

RESEARCH ARTICLE

Predicting microbe organisms using data of living micro forms of life and hybrid microbes classifier

Ali Raza¹, Furqan Rustam², Hafeez Ur Rehman Siddiqui¹, Isabel de la Torre Diez^{3*}, Imran Ashraf^{4*}

1 Department of Computer Science, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan, Pakistan, **2** School of Computer Science, University College Dublin, Dublin, Ireland, **3** Department of Signal Theory and Communications and Telematic Engineering, University of Valladolid, Valladolid, Spain, **4** Information and Communication Engineering, Yeungnam University, Gyeongsan, Korea

* isator@tel.uva.es (ITD); imranashraf@ynu.ac.kr (IA)



OPEN ACCESS

Citation: Raza A, Rustam F, Siddiqui HUR, Diez IdIT, Ashraf I (2023) Predicting microbe organisms using data of living micro forms of life and hybrid microbes classifier. PLoS ONE 18(4): e0284522. <https://doi.org/10.1371/journal.pone.0284522>

Editor: Muhammad Fazal Ijaz, Sejong University, KOREA, REPUBLIC OF

Received: November 20, 2022

Accepted: April 2, 2023

Published: April 20, 2023

Copyright: © 2023 Raza et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its [Supporting information files](#).

Funding: This research was supported by the European University of Atlantic. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Microbe organisms make up approximately 60% of the earth's living matter and the human body is home to millions of microbe organisms. Microbes are microbial threats to health and may lead to several diseases in humans like toxoplasmosis and malaria. The microbiological toxoplasmosis disease in humans is widespread, with a seroprevalence of 3.6-84% in sub-Saharan Africa. This necessitates an automated approach for microbe organisms detection. The primary objective of this study is to predict microbe organisms in the human body. A novel hybrid microbes classifier (HMC) is proposed in this study which is based on a decision tree classifier and extra tree classifier using voting criteria. Experiments involve different machine learning and deep learning models for detecting ten different living micro-forms of life. Results suggest that the proposed HMC approach achieves a 98% accuracy score, 98% geometric mean score, 97% precision score, and 97% Cohen Kappa score. The proposed model outperforms employed models, as well as, existing state-of-the-art models. Moreover, the k-fold cross-validation corroborates the results as well. The research helps microbiologists identify the type of microbe organisms with high accuracy and prevents many diseases through early detection.

Introduction

Microorganisms are the living organisms present on earth. Microorganisms are vital in medical industries to cure many diseases and maintain environmental balance [1]. The microorganism has many forms, some are beneficial, while others are harmful. The harmful microbes cause many infectious diseases and spoil other materials such as food [2]. The microbes are tiny and cannot be seen by the naked eye. A microscope is required to analyze the microorganisms. Microorganisms live everywhere, such as soil, water, and air. Scientists identified that the human body is home to millions of microorganisms. The microorganisms are of numerous

types and species [3]. Each microorganism has its significant purpose. The microorganism can be detected and classified based on its shape, size, and color. The microbe's shape can be the type of rods, spheres, and corkscrew. The microorganism has common types such as fungi [4], viruses, archaea or protists [5], algae, and bacteria [6]. The other ten most important living microforms of life are Volvox, Spirogyra, Yeast, Pithophora, Penicillium, Raizopus, Protozoa, Aspergillus sp, Ulothrix, and Diatom. These microorganisms can be identified based on microscopic data.

The microbe organisms cause many infections and diseases such as toxoplasmosis [7] and malaria [8]. According to a 2019 report, the microbiological toxoplasmosis disease is widespread in humans, with a seroprevalence of 3.6–84% in sub-Saharan Africa [9]. According to the 2020 report of the world health organization (WHO), 241 million malaria cases are found worldwide, and the number of malaria deaths is 627000 [10]. In this regard, an automatic tool for microbe organism detection would be very beneficial to save lives through the early detection of microbiological diseases.

Machine learning and deep learning have witnessed widespread use over the past decade. Artificial intelligence-based tools and techniques are widely used to process, and analyze massive amounts of medical data [11]. Artificial intelligence helps in bioinformatics for decisions making in numerous diseases using predictive analysis. Disease prediction and medical image processing [12] are the primary applications of artificial intelligence. Artificial intelligence algorithms provide the best performance on large-scale data such as the data of microorganisms [13]. With their wide deployment and superior performance, machine learning models have been adopted in disease prediction and biomedical data analytics. Researchers mostly used classical machine learning models for predicting the microorganisms in previously published studies. The prediction performance of previous studies is low using the classical machine learning models. The ensemble learning techniques were applied to enhance the prediction performance task. Keeping in view their outstanding results, this study follows a machine learning-based approach and makes the following primary contributions toward the prediction of the microbe organisms

- Microbe exploratory data analysis (MEAA) is applied to determine the dataset patterns and valuable insights for predicting the microbe organisms. The MEAA is based on the data graphs and charts representing the relations of dataset features.
- A novel hybrid microbes classifier (HMC) is proposed based on a decision tree classifier (DTC) and extra tree classifier (ETC) techniques for predicting microbe organisms. The final prediction is made using the voting criterion. Experiments involve multi-class classification with ten classes including Aspergillus sp, Diatom, Penicillium, Pithophora, Protozoa, Raizopus, Spirogyra, Ulothrix, Volvox, and Yeast.
- Ten machine learning and deep learning-based models are applied in comparison to the proposed approach for predicting microbe organisms. The multi-layer perceptron classifier (MLP), DTC, random forest classifier (RFC), logistic regression (LR), k-nearest neighbors (KNN), gradient boosting classifier (GBC), ETC, and support vector machines (SVM) are employed in this regard. Also, long short-term memory (LSTM) and gated recurrent unit (GRU) is used as the deep learning models. The performance is analyzed with respect to the accuracy, precision, recall, F1 score, and k-fold cross-validation

The remainder of this study is organized as follows. Section 2 is based on the related literature analysis. The methodology and proposed approach are discussed in Section 3. Experimental results and discussions are given in Section 4. Finally, the study is concluded in Section 5.

Related work

The identification of microbial contaminants in the pharmaceutical industry using a deep learning-based approach is studied in [14]. The Raman spectroscopy dataset is utilized to build the deep learning model. The dataset target microbial contaminants are gram-positive bacteria, gram-negative bacteria, and fungi. The convolution neural network (CNN) is used for experiments which achieve a 95% accuracy score for microbial contaminants prediction. The prediction of personalized antibiograms in microbiology using machine learning is carried out in [15]. The electronic health record data of 8342 infections and 15806 uncomplicated urinary tract infections is utilized for the model building. The gradient boosted tree (GBT) shows outstanding results among the employed machine learning models. The personalized antibiograms performance coverage rate is 90% using the proposed technique.

The generation and classification of microbial colonies images using deep learning-based models is studied in [16]. The synthetic microbial colonies dataset of Petri dishes [17] is utilized. The multi-class data of five different microbial species are utilized for classification. The R-CNN model is employed for generating and detecting microbial colonies. The proposed approach achieved a mean squared error score of 4.49 and a mean average precision accuracy score of 0.520.

The study [18] performs the detection of candida albicans fluconazole resistance using a machine learning approach. The combined dataset based on matrix-assisted laser ionization (MALI), time-of-flight (TOF), and mass spectrometry (MS) is utilized for building machine learning models. The authors leverage the linear discriminant analysis (LDA) for the detection of candida albicans which yields an 85% accuracy. Similarly, [19] proposed the detection of carbapenem-resistant *Klebsiella pneumoniae* in microbiology using a supervised machine learning approach. The MALDI-TOF MS data is utilized in this research. The study proposes a modified random forest (RF) technique that achieves an accuracy score of 97% for the detection task.

The prediction of methicillin-resistant *Staphylococcus aureus* using machine learning methods is studied in [20]. The MALDI-TOF MS spectrum data is utilized with the SVM model. Results show an accuracy of 86% using the SVM. The authors study the classification of group B *Streptococcus* serotypes in [21]. The MALDI-TOF MS data is utilized with SVM and RF models. Results suggest that the RF model outperforms with an accuracy score of 87%.

Skin syndrome detection based on deep neural networks is presented in this study [22]. The deep learning-based techniques MobileNet and long short-term memory (LSTM) are utilized to classify skin disease in real time. The proposed model achieved 85% accuracy on the HAM10000 dataset. However, it can be further improved by fine-tuning different parameters. The automatic detection of Alzheimer's disease using the fusion-based approach with a heterogeneous ensemble classifier is proposed in [23]. The proposed framework is utilized to predict Alzheimer's disease based on multimodal time-series data. The dataset is based on 1371 subjects from the Alzheimer's disease neuroimaging initiative (ADNI). Experimental results show that the proposed model achieves superior results in comparison with the state-of-the-art technique for Alzheimer's prediction.

The related literature in the context of predicting microbe organisms is examined in this section. The related research proposed approach, dataset, performance score, and the main aim of the research are analyzed. The past applied state-of-the-art approaches are comparatively analyzed in Table 1.

Study methodology

The methodological analysis of the proposed approach for predicting the microbe organisms in microbiology is visualized in Fig 1. The data of different living microforms of life is utilized

Table 1. The analysis of related literature in the context of predicting the microbe organisms.

Ref.	Year	Approach	Dataset	Accuracy %	Aim
[14]	2020	CNN	Raman spectroscopy	95	Identifying microbial contaminants in the pharmaceutical industry using a deep learning-based approach.
[15]	2022	GBT	Electronic health record data	90	The prediction of personalized antibiograms in microbiology using machine learning was proposed.
[16]	2022	R-CNN	Synthetic microbial colonies	52	The generation and classification of microbial colonies images using deep learning-based models were proposed.
[18]	2020	LDA	MALI- TOF- MS	85	The detection of candida albicans fluconazole resistance in microbiology using machine learning was proposed.
[19]	2020	RF	MALI- TOF- MS	97	This study proposed the detection of carbapenem-resistant Klebsiella pneumoniae in microbiology using a supervised machine learning approach.
[20]	2018	SVM	MALDI-TOF MS spectrum	86	The prediction of methicillin-resistant Staphylococcus aureus using machine learning methods was proposed.
[21]	2019	RF	MALDI-TOF MS	87	The classification of group B Streptococcus serotypes using machine learning was proposed.

<https://doi.org/10.1371/journal.pone.0284522.t001>

for conducting the research experiments. The MEDA is applied to obtain critical insights and patterns in predicting the microbes. The target class in data is encoded to numeric form to transform the labels into machine readable form. The preprocessed data is split into train and test portions with a ratio of 0.8 to 0.2. The novel proposed HMC approach is trained with 80% of data and evaluated using 20% of unseen test data. The proposed HMC approach is fully hyper parameterized to obtain the best results.

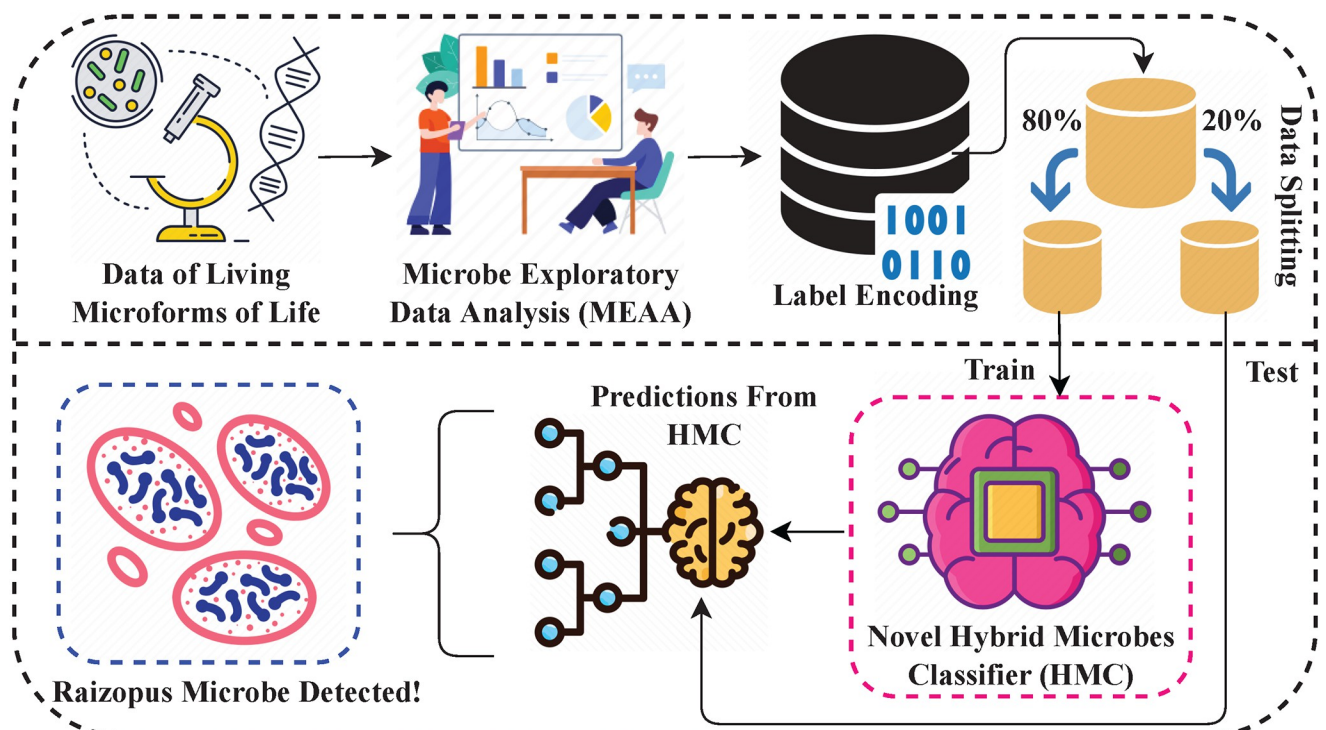


Fig 1. The architecture of the proposed approach for predicting microbe organisms. It involves data collection, exploratory data analysis, model training and testing.

<https://doi.org/10.1371/journal.pone.0284522.g001>

Microbe organisms data

The research utilizes the data of different living microforms of life that is publicly available at Kaggle [24] and used in a DPhi challenge [25]. The data contains ten different living microforms of life, which are Volvox, Spirogyra, Yeast, Pithophora, Penicillium, Raizopus, Protozoa, Aspergillus sp, Ulothrix, and Diatom. The description of the different features, types, and counts are given in Table 2. The utilized dataset features are based on the 21368 microscopic object images of different living microforms of life. The dataset is based on the 25 microscopic object features which are used to predict microbe organisms in our research study.

Microbe exploratory data analysis

MEDA is applied to the research dataset to determine patterns and valuable insights in predicting microbe organisms. The graph and chart-based MEDA are performed, representing the relations of dataset features.

The bar chart-based microorganisms target label frequency analysis is performed in Fig 2. The frequency for each label is represented in the chart's y -axis. The analysis demonstrates that the target label Ulothrix contains 5194, Volvox contains 3024, Protozoa contains 2721, Aspergillus sp contains 2721, Yeast contains 2520, Raizopus contains 1786, Diatom contains

Table 2. Description of dataset features.

Sr no.	Feature	Non-null count	Data type	Description
1	Solidity	21368	float64	The solidity is the ratio of the area of an object to the area of a convex hull of the object.
2	Eccentricity	21368	float64	The eccentricity is the ratio of the length of the major to the minor axis of an object.
3	EquivDiameter	21368	float64	The diameter of a circle with the same area as the region.
4	Extrema	21368	float64	The extrema are the points in the region.
5	FilledArea	21368	float64	The number of pixels in the filled image returned as a scalar.
6	Extent	21368	float64	The ratio of the pixel area of a region with respect to the bounding box area of an object.
7	Orientation	21368	float64	The overall direction of the shape. The value ranges from -90 degrees to 90 degrees.
8	EulerNumber	21368	float64	The number of objects in the region minus the number of holes in those objects.
9	BoundingBox1	21368	float64	Position and size of the smallest box (rectangle) which bounds the object.
10	BoundingBox2	21368	float64	
11	BoundingBox3	21368	float64	
12	BoundingBox4	21368	float64	
13	ConvexHull1	21368	float64	Smallest convex shape/polygon that contains the object.
14	ConvexHull2	21368	float64	
15	ConvexHull3	21368	float64	
16	ConvexHull4	21368	float64	
17	MajorAxisLength	21368	float64	The major axis is the endpoints of the longest line that can be drawn through the object. The length (in pixels) of the major axis is the largest dimension of the object.
18	MinorAxisLength	21368	float64	The axis perpendicular to the major axis is called the minor axis. The minor axis's length (in pixels) is the smallest line connecting a pair of points on the contour.
19	Perimeter	21368	float64	The number of pixels around the border of the region.
20	ConvexArea	21368	float64	Centre of mass of the region. It is a measure of an object's location in the image.
21	Centroid1	21368	float64	The centre point of the object.
22	Centroid2	21368	float64	
23	Area	21368	float64	The total number of pixels in a region/shape.
24	raddi	21368	float64	The radius of the object.
25	microorganisms	21368	object	The microorganism target class to where they belong.

<https://doi.org/10.1371/journal.pone.0284522.t002>

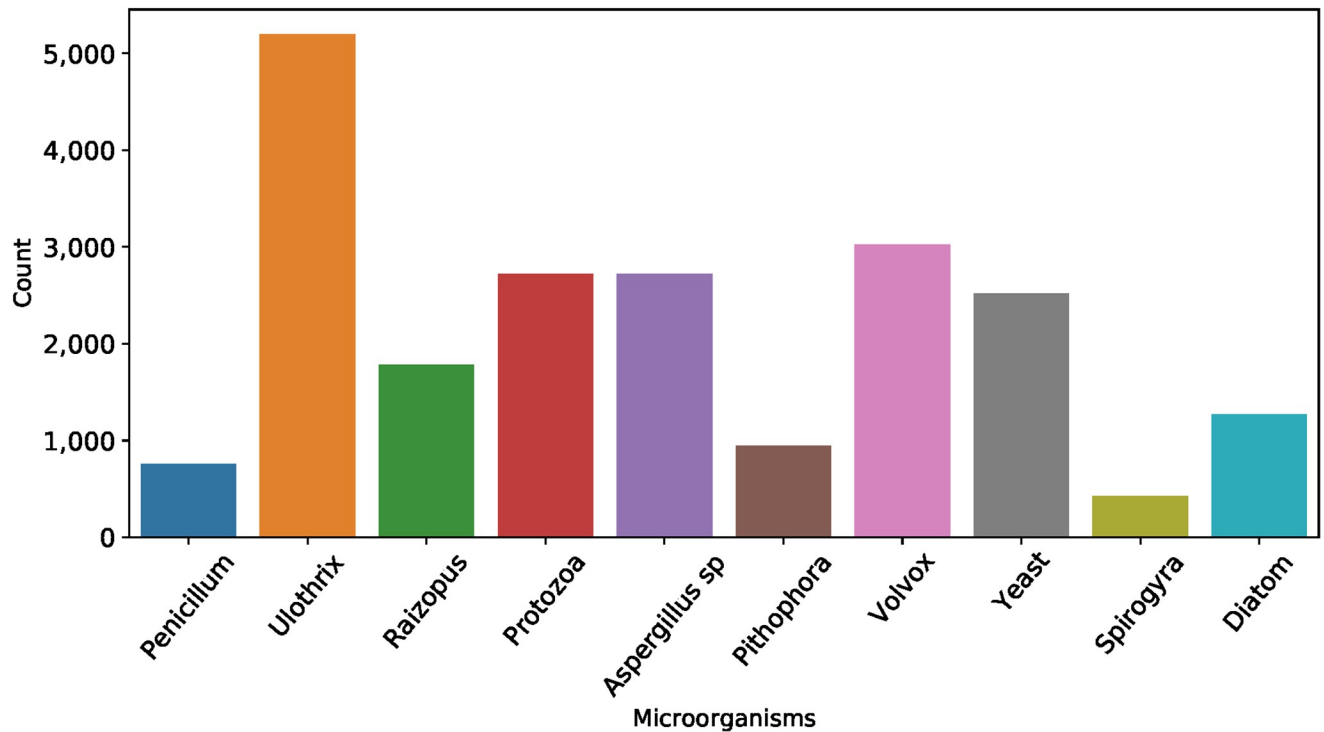


Fig 2. The bar chart-based frequency analysis of each microorganism target label showing the number of samples in each class.

<https://doi.org/10.1371/journal.pone.0284522.g002>

1273, Pithophora contains 945, Penicillium contains 756, and Spirogyra contains 428 instances. This analysis shows that Ulothrix class contains a high number of instances, and Spirogyra contains the lowest number of instances.

The statistical correlation analysis is visualized in Fig 3. The correlation is utilized to determine the linear relationship between two dataset features and analyze their association. This explains how features are related to each other. The analysis demonstrates that the features Extrema, BoundingBox, ConvexHull, and centroid have high correlation values. The features MajorAxisLength, MinorAxisLength, Perimeter, and ConvexArea also have good correlations association. The features Solidity, Extent, and EulerNumber have negative correlation values.

The scatter plot-based analysis of different data features is shown in Figs 4 and 5. The scatter plot is primarily utilized to determine the relationships between two dataset features. The dot values in the scatter plot represent the patterns involved in the prediction process. The purpose of the scatter plot is to observe the relation when the values of features change. The scatter plot analysis of features Solidity and Eccentricity along with the target class is visualized in Fig 4. The analysis demonstrates that the microorganisms have the Solidity and Eccentricity feature values in the range of 5 to 30. The analysis shows that the Raizopus microbe is identified when the Solidity values are above 15 and less than 20. All other microbes are identified when the Solidity values are less than 18 and Eccentricity values are above 5. There is a high chance of microorganism detection when the Eccentricity values are above 15 and the Solidity values are above 3.

The scatter plot analysis of features Extent and Orientation along with the target class is visualized in Fig 5. The analysis demonstrates that the microorganisms have Extent feature values in the range of 0 to 20 and Orientation feature values in the range of 0 to 30. The analysis

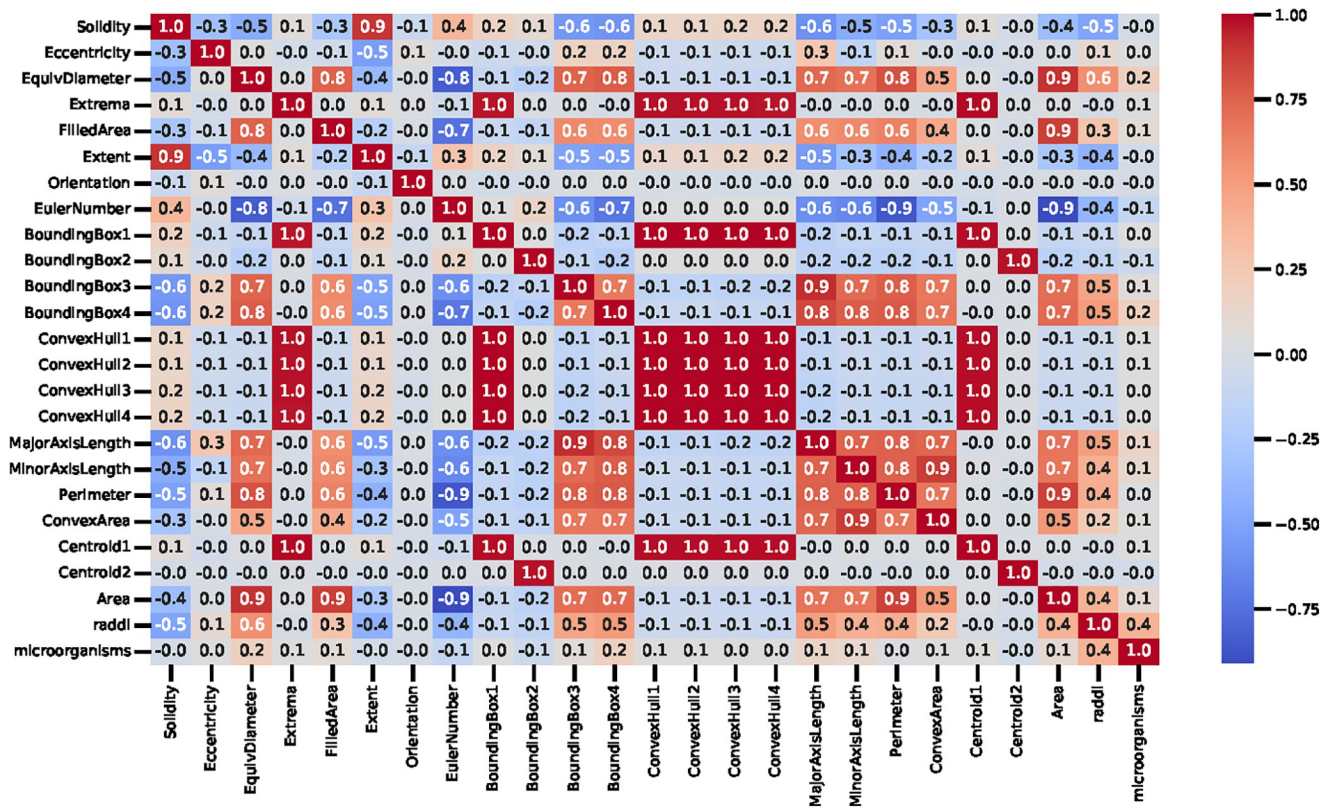


Fig 3. The correlation analysis of employed dataset features indicating the importance of features regarding the target class.

<https://doi.org/10.1371/journal.pone.0284522.g003>

shows that the microorganisms are identified when the Extent values are between 0 and 15. The high chances of Raizopus microbe detection when the Extent values are above 10.

Label encoding and data splitting

We have transformed the dataset target class labels into the machine-readable numeric form using the label encoding technique. The label encoder module from scikit-learn is utilized for the encoding process. The module encodes the target labels with a value between 0 and the total number of classes. Data splitting is a crucial part of machine learning which is applied to split the data into training and testing sets. We split the microbe dataset into 80–20 train-test splits.

Proposed hybrid classifier

A novel HMC is proposed based on a hybrid of DTC and ETC for predicting microbe organisms. The architecture of the proposed HMC approach is shown in Fig 6. The data of different living microforms of life is input to both DTC, and ETC approaches. The DTC and ETC are combined to predict the microbe organisms. The class with the majority of votes from individual predictions is taken to make the final prediction using voting. The final predictions are obtained by using ‘hard’ voting.

The proposed hybrid classifier is based on the combination of multiple supervised classifiers. The key objective of the proposed ensemble method is to reduce variance and bias thus enhancing the prediction performance. The ensemble hybrid methods are proven to show

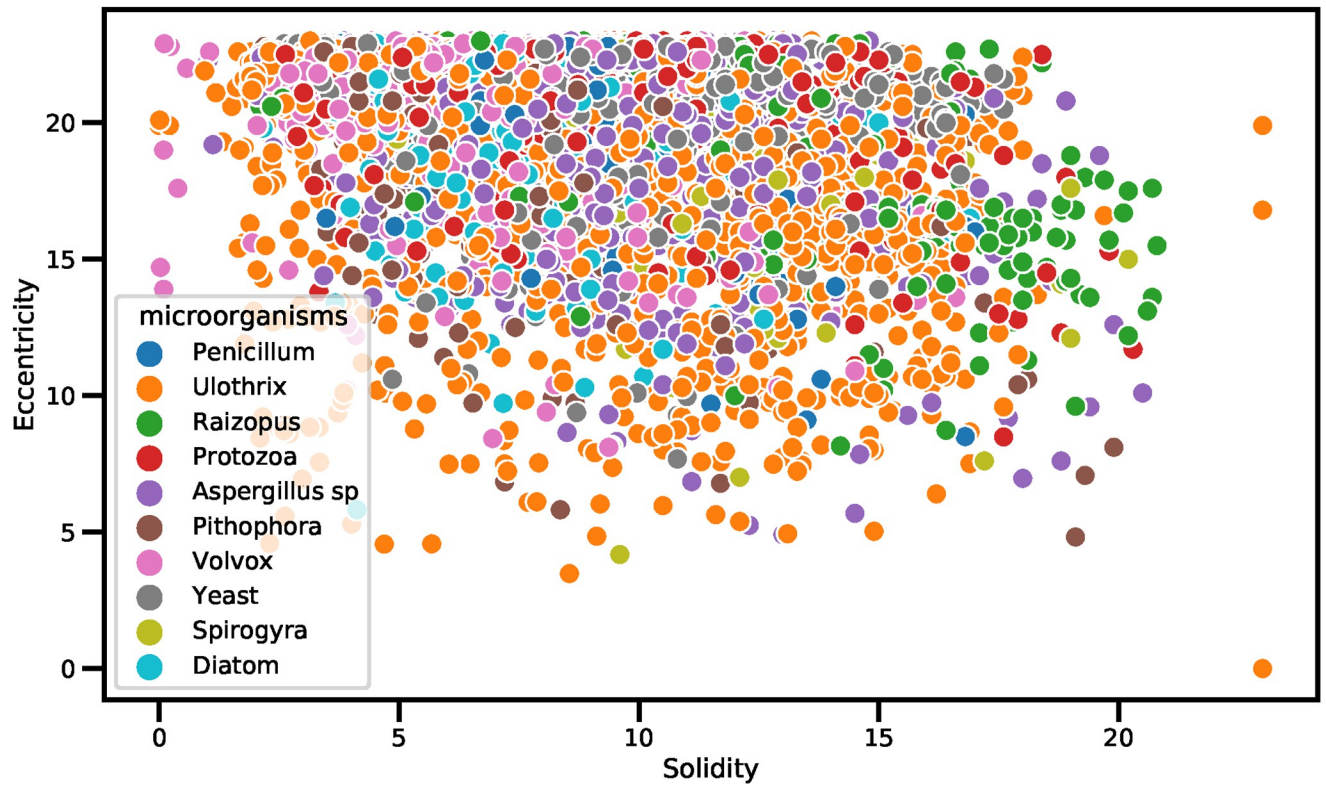


Fig 4. The scatter plot showing the distribution of features regarding Solidity and Eccentricity along with the target class.

<https://doi.org/10.1371/journal.pone.0284522.g004>

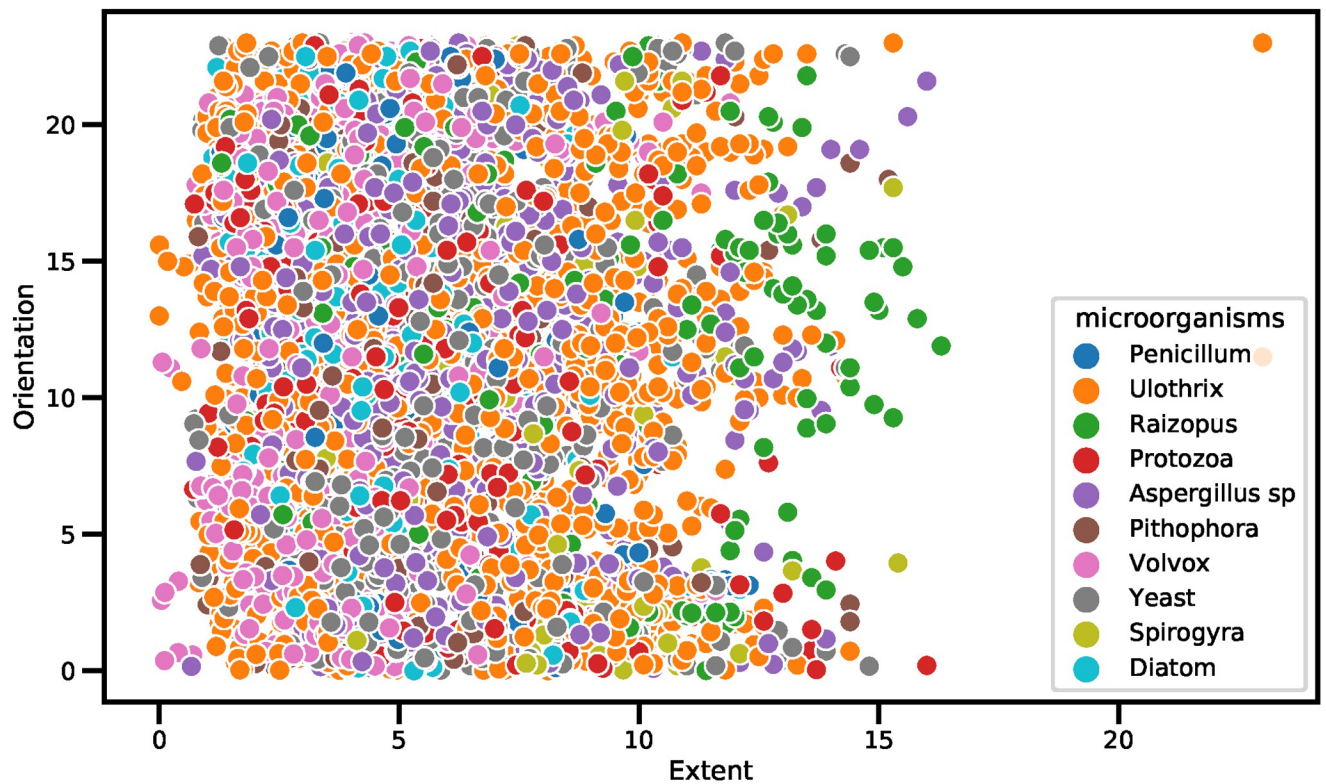


Fig 5. The scatter showing the distribution of features regarding Extent and Orientation along with the target class.

<https://doi.org/10.1371/journal.pone.0284522.g005>

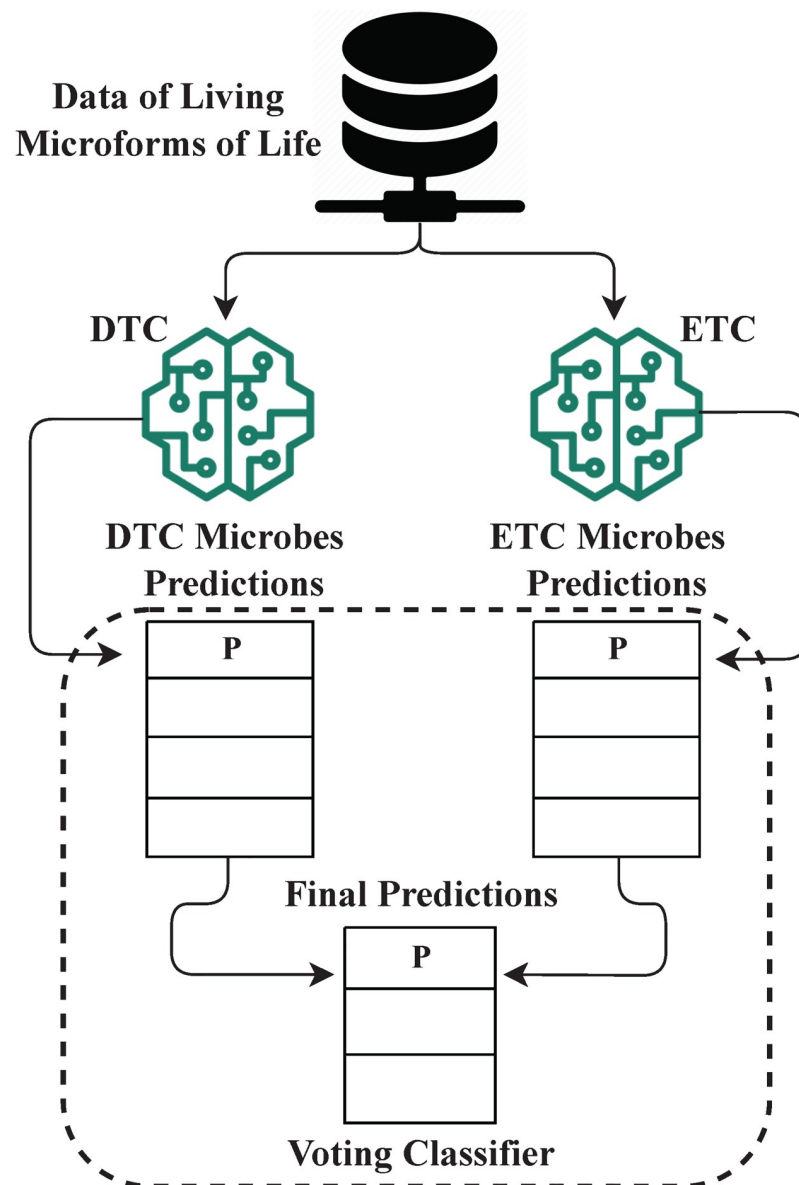


Fig 6. The architecture of the proposed HMC approach showing the voting process for the hybrid classifier.

<https://doi.org/10.1371/journal.pone.0284522.g006>

better performance where the dataset has a higher number of features. The predictions of each classifier are passed to the voting classifier to predict the output class based on the majority voting. The prediction performance is improved by resolving the error of each classifier during voting.

Employed machine learning models

The applied machine learning and deep learning models for predicting microbe organisms in microbiology are analyzed in this section.

The DTC is a supervised machine learning model commonly used to solve classification problems [26]. The DTC follows the tree structure to make a decision on data samples. The leaf nodes in the tree contain the target class labels, the tree branches represent the decision

rules, and the internal nodes contain the data attributes. The Gini index is mainly utilized in DTC to select the best data attributes during tree constructions as expressed in Eq 1, where p represents the probability of data attributes.

$$Gini\ index = 1 - \sum_j P_j^2 \quad (1)$$

RFC is an ensemble learning model which utilizes decision trees [27]. The RFC model works similarly to the DTC model. In the RFC model, multiple decision trees are created for prediction tasks instead of creating a single tree. The prediction outcomes from multiples tree are taken to make the final prediction. RFC helps to improve the prediction accuracy and control model over-fitting.

ETC is also an ensemble learning method widely used for the classification task [28]. The bagged decision trees are constructed in the ETC model for prediction. ETC is similar to the RFC model. The only difference is the tree-based forest construction in the ETC model. The predictions from multiple de-correlated decision trees are aggregated to make the final prediction.

GBC is an ensemble Learning model [29]. The GBC model combines multiple weak classifiers into a robust classifier to obtain high accuracy. During training, each weak classifier improves accuracy and reduces errors. The gradient boosting is based on the decision trees.

KNN is a non-parametric learning classifier mainly used for classification and regression problems [30]. The KNN model makes the groups of data have similar properties. The Euclidean distance metric is utilized to find the similarity between data points. For each data point, the Euclidean distance values are determined by the data points near it.

LR is another widely used supervised method primarily used to solve classification problems [31]. LR model determines the relationship between the independent and dependent variables. LR is a statistical method that utilizes a logistic sigmoid function for classification tasks. The probabilistic values lie between zero and one for using the logistic sigmoid function. Eq 2, represents the prediction process by the LR model.

$$y = \frac{e^{b_0 + b_1 * z}}{1 + (e^{b_0 + b_1 * z})} \quad (2)$$

where y is the predicted class, b_0 is the bias term, and b_1 is the coefficient for input x .

SVM is a supervised method that utilizes the support vectors to classify the data points [32]. The primary motive of the SVM model is to determine the best-fit decision boundary. The best-fit decision boundary classifies the n -dimensional feature space data into the target label. The best-fit decision boundary is also known as the hyperplane [33]. The error is minimized by the iterative process of finding the best-fit decision boundary. SVM selects the extreme support vectors to create the hyperplane. The best-fit hyperplane is represented in Eq 3.

$$\vec{w} \cdot \vec{x} + b = 0 \quad (3)$$

where w represents the weight matrix, x represents the input features and b indicates the biased values.

MLP is a feedforward artificial neural network-based supervised learning model [34]. The artificial neural network uses many representation layers to process the data. The model layers contain neuron units in the network. The layers have the graph representation between the input and output layers. The backpropagation technique [35] is utilized in the MLP model to train the network.

Table 3. The hyperparameters of employed learning techniques.

Technique	Hyperparameters
MLP	Hidden_layer_sizes = 80, Max_iter = 100, Solver = adam, Activation = relu, Alpha = 0.0001, Learning_rate_init = 0.001, Learning_rate = constant, Tol = 1e-4, Epsilon = 1e-8, Momentum = 0.9, Max_fun = 15000.
DTC	Criterion = entropy, Splitter = best, Max_depth = 20, Ccp_alpha = 0.0, Random_state = 0, Min_samples_split = 2, Min_samples_leaf = 1, Max_features = None.
RFC	N_estimators = 20, Max_depth = 20, Criterion = gini, Max_features = 1.0, Bootstrap = True, Ccp_alpha = 0.0, Random_state = 5.
LR	Random_state = 10, Solver = lbfgs, Max_iter = 100, Multi_class = auto, C = 1.0.
KNN	N_neighbors = 5, Weights = uniform, Leaf_size = 30, P = 2, Metric = minkowski.
GBC	N_estimators = 20, Max_depth = 20, Learning_rate = 0.01, Loss = log_loss, Criterion = friedman_mse.
ETC	N_estimators = 20, Random_state = 0, Max_depth = 20 Criterion = gini, Max_features = sqrt.
SVM	Random_state = 50, Max_iter = 100, Penalty = l2, Loss = squared_hinge, Tol = 1e-4, C = 1.0, Multi_class = ovr.
LSTM	Loss = categorical_crossentropy, Optimizer = adam, Metrics = accuracy, Activation = softmax.
GRU	Loss = categorical_crossentropy, Optimizer = adam, Metrics = accuracy, Activation = softmax.

<https://doi.org/10.1371/journal.pone.0284522.t003>

LSTM model is a recurrent neural network known best for learning long-term sequences [36]. The primary motive behind the LSTM model is to remember the long sequences for a long period. The LSTM model contains three gates for processing: input gate, output gate, and forget gate. The LSTM model has a high number of training parameters that use high memory.

GRU model is a recurrent neural network [37]. It contains two gates: the update gate and the reset gate which are utilized for its working mechanism. The GRU model has less complexity than the LSTM model due to a smaller number of gates. The GRU model uses fewer training parameters that use less memory and execute faster. The GRU and LSTM model benefit from overcoming the vanishing gradient problem.

The hyperparameter tuning and optimization techniques [38] is based on the iterative process of training and evaluation of learning models. In the iterative tuning process, the parameters on which the learning model gives the best performance accuracy scores are considered the best-fit hyperparameters. The best-fit hyperparameters result in higher accuracy scores for predicting the microbe organisms in this study. The final selected hyperparameters for learning models are given in Table 3.

Results and discussions

Results and discussions are presented in this section. The results of all the machine learning and deep learning models are compared. The performance evaluation is based on accuracy, error rate, precision, recall, F1, cohen kappa, and the geometric mean score.

Experimental setup

The Python 3.0 programming tool [39] is utilized to conduct all experiments. The modules Keras version 2.8.0 and TensorFlow version 2.8.2 are used for building deep learning models. Machine learning models are built using the Scikit-learn module version 1.0.2. The platform with 13GB RAM and a 2.20GHz CPU is used to complete the experiments.

Results of machine learning and deep learning models

Experimental results of all the models are given in Table 4. Results indicate that the proposed approach obtains the best results with 98% accuracy and geometric mean, 97% precision and Cohen Kappa, and 96% recall and F1 scores. Regarding the training time, propose approach

Table 4. Performance analysis of employed machine and deep learning techniques with the proposed technique.

Technique	Training time (sec)	Accuracy	Error rate	Precision	Recall	F1 score	Cohen Kappa	Geometric mean
MLP	34.665	62	0.381	61	57	58	55	73
DTC	1.242	97	0.025	97	95	96	97	97
RFC	1.437	97	0.026	98	95	96	96	97
LR	4.349	44	0.558	36	33	32	33	55
KNN	0.028	88	0.115	88	85	86	86	91
GBC	88.761	96	0.032	98	94	96	96	96
ETC	0.522	97	0.025	98	95	96	97	97
SVM	3.747	41	0.592	39	38	36	31	59
LSTM	72.297	30	0.697	65	30	41	49	44
GRU	83.966	34	0.658	74	34	47	64	50
Proposed	1.368	98	0.024	97	96	96	97	98

<https://doi.org/10.1371/journal.pone.0284522.t004>

takes 1.386 seconds which is higher than only KNN, ETC, and DTC which take 0.028, 0.522, and 1.242 seconds, respectively.

The second best accuracy is obtained jointly by the DTC, RFC and ETC which obtain 97% accuracy, as shown in Fig 7. Machine learning models tend to perform better on average, except for SVM and LR which obtains 41% and 44% accuracy, respectively. Deep learning models show poor performance and obtain the lowest accuracy scores of 30% and 34% for LSTM and GRU models. Due to the smaller dataset, the models can not get a good fit and show poor results.

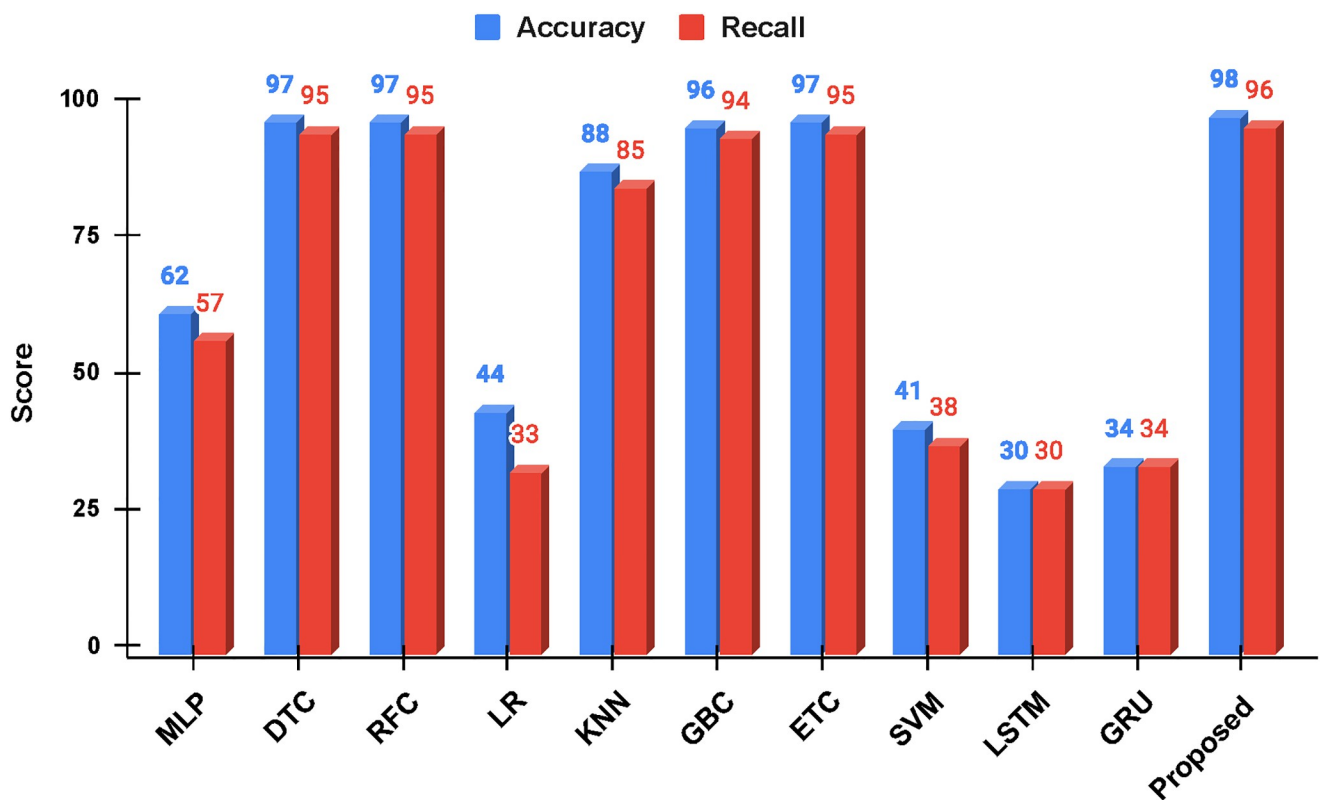


Fig 7. Comparative analysis of employed machine learning and deep learning models in terms of accuracy and recall.

<https://doi.org/10.1371/journal.pone.0284522.g007>

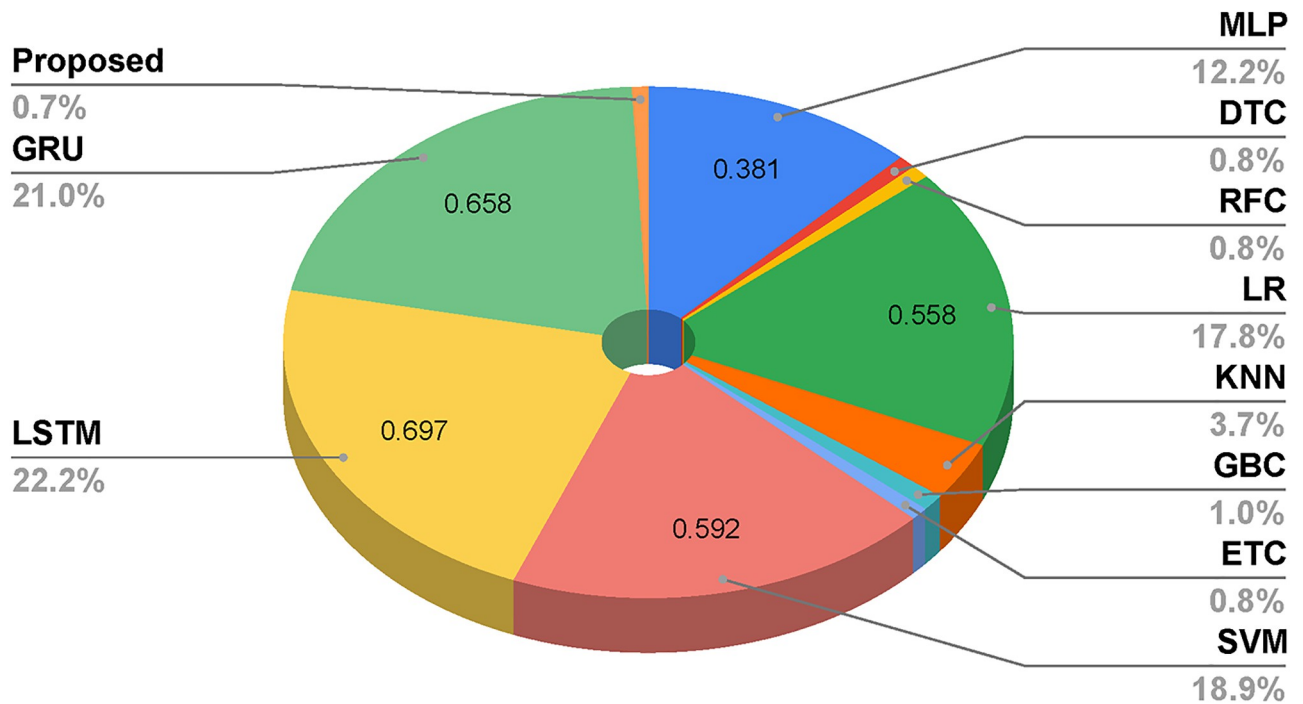


Fig 8. Comparative analysis of employed machine learning and deep learning models in terms of prediction error rate.

<https://doi.org/10.1371/journal.pone.0284522.g008>

The pie chart-based error rate comparative analysis of employed learning techniques is visualized in Fig 8. The analysis demonstrates that the proposed approach has the minimum error rate indicating high-performance accuracy scores for the microbe organism predictions. Based on this analysis, the proposed approach has a 0.7% error rate. The high error rate of 22% is achieved by the LSTM model, which indicates the low accuracy scores. The analysis shows that DTC and RFC have the same error rate of 0.8%, indicating maximum accuracy scores.

The classification report based on individual categories is given in Table 5. The analysis demonstrates that the organism’s categories Penicillum and Raizopus achieved a 100% score for all performance metrics. The categories Protozoa and Raizopus achieved 100% scores for recall and F1 score measures, respectively. The average performance metrics scores for all are

Table 5. Individual class-wise report of the proposed approach.

Category	Precision	Recall	F1-score	Support
Aspergillus sp	0.94	0.98	0.96	523
Diatom	0.93	0.97	0.95	266
Penicillum	1.00	1.00	1.00	146
Pithophora	0.91	0.87	0.89	178
Protozoa	0.99	1.00	1.00	546
Raizopus	1.00	1.00	1.00	350
Spirogyra	0.95	0.80	0.87	71
Ulothrix	0.97	0.97	0.97	1041
Volvox	0.99	1.00	0.99	636
Yeast	1.00	0.97	0.99	517
Average	0.97	0.96	0.96	4274
Accuracy	0.98			

<https://doi.org/10.1371/journal.pone.0284522.t005>

Table 6. K-fold cross-validation results of employed models.

Techniques	K-Fold	Accuracy (%)	Standard Deviation
MLP	10	61	±0.0107
DTC	10	97	±0.0038
RFC	10	97	±0.0029
LR	10	43	±0.0095
KNN	10	89	±0.0081
GBC	10	97	±0.0041
ETC	10	97	±0.0037
SVM	10	24	±0.0729
LSTM	10	81	±0.1224
GRU	10	86	±0.1119
Proposed	10	98	±0.0033

<https://doi.org/10.1371/journal.pone.0284522.t006>

between 96% to 97%. This analysis validates the proposed model results and demonstrates the high accuracy scores for the microbe's organism's predictions.

Results of k-fold cross-validation

The k-fold cross-validation results of employed learning techniques are given in Table 6. The 10-fold cross-validation results demonstrate that the proposed approach achieves a high accuracy score of 98%. The standard deviation score of the proposed approach is ±0.0033, which is the minimum compared to other techniques. The lowest accuracy score is archived by the SVM technique, which is 24% for 10-fold cross-validation. This analysis validates that the proposed model can provide generalized results for predicting microbe organisms.

Comparison with state-of-the-art approaches

The comparative performance analysis of other state-of-the-art studies is given in Table 7. The state-of-the-art studies from 2019 to 2022 are considered. These studies employ different models like RF, logit boost, KNN, and GRU. For a fair comparison, the models are implemented on the dataset used in this study. Accuracy, recall, and geometric mean scores are utilized for comparison. The analysis demonstrates that the proposed approach outperforms the state-of-the-art studies with high accuracy for predicting microbe organisms.

Discussion

The prediction of the microbe organisms using the data of different living microforms of life is presented in this study. An ensemble method based on a hybrid of DTC and ETC techniques

Table 7. Performance analysis of the proposed approach with state-of-the-art studies.

Ref.	Year	Technique	Accuracy (%)	Recall (%)	GM (%)
[40]	2019	Random Forest	97	95	97
[41]	2019	Random Forest	97	95	97
[42]	2019	logit boost	61	58	74
[43]	2021	k-nearest neighbors	88	85	95
[44]	2022	Gated recurrent units	34	34	50
[45]	2021	Random Forest	97	95	97
Proposed	2022	HMC	98	96	98

<https://doi.org/10.1371/journal.pone.0284522.t007>

is used for the prediction task. Experiments are performed using many machine learning and deep learning models for performance comparisons like DTC, RFC, LR, KNN, GBC, ETC SVM, MLP, LSTM, and GRU. These models are optimized regarding different hyperparameters to obtain the best results. For performance analysis, Cohen Kappa and geometric mean are used in addition to error rate, accuracy, recall, precision, and F1 score. Moreover, training time is also used to estimate the computational complexity of models. Results reveal that DTC, RFC, and ETC obtain the best results among machine learning models with moderate training time. On the other hand, deep learning models show poor performance and have a higher training time. The proposed approach obtains the best performance compared to both machine learning and deep learning models with 98% accuracy and geometric mean each. In addition, its error rate of 0.024 is also the lowest among all models. K-fold cross-validation proves the robustness of the proposed approach. Similarly, performance comparison with existing state-of-the-art studies shows that the results from the proposed approach are superior. The research study helps microbiologists for the identification of different types of microbe organisms with high accuracy.

Conclusions

The human body contains millions of microbe organisms that carry out both positive and negative activities. Microbe organisms can cause different infections and diseases and their prediction can be vital for the early detection of diseases. This study proposes an automatic approach for the prediction of ten types of microbe organisms like *Aspergillus* sp, Diatom, *Penicillium*, Pithophora, Protozoa, Raizopus, Spirogyra, Ulothrix, Volvox, and Yeast. The proposed hybrid approach, comprising DTC and ETC, shows better accuracy than employed machine learning and deep learning models and obtains a 98% accuracy. Similarly, the geometric mean, recall, precision, and F1 scores are the best among all the models and it obtains the lowest error of 0.024. K-fold cross-validation and performance comparison with state-of-the-art methods further validate its superior performance. Owing to the poor performance of deep learning models, we intend to incorporate a large dataset in the future. Similarly, using transfer learning and multi-class data balancing is also intended.

Supporting information

S1 Dataset.
(ZIP)

Author Contributions

Conceptualization: Ali Raza.

Data curation: Furqan Rustam.

Formal analysis: Ali Raza.

Funding acquisition: Isabel de la Torre Diez.

Investigation: Hafeez Ur Rehman Siddiqui.

Methodology: Hafeez Ur Rehman Siddiqui.

Project administration: Hafeez Ur Rehman Siddiqui.

Resources: Isabel de la Torre Diez.

Software: Furqan Rustam.

Supervision: Imran Ashraf.

Validation: Imran Ashraf.

Visualization: Isabel de la Torre Diez.

Writing – original draft: Ali Raza, Furqan Rustam.

Writing – review & editing: Imran Ashraf.

References

1. Horve PF, Lloyd S, Mhuireach GA, Dietz L, Fretz M, MacCrone G, et al. Building upon current knowledge and techniques of indoor microbiology to construct the next era of theory into microorganisms, health, and the built environment. *Journal of Exposure Science & Environmental Epidemiology*. 2020; 30(2):219–235. <https://doi.org/10.1038/s41370-019-0157-y> PMID: 31308484
2. Hou J, Pugazhendhi A, Phuong TN, Thanh NC, Brindhadevi K, Velu G, et al. Plant resistance to disease: Using biochar to inhibit harmful microbes and absorb nutrients. *Environmental Research*. 2022; p. 113883. <https://doi.org/10.1016/j.envres.2022.113883> PMID: 35835163
3. D'Abramo F, Neumeyer S. A historical and political epistemology of microbes. *Centaurus*. 2020; 62(2):321–330. <https://doi.org/10.1111/1600-0498.12300> PMID: 32834061
4. Cao J, Feng Y, Lin X, Wang J. A beneficial role of arbuscular mycorrhizal fungi in influencing the effects of silver nanoparticles on plant-microbe systems in a soil matrix. *Environmental Science and Pollution Research*. 2020; 27(11):11782–11796. <https://doi.org/10.1007/s11356-020-07781-w> PMID: 31975001
5. Gawryluk RM, Stairs CW. Diversity of electron transport chains in anaerobic protists. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*. 2021; 1862(1):148334. <https://doi.org/10.1016/j.bbabi.2020.148334> PMID: 33159845
6. Caruana JC, Walper SA. Bacterial membrane vesicles as mediators of microbe–microbe and microbe–host community interactions. *Frontiers in microbiology*. 2020; 11:432. <https://doi.org/10.3389/fmicb.2020.00432> PMID: 32265873
7. 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.
8. Joice Cordy R. Mining the human host metabolome toward an improved understanding of malaria transmission. *Frontiers in Microbiology*. 2020; 11:164. <https://doi.org/10.3389/fmicb.2020.00164> PMID: 32117175
9. Mboera LE, Kishamawe C, Kimario E, Rumisha SF. Mortality patterns of toxoplasmosis and its comorbidities in Tanzania: a 10-year retrospective hospital-based survey. *Frontiers in Public Health*. 2019; 7:25. <https://doi.org/10.3389/fpubh.2019.00025> PMID: 30838195
10. Malaria; 2022. Available from: <https://www.who.int/news-room/fact-sheets/detail/malaria>.
11. 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.
12. Gore JC. Artificial intelligence in medical imaging; 2020.
13. Zhang Y, Jiang H, Ye T, Juhas M. Deep learning for imaging and detection of microorganisms. *Trends in Microbiology*. 2021; 29(7):569–572. <https://doi.org/10.1016/j.tim.2021.01.006> PMID: 33531192
14. Maruthamuthu MK, Raffiee AH, De Oliveira DM, Ardekani AM, Verma MS. Raman spectra-based deep learning: A tool to identify microbial contamination. *MicrobiologyOpen*. 2020; 9(11):e1122. <https://doi.org/10.1002/mbo3.1122> PMID: 33063423
15. Corbin CK, Sung L, Chattopadhyay A, Noshad M, Chang A, Deresinski S, et al. Personalized antibiotics for machine learning driven antibiotic selection. *Communications medicine*. 2022; 2(1):1–14. <https://doi.org/10.1038/s43856-022-00094-8> PMID: 35603264
16. Pawłowski J, Majchrowska S, Golan T. Generation of microbial colonies dataset with deep learning style transfer. *Scientific Reports*. 2022; 12(1):1–12. <https://doi.org/10.1038/s41598-022-09264-z> PMID: 35338253
17. Wei J, Suriawinata A, Ren B, Liu X, Lisovsky M, Vaickus L, et al. A petri dish for histopathology image analysis. In: *International Conference on Artificial Intelligence in Medicine*. Springer; 2021. p. 11–24.
18. Delavy M, Cerutti L, Croxatto A, Prod'homme G, Sanglard D, Greub G, et al. Machine learning approach for *Candida albicans* fluconazole resistance detection using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Frontiers in microbiology*. 2020; 10:3000. <https://doi.org/10.3389/fmicb.2019.03000> PMID: 32010083

19. Huang TS, Lee SSJ, Lee CC, Chang FC. Detection of carbapenem-resistant *Klebsiella pneumoniae* on the basis of matrix-assisted laser desorption ionization time-of-flight mass spectrometry by using supervised machine learning approach. *PLoS One*. 2020; 15(2):e0228459. <https://doi.org/10.1371/journal.pone.0228459> PMID: 32027671
20. Wang HY, Lee TY, Tseng YJ, Liu TP, Huang KY, Chang YT, et al. A new scheme for strain typing of methicillin-resistant *Staphylococcus aureus* on the basis of matrix-assisted laser desorption ionization time-of-flight mass spectrometry by using machine learning approach. *PLoS one*. 2018; 13(3): e0194289. <https://doi.org/10.1371/journal.pone.0194289> PMID: 29534106
21. Wang HY, Li WC, Huang KY, Chung CR, Horng JT, Hsu JF, et al. Rapid classification of group B *Streptococcus* serotypes based on matrix-assisted laser desorption ionization-time of flight mass spectrometry and machine learning techniques. *BMC bioinformatics*. 2019; 20(19):1–17.
22. Srinivasu PN, SivaSai JG, Ijaz MF, Bhoi AK, Kim W, Kang JJ. Classification of skin disease using deep learning neural networks with MobileNet V2 and LSTM. *Sensors*. 2021; 21(8):2852. <https://doi.org/10.3390/s21082852> PMID: 33919583
23. El-Sappagh S, Ali F, Abuhmed T, Singh J, Alonso JM. Automatic detection of Alzheimer's disease progression: An efficient information fusion approach with heterogeneous ensemble classifiers. *Neurocomputing*. 2022; 512:203–224. <https://doi.org/10.1016/j.neucom.2022.09.009>
24. SAYAN SAHA. Microbes Dataset | Kaggle; 2022. Available from: <https://www.kaggle.com/datasets/sayansh001/microbes-dataset>.
25. DPhi. Data sprint 71—Microbes Classification | DPhi; 2022. Available from: <https://dphi.tech/challenges/data-sprint-71-microbes-classification/207/overview/about>.
26. Mahela OP, Shaik AG, Khan B, Mahla R, Alhelou HH. Recognition of complex power quality disturbances using S-transform based ruled decision tree. *IEEE Access*. 2020; 8:173530–173547. <https://doi.org/10.1109/ACCESS.2020.3025190>
27. Liu K, Hu X, Zhou H, Tong L, Widanage WD, Marco J. Feature analyses and modeling of lithium-ion battery manufacturing based on random forest classification. *IEEE/ASME Transactions on Mechatronics*. 2021; 26(6):2944–2955. <https://doi.org/10.1109/TMECH.2020.3049046>
28. Raza A, Munir K, Almutairi M, Younas F, Fareed MMS. Predicting Employee Attrition Using Machine Learning Approaches. *Applied Sciences*. 2022; 12(13):6424. <https://doi.org/10.3390/app12136424>
29. Khan MSI, Islam N, Uddin J, Islam S, Nasir MK. Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. *Journal of King Saud University-Computer and Information Sciences*. 2021;.
30. 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. <https://doi.org/10.1016/j.patrec.2020.10.005>
31. Manoharan H, Teekaraman Y, Kirpichnikova I, Kuppusamy R, Nikolovski S, Baghaee HR. Smart grid monitoring by wireless sensors using binary logistic regression. *Energies*. 2020; 13(15):3974. <https://doi.org/10.3390/en13153974>
32. Leong W, Kelani R, Ahmad Z. Prediction of air pollution index (API) using support vector machine (SVM). *Journal of Environmental Chemical Engineering*. 2020; 8(3):103208. <https://doi.org/10.1016/j.jece.2019.103208>
33. Hao PY, Kung CF, Chang CY, Ou JB. Predicting stock price trends based on financial news articles and using a novel twin support vector machine with fuzzy hyperplane. *Applied Soft Computing*. 2021; 98:106806. <https://doi.org/10.1016/j.asoc.2020.106806>
34. Zheng H, Wang G, Li X. Swin-MLP: a strawberry appearance quality identification method by Swin Transformer and multi-layer perceptron. *Journal of Food Measurement and Characterization*. 2022; p. 1–12.
35. Wright LG, Onodera T, Stein MM, Wang T, Schachter DT, Hu Z, et al. Deep physical neural networks trained with backpropagation. *Nature*. 2022; 601(7894):549–555. <https://doi.org/10.1038/s41586-021-04223-6> PMID: 35082422
36. Xayasouk T, Lee H, Lee G. Air pollution prediction using long short-term memory (LSTM) and deep autoencoder (DAE) models. *Sustainability*. 2020; 12(6):2570. <https://doi.org/10.3390/su12062570>
37. Que Z, Jin X, Xu Z. Remaining useful life prediction for bearings based on a gated recurrent unit. *IEEE Transactions on Instrumentation and Measurement*. 2021; 70:1–11. <https://doi.org/10.1109/TIM.2021.3054025>
38. Agrawal T. Hyperparameter Optimization in Machine Learning: Make Your Machine Learning and Deep Learning Models More Efficient. Springer; 2021.
39. Chandra Y, Jana A. Sentiment analysis using machine learning and deep learning. In: 2020 7th International Conference on Computing for Sustainable Global Development (INDIACom). IEEE; 2020. p. 1–4.

40. Ryan FJ. Application of machine learning techniques for creating urban microbial fingerprints. *Biology direct*. 2019; 14(1):1–13. <https://doi.org/10.1186/s13062-019-0245-x> PMID: 31420049
41. 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. <https://doi.org/10.1371/journal.pone.0215502> PMID: 31260460
42. Bang S, Yoo D, Kim SJ, Jhang S, Cho S, Kim H. Establishment and evaluation of prediction model for multiple disease classification based on gut microbial data. *Scientific reports*. 2019; 9(1):1–9. <https://doi.org/10.1038/s41598-019-46249-x> PMID: 31308384
43. Riekeles M, Schirmack J, Schulze-Makuch D. Machine learning algorithms applied to identify microbial species by their motility. *Life*. 2021; 11(1):44. <https://doi.org/10.3390/life11010044> PMID: 33445805
44. Shi H, Zhang S. Accurate Prediction of Anti-hypertensive Peptides Based on Convolutional Neural Network and Gated Recurrent unit. *Interdisciplinary Sciences: Computational Life Sciences*. 2022; p. 1–16. PMID: 35474167
45. Singh N, Bhatnagar S. Machine Learning for Prediction of Drug Targets in Microbe Associated Cardiovascular Diseases by Incorporating Host-pathogen Interaction Network Parameters. *Molecular Informatics*. 2022; 41(3):2100115. <https://doi.org/10.1002/minf.202100115> PMID: 34676983