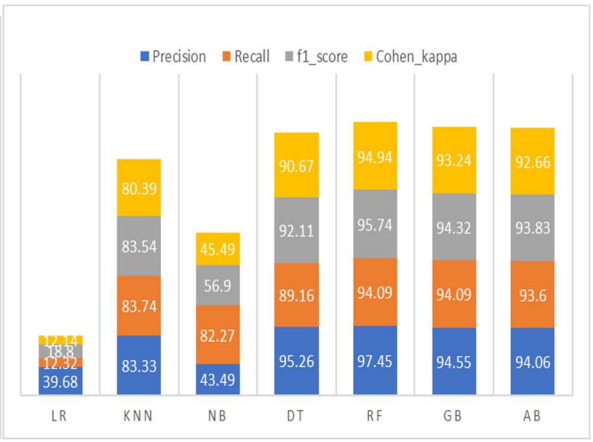
3.2 Multimodel Classifiers for Evaluating the Performance of Disease Prediction

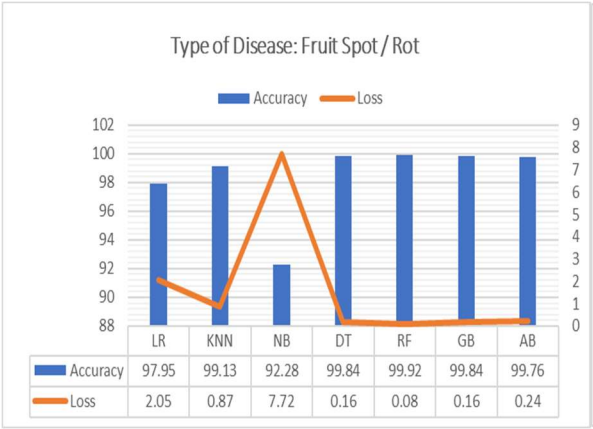
Multimodel (LR, KNN, NB, DT, RF, GB, AB) classifiers have been implemented for accurately detecting pomegranate diseases. Compared the accuracy and performance of all classification models. Experimental results demonstrate that Random Forest (97.87%) achieved a better performance in all experiment groups.

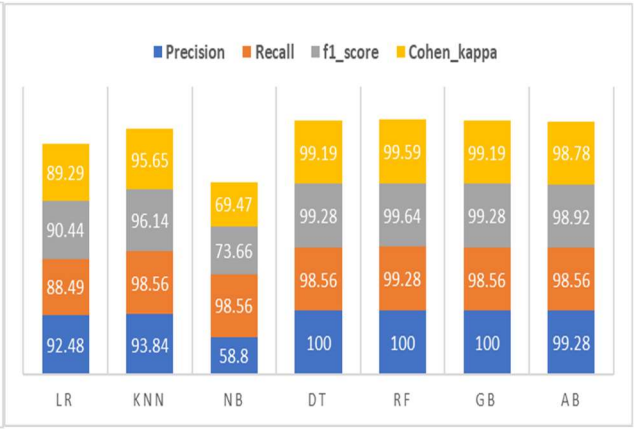

Type of Disease: Bacterial Blight / Telya
Accuracy
Loss
120
100
80
60
40
20
0
30
25
20
15
10
5
0
LR KNN NB DT RF GB AB
Accuracy 84.4 87.86 74.39 96.3 97.87 97.56 97.56
Loss 15.6 12.14 25.61 3.7 2.13 2.44 2.44


Precision
Recall
f1_score
Cohen_kappa
57.91 67.96 59.66 78.95
68.17 76.23 70.17 83.45
40.63 58.91 66.19 53.08
90.69 93.24 92.05 94.46
94.67 96.14 95.45 96.83
93.89 95.58 95.17 95.99
93.85 95.53 94.03 97.07
LR KNN NB DT RF GB AB


Type of Disease: Anthracnose
Accuracy
Loss
120
100
80
60
40
20
0
25
20
15
10
5
0
LR KNN NB DT RF GB AB
Accuracy 82.98 94.72 80.06 97.56 98.66 98.19 98.03
Loss 17.02 5.28 19.94 2.44 1.34 1.81 1.97


Precision
Recall
f1_score
Cohen_kappa
39.68 18.3 12.32
80.39 83.54 83.74 83.33
45.49 56.9 82.27 43.49
90.67 92.11 89.16 95.26
94.94 95.74 94.09 97.45
93.24 94.32 94.09 94.55
92.66 93.83 93.6 94.06
LR KNN NB DT RF GB AB


Type of Disease: Fruit Spot / Rot
Accuracy
Loss
102
100
98
96
94
92
90
88
9
8
7
6
5
4
3
2
1
0
LR KNN NB DT RF GB AB
Accuracy 97.95 99.13 92.28 99.84 99.92 99.84 99.76
Loss 2.05 0.87 7.72 0.16 0.08 0.16 0.24


Precision
Recall
f1_score
Cohen_kappa
89.29 90.44 88.49 92.48
95.65 96.14 98.56 93.84
69.47 73.66 98.56 58.8
99.19 99.28 98.56 100
99.59 99.64 99.28 100
99.19 99.28 98.56 100
98.78 98.92 98.56 99.28
LR KNN NB DT RF GB AB

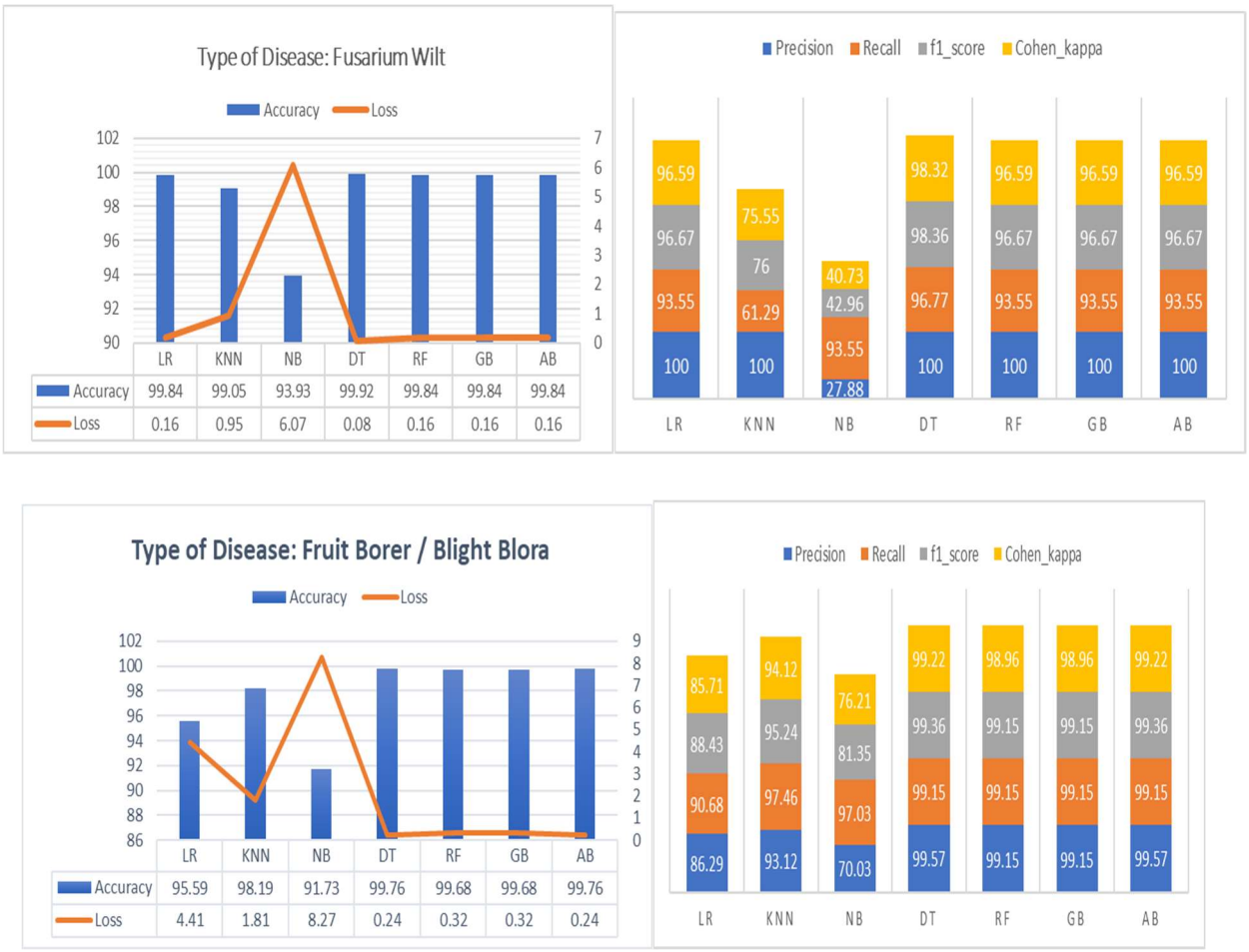


Fig. 16 Multimodel Classifiers for Evaluating the Performance of Disease Prediction

DISCUSSION

The supervised ML approach is used to perform binary pomegranate disease classification (Disease occurs or not) and multi-model classification (Predict the type of pomegranate disease) on the collected weather/soil/water dataset. As our dataset contains 32 parameters therefore to avoid overfitting and to reduce computational cost various feature selection methods have been explored. However, the single FS method is not the best method to find optimal features. The selection of the best method is either problem specific or depends on the dataset [15]. Hence, several feature selection methods can be combined known as the ensemble method where the strengths of the different methods are considered for the selection of important parameters. By applying various FS methods such as Information Gain, Correlation, Chi-square test, Variance threshold, Recursive Feature Selection, and various ML algorithms, the most important 8- features are selected. Additionally selected features were verified by a domain expert before providing to the ML models. It has been observed that reducing the number of features lessens the model's training time and overfitting problems.

CONCLUSION

A real-time pomegranate disease management system has been designed and developed using statistical, machine learning, and deep learning techniques to improve the detection, classification, and prediction accuracy of the most widespread diseases on pomegranates (Anthracnose, Bacterial Blight (Telya), Blight Borer, Rot, Fusarium wilt). Real-time micro-level parameters were collected via. agriculture drone and sensors. We have tried to analyze the hidden relationship between micro-level weather/soil/water parameters with diseases that cause economic losses. By applying various statistical methods, exploratory data analysis and pre-processing have been performed. We have tried to explore the patterns of sudden changes in climatic conditions and their impact on pomegranate diseases. The Micro Parameter Correlation Model (MPCM) has been implemented to find correlations between micro-level parameters with diseases. Optimal feature selection is done by applying various feature selection methods such as mutual information, Pearson correlation coefficient, variance threshold, P-value (Chi-square test), and recursive feature selection methods. Machine learning Binary and Multimodel classifiers have been implemented for evaluating the performance of pomegranate disease prediction. It is expected that the proposed model will help the agro-industry to correctly predict and classify the most prominent diseases of pomegranate and helps the farmers to take the right decision at the right time to avoid economical and environmental loss and crop yield.

Acknowledgments

The authors would like to thank to Dr. A.M. Navale, Professor of Plant Pathology, Mahatma Phule Krishi Vidyapeeth, Rahuri, Maharashtra for suitable guidance to proceed further in the research study, validating the dataset and results. We are also thankful to Mr. Satish Rao, Scientist / Engineer SF - ISRO for giving valuable input to formulate the research problem definition.

Conflict of interest

All the authors declare that there is no conflict of interest.

REFERENCES

1. J. DSouza and S. Velan S. (2020). Using Exploratory Data Analysis for Generating Inferences on the Correlation of COVID-19 cases, 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1-6, DOI: 10.1109/ICCCNT49239.2020.9225621.
2. Rehman, A., Belhaouari, S.B. (2021). Unsupervised outlier detection in multidimensional data. J Big Data 8, 80 (2021). https://doi.org/10.1186/s40537-021-00469-z.

3. Essam Debie, Kamran Shafi. (2019). Implications of the curse of dimensionality for supervised learning classifier systems: theoretical and empirical analyses, May 2019, Pattern Analysis and Applications, DOI: 10.1007/s10044-017-0649-0.
4. Chen, RC., Dewi, C., Huang, SW. (2020). Selecting critical features for data classification based on machine learning methods. J Big Data 7, 52 (2020). https://doi.org/10.1186/s40537-020-00327-4.
5. Pudjihartono N, Fadason T, Kempa-Liehr AW and O'Sullivan JM. (2022). A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. Front. Bioinform. 2:927312. DOI: 10.3389/fbinf.2022.927312.
6. Saba Bashir, Irfan Ullah Khattak, Aihab Khan, Farhan Hassan Khan, Abdullah Gani, Muhammad Shiraz (2022). A Novel Feature Selection Method for Classification of Medical Data Using Filters, Wrappers, and Embedded approaches, Complexity, vol. 2022, Article 8190814, 12 pages, 2022. https://doi.org/10.1155/2022/8190814.
7. Muhammad Salman Pathan, Avishek Nag, Muhammad Mohisn Pathan, Soumyabrata Dev. (2022).
   Analyzing the impact of feature selection on the accuracy of heart disease prediction, Healthcare Analytics, Volume 2,2022,100060, ISSN 2772-4425, https://doi.org/10.1016/j.health.2022.100060.
8. Feng Gu, Songhua Ma, Xiude Wang, Jian Zhao5 Ying Yu and Xinjian Song (2022), Evaluation of Feature Selection for Alzheimer's Disease Diagnosis, Front. Aging Neurosci., 24 June 2022, Sec. Alzheimer's Disease and Related Dementias, https://doi.org/10.3389/fnagi.2022.924113.
9. Ebrahim Mohammed Senan, Ibrahim Abunadi, Mukti E. Jadhav, Suliman Mohamed Fati. (2021). Score and Correlation Coefficient-Based Feature Selection for Predicting Heart Failure Diagnosis by Using Machine Learning Algorithms", Computational and Mathematical Methods in Medicine, vol. 2021, Article ID 8500314, 16 pages, 2021. https://doi.org/10.1155/2021/8500314.
10. Rajab, M. and Wang, D. (2020) Practical challenges and recommendations of filter methods for feature selection. Journal of Information & Knowledge Management, 19 (01). 2040019. ISSN 0219-6492
    https://doi.org/10.1142/s0219649220400195.
11. Pudjihartono N, Fadason T, Kempa-Liehr AW and O'Sullivan JM (2022) A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. Front. Bioinform. 2:927312.
    DOI: 10.3389/fbinf.2022.927312.
12. M. Shekhar, and N. Singh. (2021). The Impact of Climate Change on Changing Pattern of Maize Diseases in Indian Subcontinent: A Review, in Maize Genetic Resources - Breeding

Strategies and Recent Advances. London, United Kingdom: IntechOpen, 2021 [Online]. Available: https://www.intechopen.com/chapters/79453, DOI: 10.5772/intechopen.101053.

13. Nirgude, V., & Rathi, S. (2021). A Robust Deep Learning Approach To Enhance The Accuracy Of Pomegranate Fruit Disease Detection Under Real Field Condition. Journal of Experimental Biology and Agricultural Sciences, 9(6), 863–870. https://doi.org/10.18006/2021.9(6).863.870.
14. Zheng Z, Dou J, Cheng C and Gao H (2021) Correlation and Causation Analysis Between COVID-19 and Environmental Factors in China. Front. Clim. 3:619338. doi: 10.3389/fclim.2021.619338.
15. A. Abdul Rasheed (2021). Feature Selection: An Assessment of Some Evolving Methodologies, Turkish Journal of Computer and Mathematics Education, Vol. 12 No. 2(2021), 1982-1988.