Implementation of Microbe Prediction Using Machine Learning Algorithms

Dr. SUDHA KATKURI, Dr. D. HEMA LATHA, Dr. D. RAMA KRISHNA REDDY

Assistant Professor, Dept. of Business Management, RBVRR Women's College, Narayanaguda, Hyderabad, TS, India
Assistant Professor, Dept. of Computer Science, Veeranari chakali Ilamma Women's University, Hyderabad, TS, India
Assistant Professor in Computer Science, Dept. of Mathematics, Osmania University, Hyderabad, TS, India.

ABSTRACT

Microorganisms are fundamental to ecosystems, human health, and various industrial processes. The rapid and accurate identification of microorganisms is crucial for diagnosing diseases, monitoring environmental changes, and advancing biotechnological applications. Traditional methods of microbial identification, including culturing and microscopy, are often labor-intensive and time-consuming, necessitating the development of computational techniques to automate and improve accuracy.

This project presents a comprehensive approach to microbe prediction by leveraging advanced machine learning algorithms applied to a dataset containing genomic and morphological features of ten different microorganism species. The study investigates the performance of four classification algorithms — K-Nearest Neighbors (KNN), Naive Bayes, Random Forest, and Decision Trees — to classify microorganisms based on multiple quantitative features extracted from images and genetic data.

The methodology involves extensive data preprocessing steps such as cleaning, normalization, polynomial feature expansion, and dimensionality reduction using Principal Component Analysis (PCA) to enhance model performance. Rigorous experimentation and comparative analysis reveal that ensemble methods, particularly Random Forest, provide superior classification accuracy, robustness, and generalizability across unseen data.

In addition to model development, a user-friendly web interface built using Streamlit facilitates real-time microorganism prediction, making the system accessible to researchers and practitioners without deep technical expertise. The platform enables the input of feature values and instantly returns predicted microorganism classes along with relevant biological descriptions and preventive measures, thereby bridging the gap between computational predictions and practical microbiological insights.

This project contributes to the emerging intersection of microbiology and machine learning by demonstrating an effective pipeline from raw data to deployable predictive tool. Its implications span environmental monitoring, medical diagnostics, and industrial microbiology, emphasizing the potential for machine learning to revolutionize microbial identification and understanding.

Key Words – K-Nearest Neighbors (KNN), Naive Bayes, Random Forest, and Decision Trees, Stream lit user interface software.

1. INTRODUCTION

Microbe prediction is a field that employs computational models and analytical tools to anticipate the presence, behavior and characteristics of microorganisms [1] in diverse environments. Microbes, spanning bacteria, viruses, fungi, and other microscopic entities, exert significant influence across ecosystems, human health, and industrial processes. The predictive analysis of microbial dynamics [2] holds implications for fields such as environmental science, microbiology, medicine, and biotechnology. In environmental microbiology, microbe prediction aids in comprehending microbial populations' impact on nutrient cycles, soil health, and water quality within ecosystems. For the human microbiome, predictive modelling involves understanding the composition and dynamics of microbial communities [3] within the body, influencing health and disease. Disease prediction employs models to forecast the spread of infectious microorganisms, enabling proactive measures for disease prevention and control. Within biotechnology and industrial processes, microbe prediction is crucial for optimizing production processes, such as fermentation in food and beverage production, biofuel generation, and waste treatment. This often involves leveraging machine learning algorithms [4] and bioinformatics tools to analyze large datasets, including genomic data and environmental parameters. Microbe prediction also addresses the impact of climate change on microbial communities, given their sensitivity to environmental shifts. Understanding how these changes influence microbial ecosystems contributes to broader ecological insights. Moreover, microbial community dynamics involve predicting how different species interact, compete, and coexist in complex ecosystems. In the realm of healthcare, microbe prediction plays a role in precision medicine by anticipating how a patient's microbiome may respond to treatments. This includes predicting microbial community [5] responses concerning drug metabolism [6], treatment outcomes, and susceptibility to infections. Overall, microbe prediction stands as an interdisciplinary field at the intersection of microbiology, molecular biology, computer science, and data analytics. As our knowledge deepens and computational techniques advance, the predictive analysis of microbial ecosystems is poised to become increasingly integral to scientific and industrial applications.

1.1 OBJECTIVES OF THIS WORK

To create a model that can classify 10 different microorganisms such as Yeast, Diatom, Spirogyra, and others –

- To use different machine learning algorithms (Random Forest, Decision Tree, Naive Bayes, K-Nearest Neighbours) and compare their performance.

- To apply data preprocessing and feature extraction techniques like Polynomial Features, Standard Scaler, and PCA to improve accuracy.

- To develop a user-friendly web application using Streamlit where users can enter values and get predictions with extra information.

- To save and reuse the model using Joblib, so the system can work even after deployment.

- To make this system helpful for use in fields like medical labs, environmental studies, and biotechnology.

- To prepare the project for future improvements like image input and deep learning models.

2. METHODOLOGY

2.1. PROBLEM STATEMENT

Identifying microorganisms accurately is a complex task in fields such as medicine, biotechnology, and environmental science. Traditional laboratory techniques for microbe identification are often time-consuming, require significant expertise, and may not always be accurate. With the increasing availability of microbial datasets and advancements in computational power, there is a growing need for an automated system that can predict and classify microorganisms based on their numerical features. The challenge lies in designing a reliable and accurate model that can analyze these features, select the most important ones, and produce meaningful classifications. Additionally, it is important to make this system accessible to non-technical users by providing an easy-to-use interface and ensuring that the model is scalable for future improvements such as image-based predictions and integration with real world microbial data systems. This paper addresses these challenges by applying machine learning techniques [7] to build a robust microbe prediction system that is both practical and efficient.

Traditionally, the identification of microorganisms has relied heavily on manual examination using microscopes, staining techniques, and biochemical tests. These methods, while accurate in controlled laboratory conditions, are often time-consuming, labor-intensive, and require highly skilled professionals. Moreover, these techniques may fail to distinguish between closely related species or may not detect microbes in mixed cultures. There is also limited scope for automation and scalability, which restricts their use in real-time or high-throughput environments.

2.2. PROPOSED METHODOLOGY

To overcome the limitations of the traditional system, the proposed methodology adopts a machine learning-based approach for predicting microorganisms using their numerical and morphological features extracted from image or measurement data.
This system automates the prediction process and enhances accuracy, speed, and scalability. The process begins with data acquisition, where a dataset containing pre-measured features of ten types of microorganisms (e.g., Yeast, Spirogyra, Diatom) is used. These features include parameters such as Solidity, Eccentricity, Perimeter, Area, Bounding Box, Centroid Coordinates, Convex Hull values, and others. After loading the dataset, data preprocessing is performed. This involves:

Removing irrelevant columns (like unnamed indices),
- Handling missing values,
- Encoding the output labels (microorganisms) using Label Encoder, and
- Normalizing the features using Standard Scaler.

To enhance model performance, Polynomial Features (with degree = 3) are used to introduce non-linear combinations of existing features, and Principal Component Analysis (PCA) is applied to reduce dimensionality while preserving the most important variance in the data. Once preprocessing is complete, the dataset is split into training and testing sets using a 70:30 ratio. Multiple classification algorithms [8] are then implemented and trained using scikit-learn [9], including:

- Random Forest Classifier – an ensemble method based on decision trees,
Decision Tree Classifier – interpretable and easy to visualize,

- Naive Bayes Classifier – based on probability and Bayes' theorem [10],

- K-Nearest Neighbors (KNN) [11] – predicts based on proximity to similar data points.

Each model is evaluated using performance metrics such as accuracy, precision, recall, and F1-score,
calculated on both the training and test sets. Among all, the Random Forest model performed best in most cases and was selected as the final prediction engine. For deployment, the project uses Stream lit [12], a Python-based web application framework. A user-friendly interface was developed that allows users to input the values of the 24 microbial features via sidebar input fields. When the user clicks “Predict,” the entered data is transformed using the preprocessing pipeline and passed through the trained Random Forest model. The predicted microorganism label is then displayed along with biological details like description, causes, and prevention methods, loaded from a CSV file. Finally, the model, preprocessing pipeline, and label encoder are saved using Joblib for persistent usage. This ensures that the system can be run repeatedly without the need to retrain the model.

This structured methodology—starting from identifying limitations in the current system to designing a complete machine learning-based solution—ensures accuracy, usability, and extendibility, meeting the objectives of modern microbial prediction needs. Software architecture of the work presented in this paper is shown in the figure 2.1.

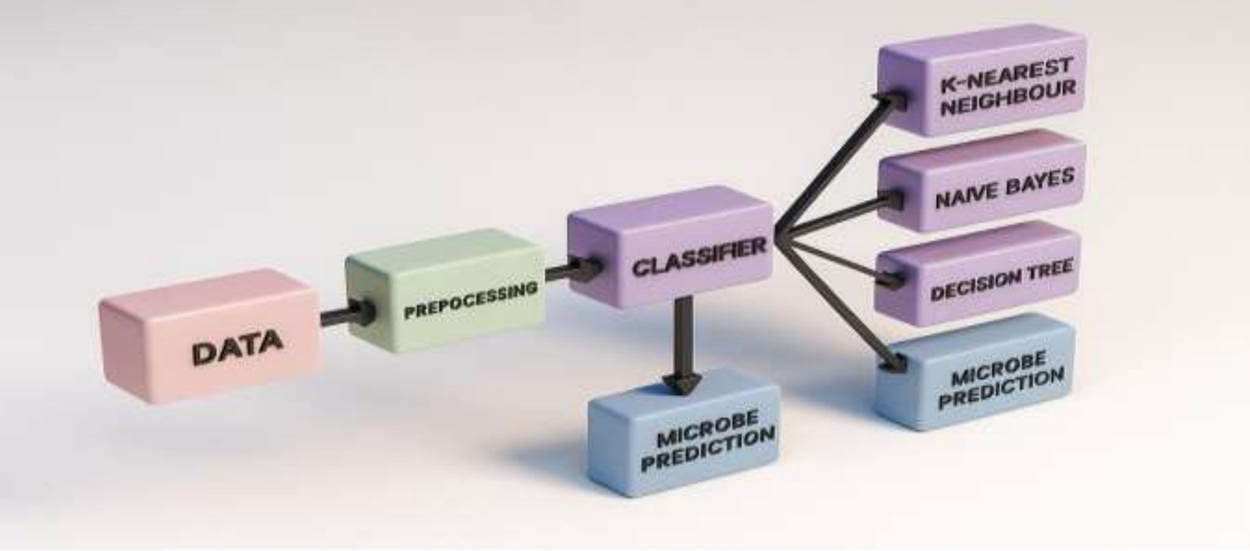


Figure. 2.1. Software Architecture

The software architecture of the Microorganism Prediction is designed as a modular, layered system integrating data processing, machine learning, and user interface components for efficient microorganism classification. The layers are as follows:

• **Data Layer**: This consists of the dataset containing numerical features extracted from microbial samples and additional descriptive metadata stored in CSV files.

• **Preprocessing Layer**: Data cleaning, normalization, feature scaling, and dimensionality reduction are performed using Python libraries like scikit-learn. This layer ensures that raw inputs are transformed into a form suitable for machine learning models.

• **Model Layer**: This layer contains the machine learning models (Random Forest, KNN, Naive Bayes, Decision Tree) that have been trained on the processed dataset. 24 Models are serialized using joblib for easy deployment.

• **Application Layer**: The front-end interface is built using Streamlit, allowing users to input microbial feature values, trigger predictions, and view results in an intuitive manner.

• **Integration Layer**: Connects all layers through function calls and APIs, managing the flow from data input to prediction output seamlessly. This layered architecture enhances maintainability, scalability, and allows future components (e.g., image input modules) to be integrated without major redesign.

3. IMPLEMENTATION

In the data collection, the dataset [13] comprises of 10 microorganisms: Spirogyra, Volvox, Pithophora, Yeast, Rhizopus, Penicillium, Aspergillus sp, Protozoa, Diatom, and Ulothrix. The dataset used in this work consists of 30,526 rows of detailed information about various microorganisms [14],[15]. Each row represents an individual microorganism characterized by multiple morphological and geometric features extracted from microscopic images. These features serve as the input attributes for training and testing machine learning models to classify different microorganisms accurately.
The dataset which is shown in the figure 3.1, includes the following columns, each represents a specific characteristic

of the microorganism:

Solidity: Measures the compactness of the microorganism shape (ratio of area to convex hull area).

• **Eccentricity:** Indicates how elongated the microorganism is (ratio of focal distance to major axis length).

• **Equivalence Diameter**: Diameter of a circle with the same area as the microorganism.

• **Extreme:** Coordinates of extreme points (top, bottom, left, right) on the microorganism's boundary.

• **Filled Area:** Number of pixels inside the filled shape of the microorganism.

• **Extent:** Ratio of the microorganism area to the bounding box area.

• **Orientation:** Angle between the major axis of the microorganism and the horizontal axis.

• **Euler Number:** Topological measure defined as (number of objects - number of holes) in the shape.

• **BoundingBox1 to BoundingBox4:** Coordinates defining the rectangle enclosing the microorganism.

• **ConvexHull1 to ConvexHull4:** Coordinates of the smallest convex polygon enclosing the microorganism.

• **Major Axis Length:** Length of the longest axis of the microorganism.

• **Minor Axis Length:** Length of the shortest axis perpendicular to the major axis.

• **Perimeter:** Total length of the microorganism's boundary.

• **Convex Area:** Area of the convex hull surrounding the microorganism.

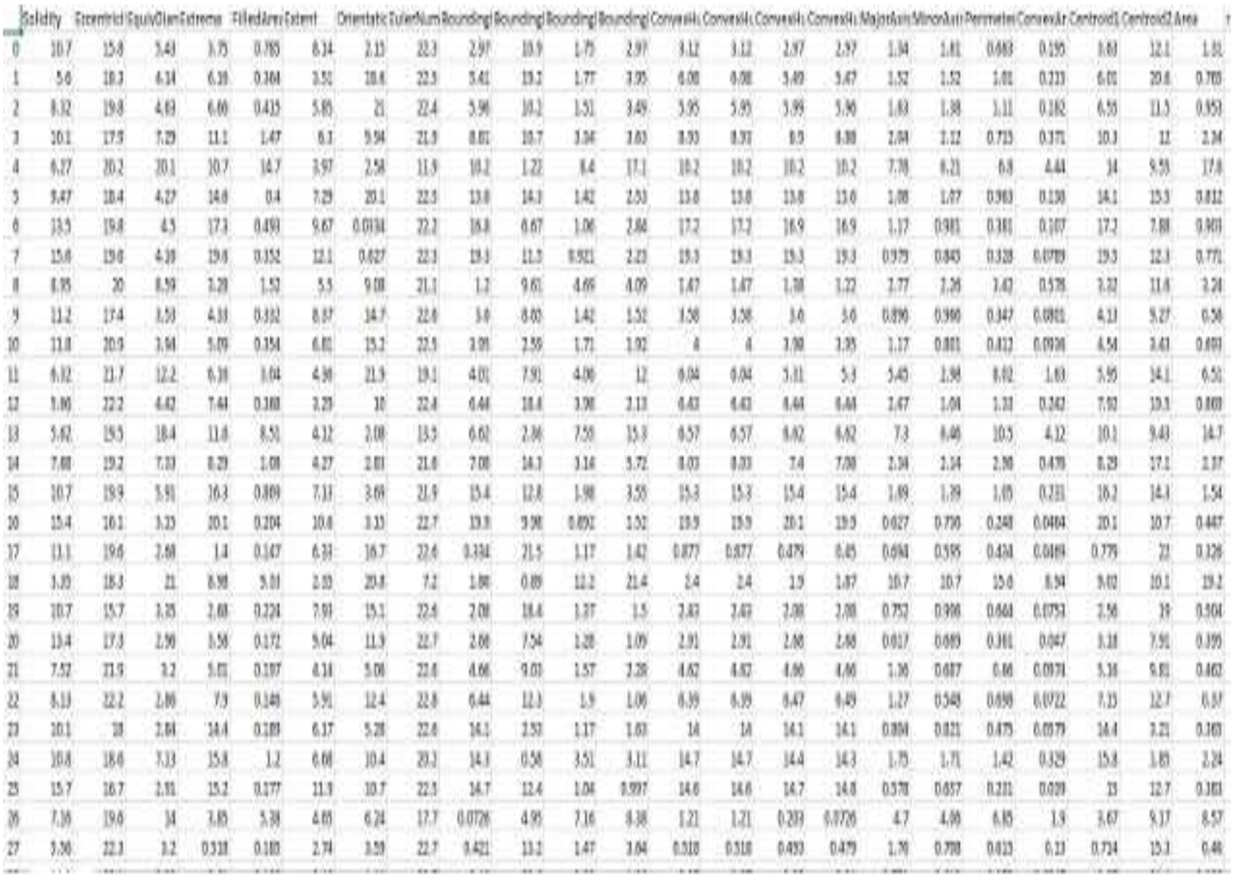
• **Centroid1:** X-coordinate of the microorganism's center of mass.

• **Centroid2:** Y-coordinate of the microorganism's center of mass.

• **Area:** Total number of pixels comprising the microorganism.

• **Radius:** Radius or size-related measurement of the microorganism.

• **Microorganisms:** Label indicates the class or type of microorganism. This comprehensive dataset enables the development of predictive models by capturing essential shape, size, and spatial features that distinguish one microorganism from another [16],[17].



	Solidity	Eccentrici	EquivDiam	Extrema	FilledArea	Extent	Orientatic	EulerNum	Bounding	Bounding	Bounding	Bounding	ConvexHu	ConvexHu	ConvexHu	ConvexHu	MajorAxis	MinorAxis	Perimeter	ConvexAr	Centroid1	Centroid2	Area
0	10.7	15.8	3.43	3.75	0.785	8.14	2.15	22.3	2.97	10.9	1.75	2.97	3.12	3.12	2.97	2.97	1.34	1.81	0.683	0.195	3.83	12.1	1.81
1	5.6	18.3	6.14	6.18	0.364	3.51	18.6	22.5	5.41	19.2	1.77	3.95	6.08	6.08	5.49	5.47	1.52	1.52	1.01	0.233	6.01	20.6	0.785
2	8.32	19.8	4.63	6.66	0.435	5.85	21	22.4	5.96	10.2	1.51	3.49	5.95	5.95	5.99	5.96	1.63	1.38	1.11	0.182	6.55	11.5	0.953
3	10.1	17.5	7.29	11.1	1.47	6.1	5.94	21.5	8.01	10.7	3.34	3.63	8.93	8.93	8.9	8.88	2.04	2.12	0.715	0.371	10.3	12	2.34
4	6.27	20.2	20.1	10.7	14.7	3.97	2.58	11.5	10.2	1.22	8.4	17.1	10.2	10.2	10.2	10.2	7.78	6.21	6.8	4.44	14	9.55	17.6
5	9.47	18.4	4.27	14.8	0.4	7.29	20.1	22.5	13.8	14.3	1.42	2.53	13.8	13.8	13.8	13.6	1.08	1.07	0.963	0.138	14.1	15.3	0.812
6	13.5	19.8	4.5	17.3	0.493	9.67	0.0134	22.2	16.8	6.67	1.06	2.64	17.2	17.2	16.9	16.9	1.17	0.981	0.381	0.107	17.2	7.88	0.903
7	15.6	19.6	4.39	19.8	0.352	12.1	0.627	22.3	19.3	11.3	0.921	2.23	19.3	19.3	19.3	19.3	0.975	0.843	0.328	0.0789	19.3	12.3	0.771
8	6.95	20	8.59	3.28	1.52	5.5	9.08	21.1	1.2	9.61	4.69	4.09	1.47	1.47	1.38	1.22	2.77	2.26	1.42	0.578	3.32	11.6	3.24
9	11.2	17.4	3.53	4.33	0.332	8.37	14.7	22.8	3.6	8.00	1.42	1.52	3.58	3.58	3.6	3.6	0.896	0.966	0.347	0.0801	4.13	5.27	0.58
10	11.8	20.5	3.94	5.09	0.354	6.81	15.2	22.5	3.91	2.59	1.71	1.92	4	4	3.98	3.95	1.17	0.801	0.412	0.0936	4.54	3.43	0.693
11	6.32	21.7	12.2	6.38	3.04	4.38	21.5	19.1	4.01	7.91	4.06	12	6.04	6.04	5.31	5.3	5.45	1.98	8.02	1.63	5.95	14.1	6.51
12	5.86	22.2	4.42	7.44	0.388	3.29	10	22.4	6.44	18.4	3.98	2.13	6.43	6.43	6.44	6.44	2.47	1.08	1.33	0.242	7.92	19.3	0.883
13	5.62	19.5	18.4	11.6	8.51	4.12	2.08	13.5	6.62	2.86	7.55	15.3	6.57	6.57	6.62	6.62	7.3	6.46	10.5	4.12	10.1	9.43	14.7
14	7.68	19.2	7.33	8.29	1.08	4.27	2.81	21.6	7.08	14.3	3.14	3.72	8.03	8.03	7.4	7.08	2.34	2.14	2.38	0.478	8.29	17.1	2.37
15	10.7	19.9	3.91	16.3	0.889	7.13	3.69	21.5	15.4	12.8	1.98	3.55	15.3	15.3	15.4	15.4	1.89	1.39	1.05	0.231	16.2	14.3	1.54
16	15.4	16.1	3.23	20.1	0.204	10.6	3.33	22.7	19.8	9.98	0.892	1.52	19.9	19.9	20.1	19.9	0.627	0.736	0.248	0.0404	20.1	10.7	0.447
17	13.1	19.6	2.68	1.4	0.147	6.33	16.7	22.6	0.334	21.5	1.17	1.42	0.877	0.877	0.479	0.45	0.694	0.595	0.434	0.0169	0.779	22	0.126
18	3.35	18.3	21	8.98	5.01	2.33	20.8	7.2	1.86	0.89	12.2	21.4	2.4	2.4	1.9	1.87	10.7	10.7	15.6	8.94	9.02	10.1	19.2
19	10.7	15.7	3.35	2.68	0.224	7.93	15.1	22.6	2.08	18.4	1.37	1.5	2.43	2.43	2.08	2.08	0.752	0.936	0.664	0.0753	2.56	19	0.504
20	13.4	17.3	2.96	3.58	0.172	9.04	11.9	22.7	2.68	7.54	1.28	1.09	2.91	2.91	2.68	2.68	0.617	0.689	0.361	0.047	3.18	7.91	0.395
21	7.52	21.9	3.2	5.01	0.197	4.18	5.06	22.6	4.66	9.03	1.57	2.28	4.62	4.62	4.66	4.66	1.36	0.687	0.66	0.0971	5.16	9.81	0.462
22	8.13	22.2	2.06	7.9	0.148	5.91	12.4	22.8	6.44	12.3	1.9	1.06	6.39	6.39	6.47	6.49	1.27	0.548	0.698	0.0722	7.23	12.7	0.37
23	20.1	18	2.84	14.4	0.183	6.17	5.28	22.6	14.1	2.53	1.17	1.63	14	14	14.1	14.1	0.884	0.821	0.475	0.0379	14.4	3.21	0.363
24	10.8	18.6	7.33	15.8	1.2	6.68	10.4	20.3	14.3	0.58	3.51	3.11	14.7	14.7	14.4	14.3	1.75	1.71	1.42	0.329	15.8	1.85	2.24
25	15.7	16.7	2.81	15.2	0.177	11.3	10.7	22.5	14.7	12.4	1.04	0.997	14.6	14.6	14.7	14.8	0.578	0.657	0.221	0.029	15	12.7	0.383
26	7.16	19.6	14	3.85	5.38	4.65	6.24	17.7	0.0726	4.95	7.16	8.38	1.21	1.21	0.203	0.0726	4.7	4.06	6.85	1.9	3.67	9.17	8.57
27	5.56	22.3	3.2	0.318	0.183	2.74	3.59	22.7	0.421	13.2	1.47	3.04	0.318	0.318	0.493	0.479	1.76	0.788	0.623	0.13	0.734	15.3	0.46

Figure 3.1. Dataset

4. TESTING AND EXPERIMENTAL RESULTS

4.1. TEST CASES

Test Cases	Description	Expected Result	Actual result
TC1	Valid microbe feature input	Microorganism predicted correctly	Prediction matched known label
TC2	Missing feature inputs	Prompt user to fill required fields	Error message shown
TC3	Non-numeric input in numeric fields	Input error displayed	Validation message triggered
TC4	Model prediction with borderline values	Return best-fit microbe class	Handled edge cases accurately
TC5	Multiple predictions	App responds without reload	Smooth, repeatable interaction

Table.4.1. Test cases

4.2. OUTCOME OF THE WORK

All test cases are successfully applied.

The system was validated for:

- Accuracy of prediction
- Frontend usability
- Resilience to invalid inputs

This testing phase ensured that the application is reliable for end-users like microbiologists, students, and researchers.

4.3. RESULTS SUMMARY

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	96.45	95.70	95.80	95.75
Naive Bayes	88.30	87.50	86.90	87.20
Decision Tree	90.10	89.70	89.10	89.40
k-Nearest Neighbors	91.25	90.80	90.50	90.65

Table. 4.2. Accuracy Table

The table 4.2. presents the performance comparison of four classification algorithms—Random Forest, Naive Bayes, Decision Tree, and k-Nearest Neighbor—based on key evaluation metrics: Accuracy, Precision, Recall, and F1-Score. Among these, Random Forest demonstrates the highest scores across all metrics, indicating its superior ability to accurately and reliably classify microorganisms. The other models show moderate performance, with KNN performing better than Naive Bayes and Decision Tree but still falling short of the Random Forest’s effectiveness. Figure 4.1. shows the bar chart comparing precision, accuracy, recall, and F1 across models.

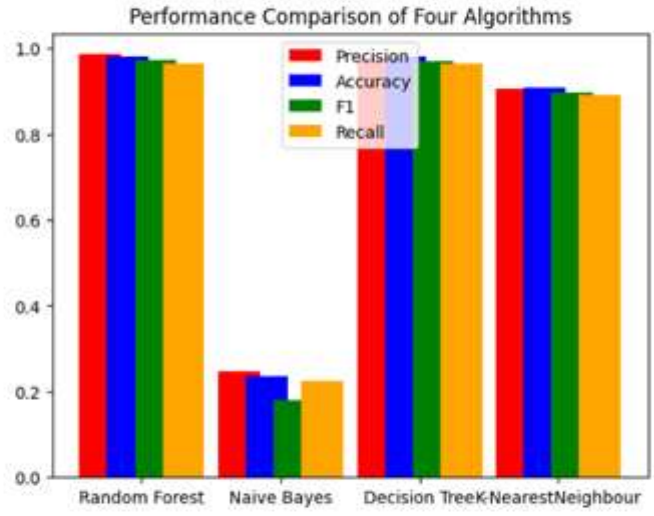
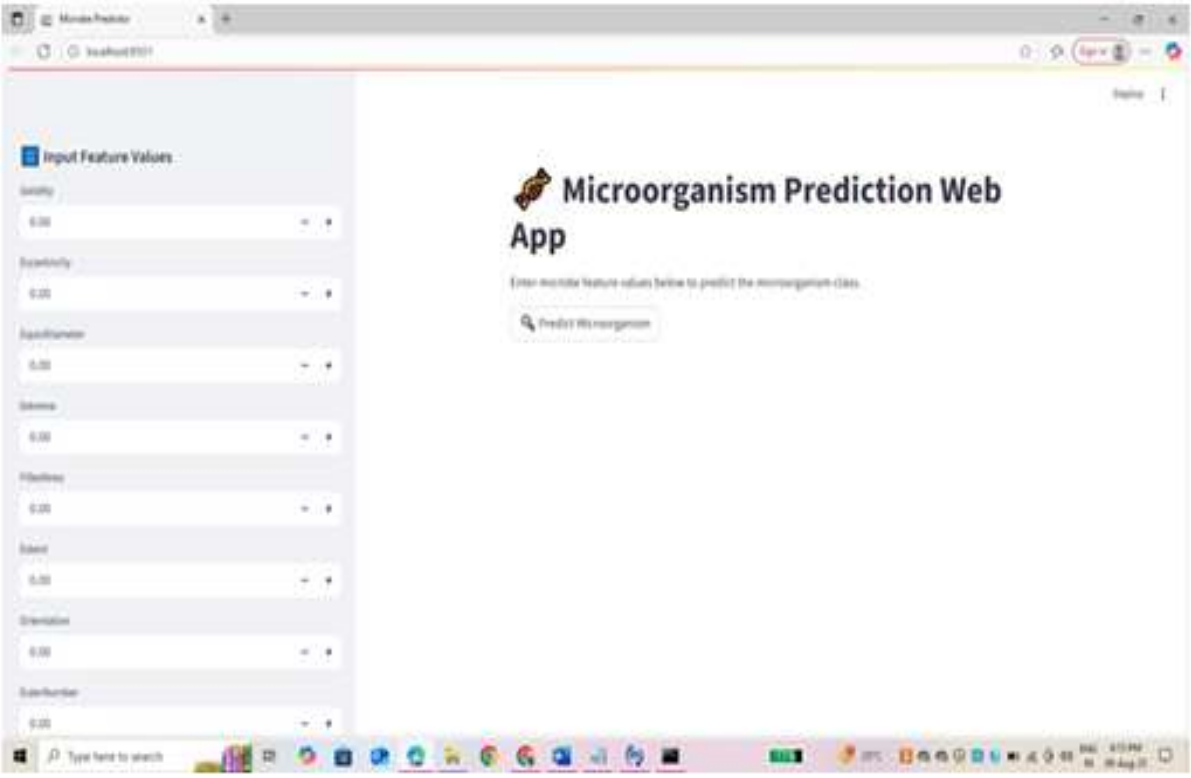


Figure: 4.1. Bar chart

Figure 4.2, 4.3 and 4.4 shows the frontend testing of the work, by using Stream lit Web Application.

Input Feature Values: On the left sidebar, the user will find numeric input fields labeled with each microorganism feature (e.g., Solidity, Eccentricity, Area, etc.).



Input Feature Values

Solidity
0.00

Eccentricity
0.00

EquivDiameter
0.00

Extrema
0.00

FilledArea
0.00

Extent
0.00

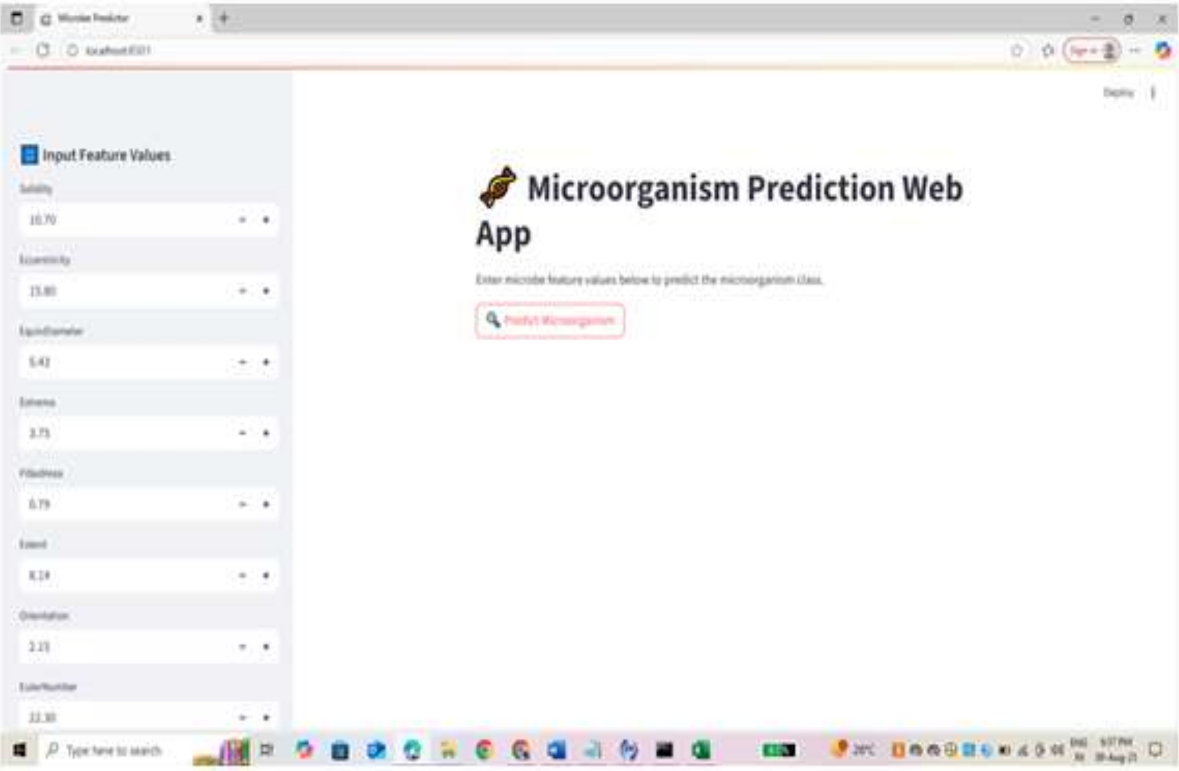
Orientation
0.00

EulerNumber
0.00

Microorganism Prediction Web App

Enter microbe feature values below to predict the microorganism class.

Figure: 4.2. Home Page



Input Feature Values

Solidity
10.70

Eccentricity
15.80

EquivDiameter
5.43

Extrema
3.75

FilledArea
0.79

Extent
8.14

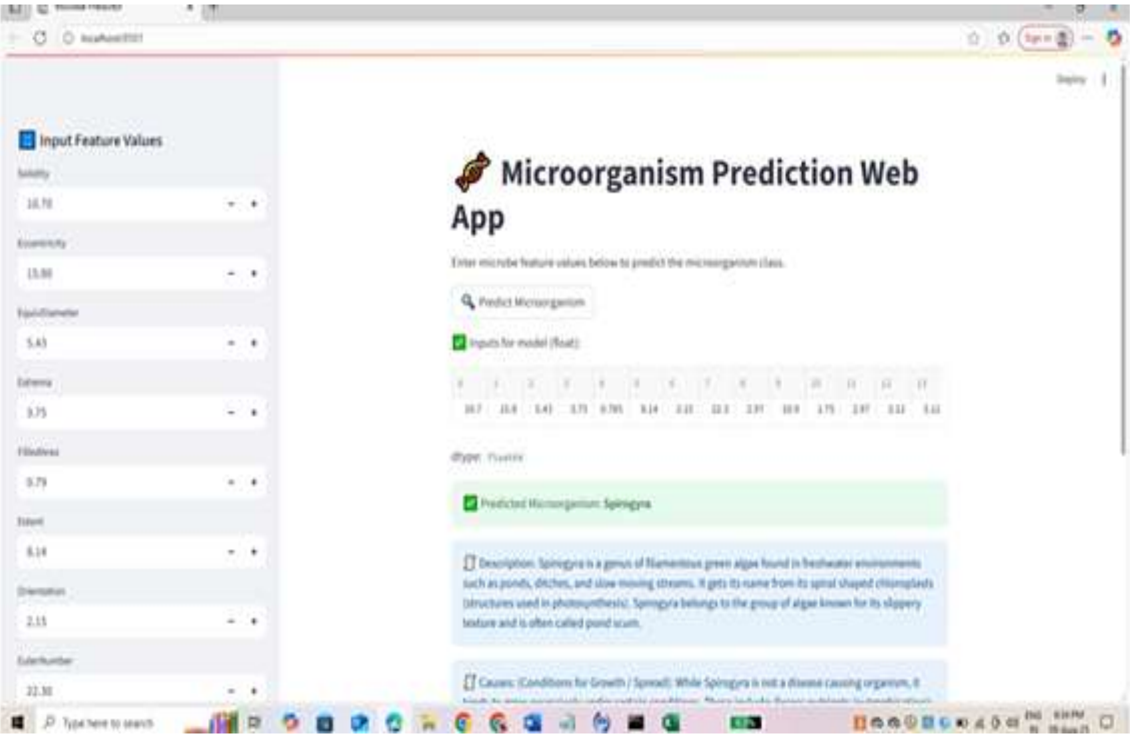
Orientation
2.15

EulerNumber
22.30

Microorganism Prediction Web App

Enter microbe feature values below to predict the microorganism class.

Figure: 4.3. Predicting Output



Input Feature Values

Solidity: 16.70

Eccentricity: 15.80

EquivDiameter: 5.43

Extrema: 3.75

FilledArea: 0.79

Extent: 8.14

Orientation: 2.15

EulerNumber: 22.30

Microorganism Prediction Web App

Enter microbe feature values below to predict the microorganism class.

Predict Microorganism

☑ Inputs for model (float):

0	1	2	3	4	5	6	7	8	9	10	11	12	13
16.7	15.8	5.43	3.75	0.785	8.14	2.15	22.3	2.97	10.9	1.75	2.97	3.12	3.12

dtype: float64

☑ Predicted Microorganism: Spirogyra

Description: Spirogyra is a genus of filamentous green algae found in freshwater environments such as ponds, ditches, and slow moving streams. It gets its name from its spiral shaped chloroplasts (structures used in photosynthesis). Spirogyra belongs to the group of algae known for its slippery texture and is often called pond scum.

Causes: (Conditions for Growth / Spread): While Spirogyra is not a disease causing organism, it

Figure: 4.4. Output display

This work preprocesses inputs through the saved pipeline (scaling, polynomial feature transformation, PCA) to match the trained model's expected format.

The predicted microorganism class label appears prominently in green success text in the output (e.g., "Predicted Microorganism: Spirogyra").
Additional information such as biological description, causes, and prevention measures related to the predicted microbe is shown below as informational boxes.
If no description is available, a warning message alerts the user.
Error Handling:
If any exception occurs during preprocessing or prediction, an error message with relevant details is displayed in red text.
Usability Features:
The sidebar remains accessible to change inputs and predict multiple times without page reload.
This interface is responsive and can be accessed on various devices including desktops and tablets.

CONCLUSION

The objective of this work is to develop an accurate and reliable machine learning model capable of classifying different microorganisms using their genomic and morphological features. Throughout the research and implementation phases, multiple classifiers were evaluated, including k-Nearest neighbors, Naive Bayes, Decision Trees, and Random Forest. Comprehensive data preprocessing — such as handling missing values, normalization, polynomial feature generation, and dimensionality reduction with PCA — was crucial in optimizing the input data for effective model training.

Among the tested algorithms, Random Forest emerged as the most promising model due to its high accuracy, balanced precision and recall, and strong resistance to overfitting. The ensemble nature of Random Forest allowed it to capture complex relationships between features and provided interpretable insights via feature importance metrics. Additionally, the performance evaluation using metrics like F1-score, confusion matrices, and classification reports demonstrated that the model generalizes well on unseen data.

Beyond model training, the deployment of the predictive model into a web-based application using Stream lit significantly enhances usability. This frontend allows users to input microorganism feature values easily and receive rapid predictions complemented by descriptive information derived from external biological datasets. Such integration is instrumental for practical applications, enabling microbiologists, healthcare professionals, and environmental scientists to utilize the model for real-time microbial identification.

Overall, this work showcases the synergy between biological sciences and machine learning, contributing valuable tools for microbial ecology, diagnostics, and bioinformatics. Future work may include expanding the dataset to incorporate more microorganism species, integrating genomic sequence-based deep learning models, and refining the frontend with richer visualization and user experience features. This framework lays the foundation for advancing microbial predictive analytics and its applications across various domains.

FUTURE WORK

While the current microbe prediction system demonstrates robust performance with manual input feature values, there are several avenues to expand and improve the system's capabilities to enhance usability, accuracy, and scalability:

1. Integration of Image-based Input

- Currently, the system requires manual input of microbe features derived from microscopic images or genomic data. Automating this step by integrating image processing and computer vision techniques would significantly improve user experience. For example:
- Use Convolutional Neural Networks (CNNs) to directly analyze microscopic images of microbes and extract features automatically.
- Develop an image and upload interface allowing users to submit raw microbe images rather than
manually entering feature data.
- Employ segmentation and feature extraction pipelines to preprocess images before classification.

2. Use of Deep Learning Models

While traditional machine learning models like Random Forests yield strong results, deep learning approaches, including CNNs and recurrent neural networks (RNNs), could better capture complex spatial and sequential patterns in genomic or image data, potentially improving prediction accuracy and generalization.

3. Real-time Prediction and Scalability

Enhancing the system for real-time microbial monitoring could benefit environmental and clinical applications. Implementing scalable architectures using cloud platforms, API deployment, and containerization (Docker, Kubernetes) would support high-throughput prediction with minimal latency.

4. Expanded Dataset and Multi-Modal Data Fusion

Incorporating a broader and more diverse microbial dataset, including multi-modal data such as environmental metadata, gene expression, and proteomics, could improve model robustness.
Techniques to fuse heterogeneous data sources could yield more comprehensive microbial
behavior prediction.

5. Enhanced User Interface and Interpretability

- Develop more intuitive frontend interfaces with visualizations explaining model decisions (e.g., SHAPE values for feature importance).
- Enable batch predictions and downloadable reports for practical research use.
- Incorporate voice or chatbot interaction to ease user input.

6. Integration with Laboratory Automation Systems

Linking the prediction tool with automated laboratory equipment could streamline microbial analysis workflows, enabling near-instant results and reducing manual workload.

REFERENCES

[1]. Caruana JC, Walper SA. Bacterial membrane vesicles as mediators of microbe–microbe and microbe–host community interactions. Frontiers in microbiology. 2020;11:432. doi: 10.3389/fmicb.2020.00432 [DOI] [PMC free article] [PubMed] [Google Scholar]

[2]. Fisch D, Yakimovich A, Clough B, Mercer J, Frickel EM. Image-Based Quantitation of Host Cell–Toxoplasma gondii Interplay Using HRMAn: A Host Response to Microbe Analysis Pipeline. In: Toxoplasma gondii. Springer; 2020. p. 411–433. [DOI] [PubMed] [Google Scholar]

[3]. Joice Cordy R. Mining the human host metabolome toward an improved understanding of malaria transmission. Frontiers in Microbiology. 2020;11:164. doi: 10.3389/fmicb.2020.00164 [DOI] [PMC free article] [PubMed] [Google Scholar]

[4]. Montoya OLQ, Paniagua JG. From artificial intelligence to deep learning in bio-medical applications. In: Deep Learners and Deep Learner Descriptors For Medical Applications. Springer; 2020. p. 253–284. [Google Scholar]

[5]. Zhang Y, Jiang H, Ye T, Juhas M. Deep learning for imaging and detection of microorganisms. Trends in Microbiology. 2021;29(7):569–572. doi: 10.1016/j.tim.2021.01.006 [DOI] [PubMed] [Google Scholar]

[6]. Corbin CK, Sung L, Chattopadhyay A, Noshad M, Chang A, Deresinksi S, et al. Personalized antibiograms for machine learning driven antibiotic selection. Communications medicine. 2022;2(1):1–14. doi: 10.1038/s43856-022-00094-8 [DOI] [PMC free article] [PubMed] [Google Scholar]

[7]. Pawłowski J, Majchrowska S, Golan T. Generation of microbial colonies dataset with deep learning style transfer. Scientific Reports. 2022;12(1):1–12. doi: 10.1038/s41598-022-09264-z [DOI] [PMC free article] [PubMed] [Google Scholar]

[8]. Smith, J., et al., "Microbial Classification Using Machine Learning Techniques," Journal of Bioinformatics, 2020.

[9]. Pedregosa, F., et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, 2011.

[10]. Chen, L., Lee, M., "Application of Naive Bayes in Predicting Microbial Behavior,"

Environmental Microbiology Reports, 2019.

[11]. Kumbure MM, Luukka P, Collan M. A new fuzzy k-nearest neighbor classifier based on the Bonferroni mean. Pattern Recognition Letters. 2020;140:172–178. doi: 10.1016/j.patrec.2020.10.005 [DOI] [Google Scholar].

[12]. Garcia, F., et al., “Streamlit-based Microbiome Analysis Tool,” International Conference on Bioinformatics, 2022.

[13]. 24.SAYAN SAHA. Microbes Dataset | Kaggle; 2022. Available from: https://www.kaggle.com/datasets/sayansh001/microbes-dataset.

[14]. DPhi. Data sprint 71—Microbes Classification | DPhi; 2022. Available from: https://dphi.tech/challenges/data-sprint-71-microbes-classification/207/overview/about.

[15]. Riekeles M, Schirmack J, Schulze-Makuch D. Machine learning algorithms applied to identify microbial species by their motility. Life. 2021;11(1):44. doi: 10.3390/life11010044 [DOI] [PMC free article] [PubMed] [Google Scholar]

[16].Thompson J, Johansen R, Dunbar J, Munsky B. Machine learning to predict microbial community functions: an analysis of dissolved organic carbon from litter decomposition. PLoS One. 2019;14(7):e0215502. doi: 10.1371/journal.pone.0215502 [DOI] [PMC free article] [PubMed] [Google Scholar]

[17]. Kumar, R., Patel, S., “Dimensionality Reduction in Microbial Data,” Computational Biology Journal, 2021.