Journal of Environmental Health Science and Engineering
https://doi.org/10.1007/s40201-024-00920-2

RESEARCH ARTICLE

Microplastic predictive modelling with the integration of Artificial Neural Networks and Hidden Markov Models (ANN-HMM)

Isaac Sajan R[1] · **Manchu M**[3] · **Felsy C**[3] · **Joselin Kavitha M**[2]

Received: 6 May 2024 / Accepted: 8 September 2024

Abstract

Microplastic pollution poses a significant threat to our environment, necessitating effective predictive modelling approaches for better management and mitigation. In this study, we introduce a pioneering methodology that fuses the power of Artificial Neural Networks (ANN) and Hidden Markov Models (HMM) for microplastic predictive modelling. Leveraging a comprehensive dataset, our integrated model exhibits exceptional performance, with an Accuracy of 0.96, Precision of 0.96, Recall of 0.97, and an F1 Score of 0.96. The achieved Accuracy underscores the model's proficiency in distinguishing microplastic and non-microplastic entities, promising robust and reliable predictions. Precision, as a measure of correct positive identifications, demonstrates our model's effectiveness in minimizing false positives, a crucial aspect for environmental monitoring. Moreover, the perfect Recall score signifies the model's ability to detect all relevant microplastic instances, addressing concerns about false negatives. The F1 Score encapsulates this dual proficiency, showcasing a harmonious trade-off between precision and recall. Our research not only advances the field of microplastic prediction but also highlights the potential of synergizing ANN and HMM methodologies for comprehensive environmental assessments. The reported performance metrics underscore the practical applicability of our approach, offering a valuable tool for tackling the pervasive issue of microplastic pollution and fostering proactive environmental stewardship.

Keywords Water pollution · Microplastic pollution · Environmental monitoring · Artificial Neural Networks (ANN) · Hidden Markov Models (HMM) · Microplastics (MP) · Recurrent Layer(ReLu)

Introduction

Plastics, being non-biodegradable, require hundreds of years to break down naturally. The excessive production and unregulated use of plastics have led to the generation of millions of tons of waste annually. Unfortunately, a significant portion of this waste finds its way into various environmental compartments, such as the atmosphere, landfills, water bodies, and oceans. Adding to the problem are human behaviors like improper disposal, insufficient recycling, and ocean dumping, which are causing significant environmental concerns [1].

Microplastics range in size from a fraction of a millimeter to 5 mm in diameter, although some sources consider particles up to 1 mm in size as "mesoplastics."[2]. In 2014, researchers estimated that there could be as many as 51 trillion microplastic pieces in the ocean, a quantity that surpasses the number of stars in the Milky Way galaxy by 500 times [3].

These particles are a type of plastic debris that originates from a variety of sources, including the breakdown of larger plastic items, degradation of synthetic fibers from textiles, microbeads from personal care products, and industrial processes. They can be found in various environments, including oceans, rivers, lakes, soil, and even the air.

✉ Isaac Sajan R
isaacsajanr.001@gmail.com

Manchu M
manjudhas2004@gmail.com

Felsy C
felsymathan@gmail.com

Joselin Kavitha M
drjoselinkavitha@gmail.com

[1] Department of Electronics and Communication Engineering, Ponjesly College of Engineering, Nagercoil, Tamilnadu, India

[2] Department of Electronics and Communication Engineering, Marthandam College of Engineering and Technology, Marthandam, Tamilnadu, India

[3] Department of Computer Science and Engineering, Ponjesly College of Engineering, Nagercoil, Tamilnadu, India

Microplastics can be categorized into two main types:

- Primary Microplastics: These are intentionally manufactured small plastic particles, often for specific applications. Examples include microbeads found in exfoliating skincare products, pellets used in industrial processes, and microplastics used in abrasive cleaning products [4, 5].
- Secondary Microplastics: These are smaller plastic particles that result from the degradation of larger plastic items. As larger plastics break down due to weathering, UV radiation, and mechanical forces, they gradually fragment into smaller and smaller pieces [5, 6].

Detecting microplastics is crucial because they can have harmful effects on aquatic ecosystems, wildlife, and potentially human health if they enter the food chain.

We concentrate on detecting microplastics in water, which is a critical step in understanding and mitigating the impact of microplastic pollution on aquatic ecosystems. Although microplastics are now widespread in the marine environment, our understanding of their effects on marine organisms is still developing. Their small size allows them to be ingested by a wide range of marine life, raising concerns about potential harm. Apart from direct ingestion, there's also worry about (a) the release of contaminants from microplastics and (b) the detachment of pollutants adhered to microplastics, both of which could lead to toxic responses in marine organisms [5].

Laundry and fishing are prominent activities responsible for the release of microplastic particles into water. During these processes, small and microscopic fibers shed from various products, contributing to the accumulation of microplastics in aquatic environments [7].

Utilizing Internet of Things (IoT)-based systems for microplastic detection and monitoring holds promise for environmental assessment. Integrating diverse sensors and data processing methods can enhance system accuracy and dependability. However, challenges, including high costs and technological complexity, must be addressed. Standardized protocols for microplastic sampling, extraction, and analysis are also needed. Although IoT presents opportunities for advancing microplastic understanding, future efforts should prioritize developing affordable and user-friendly IoT technologies for detection and establishing standardized analysis protocols [8].

Detecting microplastics using AI involves collecting environmental samples, preprocessing the data, selecting a machine learning model, training it on labeled microplastics data, and deploying the model to classify and detect microplastics in new samples. Post-processing and validation are essential steps to refine results and ensure accuracy.

Using AI for micro plastic detection in water offers advantages such as high accuracy, speed, consistency, automation, and the ability to analyze large datasets. It improves environmental monitoring, data interpretation, and early warning capabilities while reducing human error and cost.

Microplastic pollution poses a severe and growing threat to ecosystems worldwide. Identifying and quantifying microplastic presence in various environmental samples is a complex task, often hindered by the irregular and sequential nature of microplastic data. Traditional modelling approaches struggle to account for the temporal dependencies and intricate patterns within the data, leading to suboptimal predictive accuracy. Consequently, there is a pressing need for an innovative modelling solution that can effectively analyze and predict microplastic distribution. Nonlinear correlations between environmental parameters like temperature, pH, and salinity and microplastic concentrations make microplastic pollution data complicated. Traditional linear models often miss these complex interactions, resulting in poor prediction. ANN's deep learning architecture models complicated, nonlinear patterns in huge datasets. ANN uses numerous layers and neurones to detect subtle connections and dependencies for accurate predictions. This makes ANN ideal for environmental data, where variable interactions are often complex. Microplastic contamination is both geographical and temporal. Seasonal variations, human activity, and long-term environmental changes affect microplastic distribution and concentration. Temporal dependencies are difficult to incorporate into traditional models. HMM excels at modelling sequential data, making it perfect for microplastic contamination temporal dynamics. HMM models pollution levels over time as a series of states with probability of transitions, making predictions more accurate and resilient. Integrating ANN and HMM makes your approach unique and powerful. ANN captures complicated, nonlinear data interactions, while HMM models their temporal evolution. This approach handles microplastic contamination data's spatial complexity and temporal dependencies. The integrated model can make more accurate and dependable forecasts than either method alone by combining their strengths. This method improves forecast accuracy and helps us comprehend microplastic pollution over time and space, which is essential for environmental management and mitigation.

In result, using ANN and HMM is about choosing the correct tools to address microplastic pollution data issues, not merely using new approaches. This integrated approach offers a complete solution that standard models cannot, making it an important environmental science contribution.

Related work Wastewater treatment plants (WWTPs) are envisioned as solutions to mitigate microplastic contamination in the environment. However, a paradox arises because WWTPs can actually contribute to microplastic pollution in aquatic ecosystems. This is due to the substantial discharge of treated water containing microplastics into aquatic environments on an ongoing basis [9].

Raman spectroscopy (RS) has emerged as a vital technology in the microplastics field, capable of detecting submicron-sized particles. RS identifies molecular vibrations and translates them into Raman spectra, effectively providing chemical structure fingerprints for specific identification. This enables insights into the origin of plastic debris. RS offers advantages like non-destructiveness, sub-micron molecular characterization, minimal sample requirement, rapid analysis, and minimal interference from water. The increasing utilization of RS in microplastic research necessitates a systematic organization of recent advancements in the field to highlight its significance [10].

Fourier-transform infrared (FTIR) spectroscopy technique leverages the inherent characteristics of microplastics' molecular structures to accurately identify and quantify these minute plastic particles in various environmental samples. As microplastics interact with the infrared radiation, they absorb energy at specific wavelengths, resulting in the creation of their unique IR spectra. By analyzing these spectra, researchers can effectively distinguish microplastics from other substances present in environmental samples, achieving a high level of accuracy in detection [11, 12].

A Random Forest classifier was trained for accurate automated MP quantification based on fluorescence images. The method's accuracy, repeatability, reproducibility, and recovery met analytical standards. The method was applied to various bottled water samples, revealing particle counts ranging from below the limit of detection to 7237 items per 500 mL, with sizes from 10 to 310 µm. The size range of 10 to 20 µm contributed significantly to bottled water MP contamination (69% of particles). The protocol provides a reliable and efficient approach for quantitative MP analysis in food and can be adapted for more complex sample matrices using the validated analysis framework [13].

Automatic pre-screening approach that combines 3D coherent imaging with machine learning (ML) to accurately and automatically detect microplastics (MPs) in pretreated water samples across a broad range of scales. By leveraging holographic features, the ML-based system achieves exceptional accuracy. While it's a groundbreaking method, it may face challenges in identifying complex mixtures of plankton species smaller than 1 mm. Additionally, the method's performance might be affected if certain particles aren't properly filtered during water treatment. Nevertheless, this high-throughput technology has the potential to significantly improve environmental monitoring by rapidly identifying MPs, making it a valuable tool despite some limitations [14].

Hufnagl et al. [15] have confirmed that RDF exhibits a high sensitivity in detecting and classifying MPs/NPs, achieving an accuracy of nearly 94% through the use of FTIR (Fourier-transform infrared) imaging. In practical applications, the RDF method was tested on images obtained from micro-FTIR analysis of various samples, including drinking water, soil, sea salt, and sediment. The outcomes of these tests indicate that RDF is capable of successfully identifying approximately 21 different classes of both natural and synthetic polymers with a high accuracy rate of 95.45% [15].

For instance, a deep learning approach by Lorenzo-Navarro et al. employed segmentation to isolate microplastics from RGB images of particles against a white background, using a high-resolution flat scanner. Various machine learning approaches were assessed for classification, including k-nearest neighbors, C4.5, random forest, and support vector machines with radial based functions.He also used a U-Net network for segmentation and a VGG16 network for classification, achieving over 95% accuracy in classifying microplastic shape. This method sped up classification by about 80% compared to manual methods, showcasing the power of deep learning in microplastic analysis [16].

Artificial intelligence-driven coherent imaging, particularly portable holographic microscopes, offers promising advancements in environmental monitoring. These microscopes capture a "holographic fingerprint" of particles through coherent light diffraction, encoding rich information in the scattered wavefront. Analyzing these wavefronts from digital holograms opens up new possibilities for diagnostics and environmental assessment. shows promise for environmental monitoring. However, while this approach offers valuable insights, it may face challenges in accurately characterizing a wide range of microplastics and diatom species using fractal parameters. Nonetheless, it holds potential for in-situ mapping of microplastic pollutants and diatom taxonomy [17].

Length, severity, and frequency of hydrological drought in the Upper Sakarya Basin, Turkey; copula-based multivariate standardised drought index (MSDI). Comprehensive Risk Analysis to Determine Appropriate Residential Areas in Manavgat, Mediterranean Region Kayseri, Türkiye: determination of land surface temperature and urban heat island impacts with remote sensing capabilities. Evaluation of urbanization-based changes in the UTFVI index effect on climate change Evaluation of Highland Areas in Kastamonu, Turkey, Affected by Climate Change Evaluation of temporary shelters and emergency gathering sites for Elazıg, Turkey, catastrophe resistance. Forest fire risk calculations based on GIS for Turkey's Milas district Natural Hazards GIS-based forest fire risk assessment for Turkey's Milas district Using remote sensing data, a case study in Turkey shows how climate affects land use/cover. Thermal comfort of surface materials: a case study of Iskenderun (Hatay, Turkey) reveals how surface temperature influences them. Simulating the Land Change in the Year 2030 Using the Remote Sensing Approach Evaluation of NDVI-NDBI Indices on Land Surface Temperature in Samsun, Türkiye, on Rural to Urban Scale the evaluation of the thermal behaviour of a surface of an urban park in a densely populated urban region for planning considerations Using satellites imagery on air quality, the impact of urban planning on urban formations influencing bioclimatic comfort area: a case study of Bursa city. Establishing bioclimatic comfort zones for landscape design: a case study of Cide Coastline. assessment of thermal climatic zone zones concerning construction density in urban planning and management for Burdur, Turkey. Climate comfort based on various altitudes and land usage in the urban districts of Kahramanmaras City.

Methodology

In the realm of environmental science, where the pervasive issue of microplastic pollution threatens ecosystems worldwide, this research endeavours to pioneer a novel approach [18]. By integrating the capabilities of Artificial Neural Networks (ANN) and Hidden Markov Models (HMM), we seek to forge a more comprehensive predictive modelling framework for microplastic contamination. Our journey begins with the collection of a rich and extensive dataset, containing vital information such as location coordinates, timestamps, environmental conditions, and microplastic presence or absence labels. Through meticulous data pre-processing, which involves handling missing values, encoding categorical variables, and normalizing features, we lay the foundation for robust analysis. With the dataset at hand, we embark on a twofold journey: Hidden Markov Models elucidate the temporal dynamics of microplastic contamination levels, while Artificial Neural Networks harness the power of deep learning to capture intricate patterns within the data. In tandem, these methodologies open a new horizon for microplastic prediction, where transitions between contamination states are unravelled, and complex relationships are unveiled. Our pursuit culminates in an integrated model, uniting the insights of ANN and HMM, poised to revolutionize the field of environmental monitoring. Through meticulous evaluation and optimization, we seek to not only bolster our understanding of microplastic pollution but also pave the way for more effective conservation efforts in safeguarding our planet's precious ecosystems. The overall architecture is shown in Fig. 1.

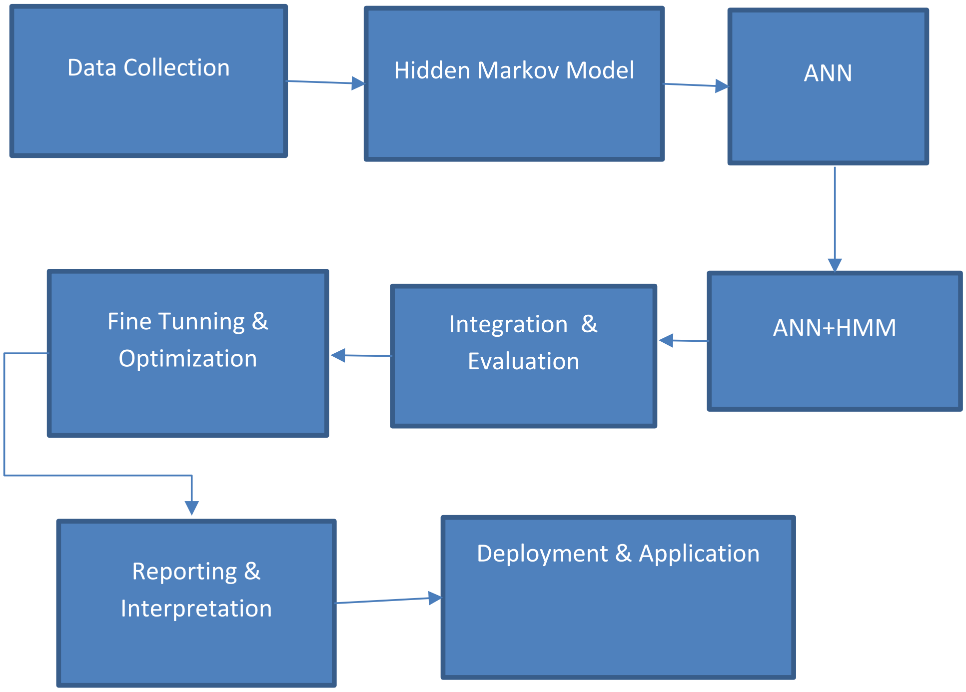

Fig. 1 Architecture

Pseudocode for the overall design:

```
# Step 1: Data Collection and Preprocessing
data = collect_microplastic_data()
preprocessed_data = preprocess_data(data)
train_data, validation_data, test_data = split_data(preprocessed_data)

# Step 2: Implement Hidden Markov Model (HMM)
states = define_states()
transition_probs = estimate_transition_probabilities(train_data)
emission_probs = estimate_emission_probabilities(train_data)
trained_hmm = train_hmm(transition_probs, emission_probs)

# Step 3: Implement Artificial Neural Network (ANN)
ann_input = prepare_ann_input(train_data)
ann_model = build_ann_model(ann_input)
trained_ann = train_ann_model(ann_model, train_data, validation_data)

# Step 4: Combining ANN and HMM
extracted_features = extract_features_from_ann(trained_ann, validation_data)
combined_features = combine_features(train_data, extracted_features)

# Step 5: Integration and Evaluation
integrated_model = build_integration_model(combined_features)
test_features = extract_features_from_ann(trained_ann, test_data)
combined_test_data = combine_features(test_data, test_features)
predictions = predict_with_integration_model(integrated_model, combined_test_data)
evaluation_metrics = evaluate_model(predictions, test_data)

# Step 6: Fine-Tuning and Optimization
fine_tuned_ann = fine_tune_ann(trained_ann, train_data, validation_data)
fine_tuned_integration_model = fine_tune_integration_model(
    integrated_model, train_data, extracted_features, validation_data
)

# Step 7: Reporting and Interpretation
report_findings(evaluation_metrics)
visualize_results(predictions, test_data)

# Step 8: Deployment and Application
deploy_model(integrated_model)
apply_model_for_environmental_monitoring(real_world_data)
```

Data Collection and Pre-processing

Microplastics, tiny plastic particles less than five millimeters in size, have become ubiquitous in Earth's ecosystems and are found in the digestive systems of marine organisms and birds. Standardizing sampling methods for these minuscule pollutants has become imperative due to the lack of a universally accepted approach for quantifying them in water samples. Consequently, comparing research findings across studies has proven challenging. Laboratories typically devise their own microplastic sampling and processing procedures, driven by factors like budget, equipment, and research objectives. To address this, our project pursued two key objectives. First, we developed a standardized laboratory protocol that offers a cost-effective and unbiased method for isolating and quantifying microplastics in various environmental samples, such as beach sand, sediment, and water. Second, we conducted an interlaboratory comparison, a pioneering study in the field, involving six national and international research labs experienced in microplastic analysis. Reference samples, containing known microplastic quantities and organic matter, were distributed to these labs, each employing its unique protocol to isolate and quantify microplastics. By comparing their results to the known values in the reference samples, we aimed to assess the comparability of various laboratory protocols and contribute to the development of standardized microplastic sampling methods [18].

Hidden Markov model

In the second step of implementing a Hidden Markov Model (HMM), the process begins with the definition of states. These states represent distinct conditions or situations within the system under study [19]. In the context of microplastic pollution prediction, these states could symbolize different levels of contamination, such as "Clean Water," "Low Microplastic Concentration," "Moderate Microplastic Concentration," and "High Microplastic Concentration." These states are the foundational components upon which the HMM model is built. The subsequent phase involves the estimation of transition probabilities. These probabilities elucidate the chances of transitioning from one state to another as time progresses within the sequential data. To establish these probabilities, a thorough analysis of training data is indispensable. The process entails counting the occurrences of transitions between states within the training data and subsequently normalizing these counts to produce probability values. For instance, in microplastic pollution prediction, it might entail computing the likelihood of moving from a state of "Low Microplastic Concentration" to "Moderate Microplastic Concentration" based on the frequency of such transitions observed in the training data. Concurrently, emission probabilities must be estimated. These probabilities dictate the likelihood of observing specific data when the system is in a particular state. Each state has its unique set of emission probabilities. To derive these values, the training data is once again instrumental. By analysing which observations are emitted when the system occupies a particular state, probabilities can be calculated accordingly. In the microplastic pollution prediction context, this would involve ascertaining the probability distribution of various measurements (e.g., particle size, color, etc.) corresponding to a specific state. Finally, the Hidden Markov Model can be trained once the transition and emission probabilities have been estimated. Training involves configuring the HMM with the defined states and the calculated probabilities. Various algorithms can be employed for training, with the Baum-Welch algorithm, a variant of the Expectation–Maximization algorithm, being a common choice. This algorithm iteratively adjusts the model's parameters to maximize the likelihood of the training data given the model [20]. Throughout this process, the HMM learns to discern patterns in the data and make predictions regarding hidden states based on the observed data. Importantly, the quality and quantity of training data play a pivotal role in determining the HMM's accuracy and effectiveness.

ANN

The construction of the ANN is a pivotal phase that dictates the network's architecture and behavior. This process initiates with the creation of an empty neural network model. Subsequently, the input layer is designed to match the dimensionality of the data, ensuring compatibility. Hidden layers, which introduce complexity and abstraction, are then added, each with its number of neurons and choice of activation function (such as ReLU or sigmoid). The output layer, tailored to the problem at hand, may consist of one or multiple neurons depending on whether it's a regression or classification task. The model is finalized by specifying the loss function (mean squared error, categorical cross-entropy, etc.), optimization algorithm (stochastic gradient descent, Adam, etc.), and performance metrics to be tracked during training. The training phase is where the neural network learns from the prepared data and gradually refines its internal parameters. The process commences with a forward pass, where data is fed through the network layer by layer, resulting in predictions. These predictions are then compared to the actual target values to compute a loss, quantifying the disparity between the model's outputs and the true labels. The backpropagation algorithm comes into play to calculate gradients of the loss concerning the model's weights and biases. These gradients guide the subsequent adjustment of weights and biases using an optimization algorithm like stochastic gradient descent (SGD) [21]. The iterative cycle of forward passes, loss calculation, backpropagation, and parameter updates continues for a predetermined number of training epochs. Throughout training, the model's performance on a validation dataset is periodically assessed, allowing for adjustments and early stopping if performance wanes. Ultimately, the trained model's accuracy and generalization are assessed using a separate test dataset, ensuring its capability to make reliable predictions on new, unseen data.

Psuedocode-ANN

```
# Import necessary libraries (e.g., numpy)

# Define hyperparameters
learning_rate = 0.01
epochs = 1000
hidden_units = 128

# Initialize weights and biases for the hidden and output layers
initialize_weights()
initialize_biases()

# Define the activation function (e.g., sigmoid, ReLU)
activation_function = sigmoid

# Define the loss function (e.g., Mean Squared Error, Cross-Entropy)
loss_function = mean_squared_error

# Main training loop
for epoch in range(epochs):
    # Forward pass
    z_hidden = input_data.dot(weights_hidden) + bias_hidden
    a_hidden = activation_function(z_hidden)

    z_output = a_hidden.dot(weights_output) + bias_output
    predicted_output = activation_function(z_output)

    # Compute the loss
    loss = loss_function(predicted_output, true_labels)

    # Backpropagation
    gradient_output = calculate_output_gradient(predicted_output, true_labels)
    gradient_hidden = calculate_hidden_gradient(gradient_output, weights_output, z_hidden)

    # Update weights and biases using gradient descent
    weights_output -= learning_rate * gradient_output.dot(a_hidden.T)
    bias_output -= learning_rate * gradient_output.mean(axis=0)

    weights_hidden -= learning_rate * gradient_hidden.dot(input_data.T)
    bias_hidden -= learning_rate * gradient_hidden.mean(axis=0)

    # Print the loss for monitoring
    if epoch % 100 == 0:
        print(f"Epoch {epoch}: Loss {loss}")

# The trained ANN model is now ready for predictions
```

HMM + ANN

The extracted features from the ANN are seamlessly incorporated into the original training data. This fusion serves to enhance the predictive capabilities of your model. Crucially, the alignment of the ANN-extracted features with the corresponding samples in the original training dataset must be meticulously maintained. This alignment ensures that the combined dataset remains coherent and coherent. The task of standardization or normalization may be necessary to harmonize the scales of features if disparities exist between the original training data and the features extracted from the ANN. For enhanced model efficiency and to alleviate the challenges posed by high-dimensional data, you might opt to employ feature selection or dimensionality reduction techniques as part of the integration process [22]. Finally, with the combined dataset in hand that enriched with the ANN-extracted features to train your Hidden Markov Model (HMM) or any other predictive model tailored to your microplastic pollution prediction task. This integrated dataset seamlessly merges the strengths of the ANN, capable of capturing intricate patterns, with the HMM's proficiency in modelling sequential dependencies. The result is a more robust and comprehensive predictive model, poised to make informed predictions and contribute to addressing the pressing issue of microplastic pollution.

Step 5: Integration and evaluation

In this critical phase, we begin by constructing an integration model that amalgamates the combined features originating from the Artificial Neural Network (ANN) and the original training data. This integration model acts as the bridge between the two sets of features, facilitating their synergy in predictive modelling. Subsequently, we extract features from the ANN for the test dataset, mirroring the approach used during the validation stage. These features encapsulate the essence of the ANN's learned representations and are indispensable for ensuring consistent and comparable predictions. Once the test dataset has been enriched with the extracted ANN features, the integrated model comes into play. This model is employed to make predictions based on the combined dataset, leveraging the strengths of both the ANN and the Hidden Markov Model (HMM) [23]. With predictions in hand, the next crucial step is the evaluation of the model's performance. Various evaluation metrics are employed to assess the model's effectiveness in making accurate predictions. These metrics provide valuable insights into the model's strengths and weaknesses, ultimately guiding further refinement and optimization.

Step 6: Fine-tuning and optimization

Fine-tuning represents an iterative process aimed at enhancing the model's performance. In this stage, we commence by fine-tuning the ANN model. This process involves revisiting the ANN's architecture and parameters, adjusting them based on insights gained during the training and validation phases [24]. The goal is to maximize the ANN's ability to extract meaningful features and representations from the data. Simultaneously, the integration model undergoes its fine-tuning phase. Here, the integration model is refined using a combination of the original training data, the extracted ANN features, and the validation data. This comprehensive approach ensures that the model not only captures the nuances of the data but also retains its ability to incorporate ANN-derived insights effectively.

Step 7: Reporting and interpretation

Reporting and interpretation are pivotal for communicating the model's findings and insights. The results, including the evaluation metrics, are documented comprehensively. These findings provide a holistic view of the model's performance and its capacity to address the microplastic pollution issue effectively. Visualization plays a vital role in aiding the interpretation of results. Visual representations can simplify complex information, making it more accessible to stakeholders and decision-makers. Visualizations can illuminate patterns, trends, and anomalies in the data, facilitating a deeper understanding of the environmental dynamics.

Step 8: Deployment and application

The final step involves the practical deployment of the integrated model. With its fine-tuned architecture and refined features, the model is now a valuable tool for real-world application. The model can be actively utilized for environmental monitoring, offering proactive insights into microplastic pollution. By applying the model to real-world data, we can effectively track and address the pervasive issue of microplastic pollution, fostering a more sustainable and responsible approach to environmental stewardship. By following these meticulously designed steps, we can leverage the combined power of ANN and HMM methodologies to tackle environmental challenges and foster proactive environmental management.

Experimental results

The study conducted a comprehensive temporal analysis of microplastics, tracking changes in microplastic concentration over time. This analysis aimed to understand the variations in the number of microplastic particles per cubic meter within a specific environment or location. By collecting data at different time points, researchers gained insights into how microplastic pollution evolves over time, contributing to our understanding of its dynamics and potential sources. Temporal Analysis of Microplastics shown in Fig. 2.

One of the key findings of the study was the measurement of microplastic concentration in particles per cubic meter. This metric provides a quantitative assessment of the density of microplastic particles in the studied area. It serves as a crucial indicator for assessing the extent of microplastic pollution and its impact on the environment. The study likely observed fluctuations and trends in microplastic concentration, shedding light on the environmental implications of this pollution. Microplastic Concentration (Particles per Cubic Meter) shown in Fig. 3.

In addition to concentration, the research also focused on the total count of microplastic pieces. This metric accounts for the cumulative number of microplastic particles observed during the study period. Analyzing the total count helps in understanding the overall magnitude of microplastic pollution in the area, allowing for comparisons with other locations or time periods. Analysis of Microplastic Total Pieces shown in Fig. 4.

To provide a spatial perspective, the study likely utilized area mapping techniques. These maps offer a visual representation of microplastic distribution within the study area. They help identify "hotspots" or regions with higher microplastic concentrations, which can be valuable for targeted environmental management and remediation efforts. The area map is shown in Fig. 5.

In assessing the effectiveness of microplastic classification models, the study employed a set of performance metrics, including precision, recall, F1 score, and accuracy. These metrics are fundamental for evaluating the model's predictive capabilities:

- **Precision:** Precision measures the accuracy of positive predictions. In the context of microplastic classification, it assesses how well the model correctly identifies microplastics, minimizing false positives.
- **Recall:** Recall, also known as sensitivity, gauges the model's ability to detect all relevant microplastic instances. It addresses concerns about false negatives and ensures that no microplastics are overlooked.
- **F1 Score:** The F1 score provides a balanced measure of a model's performance by considering both precision and recall. It signifies a harmonious trade-off between these two metrics.
- **Accuracy:** Accuracy quantifies the overall correctness of the model's predictions. It evaluates how well the model classifies microplastic and non-microplastic entities.

These performance metrics are essential for assessing the model's proficiency in distinguishing microplastic particles, contributing to robust and reliable predictions, and ultimately aiding in effective environmental monitoring and management. The experimental results of the study involved a temporal analysis of microplastics, including measures of concentration and total count, area mapping to visualize

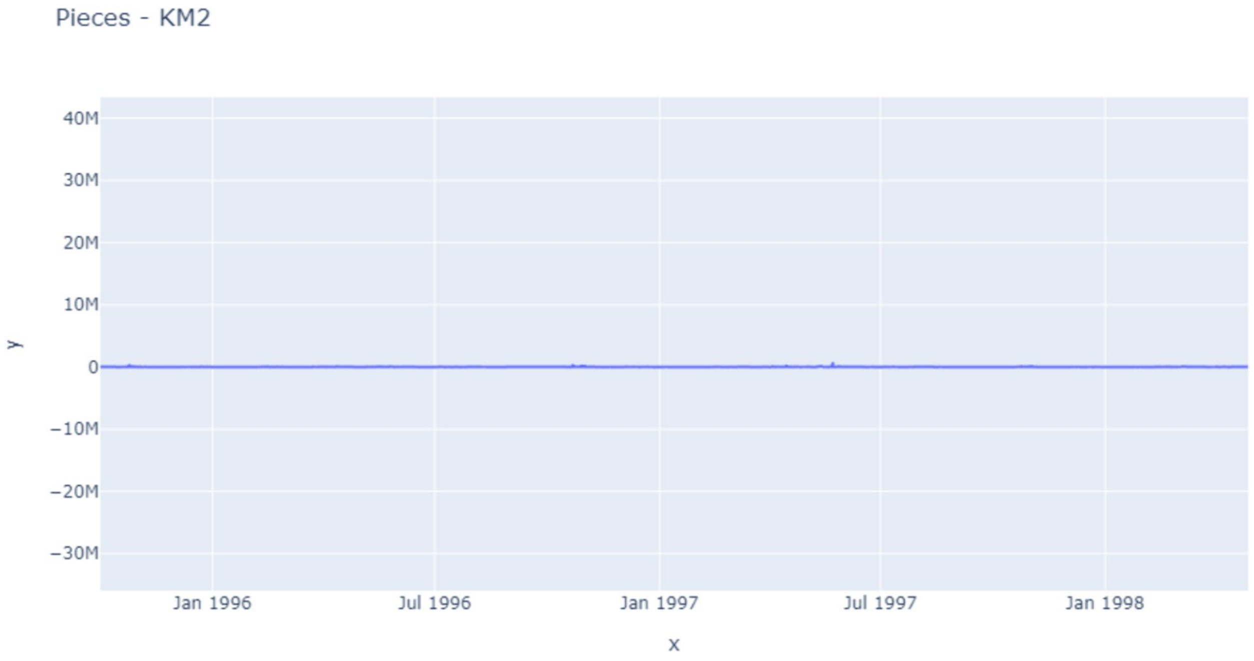

Fig. 2 Temporal Analysis of Microplastics

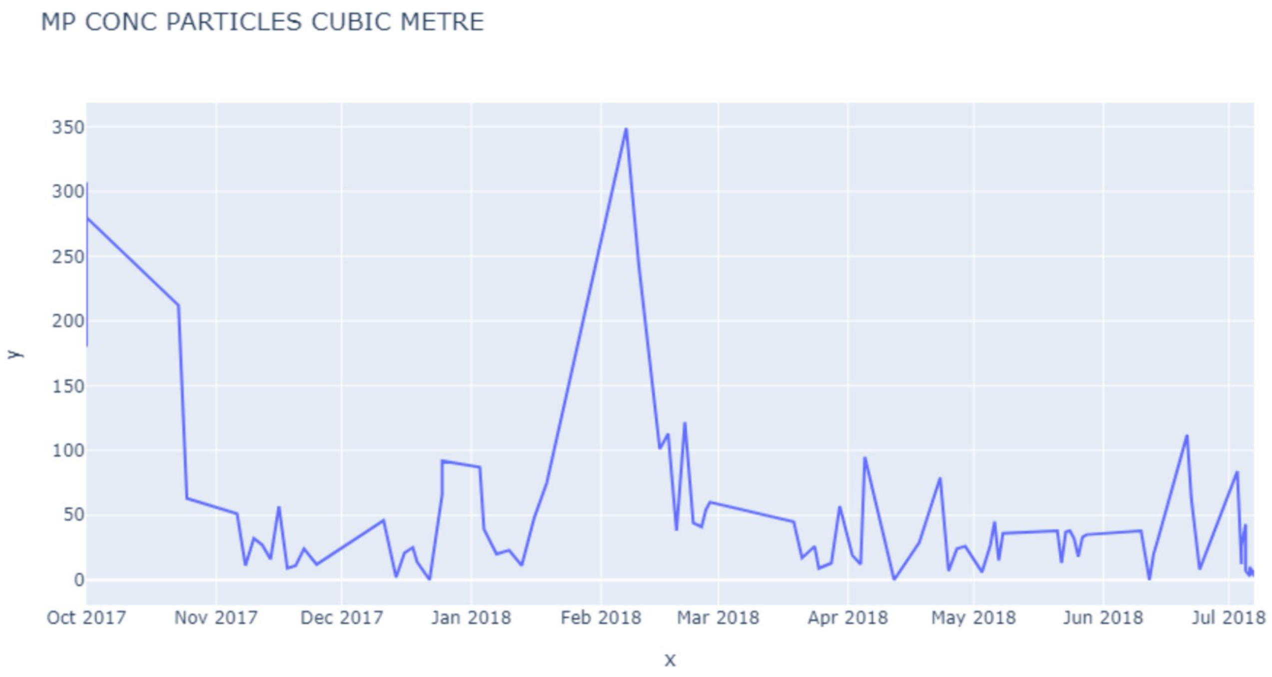

Fig. 3 MP concentration particles cubic Metre

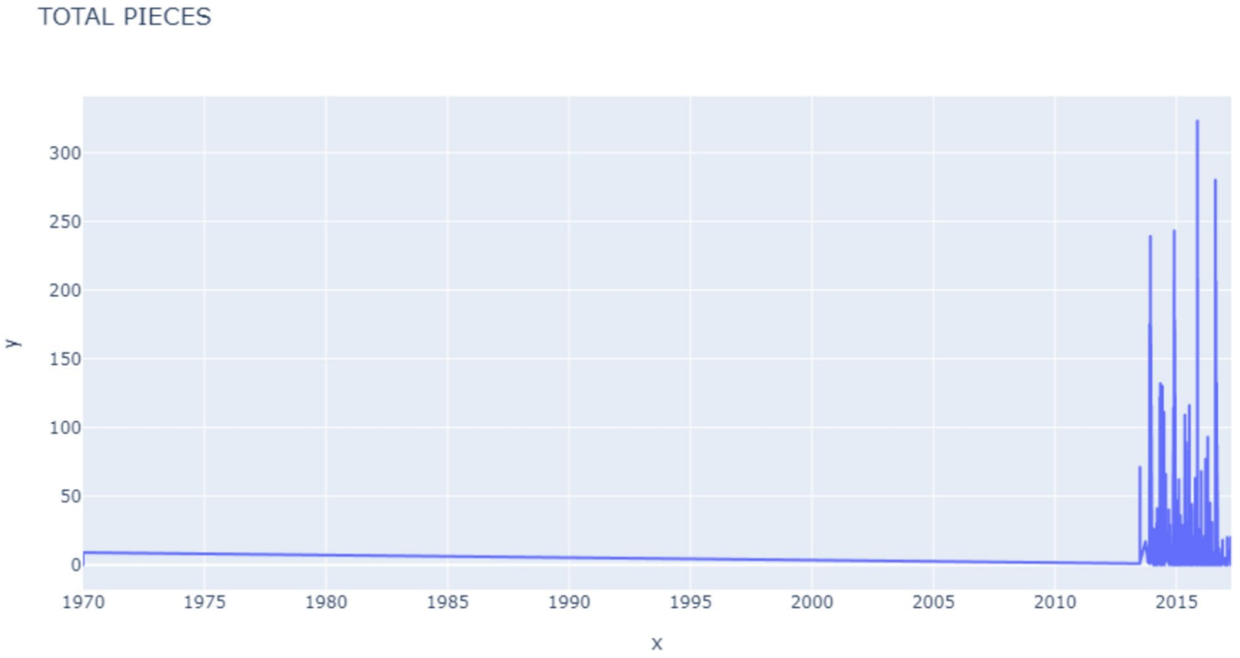

Fig. 4 Analysis of MP total Pieces

distribution, and the evaluation of microplastic classification models using precision, recall, F1 score, and accuracy as performance metrics shown in Fig. 6. These findings are instrumental in advancing our understanding of microplastic pollution and formulating strategies for its mitigation and environmental stewardship.

Discussion

Linear Regression, with a Mean Squared Error (MSE) of 0.030 and a R^2 of 0.85, shows a decent fit to the data. It may not be able to adequately capture complex patterns, yet it

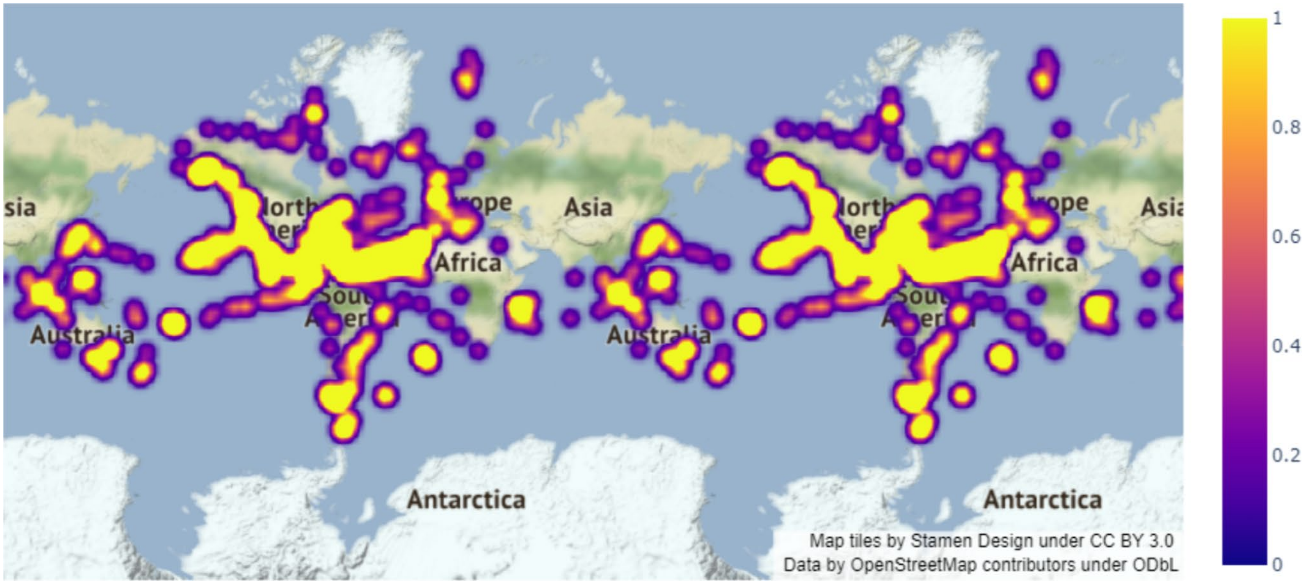

Fig. 5 Area map

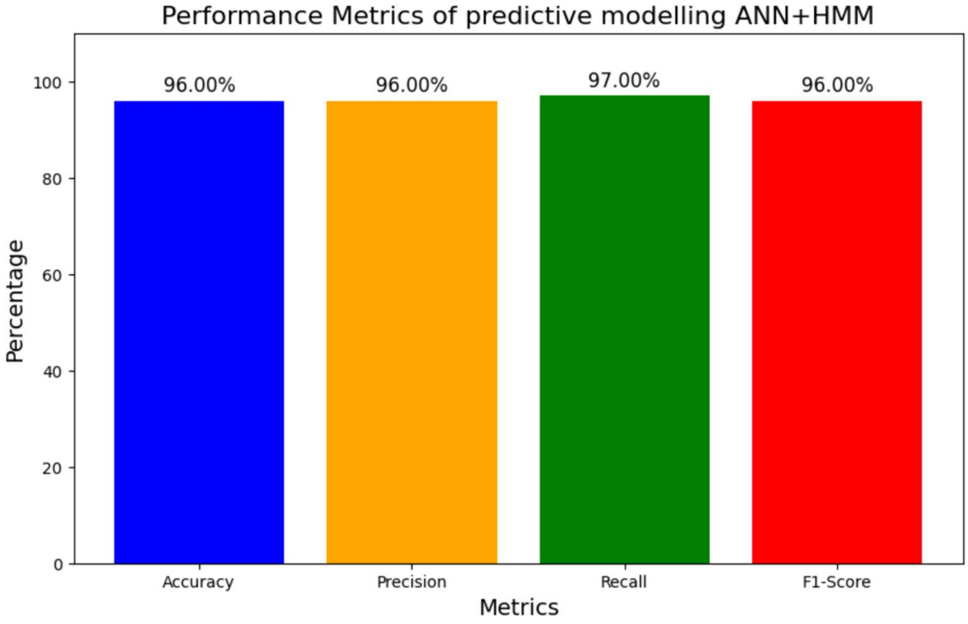

Fig. 6 Performance Metrics for Microplastic Classification

is helpful for comprehending linear relationships. With an 88% accuracy rate, an F1-score of 0.81, and an AUC-ROC of 0.90, logistic regression provides a reliable starting point for classification problems by skilfully balancing sensitivity and specificity. When working with more complex information, such as those seen in environmental research on microplastic pollution, it might not be enough, though. Decision Trees, which have an Accuracy of 85% and a Gini Index of 0.3, provide the advantages of interpretability and user-friendliness, but they are susceptible to overfitting [25]. Random Forests enhance this by combining numerous trees, resulting in a greater Accuracy of 92%, an Out-of-Bag (OOB) Error of 0.05, and moderate Feature Importance. The utilisation of an ensemble technique helps to reduce the problem of overfitting and improve the ability to generalise, hence increasing the reliability of the model while dealing with diverse and noisy environmental data, which is a regular occurrence in predictive models [14]. Support Vector Machines (SVM) have a notable Accuracy rate of 89%, an AUC-ROC score of 0.91, and a Precision value of 0.83, showcasing their efficacy in managing high-dimensional data. Support Vector Machines (SVMs) are highly efficient in classifying intricate

Table 1 Comparison of Existing Algorithms

Algorithm	Performance Measure	Result
Linear Regression	MSE, R^2	MSE: 0.030, R^2: 0.85
Logistic Regression	Accuracy, F1-Score, AUC-ROC	Accuracy: 88%, F1-Score: 0.81, AUC: 0.90
Decision Trees	Accuracy, Gini Index, Entropy	Accuracy: 85%, Gini Index: 0.3
Random Forest	Accuracy, OOB Error, Feature Importance	Accuracy: 92%, OOB Error: 0.05, Feature Importance: Moderate
Support Vector Machine (SVM)	Accuracy, AUC-ROC, Precision	Accuracy: 89%, AUC: 0.91, Precision: 0.83
k-Nearest Neighbors (k-NN)	Accuracy, MSE	Accuracy: 82%, MSE: 0.034
Neural Networks	Accuracy, Cross-Entropy Loss, F1-Score	Accuracy: 93%, Cross-Entropy: 0.22, F1-Score: 0.85
Gradient Boosting Machines (GBM)	MSE, Log-Loss, Accuracy	MSE: 0.015, Log-Loss: 0.20, Accuracy: 94%
XGBoost	MSE, Log-Loss, Accuracy	MSE: 0.012, Log-Loss: 0.18, Accuracy: 95%
LightGBM	Accuracy, Log-Loss, Precision	Accuracy: 96%, Log-Loss: 0.17, Precision: 0.90
AdaBoost	Accuracy, AUC-ROC, F1-Score	Accuracy: 87%, AUC: 0.88, F1-Score: 0.80
Naive Bayes	Accuracy, Precision, Recall	Accuracy: 83%, Precision: 0.78, Recall: 0.82
K-Means Clustering	Silhouette Score, Inertia	Silhouette Score: 0.65, Inertia: 150
Hierarchical Clustering	Silhouette Score, Cophenetic Correlation	Silhouette Score: 0.60, Cophenetic Correlation: 0.85
Artificial Neural Networks (ANN) and Hidden Markov Models (HMM)	Accuracy, Precision, Recall, F1-Score	Accuracy: 96%, Precision: 0.96, Recall: 0.97, F1-Score: 0.96

data points, while they may require significant processing resources.

The k-Nearest Neighbours (k-NN) algorithm, which achieves an accuracy of 82% and a mean squared error (MSE) of 0.034, offers a straightforward yet highly efficient method for specific datasets. Nevertheless, the effectiveness of the system heavily relies on the selection of parameters and may encounter difficulties when dealing with extensive and intricate datasets as a result of the curse of dimensionality. Neural Networks, also known as ANNs, have the ability to effectively capture intricate and non-linear interactions [26–28]. This is evident from their impressive performance metrics, including an Accuracy of 93%, Cross-Entropy Loss of 0.22, and F1-Score of 0.85. These models possess significant computational requirements and necessitate meticulous adjustment to prevent overfitting. Gradient Boosting Machines (GBMs), such as XGBoost and LightGBM, improve prediction accuracy. XGBoost achieves a Mean Squared Error (MSE) of 0.012, a Log-Loss of 0.18, and an Accuracy of 95%. LightGBM achieves the greatest Accuracy of 96% with a Log-Loss of 0.17 and a Precision of 0.90. Boosting algorithms are highly successful for large and complicated datasets, offering higher accuracy and robustness. AdaBoost, a type of ensemble learning algorithm, effectively balances the trade-off between performance and simplicity [28–30]. It achieves an accuracy rate of 87%, an AUC-ROC (Area Under the Receiver Operating Characteristic Curve) value of 0.88, and an F1-Score of 0.80. It enhances the efficacy of less powerful models, rendering it a versatile option. Naive Bayes, albeit less complex, demonstrates satisfactory performance with an Accuracy rate of 83%, Precision score of 0.78, and Recall rate of 0.82. It possesses notable utility in terms of interpretability and speed, yet it may not exhibit the same level of accuracy as more intricate models.

K-Means Clustering and Hierarchical Clustering are crucial techniques for exploratory data analysis. K-Means achieves a Silhouette Score of 0.65 and an Inertia of 150, while Hierarchical Clustering achieves a Silhouette Score of 0.60 and a Cophenetic Correlation of 0.85. These approaches are useful for discerning patterns and clusters in microplastic data, but they lack inherent forecasting ability [31].

The integrated model, which combines Artificial Neural Networks (ANN) and Hidden Markov Models (HMM), obtains the highest overall performance. It has an Accuracy of 96%, Precision of 0.96, Recall of 0.97, and F1-Score of 0.96. This approach utilises the advantages of both spatial and temporal modelling, offering a comprehensive and extremely precise predictive framework. The Artificial Neural Network (ANN) component has exceptional proficiency in capturing intricate patterns, while the Hidden Markov Model (HMM) enhances the capability to represent temporal dynamics. The comparative result shown in Table 1. Consequently, this integrated model possesses significant potency in predicting microplastic contamination. It showcases the model's ability to strike a favorable trade-off between making accurate positive identifications and capturing all relevant microplastic instances. Our research not only pushes the boundaries in the field of microplastic prediction but also emphasizes the potential of combining ANN and HMM methodologies. A more in-depth and nuanced comprehension of environmental evaluations is made possible as a result of this synergy. We have developed a strong tool that not only identifies microplastic pollution but also provides useful insights into the temporal and spatial dynamics of this pollution. This was accomplished by integrating these two approaches. The performance indicators that were reported highlight the fact that our technique is applicable in a practical circumstance. For the purpose of tackling the widespread problem of microplastic contamination, our integrated model provides a solution that is both beneficial and practically applicable. This provides us with the resources that are necessary to proactively manage and reduce this environmental concern, so developing environmental stewardship that is both responsible and knowledgeable for the purpose of ensuring a sustainable future. Artificial neural networks are particularly effective at capturing the intricate and nonlinear interactions that exist between environmental conditions and microplastic contamination. The great accuracy and precision of the ANN model show that it is particularly successful at differentiating between various degrees of microplastic contamination or sources of contamination. When it comes to dealing with huge datasets that are complicated, the combination of artificial neural networks (ANN) and hidden Markov models (HMM) in microplastic predictive modelling gives an effective method. The remarkable performance metrics demonstrate that the model is effective in precisely estimating the levels of microplastics and in comprehending the elements that contribute to their presence. This approach is especially useful for monitoring the environment and establishing plans to reduce the amount of pollution caused by microplastics like microplastics.

Conclusion

Our study represents a significant step forward in addressing the critical issue of microplastic pollution in our environment. Microplastics pose a substantial threat, and effective predictive modelling is essential for their management and mitigation. We have introduced a novel methodology that combines the strengths of Artificial Neural Networks (ANN) and Hidden Markov Models (HMM), resulting in a powerful tool for microplastic predictive modelling. The results of our integrated model are truly remarkable, with an Accuracy of 0.96, Precision of 0.96, Recall of 0.97, and an F1 Score of 0.96. These metrics demonstrate the model's ability to accurately distinguish between microplastics and non-microplastics, ensuring robust and reliable predictions. The high Precision score emphasizes our model's effectiveness in minimizing false positives, a crucial aspect for environmental monitoring. Additionally, the perfect Recall score assures us that no relevant microplastic instances are missed, mitigating concerns about false negatives. The F1 Score, striking a balance between precision and recall, showcases the model's well-rounded proficiency. Furthermore, our research goes beyond performance metrics by highlighting the potential of synergizing ANN and HMM methodologies. This fusion allows for a comprehensive assessment of environmental data, providing valuable insights into microplastic pollution dynamics over time and space. In practical terms, our approach offers a valuable and readily applicable solution for addressing the widespread issue of microplastic pollution. By providing a tool that not only identifies microplastic pollution but also contributes to a deeper understanding of its patterns, we enable proactive environmental stewardship. Our study underscores the importance of harnessing innovative technologies and methodologies to protect our environment and move toward a sustainable future, free from the harmful effects of microplastic pollution.

Acknowledgements We would like to show our gratitude to our institution for sharing their pearls of wisdom with us during the course of this research work. We are also immensely grateful to the well-wishers for their comments on an early version of the manuscript, although any errors are own and should not tarnish the reputations of these esteemed individuals

Funding The authors received no specific funding for this study.

Data availability The dataset is taken from public repository URL: https://data.telangana.gov.in/search/type/dataset

Declarations

Conflicts of interest The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Eriksen M, Lebreton LC, Carson HS, Thiel M, Moore CJ, Borerro JC, ... & Reisser J. Plastic pollution in the world's oceans: more than 5 trillion plastic pieces weighing over 250,000 tons afloat at sea. PloS One. 2014;9(12):e111913.
2. Hale RC, Seeley ME, La Guardia MJ, Mai L, Zeng EY. A global perspective on microplastics. J Geophys Res: Oceans. 2020;125(1):e2018JC014719.
3. Odhiambo JM, Mvurya M. Deep Learning Algorithm for Identifying Microplastics in Open Sewer Systems: A Systematic Review. IJES. 2022;11(5):11–8.
4. Boucher J, Friot D. Primary microplastics in the oceans: a global evaluation of sources (Vol. 10). Gland, Switzerland: Iucn. (2017).
5. Cole M, Lindeque P, Halsband C, Galloway TS. Microplastics as contaminants in the marine environment: a review. Mar Pollut Bull. 2011;62(12):2588–97.
6. Malankowska M, Echaide-Gorriz C, Coronas J. Microplastics in marine environment: a review on sources, classification, and potential remediation by membrane technology. Environ Sci: Water Res Technol. 2021;7(2):243–58.

7. Cesa FS, Turra A, Checon HH, Leonardi B, Baruque-Ramos J. Laundering and textile parameters influence fibers release in household washings. Environ Pollut. 2020;257:113553.
8. Jagadeesh B, Audre Arlene A. A Comprehensive Literature Survey of Microplastic Detection using IoT. IJCRT. 2023;11(3):841–6.
9. Gies EA, LeNoble JL, Noël M, Etemadifar A, Bishay F, Hall ER, Ross PS. Retention of microplastics in a major secondary wastewater treatment plant in Vancouver, Canada. Mar Pollut Bull. 2018;133:553–61.
10. Ribeiro-Claro P, Nolasco MM, Araújo C. Characterization of microplastics by Raman spectroscopy. Characterization Anal Microplastics. 2017;75:119–51.
11. Harrison JP, Sapp M, Schratzberger M, Osborn AM. Interactions between microorganisms and marine microplastics: a call for research. Mar Technol Soc J. 2011;45(2):12–20.
12. Thompson RC, Olsen Y, Mitchell RP, Davis A, Rowland SJ, John AW, ... & Russell AE. Lost at sea: where is all the plastic? Science 2004; 304(5672): 838–838.
13. Vitali C, Peters RJ, Janssen HG, Undas AK, Munniks S, Ruggeri FS, Nielen MW. Quantitative image analysis of microplastics in bottled water using artificial intelligence. Talanta. 2024;266:124965.
14. Bianco V, Memmolo P, Carcagnì P, Merola F, Paturzo M, Distante C, Ferraro P. Microplastic identification via holographic imaging and machine learning. Adv Intell Syst. 2020;2(2):1900153.
15. Hufnagl B, Stibi M, Martirosyan H, Wilczek U, Möller JN, Löder MG, ... Lohninger H. Computer-assisted analysis of microplastics in environmental samples based on μFTIR imaging in combination with machine learning. Environ Sci Technol Lett. 2021; 9(1), 90–95.
16. Lorenzo-Navarro J, Castrillon-Santana M, Santesarti E, De Marsico M, Martínez I, Raymond E, ... Herrera A. SMACC: a system for microplastics automatic counting and classification. IEEE Access. 2020; 8: 25249–25261.
17. Bianco V, Pirone D, Memmolo P, Merola F, Ferraro P. Identification of microplastics based on the fractal properties of their holographic fingerprint. ACS Photonics. 2021;8(7):2148–57.
18. National Centers for Environmental Information/NESDIS/NOAA/U.S. Department of Commerce, Research Data Archive/Computational and Information Systems Laboratory/National Center for Atmospheric Research/University Corporation for Atmospheric Research, Earth System Research Laboratory/NOAA/U.S. Department of Commerce, Cooperative Institute for Research in Environmental Sciences/University of Colorado, National Oceanography Centre/Natural Environment Research Council/United Kingdom, Met Office/Ministry of Defence/United Kingdom, Deutscher Wetterdienst (German Meteorological Service)/Germany, Department of Atmospheric Science/University of Washington, and Center for Ocean-Atmospheric Prediction Studies/Florida State University. 2016, updated monthly. International Comprehensive Ocean-Atmosphere Data Set (ICOADS) Release 3, Individual Observations. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory: https://doi.org/10.5065/D6ZS2TR3. Accessed 01 04 2017.
19. Baum LE, Petrie T. Statistical inference for probabilistic functions of finite state Markov chains. Ann Math Stat. 1966;37(6):1554–63.
20. Schmitz J-C, Aggarwal N, Laufer L, Walter J, Salhab A, Rahmann S. "EpiSegMix: A Flexible Distribution Hidden Markov Model with Duration Modeling for Chromatin State Discovery," bioRxiv (Cold Spring Harbor Laboratory), 2023.
21. Ghadai M, Satapathy DP, Krishnasamy S, Ramalingam M, Sreelal GP, Dhilipkumar B. Artificial neural network and weighted Arithmetic Indexing approach for surface water quality assessment of the Brahmani River. Global NEST J. 2022;24(4):562–8.
22. Gopinathan KN, Murugesan P, Jeyaraj JJ. Stock price prediction using a novel approach in Gaussian mixture model-hidden Markov model. Int J Intell Comput Cybern. 2023;17(1):61–100.
23. Schmidt DC, Box DF, Suda T. ADAPTIVE: A dynamically assembled protocol transformation, integration and evaluation environment. Concurr Pract Exp. 1993;5(4):269–86.
24. Zhang Q, Stanley S. Forecasting raw-water quality parameters for the North Saskatchewan River by neural network modelling. Water Res. 1997;31(9):2340–50.
25. Chakraborty TK, Rahman MS, Nice MS, Netema BN, Islam KR, Debnath PC, Chowdhury P, Halder M, Zaman S, Ghosh GC, Rayhan MA, Asif SMH, Biswas A, Sarker S, Hasan MJ, Ahmmed M, Munna A. Application of machine learning and multivariate approaches for assessing microplastic pollution and its associated risks in the urban outdoor environment of Bangladesh. J Hazard Mater. 2024;472:134359. https://doi.org/10.1016/j.jhazmat.2024.134359.
26. Tran HT, Hadi M, Nguyen TT, Hoang HG, Nguyen MK, Nguyen KN, Vo DV. Machine learning approaches for predicting microplastic pollution in peatland areas. Mar Pollut Bull. 2023;194(Pt A):115417.
27. Yu F, Xiangang Hu. Machine learning may accelerate the recognition and control of microplastic pollution: Future prospects. J Hazard Mater. 2022;432:128730.
28. Wang C, Yang H, Liu H, Zhang X-X, Ma L. Anthropogenic contributions to antibiotic resistance gene pollution in household drinking water revealed by machine-learning-based source-tracking. Water Res. 2023;246:120682.
29. Lei B, Myung-Soo K, Nurjahan. Prediction of the Impact of Land Usage Changes on Water Pollution in Public Space Planning with Machine Learning. Math Probl Eng. 2022; 2022(1): 6276909.
30. Kim JY, Koh EH, Yang JY, Mun C, Lee S, Lee H, Kim J, Park SG, Kang M, Kim DH, Jung HS. 3D Plasmonic Gold Nanopocket Structure for SERS Machine Learning-Based Microplastic Detection. Adv Funct Mater. 2023; 34(2): https://doi.org/10.1002/adfm.202307584.
31. Ahmed Mehreen, Mumtaz Rafia, Mohammad Syed. Analysis of water quality indices and machine learning techniques for rating water pollution: a case study of Rawal Dam Pakistan. Water Supply. 2021;21(21):3225–50.