Microbial Soil Health Restoration

Christy Jeba Malar A[1*], Bhuvanesh S[2], Heeranya R[3], Janani M[4]

[1,2,3,4]Sri Krishna College of Technology, Coimbatore, India
[1*]a.christyjebamalar@skct.edu.in
[2]727822tuit029@skct.edu.in
[3]727822tuit060@skct.edu.in
[4]727822tuit064@skct.edu.in

Abstract. Proper detection of microorganisms in microbiology, agriculture and medical services is important as it facilitates the prevention of diseases, environmental pollution, and research. In the project, Microbes Prediction and Recommendation is a web application based on machine-learning that was developed to predict microorganisms by using quantitative features of images. Various models are incorporated in the system which consist Artificial Neural Networks (ANN), XGBoost, Support Vector Machine (SVM) and Random Forest (RF). In order to make the performance better, feature scaling and Principal Component Analysis (PCA) are implemented and allow users to upload datasets, view predictions of all models, examine training metrics, confusion matrices, and feature importances and be presented with automated recommendations basing on the Random Forest model. The experimental findings indicate that ensemble models like XGBoost and the Random Forest had the highest accuracy, ANN had an intermediate result, and SVM had poorer generalization. On the whole, such a platform provides an interactive and comprehensible method of microorganism classification assisting researchers and practitioners in making adequate decisions and conducting additional analysis.

Keywords: Smart agriculture, Internet of Things(IoT), remote monitoring, soil moisture control, precision farming, wireless sensor network, automated irrigation systems, water conservation, climate change adaptations, and sustainable agriculture.

1 Introduction

Combining the best machine learning methods with modern sensor technology, the proposed system creates a promising strategy of monitoring the soil parameters. The system collects big amounts of information on vital soil parameters like water and mineral content with the help of modern sensors and remote sensing technologies. The use of the Artificial Neural Networks (ANN) to find the meaningful patterns in the data and Support Vector Machines (SVM) to deal with the complex and non-linear relationships lead to the fact that the real-time soil health predictions are more accurate. This integrated method aids in sustainable land management with timely information as it assists the user in making an informed choice and with the growing need to adopt precision agriculture. It entails the process of measuring all the different characteristics of

B. Singh et al. (eds.), *Proceedings of the International Conference on Advances in Computing Technology and Artificial Intelligence (COMPUTATIA 2026)*, Atlantis Highlights in Intelligent Systems 18,
https://doi.org/10.2991/978-94-6239-713-2_23

soil, moisture content, nutrients, and temperature. Effective soil monitoring is useful in ensuring the production of crops is enhanced, sustainable farming practices are encouraged and environmental impact is minimized. Due to ongoing technological development, novel technologies like machine learning models and modern sensors are turning the conventional soil monitoring process into a quicker and more precise one. Modern soil monitoring systems are based on sensor technology. These sensors gather real time data on the status of the soil starting with simple soil probes to the high-tech remote sensing equipment. With the advancement of sensor technology, monitoring systems can be more specific and smart to enable farmers, researchers, and environmental experts to make more decisions using the accurate information regarding the soil.

2 Literature Review

Raiten et al. [1] covered the term nutritional ecology where the variability of climate has an impact on the food systems and the population health in terms of the necessity to implement integrated approaches to agriculture. Cavicchioli et al. [2] also emphasized the susceptibility of microbial ecosystems to climate change by confirming that the disturbance of microbial balance has a direct effect on soil fertility and equilibrium in the environment. It was Randall and van Woesik [3] who showed the impacts of climate oscillations on biological systems, which could be used in terrestrial ecosystems including soil environments. Huong et al. [4] measured the economic effects of the climate change on the agricultural sector and found that this reduces farm productivity by a substantial percentage with shifts in the temperatures and rainfall, which led to the adoption of adaptive farming practices. The spatial aspect of climate-smart agriculture is frequently given little notice, and Prestele and Verburg [5] emphasized it as imperative to find local solutions to support effective implementation. H. As explained by Jat et al.[6], climate-smart agricultural practices have shown a high level of improvement of soil microbial activity and nutrient dynamics throughout the ages, which lead to the increase of crop performance. Guo et al. [7] suggested to combine smart sensing with phytoremediation to make the soil microbial restoration possible in real-time to detect the initial soil degradation. Boursianis et al. [8] conducted a systematic review of IoT and UAV-based smart farming technologies, which proved to be effective in crop and soil monitoring, whereas Koduru et al. [9] and Premkumar et al. [10] designed frameworks of automated irrigation systems that use IoTs to optimise the use of water and minimise the reliance on manual control. Wilhelm et al.[11] used machine learning to work with microbiome and forecast soil health, which demonstrates the development of artificial intelligence in precision agriculture. Singh et al. [12] described the use of advanced microbial devices to increase soil organic matter and ecosystem restoration in degraded soil, and the UNL CropWatch Team[13] discovered recent progress of real-time monitoring of soil with sensor-based technologies. Iqbal et al. [14] outlined the importance of the microbial community in renewing the soil resilience to the climate stress, and Li et al. [15] provided the integrated system monitoring-modeling strategies to manage the precision of the soil-plant system. Together, this body of research shows that the integration of climate-smart agriculture, IoT, machine learning, and microbial restoration offers a sustainable framework of managing the soil health under changing climatic conditions.

3 Methodology

The Microbes Prediction and Recommendation system is a suggested interactive machine learning system, which is supposed to identify microorganisms on the basis of quantitative features obtained by image processing and provide practical advice. The process starts with the preprocessing of the file of the microbes.csv where the raw features are standardized, the encoding of the class labels, and the dimensionality is reduced with the help of the Principal Component Analysis in order to represent the most significant information. The processed data are trained using various machine learning models, such as Artificial Neural Networks, XGBoost, Support Vector Machines, and Random Forest, and processed to predict the microorganism classes. The standard metrics that are used in evaluating model performance include accuracy, precision, recall, F1-score, and confusion matrices. Moreover, the analysis of the feature importance helps to identify the PCA components that the most significantly contribute to the predictions. The models are trained and deployed in a web application, which is built on Streamlit. By using this interface, users can upload their own datasets, obtain predictions of each of the models, see performance measures and confusion matrices, discover key features, and receive automated recommendations they obtained using the Random Forest model. In general, the system offers a convenient and integrated system that integrates data preprocessing, multi-model prediction, result visualization, and recommendation, which will allow researchers and practitioners to effectively classify microorganisms and make the correct decisions.

3.1 Data Collection and Preparation.

The primary dataset, microbes.csv, is composed of 30,527 samples and 24 numerical measurements derived out of images of microorganisms. These characteristics characterize various morphological and geometrical characteristics like solidity, eccentricity, length of major and minor axis, perimeter, area, coordinates of bounding box, and measurements of the convex hull. The target variable is the category of microorganisms in each sample.The unwanted columns including index values are eliminated in the process of preprocessing. Numerical variables are standardized with the help of Standard Scaler whereas the labels are transformed to numerical values with the help of Label Encoder. Principal Component Analysis PCA is done to reduce the size of the dataset and increase the efficiency of the model, keeping 95 percent of the total variance. The result of this step is a minimum set of transformed features that is further used to train the models.

3.2 Model Training

The dataset is then preprocessed and divided into a training and testing set to test the accuracy of the models. Various machine learning methods are used such as Artificial Neural Networks (ANN), XGBoost, Support Vector Machine (SVM) with RBF kernel, and Random Forest (RF). Training The individual models are trained on the features that have undergone PCA transformation to classify microorganisms.The results of the training are measured in terms of accuracy, precision, recall and F1-score. Confusion matrices are also created to represent all the correct and incorrect prediction under the different classes of microorganisms which will give a better realization on the workings of each model.

3.3 Model Evaluation and Interpretation

Once the models have been trained they are tested to determine the level of effectiveness with which the models classify microorganisms. Random Forest model is further investigated on the basis of mean decrease impurity to determine the most effective PCA components that play a role in its predictions. Confusion matrices are also used to identify the trends of misclassification which help to comprehend the areas where each of the models is effective and what requires improvement. This assessment test is useful in identifying the most effective models besides giving a more insightful view of the underlying framework of the data.

3.4 Web Application Deployment

It creates a web application with the help of Streamlit in order to make the models easily available to users. The software has easy interface where the users can post their datasets in the form of a CSV file. The data is further preprocessed automatically after uploading it with the earlier saved scaler and PCA models. The system then provides predictions of all models that are trained such as ANN, XGBoost, SVM, and random forest with automated recommendations that are provided by the random forest model itself. The final outputs are also available as CSV files so that the users can download them and further analyze the results in the form of prediction results, see performance metrics, confusion matrices, and feature importance.

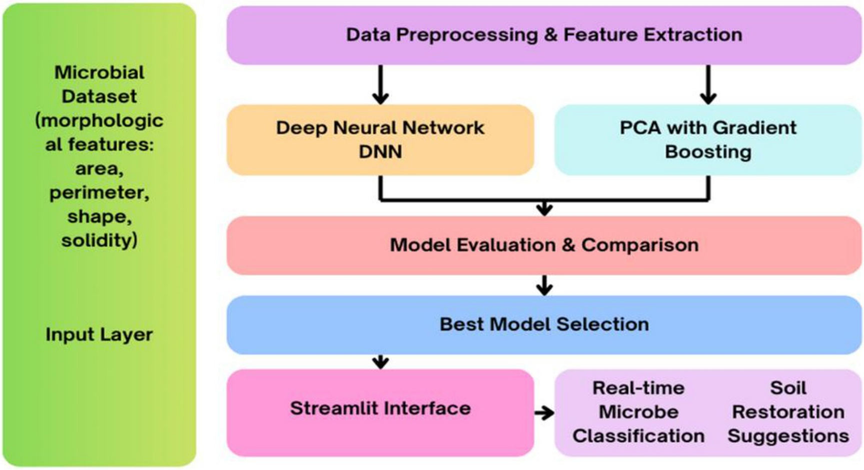


Figure 1.Architecture Diagram

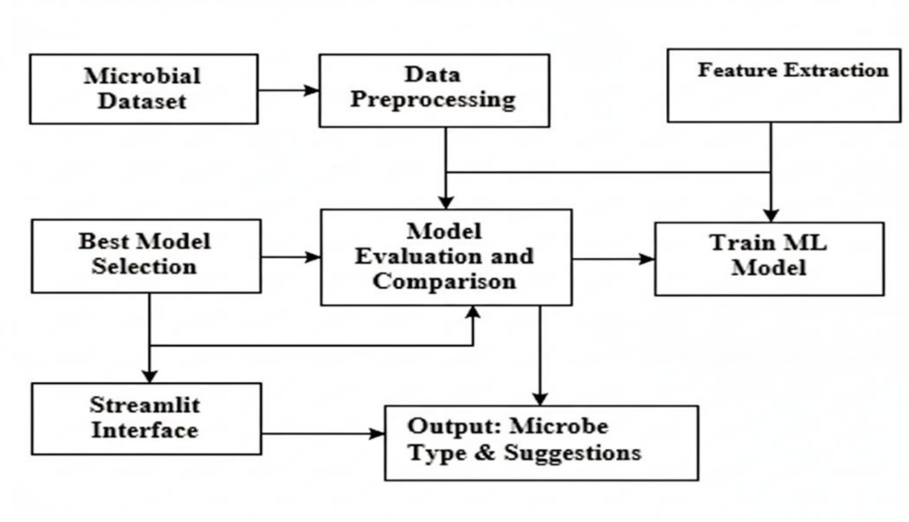


Figure 2. system workflow

4 Experimental Result and Analysis

The Microbes Prediction and Recommendation system works differently with the various machine learning models. XGBoost and Random Forest were the two that had both achieved high training accuracies of 99.64% and 99.80, respectively, and this means that the two have a high learning capacity with regard to the data provided. Their confusion matrices also indicate that there were high levels of correct classification with few errors and this indicates the ANN model was successful at picking patterns in the PCA- transformed features. Although it demonstrated reasonable generalization, its performance did not match in consistency with the results of the tree-based models. The SVM model was found to be lowest in accuracy with 62.1 percent, and high in misclassification in many classes thus showing difficulty in the separation of two or more classes in a high dimensional feature space.Further examinations of the random forest model revealed that PCA components, PCA7, PCA6, and PCA A1 had the highest contribution to the prediction results. This implies that these taxonomic characteristics are crucial to classification of microorganisms. Altogether, the results indicate that tree-based ensemble models are the most predictable using this dataset, and ANN may be considered as the compromise between performance and interpretability. Conversely, SVM does not perform well with multi-class classification.Lastly, the system allows one to retrieve prediction table and CSV files, which can be downloaded anonymously and allow users to check the results and process additional analysis, therefore, improving the usability of the application in a more practical way.

Table 1. Training Accuracy and Performance Metrics

Model	Accuracy	Precision	Recall	F1-Score
XGBoost	0.9964	0.9964	0.9964	0.9964
Radom Forest	0.9980	0.9964	0.9980	0.9980
ANN	0.7154	0.7183	0.7154	0.7129
SVM	0.6212	0.6255	0.6212	0.6083

XGBoost and random forest demonstrated almost 100 percent accuracy on the training data, and it indicates that the models have acquired the training data well but this also indicates the possibility of overfitting. Results of ANN model were quite satisfactory, and generalized better compared to SVM but still were worse than the accuracy of the ensemble methods. SVM, in its turn, performed worse, primarily because of the challenges with the multi-class classification task, which resulted in lower values of both accuracy and F1-score shown in table 1

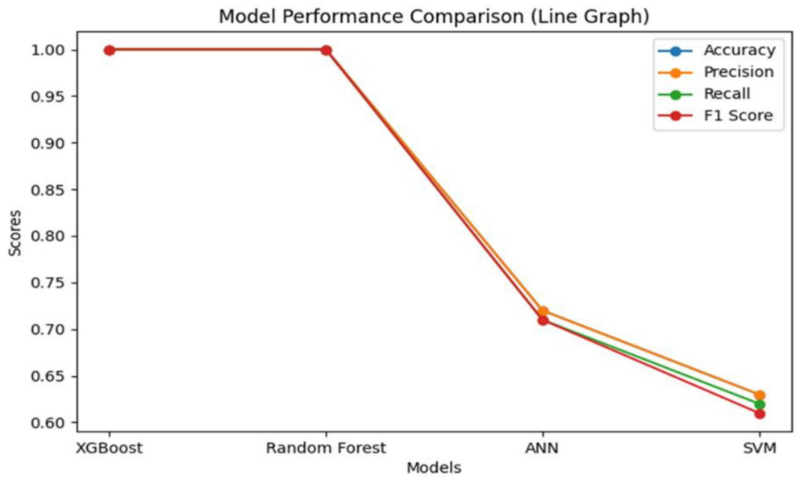


Figure 3. Training Accuracy and Performance Metrics.

Fig 3 shows XGBoost and random forest are evidently among the best with practically perfect accuracy. ANN presents average performance (71.5%) and is fairly superior to

SVM in generalization. SVM has the worst performance (62 percent) and is totally devoid of multi-class classification.

4.1 Evaluation Metrics

4.1.1 Accuracy

Accuracy is an expression of the frequency of appropriate predictions of the system by the number of correctly classified microorganisms versus total samples assessed. The existence of larger values of accuracy means that the model is more efficient to identify various microbial species.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

4.1.2 Precision

Positive Predictive Value is the percentage of the microorganisms that are predicted by a given class that are correct. Greater accuracy translates to less false positives by the model which makes its classifications more dependable.

$$\text{Precision} = \frac{TP}{TP + FP}$$

4.1.3 Recall (Sensitivity / True Positive Rate)

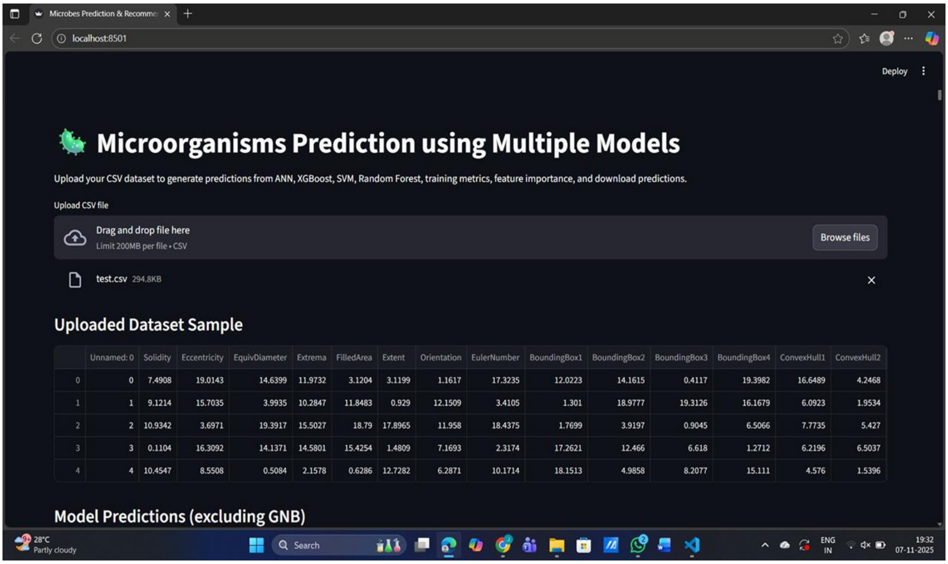
The recall is the indication of the ability of the system to recognize microorganisms that actually belong to a particular classification. With a higher recall, fewer microbes will be missed during the classification process enhancing the completeness of the results.

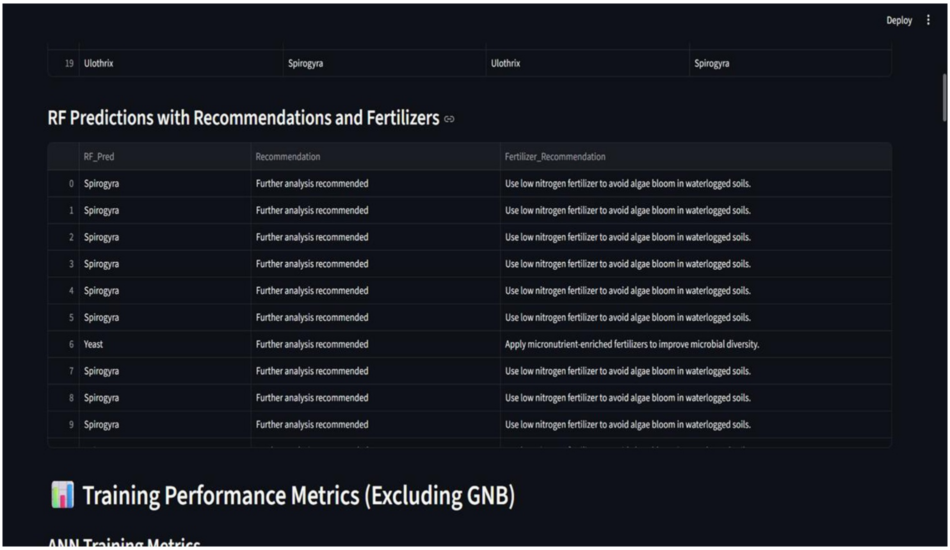
$$\text{Recall} = \frac{TP}{TP + FN}$$

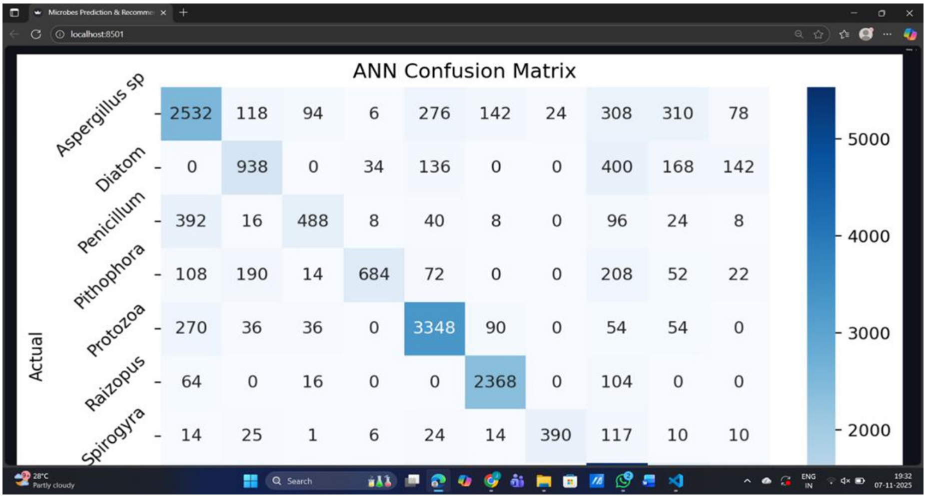
4.1.4 F1-Score

The F1-score is a combination of precision and recall in a single measure where it provides a balanced perspective of the model performance. It is particularly applied in the classification of microorganisms, the false positives and false negatives of which may prove to be critical.

$$\text{F1-Score} = \frac{2 * \text{Precision} + \text{Recall}}{\text{Precision} + \text{Recall}}$$


Microbes Prediction & Recomme
localhost:8501
Deploy
Microorganisms Prediction using Multiple Models
Upload your CSV dataset to generate predictions from ANN, XGBoost, SVM, Random Forest, training metrics, feature importance, and download predictions.
Upload CSV file
Drag and drop file here
Limit 200MB per file • CSV
Browse files
test.csv 294.8KB
Uploaded Dataset Sample
Unnamed: 0 Solidity Eccentricity EquivDiameter Extrema FilledArea Extent Orientation EulerNumber BoundingBox1 BoundingBox2 BoundingBox3 BoundingBox4 ConvexHull1 ConvexHull2
0 0 7.4908 19.0143 14.6399 11.9732 3.1204 3.1199 1.1617 17.3235 12.0223 14.1615 0.4117 19.3982 16.6489 4.2468
1 1 9.1214 15.7035 3.9935 10.2847 11.8483 0.929 12.1509 3.4105 1.301 18.9777 19.3126 16.1679 6.0923 1.9534
2 2 10.9342 3.6971 19.3917 15.5027 18.79 17.8965 11.958 18.4375 1.7699 3.9197 0.9045 6.5066 7.7735 5.427
3 3 0.1104 16.3092 14.1371 14.5801 15.4254 1.4809 7.1693 2.3174 17.2621 12.466 6.618 1.2712 6.2196 6.5037
4 4 10.4547 8.5508 0.5084 2.1578 0.6286 12.7282 6.2871 10.1714 18.1513 4.9858 8.2077 15.111 4.576 1.5396
Model Predictions (excluding GNB)
28°C
Partly cloudy
Search
ENG
IN
19:32
07-11-2025


Deploy
19 Ulothrix Spirogyra Ulothrix Spirogyra
RF Predictions with Recommendations and Fertilizers
RF_Pred Recommendation Fertilizer_Recommendation
0 Spirogyra Further analysis recommended Use low nitrogen fertilizer to avoid algae bloom in waterlogged soils.
1 Spirogyra Further analysis recommended Use low nitrogen fertilizer to avoid algae bloom in waterlogged soils.
2 Spirogyra Further analysis recommended Use low nitrogen fertilizer to avoid algae bloom in waterlogged soils.
3 Spirogyra Further analysis recommended Use low nitrogen fertilizer to avoid algae bloom in waterlogged soils.
4 Spirogyra Further analysis recommended Use low nitrogen fertilizer to avoid algae bloom in waterlogged soils.
5 Spirogyra Further analysis recommended Use low nitrogen fertilizer to avoid algae bloom in waterlogged soils.
6 Yeast Further analysis recommended Apply micronutrient-enriched fertilizers to improve microbial diversity.
7 Spirogyra Further analysis recommended Use low nitrogen fertilizer to avoid algae bloom in waterlogged soils.
8 Spirogyra Further analysis recommended Use low nitrogen fertilizer to avoid algae bloom in waterlogged soils.
9 Spirogyra Further analysis recommended Use low nitrogen fertilizer to avoid algae bloom in waterlogged soils.
Training Performance Metrics (Excluding GNB)


Microbes Prediction & Recomme
localhost:8501
ANN Confusion Matrix
Actual
Aspergillus sp 2532 118 94 6 276 142 24 308 310 78
Diatom 0 938 0 34 136 0 0 400 168 142
Penicillum 392 16 488 8 40 8 0 96 24 8
Pithophora 108 190 14 684 72 0 0 208 52 22
Protozoa 270 36 36 0 3348 90 0 54 54 0
Raizopus 64 0 16 0 0 2368 0 104 0 0
Spirogyra 14 25 1 6 24 14 390 117 10 10
5000
4000
3000
2000
28°C
Partly cloudy
Search
ENG
IN
19:32
07-11-2025

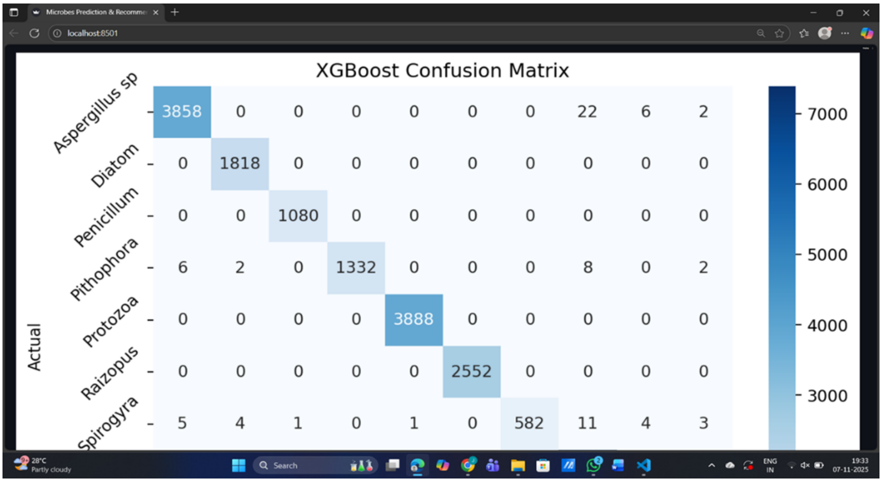


Figure 4. Output images

The dashboard starts by showing a preview of the uploaded dataset that is made up of various numerical features in a number of samples. This data is the input of all prediction models. The Model Predictions and the Random Forest (RF) Predictions and Recommendations are represented below and indicate the results of the training by each trained model, and RF also suggests the recommendations automatically.Training performance measures and confusion matrices can also be found on the dashboard. The performance values (0.71254 and 0.71383) presented in ANN model probably depict some evaluation measures like accuracy, precision or recall. The confusion matrices of ANN and SVM are the most detailed parts of the visual elements and demonstrate the level of accuracy at which each model is able to classify the microorganisms. These matrices indicate the right and wrong prediction of various classes like Aspergillus sp., Diatom, Penicillium, Pithophora and Protozoa and provide an understandable perspective of the advantages and weakness of each model shown in fig 4.

5 Conclusion

The Microbes Prediction and Recommendation System is an ensemble of high-level machine learning techniques and a convenient web interface that is designed to assist with an effective and precise classification of microorganisms. Using data preprocessing, dimensionality reduction by PCA and using several predictive models like ANN, XGBoost, SVM, and Random Forest, the system can perform well, as well as give a clear understanding of the role of features and the behavior of the trained model. Suggestions based on the use of random forests, downloadable, and well-organized results also contribute to the increased practical use of the system as they allow users to use information and make well-informed decisions supported by data.Future consideration may be to introduce the use of datasets representing a wider variety of microbial sources that can contribute to model generalization and eliminate class imbalance. Furthermore, it is possible to investigate the concept of deep learning like convolutional neural networks (CNNs) used on raw images, which can potentially bring more informative features and enhance the classification performance. Generally, the suggested system provides a versatile and stable platform, which links the complicated machine learning operations with a simple end-user interface, which offers an excellent foundation to the further evolution and study.

References

1. Raiten, D.J., Combs, G.F., et al.: Nutritional ecology: Understanding the intersection of climate/environmental change, food systems and health. In: Agriculture for Improved Nutrition: Seizing the Momentum, p. 68 (2024)
2. Cavicchioli, R., Ripple, W.J., Timmis, K.N., Azam, F., Bakken, L.R., Baylis, M., Behrenfeld, M.J., Boetius, A., Boyd, P.W., Classen, A.T., et al.: Scientists' warning to humanity: Microorganisms and climate change. Nat. Rev. Microbiol. (2023)
3. Randall, C.J., van Woesik, R.: Some coral diseases track climate oscillations in the Caribbean. Sci. Rep. 7, 5719 (2022)
4. Huong, N.T.L., Bo, Y.S., Fahad, S.: Economic impact of climate change on agriculture using Ricardian approach: A case of northwest Vietnam. J. Saudi Soc. Agric. Sci. 18(4), 449–457 (2024)
5. Prestele, R., Verburg, P.H.: The overlooked spatial dimension of climate-smart agriculture. Glob. Change Biol. 26(3), 1045–1054 (2022)
6. Jat, H., Choudhary, M., Datta, A., Yadav, A., Meena, M., Devi, R., Gathala, M., Jat, M., McDonald, A., Sharma, P.: Changes in soil microbial properties and nutrient dynamics over time with climate-smart agriculture technologies. Soil Tillage Res. 199, 104595 (2023)
7. Guo, et al.: Integration of smart sensors and phytoremediation for real-time soil microbial restoration in agricultural waste management. Front. Plant Sci. 16, 1550302 (2025)
8. Boursianis, M., Papadopoulou, M.S., Diamantoulakis, P., Liopa-Tsakalidi, A., Barouchas, P., Salahas, G., Karagiannidis, G., Wan, S., Goudos, S.K.: Internet of things (IoT) and agricultural unmanned aerial vehicles (UAVs) in smart farming: A comprehensive review. Internet Things (2024)
9. Koduru, S., Padala, V.P.R., Padala, P.: Smart irrigation system using cloud and Internet of Things. In: Proceedings of the 2nd International Conference on Communication, Computing and Networking, pp. 195–203. Springer (2023)
10. Premkumar, K., Thenmozhi, P., Praveenkumar, P., Monishaa, P., Amirtharajan, R.: IoT-assisted automatic irrigation system using wireless sensor nodes. In: Proceedings of the International Conference on Computer Communication and Informatics (ICCCI 2018), pp. 1–4. IEEE (2022)
11. Wilhelm, R.C., et al.: Predicting measures of soil health using the microbiome and machine learning. Soil Biol. Biochem. 166, 108569 (2022)
12. Singh, et al.: New microbial tools to boost restoration and soil organic matter in degraded ecosystems. Environ. Microbiol. 25(9), 2156–2168 (2023)
13. UNL CropWatch Team: Technological advancements in soil health monitoring and management using real-time sensors. CropWatch Reports, University of Nebraska–Lincoln (2024)
14. Iqbal, S., et al.: Microbial communities and their transformative role in soil health restoration under climate stress. Front. Microbiol. 16, 1345678 (2025)
15. Li, et al.: Monitoring and modeling the soil–plant system toward precision microbial health restoration. Rev. Geophys. 63(1), e2024RG000836 (2025)