Functional Group Identification

for FTIR Spectra of Interstellar Ice Analogs Using Image-Based Machine Learning

640510732 Shinya Wongyai
650510695 Jarintorn Kanthajan
650510696 Jaksana Chaiyong
650510697 Jarupong Nantapokin
Assoc. Prof. Dr. Jakramate Bootkrajang

Data Science Research Center, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand

INTRODUCTION

Infrared (IR) spectroscopy is a widely used analytical technique for **identifying functional groups** in chemical compounds. However, **manual peak assignment is time-consuming**, requires **expert knowledge**, and may lead to inter-analyst inconsistency. While **machine learning** has been applied to spectral analysis, most approaches analyze the **entire spectrum** rather than focusing on **individual peak-level classification** with **interpretable features**.This study presents an **end-to-end pipeline** integrating **automated peak detection**, **physics-informed feature extraction**, and ensemble machine learning to **classify 46 functional group** classes using real laboratory IR spectra from the LIDA database [1].

METHODS

DATASET

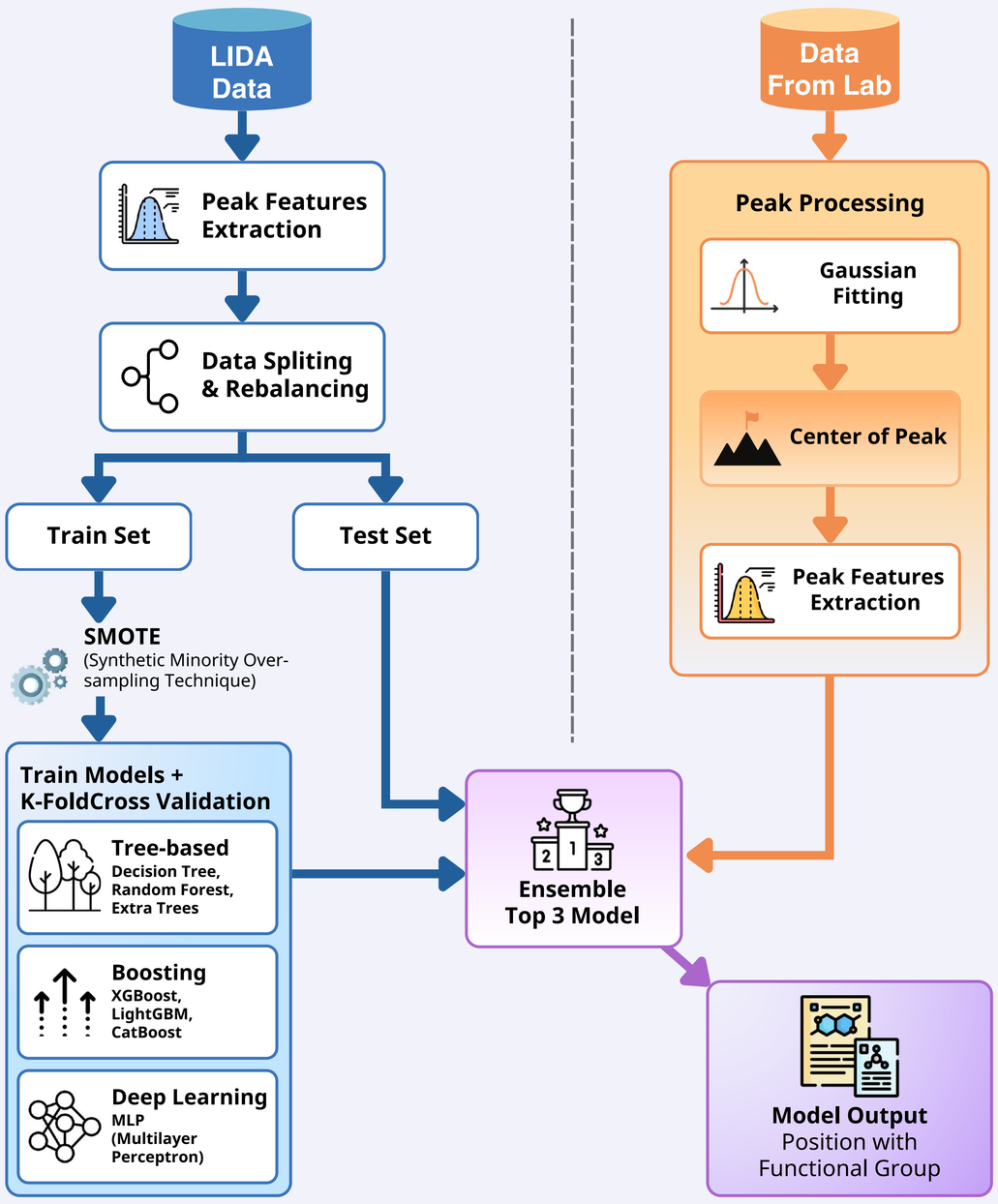
21 Chemical Compounds

46 Functional group

8,627 Peaks Labeled

1 MODEL TRAINING PIPELINE

2 PEAKS DETECTION



OBJECTIVES

- Develop a robust automated peak detection algorithm for laboratory IR spectra.
- Extract physics-informed peak features to enable interpretable and chemically meaningful analysis.
- Build and evaluate an ensemble machine learning model to accurately classify 46 functional group classes, with performance assessed using k-fold cross-validation [2]

RESULT

1 Cross-Validation Results (7 Models, 5-Fold Stratified

Model Performance (5-Fold Cross-Validation)

MODEL	ACCURACY(CV)	MACRO PRECISION (CV)	MACRO RECALL (CV)	MACRO F1 (CV)	ROC-AUC (CV)
EXTRATREES	**0.9957**	**0.9958**	**0.9964**	**0.9960**	**0.9994**
RANDOMFOREST	0.9945	0.9950	0.9957	0.9953	0.9999
XGBOOST	0.9935	0.9931	0.9942	0.9936	1.0000
DECISIONTREE	0.9894	0.9901	0.9905	0.9902	0.9952
CATBOOST	0.9849	0.9850	0.9888	0.9868	0.9996
LIGHTGBM	0.8228	0.8214	0.8138	0.8117	0.9073
MLP	0.6907	0.7080	0.6612	0.6490	0.9903

Test Set Performance (n = 1,725)

Model	Accuracy	Macro F1	ROC-AUC
Ensemble	**0.9884**	**0.9645**	**0.9995**
ExtraTrees	0.9861	0.9600	0.9917
RandomForest	0.9843	0.9567	0.999
XGBoost	0.9826	0.9469	0.9976

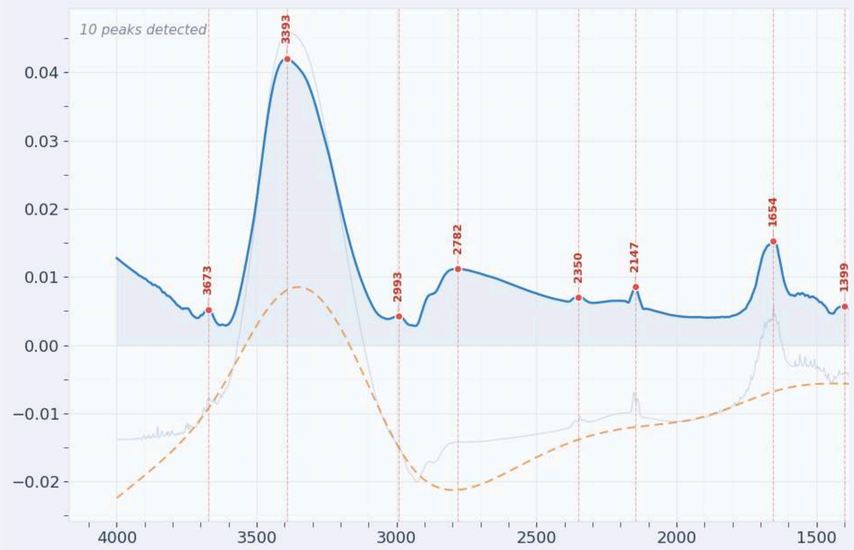
Ensemble Model
Accuracy 98.84%
Macro F1 96.45%
ROC-AUC 0.9995

Among 7 evaluated models, tree-based algorithms consistently achieved the best performance. ExtraTrees obtained the highest Macro F1 = 0.9960, followed by RandomForest (0.9953) and XGBoost (0.9936). These three models were selected for the ensemble model construction.

Models such as LightGBM and MLP showed substantially lower performance, likely due to feature scale sensitivity and class distribution imbalance in the dataset.

2 Peaks Detection

Reference : 3,061
Total peaks detected : 4,057
True Positives : 2,246
False Positives : 1,811
False Negatives : 815
Precision : 0.5536
Recall : 0.7337
Missing Ratio : 0.2663



DISCUSSION

Dominance of Position Features

Feature importance analysis shows that peak position (wavenumber) is the most important factor for classification, consistent with the spectroscopic principle that chemical bonds absorb at characteristic frequencies [4]. Additional width features (peak_width, FWHM, FWTM) further improve discrimination by capturing peak shape differences, especially when functional groups overlap in the same spectral region.

Chemically Interpretable Errors

Most misclassifications have clear chemical explanations. For example, confusion between C=O (str.) and C=O (CO_2, str.) occurs due to overlapping absorption around 2300–2400 cm^{-1}, while NH_3 (def.) and H_2O (bend.) overlap near 1600–1650 cm^{-1}. These patterns indicate that errors reflect true spectroscopic ambiguity rather than model limitations[3], [4].

CONCLUSION

This work presents a complete pipeline for automated IR peak assignment, integrating a five-step peak detection algorithm, 28 physically meaningful features, and an ensemble classifier achieving 98.84% accuracy and 99.95% ROC-AUC across 46 functional group classes.
Among seven evaluated algorithms, tree-based models performed best, with Extra Trees, Random Forest, and XGBoost forming the final ensemble.
Feature importance analysis confirms that the model follows established IR spectroscopy principles, where peak position is the dominant predictor and peak shape features provide additional discrimination[3], [4].

Future work: expand minority class data and develop adaptive feature extraction windows to improve classification of spectrally overlapping functional groups.

REFERENCES

[1] W. R. M. Rocha, M. G. Rachid, B. Olsthoorn, E. F. van Dishoeck, M. K. McClure, and H. Linnartz, "LIDA: The Leiden Ice Database for astrochemistry," Astronomy & Astrophysics, vol. 668, A63, 2022. Available: https://doi.org/10.48550/arXiv.2208.12211
[2] J. Brownlee, "A Gentle Introduction to k-fold Cross-Validation," Machine Learning Mastery, 2018. Available: https://machinelearningmastery.com/k-fold-cross-validation/
[3] A. A. Enders, N. M. North, C. M. Fensore, J. Velez-Alvarez, and H. C. Allen, "Functional group identification for FTIR spectra using image-based machine learning models," Analytical Chemistry, vol. 93, no. 28, pp. 9711–9718, 2021. https://doi.org/10.1021/acs.analchem.1c00867
[4] E. S. Thrall, S. E. Lee, J. Schrier, and Y. Zhao, "Machine learning for functional group identification in vibrational spectroscopy: A pedagogical lab for undergraduate chemistry students," Journal of Chemical Education, vol. 98, no. 10, pp. 3269–3276, 2021. https://doi.org/10.1021/acs.jchemed.1c00693