CLASSIFICATION OF BACTERIA USING FTIR WITH LDA, SVC, AND LR: EFFECTS OF PCA AND ANOVA PREPROCESSING

Thanh An Ngo[1], Bui Thi Phuong Quynh[1], Le The Nhan[2], Bui Thu Ha[1*]

[1]*Faculty of Chemical Engineering, Ho Chi Minh City University of Industry and Trade*

[2]*Division for Water Resources Planning and Investigation for the South of Vietnam*

*Email: *habt@huit.edu.vn*

Received: 2 May 2025; Revised: 20 May 2025; Accepted: 31 May 2025

ABSTRACT

This study aimed to apply three supervised models, including LDA, SVC, and LR, to classify 15 different bacterial species based on FTIR. These models were directly applied to the preprocessed data and the filtered spectral data. ANOVA was used to select key features, while PCA helped retain principal components. The results indicated that selecting feature values with an F-value threshold of 3 achieved the highest accuracy with the LDA model (94%), followed by the SVC (80%) and LR (74%). In addition, applying PCA to retain 300 principal components offered the accuracies of LDA, SVC, and LR of 93.1, 76.7, and 72.3%, respectively. Both feature selection methods were demonstrated to be informative and yielded higher accuracies than the unfiltered data in the classification of the bacterial species studied.

Keywords: FTIR, bacterial classification, PCA, ANOVA, SVC, LR.

1. INTRODUCTION

Fourier transform infrared (FTIR) spectroscopy has recently emerged as a powerful and rapid analytical technique for the identification and classification of microorganisms [1, 2]. By capturing the unique biochemical fingerprint of a bacterial cell, FTIR spectroscopy provides detailed spectral data that can be used to differentiate between species and even strains. This method allows the sample to be scanned for whole bacterial cells or cell fragments at infrared frequencies between 4,000 and 600 cm^{-1}. Over the past decades, significant advances in instrumentation and data processing have increased both the throughput and quality of FTIR measurements. Despite these improvements, the complexity of bacterial spectra, which is influenced by variations in culture conditions, sample preparation, and inherent biodiversity, continues to pose challenges to accurate classification [3]. FTIR spectroscopy offers several key advantages over many other chemical analysis methods. It provides rich molecular information, reflects the chemical composition of the sample, and often provides a high signal-to-noise ratio. Furthermore, the technique is relatively fast, low-cost, and can be automated for high-throughput laboratories. With fast processing times, no reagents are required, and there is virtually no sample preparation required from a sample analysis perspective. Another important benefit is that FTIR is a non-destructive method, allowing for subsequent analysis of the same sample. However, FTIR does have some limitations. First, spectral analysis is limited to specific regions of the infrared spectrum, and the presence of water can interfere with these regions, potentially complicating interpretation. Second, specialized knowledge is required to accurately analyze and interpret the complex

spectra produced by FTIR. This technique does not detect some elements or diatomic molecules (e.g., N_2 or O_2), and the overlap of spectral regions from different components can sometimes lead to ambiguous results [4]. Third, due to the complex composition of biofilms, there are challenges in identifying specific components due to overlapping spectral bands and heterogeneous mixtures of biomolecules and extracellular polymers [5]. Finally, variations in microbiological environmental parameters – such as culture medium composition, growth temperature, or incubation time – can cause differences in the spectra obtained [6]. Although such variations can provide useful information about bacterial physiology, they also add a layer of complexity to the analysis.

To overcome these challenges, chemometric methods are increasingly being applied to FTIR data analysis. Chemical analysis of FTIR spectra is a fundamental technique for bacterial differentiation at the genus, species, and clone levels [7]. Recently, many researchers have applied many data analysis algorithms, such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Analysis of Variance (ANOVA), etc., to FTIR spectral data analysis to enable the extraction of major chemical components that differentiate between different bacterial strains [8]. Unsupervised techniques such as PCA are widely used to reduce the dimensionality of large mass spectral data while preserving the most informative variance. Meanwhile, statistical methods such as ANOVA have proven effective in identifying important spectral features that differentiate between bacterial groups. Although there have been many studies related to this field of chemometrics in FTIR spectral analysis for microbial identification, there has been little reported on the role of data preprocessing using PCA or using ANOVA for feature selection, which then serves the classification models such as LDA, Support Vector Classifier (SVC) and Logistic Regression (LR). In addition, there has been no publication comparing the accuracy performance of microbial classification models using FTIR spectral analysis.

Based on the above analysis, in this study, we proposed an approach using PCA and ANOVA as preprocessing steps before applying supervised learning algorithms such as LDA, SVC, and LR for FTIR-based bacterial classification. PCA is used to distill the high-dimensional spectral data into a set of concise principal components that capture the dominant variance, while ANOVA is used to select the most discriminative features. All models will be compared through predictive accuracy determined through a confusion matrix.

2. MATERIALS AND METHODS

2.1. Dataset and Software

The original dataset used for this study is a publicly available FTIR spectral dataset made available by Smirnova et al. (2020) [9]. It consists of 795 samples, each represented by 1,820 columns, while information pertaining to the FTIR starts from the 10th column. The spectral data range of wavenumbers has been extended to capture broad absorption patterns for characterizing bacterial species. The classification task was aimed at differentiating 15 bacterial species represented in the 'Species' column (column 3). These include spp, ant, cry, ory, arc, sol, kaf, rub, flu, ext, ver, ura, ery, yun, and gla, corresponding to the number of samples as follows: 263, 150, 19, 18, 16, 82, 18, 52, 15, 15, 16, 37, 41, 35, and 18, respectively. Owing to the complexity of spectral data, where there may be overlapping features across different species, appropriate preprocessing and dimensionality reduction techniques were used for better classification.

All data preprocessing, feature selection, dimensionality reduction, and classification functions were written in Python version 3.9.6. All calculations were performed on a personal

workstation with an Intel Core i7 processor and 32GB of RAM under a Windows 10 operating system.

2.2. Data analysis method

Principal Component Analysis: PCA is a widely used unsupervised dimensionality reduction technique that transforms high-dimensional datasets into a lower-dimensional space by identifying principal components that maximize variance. It is particularly valuable for data visualization, noise reduction, and preprocessing in machine learning, leveraging linear relationships within the data to simplify complex datasets [10].

Linear Discriminant Analysis: LDA is a supervised classification approach that optimizes class separability by projecting data onto a reduced-dimensional space where the ratio of between-class variance to within-class variance is maximized. This makes it an effective method for classifying labeled data, especially when linear boundaries are sufficient to distinguish classes [11].

Support Vector Classifier: SVC is a powerful supervised learning algorithm that constructs optimal hyperplanes to separate classes with the maximum margin in a high-dimensional space. Utilizing kernel functions (e.g., linear, poly, rbf, etc), it can handle both linear and non-linear classification tasks, offering flexibility across diverse datasets [12].

Logistic Regression: LR is a supervised statistical method that predicts the probability of a categorical outcome using the logistic function. It models the relationship between input features and the likelihood of a specific class, making it a robust choice for binary or multi-class classification with linearly separable data [13].

ANOVA Feature Selection: ANOVA feature selection is a statistical technique that identifies the most discriminative features for classification by assessing the significance of mean differences across groups using *F*-values. This supervised method reduces dimensionality by retaining features with the strongest statistical relevance, enhancing model performance [14].

2.3. Data Processing Sequence

Two preprocessing pipelines were generated in this study to ensure robust classification of the different groups of fast-growing bacteria with the use of FTIR spectra data. In the first method, PCA was employed to perform the dimensionality reduction in extracting the most important spectral data. The second method carried out Feature Selection on the data using ANOVA for the detection of the most informative spectral variables, while still preserving their original meanings. After dimensionality reduction or selection of features, Savitzky-Golay smoothing was carried out on the selected spectral data to minimize the white noise while maintaining the characteristics of the spectra, where the parameters are selected as follows: window length of 3 and poly order of 2. The smoothed data was standardized to ensure proper distribution of feature values. The dataset was then split into 70% training and 30% testing sets. SMOTE was applied to the training set to balance class distribution. Subsequently, three classification models - LDA, SVC with a linear kernel, and multinomial LR - were trained on the training set using Stratified K-Fold cross-validation, a technique that divides the data into folds while preserving the class distribution to ensure robust evaluation. Their performance was then evaluated on the test set.

These two methodologies allow for a comparative analysis of dimensionality reduction via PCA versus feature selection via ANOVA, providing insights into their respective effects on classification accuracy for bacterial identification.

2.4. Model evaluation

To evaluate the performance of the classification models, two approaches are employed depending on the data splitting strategy. When using Stratified K-Fold Cross-Validation, the average accuracy across all folds is calculated to assess the models' robustness and generalization ability. Stratified K-Fold Cross-Validation ensures that each fold maintains the same class distribution as the original dataset, providing a reliable estimate of model performance, particularly for multi-class problems such as bacterial classification [15].

For the approach where the dataset is split into 70% training and 30% testing sets, a confusion matrix is used to evaluate the quality of the classification process. The confusion matrix provides a tabular summary of predicted versus actual class labels, detailing true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). This enables a comprehensive analysis of model performance across different classes, highlighting specific misclassifications in multi-class scenarios [16].

3. RESULTS AND DISCUSSION

3.1. Preprocessed data

Figure 1 displays the mean FTIR spectrum for each bacterium. By comparing with the characteristic spectral regions of bacteria listed in Table 1 [2], it is evident that Figure 1 exhibits several key peaks, including those at 1739 cm^{-1} (stretching C=O of esters and carboxylic acids), 1647 cm^{-1} (amide I band), 1548 cm^{-1} (amide II band), 1456 cm^{-1} (symmetric stretching of COO^-), 1398 cm^{-1} (symmetric stretching of COO^- and CH_2/CH_3 bending), 1242 cm^{-1} (vibrations of -COOH, C-O-H, and >P=O stretching), 1084 cm^{-1} (stretching P=O of phosphodiester), and 976 cm^{-1} (symmetric stretching of phosphoryl groups), which align with the functional group assignments in Table 1.

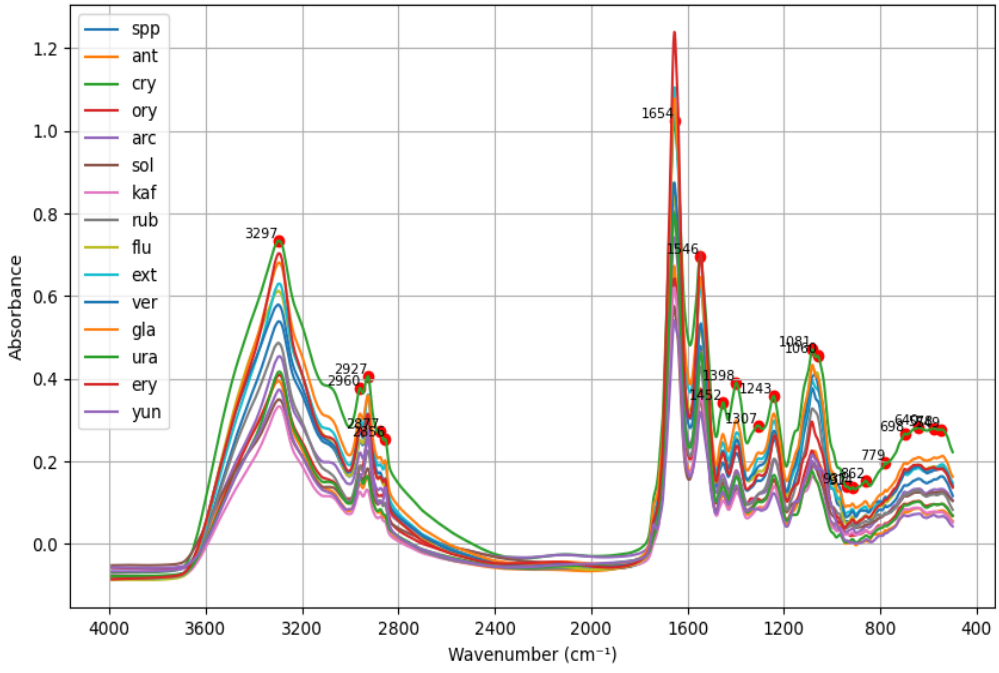

Figure 1. Mean FTIR spectrum per species

Table 1. Infrared absorption bands of the most common bacterial functional groups

Wavenumber (cm^{-1})	Functional Group Assignment
1739	Stretching C=O of ester functional groups from membrane lipids and fatty acids; stretching C=O of carboxylic acids
1647	Stretching C=O in amides (amide I band)
1548	N–H bending in amides (amide II band)
1405	Symmetrical stretching for the deprotonated COO^- group
1453	C–N stretching in amides (amide III band)
1397	Symmetric stretching of COO^-; Bending CH_2/CH_3
1305	Vibration C–N from amides
1300–1250	Vibrations of C–O from esters or carboxylic acids
1262	Vibrations of –COOH and C–O–H; Double bond stretching of >P=O of general phosphoryl groups and phosphodiester of nucleic acids
1225	Stretching of P=O in phosphates
1200–950	Asymmetric and symmetric stretching of PO_2^- and $P(OH)_2$ in phosphates; vibrations of C–OH, C–O–C and C–C of polysaccharides
1084	Stretching P=O of phosphodiester, phosphorylated proteins, or polyphosphate products
976	Symmetrical stretching vibration of phosphoryl groups

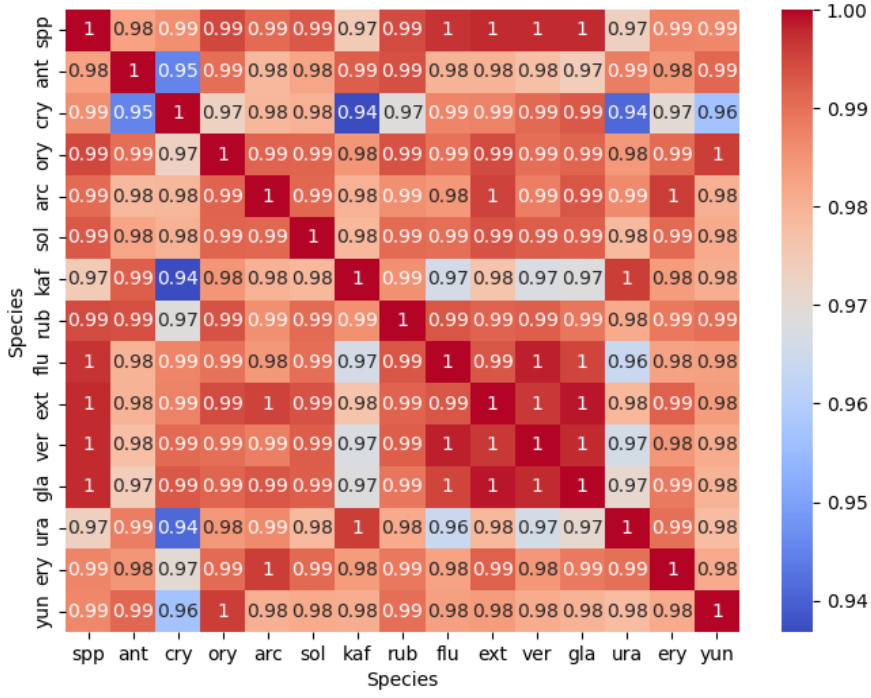

Figure 2. Cosine similarity between the mean spectra

Figure 2 presents a heatmap visualizing the Cosine similarity between the mean FTIR spectra of various bacterial species. The heatmap, displayed as a color-coded matrix, reveals that most bacterial species have highly similar spectra, with only the species "Cry", "Ura", "Kaf", and "Ant" exhibiting relatively minor differences, as indicated by Cosine similarity values of 0.96, 0.94, 0.94, and 0.95, respectively. This visualization effectively illustrates the degree of distinguishability among the mean spectra, aiding in the assessment of how easily these bacterial species can be differentiated for accurate classification in subsequent analyses.

Figure 3 displays the Scree plot from the PCA analysis of FTIR spectral data for 15 bacterial species. This plot aids in determining the optimal number of principal components for the subsequent data classification step. The results indicate that beyond approximately 10 principal components, the cumulative variance explained changes minimally. However, retaining more principal components may increase the model's R^2 value (approaching 1), but it also raises the risk of incorporating noise, which could negatively impact the performance of subsequent classification steps. Therefore, a thorough investigation is needed to assess the impact of the number of retained principal components in the PCA preprocessing step on the accuracy of the following classification models.

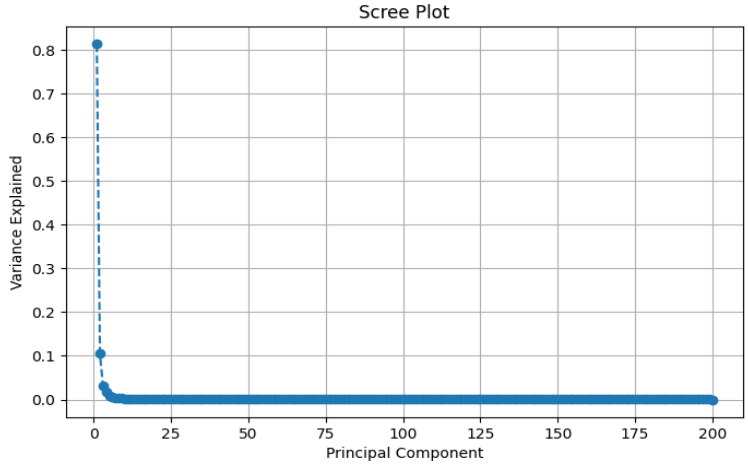

Figure 3. Scree Plot for PCA

3.2. Evaluation of classification models

After applying either PCA for dimensionality reduction or ANOVA for feature selection, classification models including LDA, SVC, and LR were employed to distinguish between different bacterial species. Model performance was evaluated using accuracy on the test set, which comprised 70% of the total data. The classification accuracies obtained with PCA preprocessing are presented in Table 2, while those using ANOVA-based feature selection are shown in Table 3. For comparison, the classification models were also applied directly to the raw, unprocessed data to highlight the impact of preprocessing on model performance.

Table 2. Accuracy of the test set for classification with preprocessed data by PCA at various components

Method	K = 100	K = 200	K = 300	K = 400
LDA	86.8	89.9	93.1	86.8
SVC	76.72	76.7	76.7	76.7
LR	72.3	72.3	72.3	72.3

These results, as presented in Table 2, suggest that the performance of the LDA model is highly sensitive to the number of principal components retained during PCA. Specifically, as the number of components (K) increases from 100 to 300, model accuracy improves, indicating that the additional components contribute relevant discriminative information. However, when K reaches 400, the accuracy begins to decline, likely due to the inclusion of noise or redundant information that hinders the model's ability to generalize.

Interestingly, both SVC and LR models appear largely unaffected by changes in the number of principal components, with their accuracies remaining relatively stable across the range of K values tested. This may imply that these models are either less sensitive to the high-dimensional variance structure captured by PCA or are not able to leverage the additional components as effectively as LDA. Moreover, the consistently lower accuracy of SVC and LR compared to LDA across all values of K highlights LDA's suitability for this particular classification task, possibly due to its assumption of normally distributed features and its focus on maximizing class separability.

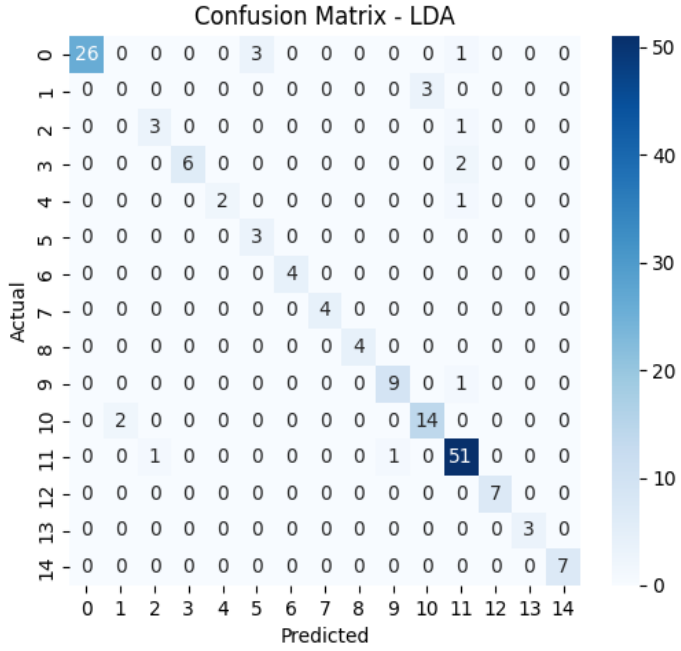

Figure 4. Confusion matrix using LDA with K = 200

Figure 4 presents the confusion matrix of the LDA model on the test set, illustrating the model's classification performance across 15 classes (labeled from 0 to 14). The results indicate that LDA performs well on several classes, notably class 11 (with 51 correctly classified samples), class 10 (14 correct predictions), class 0 (26 correct), and class 9 (9 correct). This suggests that LDA successfully learned discriminative features for these categories. However, there are some misclassifications, particularly in classes with fewer samples or overlapping features. For instance, class 10 shows minor confusion with classes 0 and 9, while samples from classes 2 and 5 are occasionally misclassified into classes like 12 or 11. These errors may reflect feature overlap or insufficiently distinctive patterns between certain classes. Overall, the dominance of values along the diagonal of the matrix confirms that LDA achieves high classification accuracy and demonstrates strong class separation in most cases.

Table 3. Accuracy of test set for classification with feature selection by ANOVA at various *F*-value thresholds

F-value threshold	Number of features	LDA accuracy (%)	SVC accuracy (%)	LR accuracy (%)
2	1599	91	75	74
3	1246	94	80	74
4	515	88	74	67
5	165	43	35	36
6	152	28	33	36

Table 3 presents the classification accuracy of three models - LDA, SVC, and LR - on the test set after applying ANOVA-based feature selection with different *F*-value thresholds. FTIR spectra typically contains thousands of wavenumber-related data points, many of which may be irrelevant, redundant, or affected by noise from the environment or equipment. Feature selection serves as an important preprocessing step to remove such noisy or uninformative regions, focusing the analysis on spectral regions with strong and meaningful signals. This reduces computational complexity, reduces the risk of overfitting, and enhances model interpretability [17]. Among the three models, LDA shows the most noticeable change in performance. It achieves the highest accuracy of 94% when the *F*-value threshold is set to 3, corresponding to retaining 1246 features. However, as the threshold increases (resulting in fewer features being selected), the performance of LDA drops sharply. This pattern highlights the dependence of LDA on a sufficiently large and informative feature set to effectively discriminate between classes. Excessive feature reduction can result in the loss of important spectral information, especially in complex FTIR datasets, thus degrading the classification performance. In contrast, the SVC and LR models show a gradual decrease in accuracy as the number of features decreases. Their accuracy peaks at lower thresholds (2–3), remains relatively stable, and is lower than that of LDA when fewer features are retained.

The results from Tables 2 and 3, as shown above, show that LDA provides higher accuracy than SVC and LR. On the other hand, it should also be noted that in machine learning, LDA is classified as a generative model while SVC and LR are classified as discriminative. The generative group is used to describe models where the classification rules are derived indirectly from the statistical model rather than directly learned from the discriminative data, as in the discriminative model. Thus, this result also implies that the generative model with the assumptions (multivariate normal distribution and classes with similar covariance matrices) is more suitable for the discriminative model.

Another notable result in this study is that the accuracy of SVC and LR is not high. Many authors in the [18, 19] reports also argue that the performance of SVM extensions varies greatly depending on the data structure, and that SVMs are not always the best choice for multiclass classification, and they can perform poorly in multiclass. Compared to the large number of classes (15 classes) in this study, it can be seen that the negative impact of the number of classes on the accuracy of the SVC model. For LR model, it is possible that due to lack of control in regularization, its performance is worse than LDA.

Figure 5 demonstrates the confusion matrix obtained from LDA combined with ANOVA feature selection at an *F*-value of 3. This confusion matrix demonstrates that the LDA model performs well in classifying labels from 0 to 14. Most samples are correctly predicted (values along the diagonal), with label 11 showing the highest accuracy, correctly identifying 72 samples. However, some misclassifications still occur - for instance, label 10 is occasionally

misclassified as labels 9 and 11, indicating that certain classes may share similar FTIR spectral characteristics.

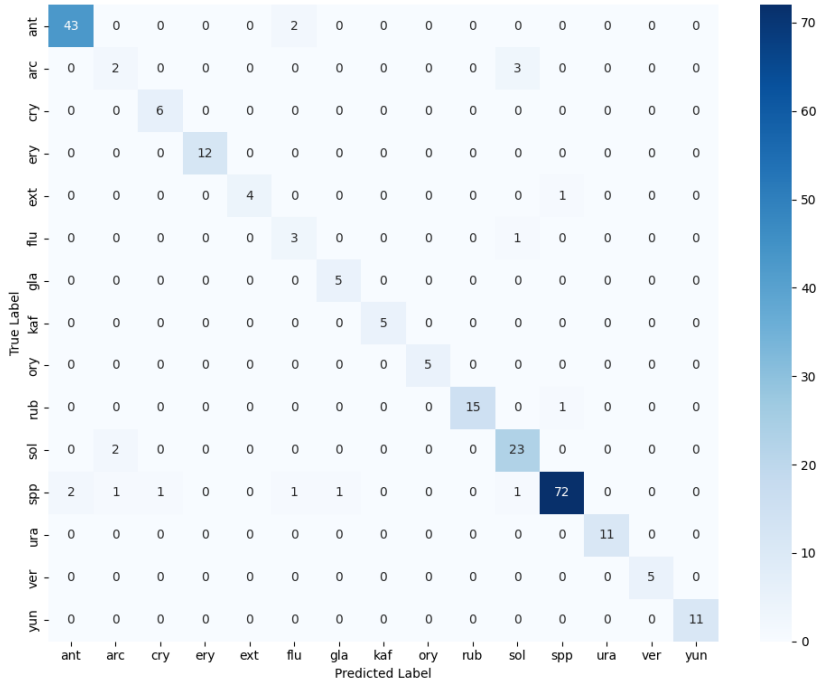

Figure 5. Confusion matrix using LDA with *F*-value = 3

Table 4. Performance metrics for multi-class bacterial classification on the test set

	Precision	Recall	F1-score	Number of samples
ant	0.96	0.96	0.96	45
arc	0.40	0.40	0.40	5
cry	0.86	1.00	0.92	6
ery	1.00	1.00	1.00	12
ext	1.00	0.80	0.89	5
flu	0.50	0.75	0.60	4
gla	0.83	1.00	0.91	5
kaf	1.00	1.00	1.00	5
ory	1.00	1.00	1.00	5
rub	1.00	0.94	0.97	16
sol	0.82	0.92	0.87	25
spp	0.97	0.91	0.94	79
ura	1.00	1.00	1.00	11
ver	1.00	1.00	1.00	5
yun	1.00	1.00	1.00	11

Table 4 reports precision, recall, and F1-score for 15 classes in a multi-class bacterial classification task, alongside test set sample counts. Classes like ery, kaf, ory, ura, ver, and yun achieve perfect scores (1.00), reflecting robust classification due to distinct features or SMOTE-augmented training data. However, arc (F1-score: 0.40, 5 samples) and flu (F1-score: 0.60, 4 samples) show poor performance, likely due to limited samples or feature overlap. The imbalanced test set (4–79 samples) highlights SMOTE's role, though minority classes remain challenging, suggesting further feature engineering is needed.

Figure 6 is a 3D Score plot for the three principal components of LDA applied to the training dataset. From this graph, the groups are well grouped, and the individual groups can be separated.

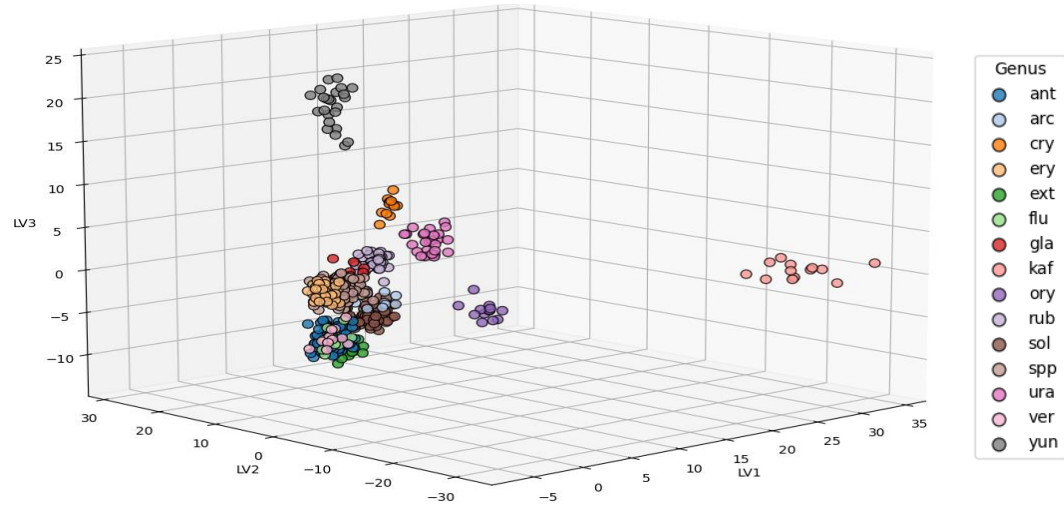

Figure 6. Score plot of LDA (*F*-value = 3)

Table 5 shows the classification accuracy on raw, unprocessed data: LDA at 89%, SVC at 77%, and LR at 74%. These results indicate that the raw data's high dimensionality and noise impact model performance. Compared to preprocessed data from Tables 2 and 3, LDA's accuracy on raw data (89%) is slightly below its peaks of 93.1% with PCA (K = 300) and 94% with ANOVA (*F*-value = 3), suggesting that preprocessing enhances LDA's performance by reducing noise. SVC and LR remain more stable, with SVC improving slightly from 77% to 80% (ANOVA, *F*-value = 3) and LR holding steady at 72–74%. This highlights LDA's greater sensitivity to noise and dimensionality, while SVC and LR are less affected but also less able to leverage additional features for improved discrimination.

Table 5. Accuracy of the test set for classification with raw data

Method	Accuracy (%)
LDA	89
SVC	77
LR	74

4. CONCLUSION

The results of this study demonstrate that applying data preprocessing techniques such as PCA and ANOVA-based feature selection significantly enhances the classification accuracy

of FTIR spectral data for different bacterial species when using models like LDA, SVC, and LR. Among these models, LDA consistently outperforms the others, achieving the highest accuracy of 94% with PCA and 93.1% with ANOVA. In contrast, SVC and LR show noticeably lower performance under the same preprocessing conditions. Importantly, for both PCA and ANOVA, the number of retained principal components or selected features plays a critical role in determining classification accuracy. Each model reaches its peak performance at an optimal number of components or features, beyond which accuracy tends to decline, highlighting the adverse effect of noise or irrelevant information on model performance.

REFERENCES

1. Kassem, A., Abbas, L., Coutinho, O., Opara, S., Najaf, H., Kasperek, D., Pokhrel, K., Li, X. & Tiquia-Arashiro, S. - Applications of Fourier Transform-Infrared spectroscopy in microbial cell biology and environmental microbiology: advances, challenges, and future perspectives, Frontiers in Microbiology **14** (2023). https://doi.org/10.3389/fmicb.2023.1304081
2. Naumann, D. - Infrared Spectroscopy in Microbiology, In Encyclopedia of Analytical Chemistry, Wiley (2000). https://doi.org/10.1002/9780470027318.a0117
3. Ramzan, M., Raza, A., un Nisa, Z., & Ghulam Musharraf, S. - Recent studies on advance spectroscopic techniques for the identification of microorganisms: A review, Arabian Journal of Chemistry **16** (3) (2023) 104521. https://doi.org/10.1016/j.arabjc.2022.104521
4. Ojeda, J. J. & Dittrich, M. - Fourier Transform Infrared Spectroscopy for Molecular Analysis of Microbial Cells (2012) 187–211. https://doi.org/10.1007/978-1-61779-827-6_8
5. Wenning, M. & Scherer, S. - Identification of microorganisms by FTIR spectroscopy: perspectives and limitations of the method, Applied Microbiology and Biotechnology **97** (16) (2013) 7111–7120. https://doi.org/10.1007/s00253-013-5087-3
6. Zarnowiec, P., Lechowicz, L., Czerwonka, G. & Kaca, W. - Fourier Transform Infrared Spectroscopy (FTIR) as a Tool for the Identification and Differentiation of Pathogenic Bacteria, Current Medicinal Chemistry **22** (14) (2015) 1710–1718. https://doi.org/10.2174/0929867322666150311152800
7. Mariey, L., Signolle, J. P., Amiel, C. & Travert, J. - Discrimination, classification, identification of microorganisms using FTIR spectroscopy and chemometrics, Vibrational Spectroscopy **26** (2) (2001) 151–159. https://doi.org/10.1016/S0924-2031(01)00113-8
8. Brito, N. M. R. de, & Lourenço, F. R. - Rapid identification of microbial contaminants in pharmaceutical products using a PCA/LDA-based FTIR-ATR method, Brazilian Journal of Pharmaceutical Sciences **57** (2021). https://doi.org/10.1590/s2175-97902020000318899
9. Margarita, S., Kohler, A. & Volha, S. - FTIR Dataset [Data set], Zenodo (2020). https://doi.org/10.5281/zenodo.4297950
10. Abdi, H. & Williams, L. J. - Principal component analysis, WIREs Computational Statistics **2** (4) (2010) 433–459. https://doi.org/10.1002/wics.101
11. Linear discriminant analysis, Nature Reviews Methods Primers **4** (1) (2024) 71. https://doi.org/10.1038/s43586-024-00357-9

12. Cortes, C. & Vapnik, V. - Support-vector networks, Machine Learning **20** (3) (1995) 273–297. https://doi.org/10.1007/BF00994018
13. Hosmer, D. W., & Lemeshow, S. - Applied Logistic Regression, Wiley (2000). https://doi.org/10.1002/0471722146
14. Mohtasham, F., Pourhoseingholi, M., Hashemi Nazari, S. S., Kavousi, K., & Zali, M. R. - Comparative analysis of feature selection techniques for COVID-19 dataset, Scientific Reports **14** (1) (2024) 18627. https://doi.org/10.1038/s41598-024-69209-6
15. Powers, D. M. W. - Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation, Journal of Machine Learning Technologies **2** (1) (2011) 37–63. https://arxiv.org/abs/2010.16061
16. Sokolova, M., & Lapalme, G. - A systematic analysis of performance measures for classification tasks, Information Processing & Management **45** (4) (2009) 427–437. https://doi.org/10.1016/j.ipm.2009.03.002
17. Boutegrabet, W., Piot, O., Guenot, D. & Gobinet, C. - Unsupervised Feature Selection by a Genetic Algorithm for Mid-Infrared Spectral Data, Analytical Chemistry **94** (46) (2022) 16050–16059. https://doi.org/10.1021/acs.analchem.2c03118
18. Chih-Wei Hsu & Chih-Jen Lin. - A comparison of methods for multiclass support vector machines, IEEE Transactions on Neural Networks **13** (2) (2002) 415–425. https://doi.org/10.1109/72.991427
19. Rifkin, R. & Klautau, A. - In Defense of One-Vs-All Classification, Journal of Machine Learning Research **5** (2004) 101-141.

TÓM TẮT

PHÂN LOẠI VI KHUẨN SỬ DỤNG FTIR KẾT HỢP LDA, SVC VÀ LR: ẢNH HƯỞNG CỦA PHƯƠNG PHÁP TIỀN XỬ LÝ BẰNG PCA VÀ ANOVA

Ngô Thanh An[1], Bùi Thị Phương Quỳnh[1], Lê Thế Nhân[2], Bùi Thu Hà[1*]

[1]*Khoa Công nghệ Hóa học, Trường Đại học Công Thương Thành phố Hồ Chí Minh*

[2]*Liên đoàn Quy hoạch và Điều tra Tài nguyên nước miền Nam*

*Email: *habt@huit.edu.vn*

Nghiên cứu này áp dụng ba mô hình có giám sát: LDA, SVC và LR để phân loại 15 loài vi khuẩn khác nhau dựa trên phổ FTIR. Các mô hình được áp dụng trực tiếp trên dữ liệu đã được tiền xử lý và dữ liệu phổ đã lọc. Phương pháp ANOVA được dùng để chọn các đặc trưng quan trọng, trong khi PCA giúp giữ lại các thành phần chính. Kết quả cho thấy việc chọn đặc trưng với ngưỡng F-value = 3 đạt độ chính xác cao nhất ở LDA (94%), tiếp theo là SVC (80%) và LR (74%). Đồng thời, khi sử dụng PCA giữ lại 300 thành phần chính, độ chính xác của LDA, SVC và LR lần lượt là 93,1%, 76,7% và 72,3%. Cả hai phương pháp chọn đặc trưng đều chứng tỏ mang lại thông tin giá trị và độ chính xác cao hơn so với dữ liệu chưa qua lọc.

Từ khóa: FTIR, phân loại vi khuẩn, PCA, ANOVA, SVC, LR.