14th U. S. National Combustion Meeting
Organized by the Eastern States Section of the Combustion Institute
March 16–19, 2025
Boston, Massachusetts

Deep Learning Algorithm for Identification of Combustion Relevant Species from FTIR Spectra

Shrey Rajesh Waghmare[1] and Nicole J. Labbe[1,]*

[1]Paul M. Rady Department of Mechanical Engineering, University of Colorado, 1111 Engineering Drive, Boulder, Colorado 80309, USA
[*]Corresponding Author Email: nicole.labbe@colorado.edu

Abstract: Fourier Transform Infrared (FTIR) Spectroscopy is a widely used analytical technique for identifying the composition of materials by analyzing their spectra and is a diagnostic used in combustion and fire experiments as an accurate chemical detection method for speciation. These spectra provide crucial information about molecular structures by revealing the presence of specific functional groups within a molecule. Traditional analysis of FTIR spectra involves interpreting absorption peaks corresponding to various bond vibrations, which are characteristic of specific functional groups. This process requires expert knowledge and a meticulous examination of the spectra to accurately associate peaks with functional groups. However, such traditional methods are not only time-consuming but also prone to inaccuracies, particularly when dealing with complex molecules where overlapping peaks complicate peak identification. Recent advances in deep learning have shown promise in automating the interpretation of FTIR spectra. Several studies have demonstrated the potential of machine learning models to identify functional groups with high accuracy. However, these models have typically been limited in scope, focusing only on a narrow range of the most common functional groups found in organic compounds. This limitation restricts their broader application, especially in analyzing more diverse or less studied chemical structures. In this work, we introduce a new deep learning framework designed to achieve accurate identification of functional groups in a wide variety of compounds, including those with complex or rare substructures. Our approach employs a combination of an autoencoder for feature extraction, a feature selection process to isolate key spectral features, and an optimized neural network for precise classification. These components work together to create a system capable of processing diverse FTIR spectra. To train and validate our model, we compiled a large dataset of infrared spectra by web scraping publicly available databases, including the National Institute of Standards and Technology (NIST) and Spectral Database (SDBS) repositories. These datasets encompass a broad range of functional groups, offering a diverse foundation for model development. The training process involved over 50,000 spectra, allowing the model to learn subtle patterns and relationships within the data. A key feature of our trained model is its ability to identify spectral patterns that human chemists typically use to recognize functional groups. This capability allows the model to achieve high accuracy over a wide range of functional groups, making it valuable tool for quick identification in combustion analysis.

Keywords: FTIR, Diagnostics, Machine Learning, Automated Chemistry

1. **Introduction:**

Functional groups in organic chemistry refer to specific arrangements of atoms that impart distinctive chemical and physical properties to molecules. These groups are the active centers in molecules, influencing their reactivity and behavior in chemical interactions and transformations [1]. Understanding functional groups is fundamental for predicting the behavior of complex

molecules. In combustion, they can help us to predict combustion properties and optimize the design of combustion systems.

The measurement of Fourier transformed infrared (FTIR) spectroscopy is one of the common techniques to identify functional groups. FTIR works on the principle of interaction of light with matter (molecules). The interactions are based on the absorption of infrared radiation which induces vibrations inside molecules in form of stretching, bending and torsional motions [2,3]. By analyzing absorption energy band position, bandwidth, and absorption coefficient one can deduce valuable information about molecule's structure and functional groups. The traditional analysis of IR spectra requires identification of intensity, position, area and width of peaks to describe molecular concentrations [2,3]. Interpreting these spectra involves manual identification of these features and correlating them with molecular substructures. However, this approach is labor-intensive and prone to inaccuracies, especially for complex molecules where overlapping peaks hinders the interpretation [4,5]. As a result, traditional methods often require significant expertise and time, limiting their practical application in large scale or complex systems.

To address these limitations, computational methods have been playing an increasing role in analyzing IR spectra, bringing advancement in machine learning. These methods are similar to the ones applied by humans mimicking pattern matching to map spectra to substructures helping in accurately predicting intricate spectra making it easier to identify substructures in molecule.

Several studies have contributed to advancement in this field. Enders et al. published first use of image-based convolutional neural networks (CNN) for functional group prediction [6]. They acquired spectral data for 8728 gas phase organic molecules from NIST database and converted this data into spectral images. Using this dataset, models were successfully trained for 15 of the most common functional group. This method demonstrated the use of functional group prediction by efficiently learning spectral patterns and facilitating the use of complex molecules for future. Fine et al. developed a hybrid machine learning model integrating an autoencoder and a three-layer neural network to predict 13 functional groups using IR and mass spectra data from 7393 compounds [7]. Their study introduced new metrics, namely molecular F1 score and molecular perfection rate, to assess model performance. Results indicated that FTIR data was consistent with functional group prediction, although MS data improved accuracy for some functional groups. Von Hoang Minh Doan et al. proposed a deep learning model based on transformer architecture, inspired by the Vision Transformer (ViT) framework [8]. This model was trained on 8677 FTIR spectra from the NIST database and good performance in predicting 17 functional groups. The use of attention mechanism in the transformer model allowed for more detailed interpretation of complex spectral data, improving prediction accuracy. Similarly, Jung et al. employed a sone dimensional CNN on large dataset of 30,000 IR spectra. Their model could identify over 37 functional groups, showcasing the potential of deep learning for large scale and automated spectral analysis [9].

These computation approaches represent a notable advancement in IR spectral analysis. However, current literature has been focusing primarily on the most common functional groups found in organic compounds. This limitation restricts their broader applicability on combustion relevant functional groups. Recognizing this limitation, we propose a new deep learning framework that addresses these gaps by accurately identifying functional groups across a wide variety of compounds, including those containing rare substructures. Our framework combines an

autoencoder for extracting essential features and an optimized fully connected neural network for precise classification. To train and validate this model, we curated a large and diverse dataset by leveraging publicly available resources such as NIST and SDBS. These datasets encompass a wide array of functional groups, providing comprehensive foundation for model development.

The model's performance was evaluated and compared with the existing literature to assess its capability. The results suggest that the framework offers a broader applicability in functional group identification while maintaining competitive performance. This study highlights the potential for further refinement and exploration of such methods for quick identification in combustion analysis.

2. **Methods:**

2.1. Collection of Data:

Data was collected from the National Institute of Standards and Technology (NIST) Chemistry WebBook and the Spectral Database for Organic Compounds (SDBS) through web scraping using Selenium [10,11]. The NIST database provided gas-phase spectra, while the SDBS database included spectra acquired in four distinct condensed-phase sample environments: KBr disc, Nujol mull, single-component liquid film, and CCl_4 solution. For the SDBS data, spectra were downloaded as image files and subsequently digitized using pixel mapping to extract transmittance values. The transmittance spectra from SDBS had a resolution of 4 cm^{-1} in the low wavenumber region (400–2000 cm^{-1}) and 8 cm^{-1} in the high wavenumber region (2000–4000 cm^{-1}). Absorbance spectra from NIST were provided in JCAMP-DX format with a resolution of 2 cm^{-1}. These JCAMP-DX files were converted into CSV format for further analysis.

2.2. Preprocessing:

To ensure uniformity across datasets, the transmittance spectra from SDBS were first converted to absorbance spectra. Absorbance data from both NIST and SDBS were then interpolated to obtain a consistent dataset of 3600 absorbance values for each molecule, spanning the frequency range of 400 cm^{-1} to 4000 cm^{-1}. All spectra were normalized to the same scale to maintain consistency across samples. A comprehensive spreadsheet was generated containing chemical identification numbers (CAS/SDBS no) along with the normalized absorbance values for each molecule, ensuring an organized format for subsequent analyses.

2.3. Functional groups assignment:

Functional group assignments were carried out through substructure matching using RDKit. For each sample, the InChI string of the molecule was obtained by resolving its name via the PubChem API. The molecule's InChI string was then matched against the SMARTS strings of functional groups using RDKit's substructure-matching functionality. Functional groups identified within a molecule were labeled as "1" (present), while absent functional groups were labeled as "0." The results, including the chemical identification number (CAS/SDBS no) and the presence or absence of each functional group, were systematically compiled into a spreadsheet. This structured labeling facilitated efficient downstream analyses of functional group distributions across the database.

3. **Results & Discussion**:

3.1 Unoptimized Neural Network (NN) outperforms traditional classifiers:

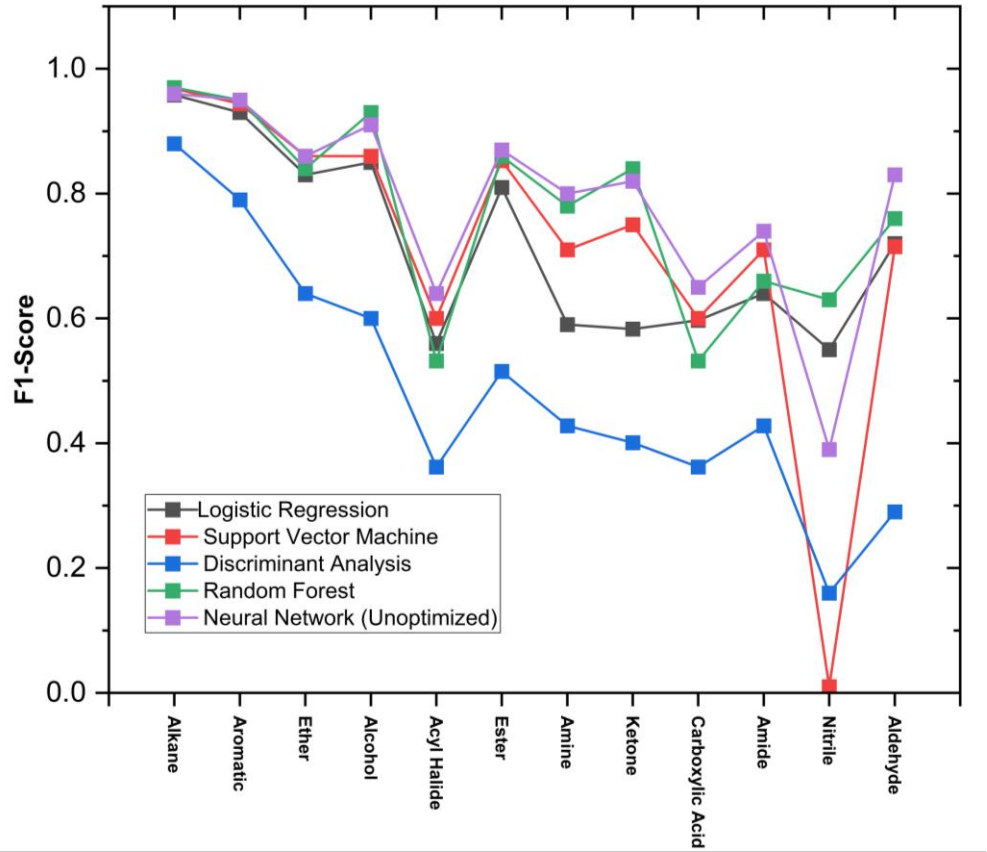

Figure 1: Comparison of traditional classifiers with unoptimized NN

An initial computational experiment was conducted to evaluate the performance of different machine learning models for predicting functional groups. This experiment compared an unoptimized 2-layer neural network to four traditional classifiers across 12 common functional groups. The primary objective was to determine whether employing a NN was justified over traditional approaches. The result demonstrated that the NN outperformed all the traditional classifiers. The average F1-score for the NN was 0.79, notable higher than average F1-score of 0.76 achieved by the best performing traditional classifier.

3.2 Neural Network Architecture:

3.2.1 Feature selection

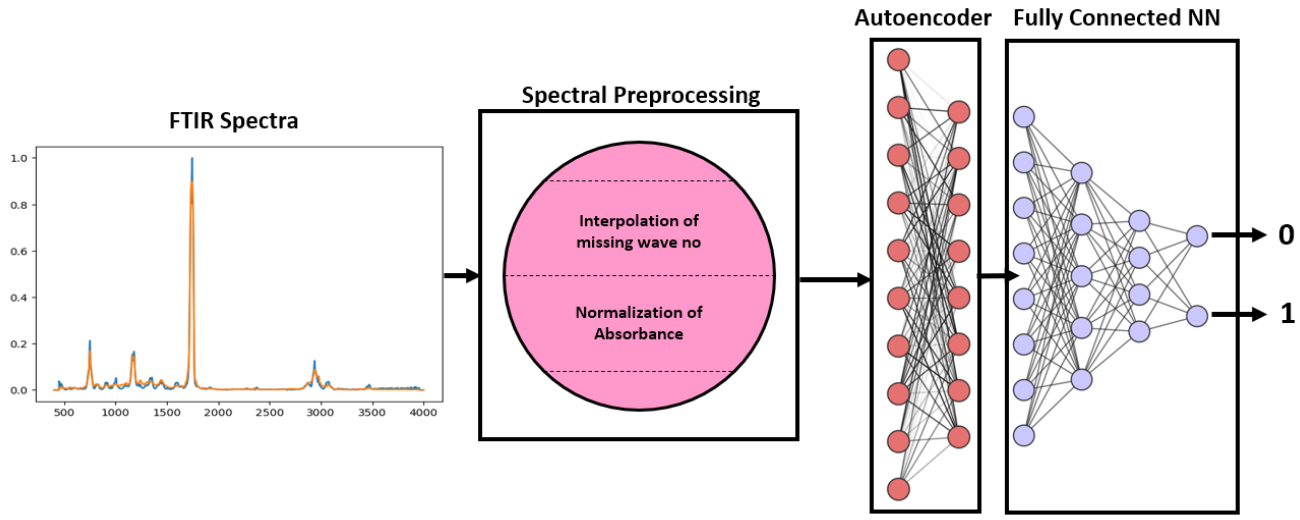

Figure 2: Overview of Deep Learning algorithm for classification of functional groups using FTIR Spectra

After confirming that the neural networks were the most suitable choice for predicting functional groups, a feature selection was undertaken to enhance model's predictive capability. Feature selection process was performed to eliminate redundant information from the spectral matrix, ensuring that the classification model focused solely on the relevant and meaningful features. An autoencoder was employed for this purpose, with bottleneck layer acting as dimensional reduction layer. After evaluating several configurations, an autoencoder with bottleneck vector length of 1050 was finalized, as it demonstrated highest information retention. The compressed vector representation was then used as a new training dataset.

3.2.2 Handling Class Imbalance

In binary classification problems with samples from two groups, class imbalance occurs when one class contains significantly fewer samples than the other class. In the current study, it was observed that many functional groups had imbalanced dataset due to limited availability of data. Learning from such imbalanced datasets is challenging as classifiers often overfit the majority class, leading to frequent misclassification of minority instances. As evident from Fig. 1, all classifiers tend to over classify the majority class due to the imbalance, resulting in poor representation of the minority group. Moreover, traditional evaluation metrics like accuracy can mislead readers by presenting inflated scores that fail to reflect the actual performance of the model on the minority class. To address this, the F1-score was used as the primary evaluation metric. To further handle class imbalance, algorithmic level modifications were implemented by redefining the loss function. A focal loss function was employed, which assigned higher weights to the minority class, thereby penalizing its misclassification more heavily. This approach ensured the model paid

greater attention to the minority instances during training, improving the overall predictions for underrepresented functional groups.

$$\text{Focal Loss} = -\sum_{i=1}^{K} y_i \log(p_i)(1-p_i)^{\gamma}\alpha_i + (1-y_i)\log(1-p_i)(p_i)^{\gamma}(1-\alpha_i)$$

where α is the weighing factor and γ is focussing parameter. Both of these parameters were computed through several iterations.

3.2.3 Model Training:

The model was built using Keras Python package. All the hidden layers were activated using rectified linear unit (ReLU), and sigmoid function is used to activate output layer. Focal loss was used as loss function during the training. The network was trained with batch size of 40 and train test split of 80-20 was used. An adjustable learning rate combined with early stopping was implemented to prevent the model from overfitting. Hyperparameter optimization was performed using genetic algorithm (GA), as detailed in the next section. A total of 12 models were trained, each corresponding to a specific functional group. The complete model architecture is illustrated in Fig 2. After training, the models were evaluated on the testing set and their performance was compared with results reported in existing literature.

3.2.4 Optimizing hyperparameters:

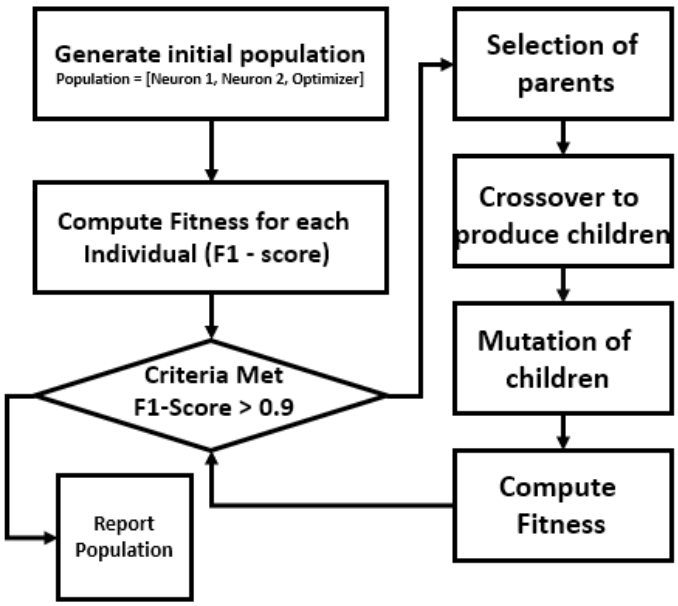

Figure 3: Flow Chart of genetic algorithm (GA)

A genetic algorithm was utilized to perform an extensive hyperparameter search for optimizing the fully connected neural network [12]. The search focused on three key parameters: the number of neurons in the first and second hidden layer, and the optimizer. To ensure diversity, the initial population consisted of 500 individuals, with neuron counts for the hidden layer randomly sampled

between 100 to 900 and optimizer randomly chosen between adam, stochastic gradient descent, and RMSprop. Each individual in the population was uniquely defined by the combination of these three parameters. The algorithm was operated using F1 score as the fitness function. The algorithm was specifically applied to amide functional group due to low imbalance ratio.

The method utilized a one-point crossover method for combining parameters, mutations were restricted to second hidden layer to increase diversity. Through successive generation, the final set of parameters was determined to be 428 neurons in the first hidden layer, 101 neurons in the second hidden layer and the adam optimizer. Using the optimized parameters, 12 classifiers were trained, each targeting a specific functional group. The F1-scores obtained for these functional groups were subsequently compared with values reported in existing literature to evaluate performance.

3.3 Model comparison with existing literature:

Functional Group	**Fine et al [7]**	**Jung et al [9]**	**Wang et al [13]**	**Present**
Alkane	0.93	0.97	0.964	0.94
Alcohols	0.91	0.95	0.95	0.9
Amines	0.85	0.93	0.89	0.81
Aromatics	0.98	0.98	0.98	0.96
Esters	0.91	0.94	0.94	0.85
Ketones	0.86	0.88	0.87	0.81
Aldehyde	0.87	0.81	0.96	0.78
Carboxylic Acid	0.91	0.93	0.93	0.82
Acyl Halides	0.73	0.84	0.82	0.73
Amides	0.55	0.79	0.64	0.84
Ethers	0.91	0.94	0.92	0.89
Nitriles	0.55	0.97	0.8	0.45

Table 1: Comparison of F1-scores with literature

In Table 1, the model demonstrated consistent performance across all functional groups, it performed averagely in some cases. This variation may stem from the fact that the hyperparameters were optimized for one functional group (amide), which may not provide optimal values for other functional groups. Despite this, the average F1 score across all functional groups was higher than 0.8, highlighting the model robustness on imbalanced datasets. This indicates that the model has

potential to handle skewed data, although slight adjustments in parameters could further enhance its performance.

The results also show that the model performs particularly well on balanced datasets, as seen in functional groups such as alkanes, alcohols, where F1 scores were comparable to those reported in literature. These findings suggest that the model architecture and training process are well suited for evenly distributed data. However, for highly imbalanced groups, there is room for improvement through better hyperparameter search.

Also, it is important to note that the dataset used in this study was not comprehensive, and the complete dataset will be explored in future work. This limitation has also contributed to the variation in performance, as a more extensive and diverse data set may provide better training for the model.

3.3.1 Extended list of Functional groups:

Functional Group	**Jung et al [9]**	**Present**
Anhydride	0.74	0.85
Imine	0.74	0.82
Thial	0.826	0.73
Azo Compound	0.859	0.76
Thiol	0.86	0.82
Phenol	0.9	0.93
Imide	0.878	0.81

Table 2: Comparison of F1 Scores with extended list of functional groups

The model was further evaluated on an extended set of functional groups, characterized by limited data availability for most functional groups. These functional groups have only been previously reported only by Jung et al [9]. All the functional groups exhibit significant class imbalance, with phenol group observed to have balanced dataset. The model demonstrated competitive performance achieving average F1 score of 0.82.

The results achieved with the current set of hyperparameters hold average performance, highlighting capability of model in handling rare functional groups. However, the performance indicates potential for further improvement.

4. **Conclusion**:

We developed a deep learning-based approach for FTIR spectral interpretation, featuring a unique architecture that combines an autoencoder for feature extraction and a fully connected neural

network for classification. The model demonstrated promising F1-scores across both commonly studied functional groups and an extended list, indicating its potential for broad applicability. Although the model was trained on a subset of the dataset with over 50,000 molecules, further improvements are anticipated with the use of a more comprehensive dataset. Future work will focus on refining the model, optimizing it with the complete dataset, and addressing any limitations to enhance its performance. These refinements are expected to expand the model's applicability and accuracy in functional group identification.

5. **References:**

[1] Visser, T. *Infrared Spectroscopy in Environmental Analysis*; John Wiley & Sons, Ltd, 2006; doi: https://doi.org/10.1002/9780470027318.a0832.

[2] Stuart, B. (2004). Infrared Spectroscopy: Fundamentals and Applications. John Wiley & Sons

[3] Bruice, P. Y. Essential Organic Chemistry, 3rd ed.; Pearson: Upper Saddle Reiver, New Jersey, 2003

[4] Nalla, R, Pinge, R, Narwaria, M., Chaudhury, B. Priority Based Functional Group Identification of Organic Molecules Using Machine Learning. Proceedings of the ACM India Joint International Conference on Data Science and Management of Data. New York, NY, USA, 2018; p 201–209, doi: https://doi.org/10.1145/3152494.3152522

[5] Wang, Z.; Feng, X.; Liu, J.; Lu, M.; Li, M. Functional groups prediction from infrared spectra based on computer-assist approaches. *Microchemical Journal* **2020**, *159*, 105395, doi: https://doi.org/10.1016/j.microc.2020.105395

[6] Enders AA, North NM, Fensore CM, Velez-Alvarez J, Allen HC. Functional Group Identification for FTIR Spectra Using Image-Based Machine Learning Models. Anal Chem. 2021 Jul 20;93(28):9711-9718. doi: https://doi.org/10.1021/acs.analchem.1c00867

[7] Fine JA, Rajasekar AA, Jethava KP, Chopra G. Spectral deep learning for prediction and prospective validation of functional groups. Chem Sci. 2020 Mar 13;11(18):4618-4630. doi: https://doi.org/10.1039/c9sc06240h

[8] Vu Hoang Minh Doan, Cao Duong Ly, Sudip Mondal, Thi Thuy Truong, Tan Dung Nguyen, Jaeyeop Choi, Byeongil Lee, and Junghwan Oh, Fcg-Former: Identification of Functional Groups in FTIR Spectra Using Enhanced Transformer-Based Model**,** Analytical Chemistry 2024 *96* (30), 12358-12369**,** doi: https://doi.org/10.1021/acs.analchem.4c01622

[9] Jung G, Jung SG, Cole JM. Automatic materials characterization from infrared spectra using convolutional neural networks. Chem Sci. 2023 Feb 23;14(13):3600-3609. doi: https://doi.org/10.1039/d2sc05892h

[10] Linstrom, P. J., & Mallard, W. G. (Eds.). (2023). NIST Chemistry WebBook. National Institute of Standards and Technology. Retrieved from https://webbook.nist.gov/

[11] National Institute of Advanced Industrial Science and Technology (AIST). (2023). Spectral Database for Organic Compounds (SDBS). Retrieved from https://sdbs.db.aist.go.jp

[12] Goldberg, D. E. (1989). Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley.

[13] Wang T, Tan Y, Chen YZ, Tan C. Infrared Spectral Analysis for Prediction of Functional Groups Based on Feature-Aggregated Deep Learning. J Chem Inf Model. 2023 Aug 14;63(15):4615-4622. doi: https://doi.org/10.1021/acs.jcim.3c00749