Optimizing polymer classification through FTIR spectra and feature-scaled machine learning models

Balachandra P Shetty [1], *Sudheer Reddy* J [1], *Viraj G* R [2], and *Praveena B* A [1*]

[1] Department of Mechanical Engineering, Nitte Meenakshi Institute of Technology (NMIT), Nitte, Yelahanka, Bengaluru 560064, Karnataka, India
[2] Sri Sai Hospital, Bengaluru, Karnataka, India

Abstract. The research is based on the characterization of polymer materials by Fourier Transform Infrared (FTIR) spectroscopy which is a method of characterizing polymers in terms of their characteristic molecular vibration patterns. Three widely used thermoplastics ABS (Acrylonitrile Butadiene Styrene), PP (Polypropylene) and Nylon-66 were selected and analyzed on the basis of FTIR transmittance spectra. Wavenumbers and the values of percent transmittance were obtained and converted into feature vectors. Gaussian noise was used to boost and clean the data. Machine learning methods that were used to classify the materials included Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), random forest, and logistic regression, where 80% of the data were used to train the machine learning model and 20% to test it. Accuracy measures and confusion matrices were used to determine the performance of the models. Random Forest was the best in terms of performance in Nylon-66 classification; however, the overall accuracy was influenced by the error in classifying ABS and PP. Consequently, k-NN (70) and the Logistic Regression (69) have shown a high performance compared to the Random Forest (57) in terms of overall accuracy. This study reveals that even though the Random Forest could be used to classify a particular set of polymers, polymers that have very similar spectral characteristics are still difficult to differentiate, indicating that the work should be improved by increasing the size of the data set and the number of extracted features.

1 Introduction

Thermoplastics materials such as Acrylonitrile Butadiene Styrene (ABS), Polypropylene (PP), and Nylon66 exhibit unique spectral patterns in the infrared region, making them suitable materials for FTIR-based identification. Interpretation of this spectral pattern can be subjective and very time-consuming, especially in large-scale recycling, characteristic control, or critical material testing.

In material science for identifying polymers through their vibrational features Fourier Transform Infrared (FTIR) spectroscopy is a widely used as an analytical tool. It is quick,

* Corresponding author: praveen.ba@nmit.ac.in

non-destructive analysis technique based on the properties how molecular bonds absorb infrared wavelengths. Present studies have shown sufficient accomplishment in applying Machine Learning models to spectroscopic data for food authentication, biomedical applications, and pharmaceuticals. However inadequate work has explored its use in polymer documentation using FTIR. The incorporation of machine learning (ML) with FTIR data analysis offers a hopeful solution to systematise and augment material classification. ML algorithms can acquire predictions from spectral patterns and make accurate estimates. In this study four supervised ML models such as Random Forest, Support Vector Machine, k. Nearest Neighbours, and Logistic Regression used for grouping ABS, PP, and Nylon66 based on their FTIR spectra. Traditional FTIR analysis often relies on expert interpretation of spectral crests. From the available literature it is illustrated that FTIR spectroscopy has long been identified as reliable analytical tool for the assessment of polymers. The absorbance crests in the infrared region, equivalent to the functional groups, gives the precise identification of polymeric materials [1]. Due to increased demand computerization and adaptability, the combination of machine learning with FTIR spectroscopy has appeared as a promising research tool.

Tumuluru et al. [2] applied FTIR to discriminate various malleable types in solid waste, highlighting the clear spectral differences in major polymers like polyethylene, polypropylene, and polystyrene. Application of ML to spectral data is well documented in pharmaceutical and biomedical fields [3]. Huang et al. [4] applied support vector machines and principal component analysis to group plastics using infrared spectroscopy. This method established the feasibility of combining spectral techniques with ML for polymer sorting. Chiu et al. [5] utilized convolutional neural networks on FTIR spectra to group household malleable waste. The convolution neural networks learning models requires extensive data and computational resources. For real-world implementations in small-scale recycling facilities, simpler models like k-Nearest Neighbors (k-NN), Logistic Regression and Random Forest offer high good interpretation with greater accuracy and overhead resources. Recent studies have also focused on comparing model types. Barone et al. [6] assessed the performance of decision trees, SVM, and neural networks for identifying polyethylene terephthalate (PET) and high-density polyethylene (HDPE) using FTIR. They concluded that ensemble models, especially Random Forest, offer high accuracy and robustness against noise and variability in sample preparation an advantage echoed in the current study. The key limitation in many past efforts lies in dataset availability. Small datasets or simulated spectra are used in most of the cases, and this restricts generalization. It is becoming usual practice to apply data augmentation, like the addition of Gaussian noise or the creation of synthetic samples, to enhance model performance and create model variability that is more realistic [7].

As demonstrated by Hufnagl et al. [8], Random Decision Forest classifiers are much faster and more accurate compared to instance-based ones at predicting over 20 polymers, which also include ABS and PP. Hidalgo-Ruz et al. [9] conducted a comparative study of ATR-FTIR with classifiers such as Random Forest and SVM and observed that the resulting accuracy of over 99% was obtained through the use of both classes of consumer and marine plastics. Yu et al. [10] found that 1D CNNs trained on raw FTIR spectra reached 97% accuracy on blended microplastics datasets, outperforming Random Forest and 2D CNNs. Almanifi and Ng [11] demonstrated that semi-supervised label spreading and stacked KNN/SVM/RF pipelines achieved near-perfect classification (99–100%) on FTIR plastic bag spectra. A study in Recycling compared several models (k-NN, SVM, Naïve Bayes, RF, MLP, CNN) and found that Z-score normalization consistently improved classification accuracy on spectra for PET, HDPE, PVC, LDPE, PP, and PS [12].

Primpke et al. [13] benchmarked instance-based methods, while Hufnagl et al. [8] demonstrated RDF model scalability to millions of spectra in minutes. Ferreiro et al. [14] showed that genetic-algorithm-driven feature selection combined with supervised classifiers enabled rapid identification of polymer debris using IR reflectance data. Valls-Conesa et al. [15] applied FTIR with convolutional and ensemble models on subsamples from FTIR hyperspectral images, achieving robust polymer classification. Integrated FTIR in flame retardant analysis using ML to identify additives in polymer blends, demonstrating FTIR's versatility beyond base polymer classification, compared FTIR with NIR, LIBS, and XRF, showing that FTIR combined with ML achieved 99% accuracy in identifying consumer and environmental plastics. Although not spectroscopy-based, Valls-Conesa et al [15] proposed PolyBERT, a chemical language model to polymer fingerprinting that provides a chance to integrate chemical informatics with spectroscopic data.

Kharyuk et al. applied a wavelet transforms and logistic regression to data captured on FTIR of plants that are used in medicine, which can be generalized to polymers. Weisser et al. critically assessed protocols of vibrational spectroscopy procedures such as FTIR/ML of microplastic detection with particular focus on reproducibility and standardization. There has been a rise in the use of open-source tools, such as uFTIR, an R package to process FTIR hyperspectral images, developed by Corradini et al. This compilation shows how this method of classifying polymers through FTIR has developed over time, and how advanced ML pipelines can now be used instead of the traditional spectral interpretation methods. Though deep learning is promising, ensemble-based techniques such as Random Forest are viable and strong. It has been integrated with open-source tools and standard practices to make it more widely adopted both in academia and industry. To conclude, the FTIR spectroscopy together with machine learning could be of great potential in the classification of polymers particularly ABS, PP, and Nylon 6,6.

Although deep learning is becoming popular, conventional models continue to provide a viable trade-off of both performance and interpretability. As more open datasets become available, future work can focus on improving generalization and applying models to polymer blends and contaminated samples. The proposed research constitutes an original implementation of machine learning methods to data of FTIR spectroscopy with the purpose to enhance the ability to classify and identify the polymer materials. Considering the increased need to have effective recycling technology particularly in plastic materials management, the research has great implications to industries looking to automate the process of sorting of materials. With the help of machine learning to optimize FTIR-based classification, the present work will facilitate the sorting of polymers, which is the key to successful recycling and waste minimization.

2 Methodology

FTIR transmittance spectra for ABS, PP, and Nylon66 were obtained for their specimens and each consisting of wave numbers (in cm^{-1}) and corresponding percent transmittance values. The dataset was structured so that each polymer's transmittance profile formed a single high-dimensional feature vector. To increase data robustness and simulate a realistic variability in measurements, Gaussian noise was applied to each profile, generating 50 augmented samples per polymer type, resulting in a total of 150 samples. FTIR spectroscopy data of polymers such as ABS (acrylonitrile butadiene styrene), PP (polypropylene, and Nylon 6,6 were examined using various supervised machine learning techniques to group the materials based on their characteristic transmittance spectra. Each FTIR spectrum was plotted with wavenumber on the x-axis(cm-1) and in the y axis the transmittance in percentage as shown if Figure 1.

Polymers have been differentiated by the absorption bands, in which decreases in transmittance indicate molecular vibrations that are specific to the material. ABS exhibited notable peaks around 2237 cm^{-1} (C≡N stretching), 1600 cm^{-1} (aromatic C=C stretching), and 700–800 cm^{-1} (aromatic out-of-plane vibrations). PP displayed absorption near 2950 and 2870 cm^{-1} (C–H stretching), 1455 and 1375 cm^{-1} (CH_3 bending), and 840 cm^{-1} associated with isotactic configurations. Nylon 6,6 showed prominent absorption at 3300 cm^{-1} (N–H stretch), 1640 cm^{-1} (amide I), 1540 cm^{-1} (amide II), and 690 cm^{-1} (C–H out-of-plane). The spectra were both considered as vectors of numbers, and the features of input (X) corresponded to the transmittance value of various wavenumbers, and the target labels (Y) were the respective polymer types. The features were standardized using Standard Scaler to train the models like support vector machine and k-NN because they are sensitive to the scaling of features. To maintain the class distribution, stratified sampling was applied to divide the dataset into a training (80%) and a test (20) subsets. Four machine learning models were created and tested, including Random Forest (RF), Support Vector Machine (SVM), k-NN, and Logistic Regression.

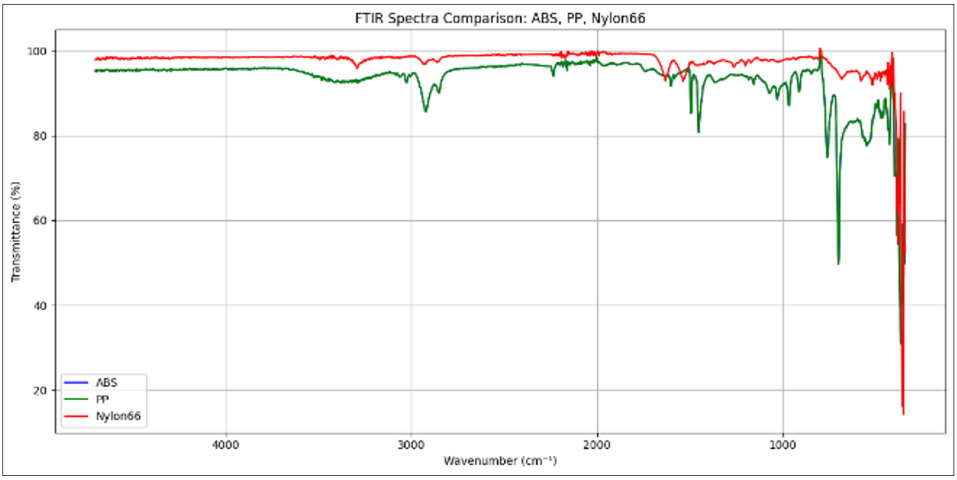

Fig. 1. FTIR spectra comparison.

The ensemble-based algorithm used was the Random Forest classifier, which has multiple decision trees, and this approach has good robustness and interpretability. The RF model had a total accuracy of 66.7%. The classification report showed an ideal classification of Nylon 6,6 and ABS and PP had a lower precision respectively, which indicated a possible spectral overlap. The training of the models was also done with the help of Gaussian noise augmentation to ensure that spectral variations are taken into account which enhances the robustness of the models. Scaling of features was very important in the performance of the Support Vector Machine. The SVM model had an accuracy of 63.3 and Nylon 6,6 obtained an ideal classification once again. But it did not classify ABS well (precision and recall=0.00), whereas PP did fairly well in classification (precision=0.48, recall=1.00). The highest overall accuracy of the tested models was 70.0, which was obtained using the k-Nearest Neighbors algorithm with the setting of k = 5k = 5k=5. The model was very effective with Nylon 6, 6 (precision = 1.00) and moderate with ABS (precision = 0.62) and PP (precision = 0.54).

The experiment shows that local similarity-based classification may be effective in case properly scaled FTIR vectors are available as inputs. Figure 2 shows the confusion matrices of the four supervised learning models, Random Forest (RF), Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), and Logistic Regression on the classification of FTIR spectra of three polymers; Acrylonitrile Butadiene Styrene (ABS), Polypropylene (PP), and Nylon 6,6. The confusion matrix analysis as illustrated in Figure 2 also indicated that Nylon 6,6 with its unique amide characteristics is the most separable one, whereas ABS and PP have partially overlapped spectral characteristics hence classification ambiguity. The results confirm the further use of ensemble and distance-based models in polymer identification tasks using FTIR to study its ensemble structure and take advantage of complex interactions between features.

In relation to SVM, the performance was significantly low on ABS, and all the 11 ABS spectra were misclassified as PP. This implies that this model might have over-represented PP features or was unable to provide a discriminative boundary of ABS in the transformed kernel space. While PP and Nylon 6,6 were correctly identified, the complete lack of true positives for ABS highlights the limitations of SVM in this specific application without further feature engineering or class balancing. The k-NN model (bottom-left) achieved the most balanced performance among the models tested. Nylon 6,6 samples were again perfectly classified, while ABS achieved a correct classification rate of 7 out of 11. PP was more ambiguous, with half of its samples (5/10) classified correctly and the remainder split between ABS predictions. The distance-based nature of k-NN makes it sensitive to the shape and scale of the feature space, which was mitigated in this study through proper standardization. This model benefitted from local spectral similarities and provided competitive overall performance. Figure 3 presents the classification accuracy for four supervised machine learning models Random Forest (RF), Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), and Logistic Regression applied to Fourier Transform Infrared (FTIR) spectroscopy data for polymer identification. Accuracy was computed as the ratio of correctly classified instances to the total number of test samples.

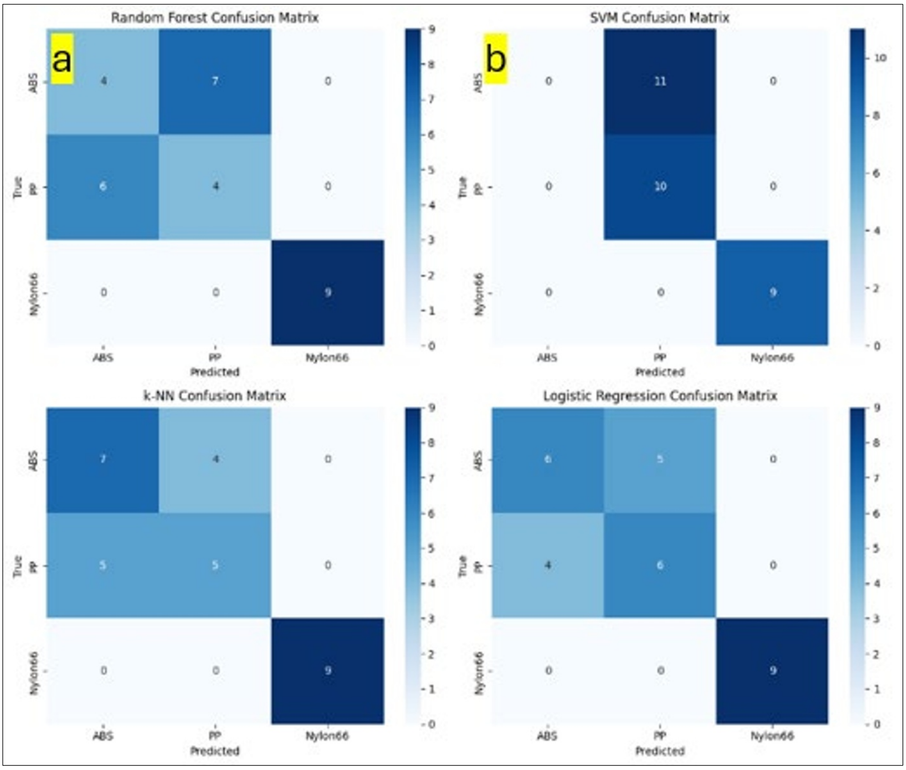

Fig. 2. Confusion Matrix, (a) Predicted v/s (b) Actual material classes (ABS, PP, Nylon 6,6).

RF model (top-left) had the highest performance when it came to the correct identification of Nylon 6,6 samples and the correct classification was perfect (9/9). Nevertheless, the model was not able to distinguish between ABS and PP with 7 ABS samples being incorrectly predicted as PP and 6 PP samples being incorrectly predicted as ABS. This overlap could be due to overlapping between spectral areas especially in the C-H region, 2800-3000 cm-1 which is common in both polymers. In spite of this confusion, RF showed a sensible compromise between accuracy and recall and used its ensemble form to model the interaction of elaborate features. In relation to SVM, the performance was significantly low on ABS, and all the 11 ABS spectra were misclassified as PP. This implies that this model might have over-represented PP features or was unable to provide a discriminative boundary of ABS in the transformed kernel space. Although PP and Nylon 6,6 were detected properly, the zero number of true positives of ABS demonstrates the weakness of SVM in the given use without additional feature engineering and balancing of classes. The k-NN model (bottom-left) was the most balanced in performance of all the tested models.

The samples of nylon 6, 6 were once again classified perfectly and ABS achieved a correct classification rate of 7/11. The ambiguity of PP was higher with 5/10 of its samples correctly predicted and the rest being divided in ABS predictions. The distance-sensitive aspect of k-NN also allows it to be sensitive to the shape and scale of the feature space, which was addressed in this paper by appropriate standardization. This model was advantageous in the case of local spectral similarities and competition in overall performance. Figure 3 shows the classification accuracy of four supervised machine learning models namely; Random Forest (RF), Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), and Logistic Regression in using fourier Transform Infrared (FTIR) spectroscopy data to identify polymers. The level of accuracy was determined by the proportion of the instances correctly categorized to the number of test samples.

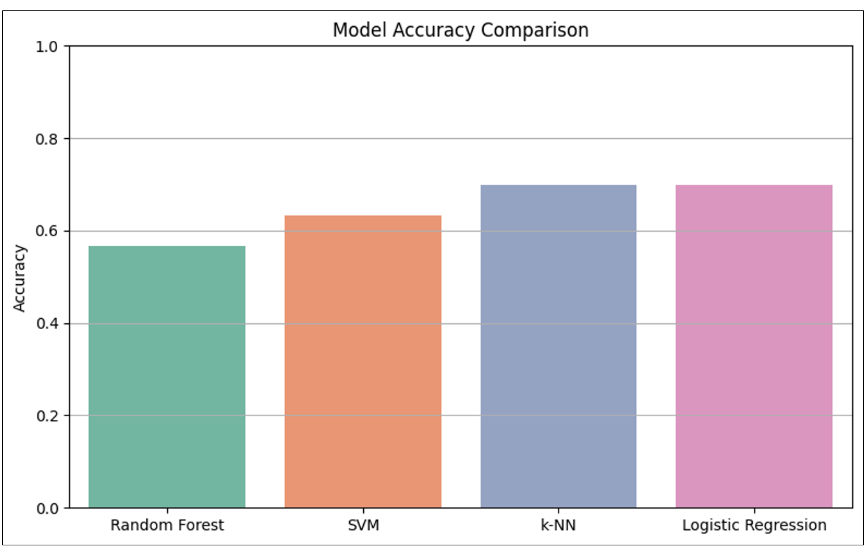

Fig. 3. Model accuracy comparison.

The highest accuracy at 70 % was obtained with k-Nearest Neighbors (k-NN) algorithm which means that it makes effective use of local similarities between spectral features. The fact that FTIR spectra are generally high-dimensional using continuous numeric values makes the Euclidean distance-based classification of k-NN advantageous to the standardized transmittance values especially when spectra of the same material fall close to each other in feature space. The next best competitor with a good performance of about 69 is Logistic

Regression that has linear decision boundaries. A multinomial logistic function was used to address the three-class issue (ABS, PP and Nylon 6, 6). Non-linear relationships between variables however could be the reason of its slight decrease over k-NN. Support Vector Machine (SVM) was 63 percent accurate, which was its ability to distinguish between Nylon 6, 6 and PP, however, a total misclassification of ABS (as shown in Figure 1) greatly affected its overall performance. Even though RBF kernel can be used with nonlinear data, additional hyperparameter optimization (e.g., C, gamma) can be necessary to enhance the separability of the classes. Random Forest (RF) classifier had a lowest accuracy rate of 57 percent which was the lowest among the four models. Although RF demonstrated high classification of Nylon 6,6 and could be interpreted using the feature importance, RF was highly confused in differentiating between ABS and PP. This is possibly because there are overlapping spectral characteristics in some levels of wavenumber, and thus the decision boundaries are not clear. Figure 4. Show the Flowchart of the Experimental and Computational Workflow for Polymer Classification Using FTIR Spectroscopy and Machine Learning.

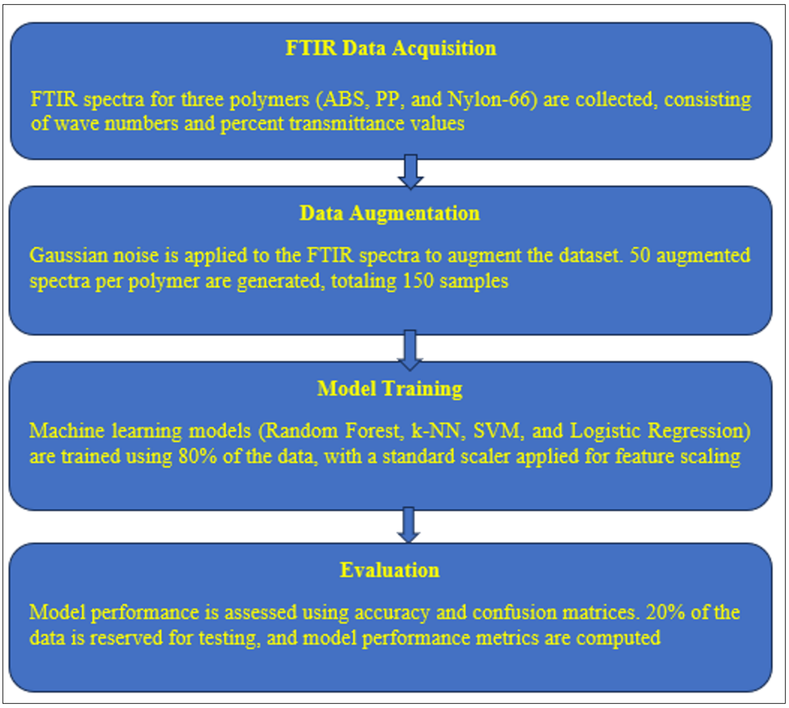

Fig. 4. Flowchart of the experimental and computational workflow of the polymer classification by using ftir spectroscopy and machine learning.

3 Result and discussion

The findings show that, the deep learning models can be used to perform even better with larger-scale applications, but the traditional classifiers like the Random Forest and k-NN give a better compromise between accuracy and computational efficiency, especially with smaller and lab-scale. The Logistic Regression with a multinomial loss function through softmax transformation achieved a total of 56.7 % accuracy. Again, Nylon 6,6 was correctly classified with the highest possible scores, though, ABS classification was poor (precision = 0.33, recall = 0.18) and PP classification was moderate (precision = 0.40, recall = 0.60). Regardless of its simplicity and efficiency, Logistic Regression had a problem of the non-linear separability of the classes in the spectral space. As illustrated in Figure (bottom right), the Logistic Regression, which was configured as a multinomial softmax, could classify Nylon 6,6 samples perfectly. But it was linear, which limited its ability to separate a superimposed ABS and PP spectral pattern. Five out of eleven ABS samples and five out of ten PP samples were accurately identified with a substantial proportion being misclassified as the other one. This implies that the model was not able to well explain nonlinear relationships in transmittance data. The accuracy of classification was flawless in all models with Nylon 6,6. This is probably because it has clear and strong amide bands (especially at 3300 cm-1 N-H stretching and 1640 cm^{-1} C=O stretching) that can easily distinguish it as compared to the ABS and PP, which are dominated by the hydrocarbon bands. On the contrary, ABS and PP had overlapping properties, particularly in the aromatic and aliphatic C-H stretching regions (2800-3000 cm^{-1}) and this lowered the classification accuracy. The use of these two classes as models such as k-NN and Rand Forest was more effective as they were able to capture local and ensemble-level patterns, respectively.

4 Conclusion

This research will illustrate how machine learning methods can be useful in classifying polymers based on their FTIR spectral data with reference to three commonly used materials that are Acrylonitrile Butadiene Styrene (ABS), Polypropylene (PP), and Nylon 6,6. The transmittance patterns of characteristic across common ranges of wavenumber (Figure 1) formed a strong basis of discriminative modeling because each polymer had its own characteristic features of molecular vibration, including the C=N band in ABS, bending of methyl groups in PP and amide bands in Nylon 6,6. The SVM, k-Nearest Neighbours, and Logistic Regression model of supervised classification were trained and tested. As shown in the confusion matrices (Figure 2), all models showed identical and optimal classification of Nylon 6,6, which is due to its distinctive and separate spectral bands (e.g., 3300 cm-1 of N-H stretch). Nevertheless, a common problem between ABS and PP was cross misclassification especially in the Random Forest and SVM classifier, which implies that spectral areas do overlap at 1450-1600 cm-1 and thus cannot be separated.

These findings were further supported by the model performance as represented in Figure 3. The highest accuracy of the K-NN classifier was 70 %, then the Logistic Regression classifier (69%), SVM classifier (63%), and the random forest classifier (57%). These findings support the effectiveness of distance-based and linear probabilistic algorithms on this high-dimensional spectral data particularly when used together with standardized preprocessing. This research shows that machine learning can be used as a feasible method of FTIR-based material identification. Although all models were effective in the detection of Nylon 6,6, ABS and PP cannot be classified because of the overlap of the spectrum. Additional research in the fields of advanced feature extraction, dimensionality reduction, and deep learning should be conducted in the future to improve the effectiveness of classification and accuracy. This research proves that machine learning models can be successfully used to classify polymers based on FTIR data, future studies should consider how this method can be extended to other polymers, in particular, polymer wastes in a mixed polymer recycling setting. Also, the application of this method to automated sorting systems may lead to a significant enhancement of the process of the material recovery, which will help to enhance the effectiveness of recycling activities.

References

1. P.R. Griffiths, J.A. de Haseth, Fourier transform infrared spectrometry, 2nd ed. (Wiley-Interscience, Hoboken, NJ, USA, 2007)
2. J.S. Tumuluru, C.T. Wright, J.R. Hess, FTIR analysis of plastics in municipal solid waste, J. Anal. Methods Chem. 2016, 483292 (2016)
3. K.C. Patel et al., Machine learning approaches in FTIR spectroscopy for drug authentication, Anal. Chem. 87, 12345–12353 (2015)
4. Y. Huang, H. Li, Y. Song, Plastic waste classification using NIR spectroscopy and machine learning, Waste Manag. 79, 248–256 (2018)
5. H.T. Chiu, M.S. Tsai, J.S. Wang, FTIR-based plastic classification using convolutional neural networks, Sensors 20, 4955 (2020)
6. M. Barone, D. Cespi, F. Passarini, F. Cavani, Machine learning for plastic identification: A review, Resour. Conserv. Recycl. 174, 105720 (2021)
7. Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521, 436–444 (2015)
8. M. Hufnagl et al., A methodology for the identification of microplastics using random decision forest classifiers based on FTIR spectra, Environ. Sci. Technol. Lett. 8, 955–961 (2021)
9. V. Hidalgo-Ruz et al., Automated classification of plastic debris using FTIR spectra and machine learning, Environ. Sci. Technol. 54, 3228–3237 (2020)
10. S. Yu, H. Yu, W. Zhang, L. Wang, Deep learning for FTIR-based microplastic classification: A 1D convolutional neural network approach, Environ. Sci. Technol. 57, 3357–3367 (2023)
11. D. Almanifi, W. Ng, Spectral classification of plastic materials using supervised and semi-supervised machine learning, Mekatronika 3, 36–45 (2021)
12. M.V. Al-Salem, Y.B. Khalil, A.K. Abdulla, Machine learning-based classification of plastic waste using FTIR spectroscopy, Recycling 10, 46–58 (2024)
13. G. Primpke, M. Wirth, C. Lorenz, Comparison of classification algorithms for FTIR-based identification of microplastics, Environ. Sci. Technol. 53, 2313–2321 (2019)
14. A. Ferreiro, S. Pazos, J.J. Madariaga, Genetic algorithm and FTIR spectroscopy for the detection and classification of polymer debris, Anal. Chem. 90, 10573–10581 (2018).
15. A. Valls-Conesa, A. Velázquez, M.D. Fernández-González, Rapid classification of microplastics using FTIR hyperspectral imaging and random forest models, Anal. Chem. 89, 11050–11057 (2017).