Priority Based Functional Group Identification of Organic Molecules using Machine Learning

Rushikesh Nalla
Group in Computational Science and HPC
DA-IICT, Gandhinagar, India

Rajdeep Pinge
Group in Computational Science and HPC
DA-IICT, Gandhinagar, India

Manish Narwaria
Group in Computational Science and HPC
DA-IICT, Gandhinagar, India
manish_narwaria@daiict.ac.in

Bhaskhar Chaudhury
Group in Computational Science and HPC
DA-IICT, Gandhinagar, India
bhaskar_chaudhury@daiict.ac.in

ABSTRACT

Functional groups in organic compounds determine the properties of the compounds/molecules. When multiple functional groups are present, the dominant functional group determines majority of the properties of the compound. Hence priority based identification of functional groups is an important problem in chemistry. Fourier-transform Infrared spectroscopy (FTIR) is a commonly used spectroscopic method for identifying the presence or absence of functional groups within a compound, and the current approach for this task mainly relies on visual inspection and analysis of the FTIR spectral data. However, such visual identification process by humans is error prone, especially when patterns in the FTIR spectrum overlap, resulting in loss of uniqueness of features which help in identification of different functional groups in the unknown sample. Therefore, the main goal of this paper is to develop a machine-learning based classification system which can perform priority based functional group identification of organic molecules. To the best of our knowledge, this is the first effort to address this problem using machine learning (ML), and a unique aspect of our study is the incorporation of domain specific information into the process of classification by employing a set of priority rules generated from expert knowledge. We have carried out extensive study on real IR spectral data, first using a rule based approach and then using ML in an effort to improve the classification accuracy. Our analysis indicates that the basic rule based method is reasonably effective in predicting the presence (or absence) of functional groups. However, such approach is practically not accurate enough for the more challenging problem of priority based identification, and ML based classification offers much higher identification accuracies in this case. The primary reason is that ML algorithm can adaptively exploit data patterns to classify the functional group unlike the rule-based approach which uses a fixed set of rules for the said purpose. Finally, we have also carried out extensive statistical analysis of the results by using confidence intervals and permutation tests, in an effort to gain more descriptive information about the learning process, and not simply treat it as a black box.

KEYWORDS

Organic Compounds, Chemical Bonds, Functional Groups, Functional Group Priority, Fourier Transform Infrared (FTIR) Spectroscopy, Pattern Identification, Machine Learning (ML)

1 INTRODUCTION

In Organic Chemistry, *Functional Groups* are specific/classified groups of bound atoms which appear together within molecules, and determine the chemical and physical properties of the compounds/molecules. Regardless of what molecule contains it, does not matter how large or small it is, same functional groups will go through similar chemical reactions and a molecule containing a particular functional group is expected to exhibit reactions characteristic of that functional group. In other words, functional groups are centers of chemical reactivity of a molecule and it is necessary to identify the functional groups within a molecule when naming it [1]. Accurate identification of functional groups has significant applications in several fields such as biochemistry, drug discovery, medicinal chemistry, toxicity assessment, pharmaceuticals, molecular biology and chemical nomenclature. There are many real-life situations where knowledge of chemical properties of a mixture of unknown organic molecules is important in order to understand its behaviour. In such cases, it is even more important to identify the functional groups present in the mixture since they represent the behaviour and properties of the mixture.

Fourier-transform Infrared (FTIR) spectroscopy is an important, commonly used spectroscopic method for identifying the presence or absence of functional groups within a compound and thereby helps in the structural identification of unknown molecules [2, 3]. It is based on the interaction of Infra-Red light with molecules present in a sample and the absorption of particular frequencies of the IR radiation. Each chemical bond within a molecule has a certain characteristic vibrational frequency and corresponding wave-number (multiplicative inverse of frequency). When IR light is incident on a sample, the frequencies of absorption correlate to the vibration of specific chemical bonds present in the molecule. Examination of the transmitted light gives a spectrum of transmittance with respect to incident frequency of IR light thereby revealing how much energy is absorbed at each frequency (or wavelength). Generally a conventional infrared instrument records spectra from an upper limit of around 4000 cm^{-1} down to 400 cm^{-1}. Characteristic absorption (or transmittance) patterns in the IR spectrum of a sample molecule helps in deducing the presence or absence of a specific functional group in the sample. The spectrum is rich in information with complex patterns and consists specific features. A spectroscopist (human) interprets the data by using the well established relationships between the molecular structure and patterns in the

obtained spectra. In addition to these rules, visual aspects of spectroscopy which includes recognizing characteristic patterns and its interpretation relative to the structure requires experience, and plays an important role in accurate interpretation. The IR spectrum of an unknown sample can be compared with previously known reference spectra patterns leading to identification of unknown functional groups in the sample molecule and this forms the basis of machine learning based spectral searching.

There are millions of organic compounds and the most important reason for classifying compounds by their functional groups is that it classifies their chemical behaviour. However, if more than one functional group is present in the compound, then the properties of the compound are determined by the most dominant functional group present in the compound [4]. The dominant functional group is determined by priority order of functional groups and plays an important role in determining the properties of complex organic molecules. Therefore the interpretation of FTIR spectra is not only simply assigning group frequencies to molecules but requires classification based on functional group priority. Function f denotes the mapping between features and corresponding functional groups.
f : Features $\rightarrow$ Functional Group of Compound.

Usually, function f is identified by humans through certain rules as well as by experience. The common features include, transmittance level or absorption level of IR light, width of peaks appearing in the spectrum, number of peaks in the given range, etc. We believe, machines can identify a better f through supervised machine learning methods which will help in more accurate identification of priority based functional groups. Most of the works in existing literature, regarding application of machine learning for interpretation of IR spectra, focuses on determining the presence or absence of a functional group in a molecule.

To the best of our knowledge, this is the first report describing the priority based functional group identification of organic molecules using machine learning on FTIR data. We first look at the rule based method in which we try to replicate how a human being would classify the spectrum and state its limitations of identifying the dominant functional group. We later see if a machine can identify better structures among the human identifiable features in the intermediate approach. Finally, we use the entire spectrum and let the machine extract the best hidden features by using different machine learning algorithms. In the end we compare the results obtained in all the three approaches and validate our models by performing statistical tests.

2 RELATED WORK

During early 1990s, a lot of research work involving basic ANN (Artificial Neural Networks) methods for identification of chemical structural features from FT-IR spectrum of organic molecules have been reported. A very detailed study of identifying functional groups from IR spectrum using ANN was published by Robb and Munk [5] in 1990. They used a one-layer architecture of ANN to classify a large number of molecules. They represented spectrum range of 4000 cm^{-1} - 400 cm^{-1} using 640 uniformly spaced data points and emphasized on identifying the presence or absence of about 128 structural features. So essentially, there were 128 binary classifiers in the working model. But the main drawback of this work was that it used a linear activation function to learn and classify molecules into different functional groups which was criticized in the book by Minsky and Pappert [6] while Fessenden and Gyorgyi [7] termed it as "oversimplification" of the problem. Robb and Munk along with Madison published another paper [8] which used one hidden layer in the architecture on the same dataset. In this case, the number of features were reduced to 36 and sigmoid activation function was used. A detailed analysis of the method was done by varying various parameters. The accuracy achieved on trained data set was nearly 80% while it was just over 60% on the testing data set. Fessenden and Gyorgyi [7] also used single hidden layer architecture but they did the experiment only on 6 major functional groups and also checked for the presence of 3 more bonds. So, in total 9 element vector predictions were done. This was insufficient since many of the important functional groups as well as their priority order was not considered during identification.

In the next few years a lot of studies were conducted by varying the number of input points, number of structural features, number of classes, output threshold, etc. [8–14]. Meyer and Hobert [15] further used Principle Component Analysis (PCA) to reduce the spectral data required for correct prediction. Visser, Luinge and van der Maas [16] focused on identification of individual bonds which have overlapping IR spectrum range. However, all these works have two major drawbacks. Firstly, they used only the standard Multilayer Perceptron (MLP) ANN with one hidden layer, and secondly they predicted only the presence/absence of each functional group and/or structural feature. Therefore the above works lacked the vital information about the priority order of functional groups, which helps in determining the name and chemical properties of the compound if it contains multiple functional groups.

In 2000, Tchistiakov and Ruckebusch [17] used different preprocessing techniques like wavelet and Fourier-transform coefficients for reduction of spectra in combination with different Artificial Neural Networks (ANNs) for non-linear hierarchical modelling to comply with scarcity of samples. Brown and Lo [18] did extensive analysis on classification using Radial Basis Function (RBF) classification technique of ANN. But these two also neglected the priority order of functional groups and focused only on presence-absence of each feature individually.

In 2001, Tanabe and group [19] used 10000 samples with 15 functional groups and more than 100 structural features. They also checked only for presence or absence of all these structural features although the scale of experiment was very large. They were able to achieve an average accuracy of 80% for classification of main functional groups and around 70% when halogens and sulphur groups were considered. They believe that it is the limitation of infrared method to identify structures from spectra. The spectral database system (SDBS) [20] constructed by the above group is an open-source website containing over 50000 infrared spectra. We have used IR spectra from this open-source database (images of IR spectra) for our study.

From the above literature review it is clear that most of the work on this topic considered only the presence/absence of individual functional groups, bonds and other structural features in a FTIR spectra of a molecule. Functional group priority order was not considered by any of these works. Furthermore, the works were done using only the basic ANN methods. Our aim is to consider

and exploit the priority order of functional groups while assigning a single label to a molecule, since the properties of the molecule are determined by the dominant functional group.

3 PROBLEM DESCRIPTION AND IR SPECTRAL DATA

3.1 Problem Description

Let $\mathbf{X}$ be an $n \times m$ data matrix (the i-th row and j-th column vectors can be denoted as $\mathbf{X}_i$ and $\mathbf{X}^j$, respectively). As we are using FTIR spectrum data, the elements of $\mathbf{X}$ typically represent the value of transmittance (feature) that is related to a given functional group. Further, a discrete value $y_i \rightarrow \mathbf{C}$ where $\mathbf{C} = [c_1, c_2, ...c_B]$ represents the discrete class label vector (in our case $B = 14$), and is associated with each data point $\mathbf{X}_i$.

The labelled data set will be $\mathbf{D} = \{(\mathbf{X}_i, \mathbf{C_i})\}_{\mathbf{i=1}}^{\mathbf{n}}$. Then the problem can be formulated as one of learning the function f which maps the data values into the corresponding class label i.e. $f : X \rightarrow \mathbf{C}$. However, unlike the traditional supervised classification problem where the class label $\mathbf{C}$ is well-defined (i.e. the given sample belongs completely to one of the classes), the classes in our case tend to be fuzzy. Specifically, a compound will almost always contain more than one functional group. Therefore, a seemingly logical approach from machine learning view point would be the use of fuzzy classification approach. However, from the view point of the application (i.e. functional group identification) and the spectroscopic data, the chemical properties of the compound are determined, to a large extent, by the dominant group, and practically less affected by the presence of other groups. In light of such context and data information, a more effective approach would be priority based classification. Such priority based approach entails exploiting the hierarchy of functional groups i.e. the more dominant group will decide the chemical properties of the given test sample. Such context-specific information is available from domain experts, and is shown in Table 1. The ordering in this table shows the priority-wise arrangement of importance of functional groups. For instance, if the IR data indicates the presence of carboxylic acid, acid halide and amide, then the corresponding sample will inherit properties of carboxylic acid (which is first in the order of priority), and needs to be classified as carboxylic acid. Therefore, the mapping $f : X \rightarrow \mathbf{C}$ needs to consider this information.

As already stated, the current approach to estimate f largely depends on visual inspection by experts. Given the peculiarities and fine patterns in IR spectroscopic data, such human-based classification is not only error prone (we provide specific motivating examples in the next sub-section) but also time-consuming. Therefore, in this paper, we consider the use of machine learning to derive the mapping f, and evaluate its effectiveness against a rule-based approach that humans typically use to classify functional groups from FT-IR data.

3.2 Peculiarities in IR Spectral Data

As shown in Figure 1, the spectrum of a molecule is a graph of Transmittance (%) vs Wave-number (cm^{-1}). A basic characterization (absence/presence of a functional group) of an unknown sample is possible by investigating the IR spectrum using first principles and well established rules for interpreting IR spectra patterns. However,

Table 1: PRIORITY ORDER OF FUNCTIONAL GROUPS

	Functional Group	**Prefix**	**Suffix**
1	Carboxylic Acid	carboxy-	-carboxylic acid -oic acid
2	Ester	(R)-oxycarbonyl	-oate
3	Acid Halide	halocarbonyl-	-oyl halide
4	Amide	carbonyl-	-carboxamide -amide
5	Nitrile	cyano-	-nitrile
6	Aldehyde	formyl-	-al -carbaldehde
7	Ketone	oxo-	-one
8	Alcohol	hydroxy-	-ol
9	Thiol	mercapto-	-thiol
10	Amine	amino-	-amine
11	Arene (cyclic arrays of C=C)	-	benzene
12	Alkene	alkenyl	-ene
13	Alkyne	alkynyl	-yne
14	Alkane	alkyl	-ane
15	Ether	alkoxy	-ane
16	Alkyl Halide	halo-	-ane
17	Nitro	nitro-	-ane

it is generally difficult to identify the dominant functional group in the spectrum because of the presence of multiple functional groups. The presence of various chemical bonds introduces specific patterns in the IR spectrum. These patterns are unique to each functional group but may not be distinguishable upon visual inspection. For example, the presence of chemical bonds causes one or more peaks ('inverted peaks' which are here-on referred to as simply 'peaks') to appear in the spectra and these peaks are generally Gaussian shaped. Various chemical bonds have different wave-number ranges and some even have overlapping ranges. Some of the peaks may superimpose on each other if the ranges are overlapping. So, it is difficult to identify which bonds are actually present from these mixture of Gaussian shaped peaks. Also, chemical effects or interactions between atoms can cause a shift in the peaks. This makes it difficult to visually identify all the unique patterns in the spectrum. Therefore to substitute for the current, heavily used and error prone visual inspection process, machine learning based methods are required which will aid in better identification of dominant functional group present in a molecule/compound.

Studying the visual identification process further, a human would be able to easily identify the functional group in some cases. For example, the graphs in Figure 2 shows how a human would observe the spectrum to extract essential features which would help in identification of dominant functional group. In the spectrum on the left, two functional groups, namely Carboxylic Acid (made of O-H,

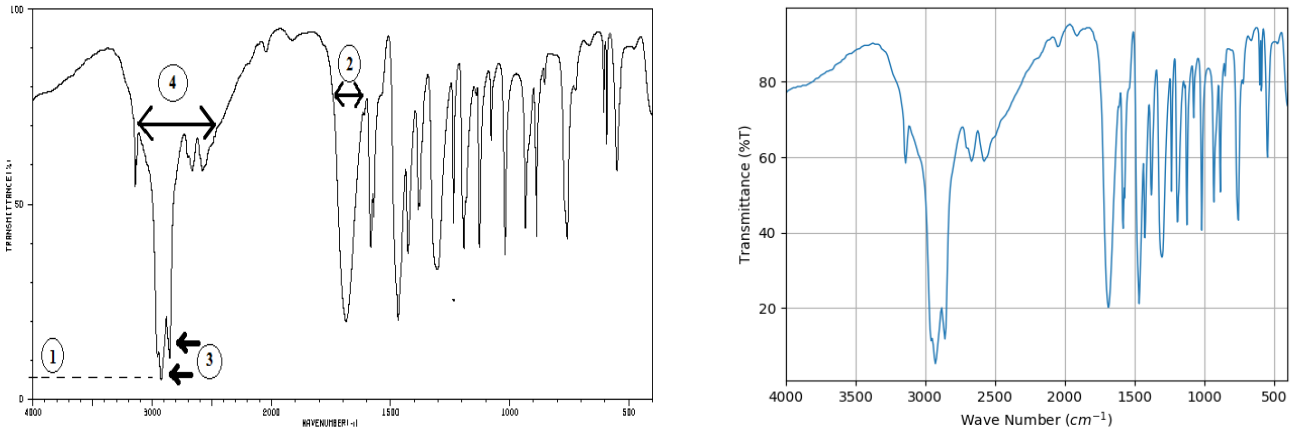

Figure 1: Examples of FTIR spectrum. The spectrum on the left is the original spectrum taken from the SDBS database [20]. It also shows the rule based features that have been considered in the rule based approach. (1) Transmittance Level, (2) Width of Peak, (3) Number of Peaks in the given Range and (4) Sum of Widths of All Peaks in the given Range. The spectrum on the right is the one regenerated after extracting the data in quantified form from the original image on the left.

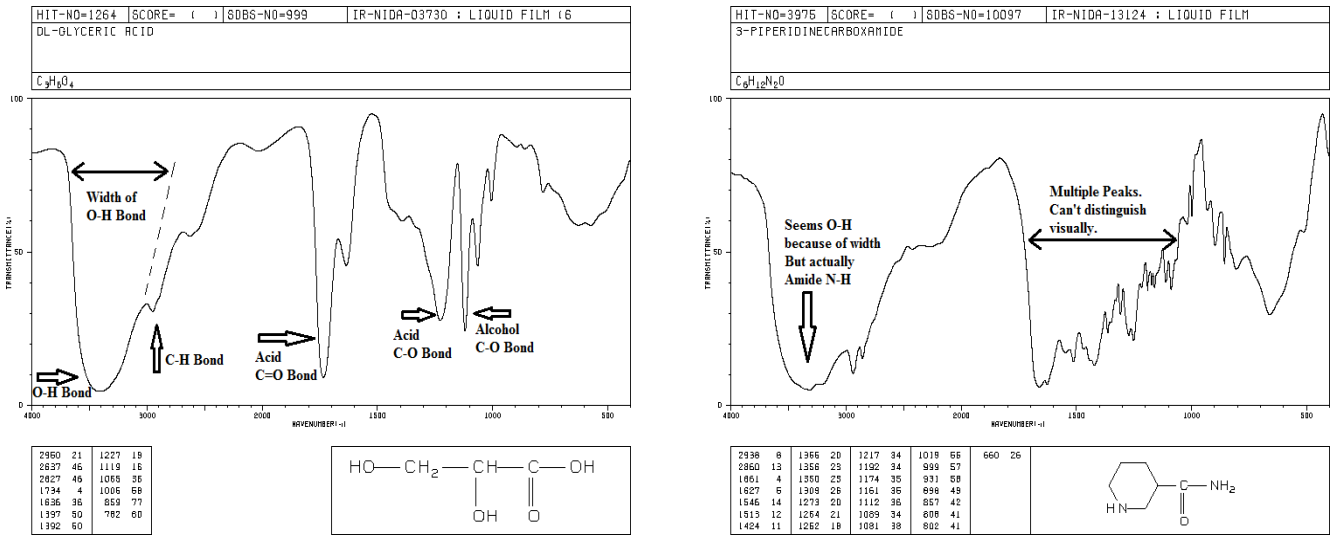

Figure 2: Steps in visual identification of functional groups. The spectrum on the left shows that visual identification of functional groups is possible due to simplicity of spectrum. The spectrum on the right shows the peculiarities and hence problems in identifications due to its complex nature. The original images are taken from SDBS database [20]

C=O, C-O and C-H bonds) and Alcohol (made of O-H, C-O and C-H) are present. But the Carboxylic Acid group is higher in the priority table given in Table 1. Hence according to the rules any average human being will be able to identify this as a Carboxylic Acid. On the contrary, for spectra on the right in Figure 2, it would be very difficult for an amateur, inexperienced human being to identify the compound because the spectrum is complex and the standard rules are unable to distinguish it.

In order to understand the problems in the widely used approach of visual identification, we first investigate the rule-based approach used by humans for classification. Secondly, we study an intermediate method where features used by humans are given to the machine learner to understand the patterns and classify the compounds. Results of both these methods are then compared. Finally, machine learning is used on whole data so that machine itself can identify its own features and patterns from the data and build mapping to identify the functional groups. Furthermore, we compare the different machine learning techniques to check whether using different learners changes the functional group accuracy and we also check the validity of the obtained results by performing certain

statistical tests. On the whole, the goal is to check whether manual feature extraction and human judgment can be replaced by machine based feature extraction and machine judgment.

3.3 Data Collection

SDBS Database [20] is a large open-source database with spectra of over 50,000 compounds. However, data on the SDBS Database is in the form of images which cannot be used directly. Spectrum on the left in Figure 1 is an example of the image of FT-IR spectrum of a molecule.

Quantification of data from the images is an essential step before data processing. Since our aim is to predict a single dominant functional group present in the molecule, we do not check for the presence or absence of any of the sub-functional groups. We rather let the machine identify these sub-features on its own and accordingly predict the functional group of the molecule. Furthermore, we can work on a small data-set because the number of labels we focus on is comparatively less. Also, we can do with less number of points than suggested by Tanabe et al. [19] which is 3600 because the broad pattern of spectrum remains same even for lesser number of points.

Since we are using the same database as used by Tanabe et al., we have taken the help of their paper to finalize the list of functional groups to be considered for our experiment. Out of the 15 functional groups used by Tanabe et al. we have used 12. Two of the three remaining groups contain Phosphorous (P) and Sulfur (S) as part of the compounds. These are minor groups and S is generally a substitute for Oxygen(O) while P is a substitute for Nitrogen(N). Hence, we have not considered them. The wave-number range for halogen bonds is from 1400 cm^{-1} to 500 cm^{-1} which is part of the fingerprint region of spectrum. In this region, there are a lot of peaks irrespective of the presence of halogen bonds because of the characteristics of IR spectrum. Thus we might encounter a lot of false-positives if we include halogens which might reduce the overall accuracy. Hence we have not included the last remaining group, the halogens. Instead, we have added two more functional groups - pure alkanes (C-C and C-H bond) and alkynes (C triple bond C, C-C, and C-H bond). The reason for this addition is that many of these pure Hydrocarbons exist in nature and from Table 2 it is clear that they have distinct ranges where the peaks may exist and which can be identified uniquely. Among the groups mentioned in Table 1, we have taken into consideration the functional groups in white background whereas the groups in gray background have not been considered. So, in effect we have taken 14 functional groups for our problem. Except for "nitro" and "alkyne" groups, all other groups have 100 samples each. Due to unavailability of samples in the SDBS Database, we could get only 92 samples for 'nitro' group and 49 samples for 'alkyne' group.

The obtained data in image form is further processed and quantified. Data cleaning, proper x-y mapping, scaling and interpolation is performed to prepare data for feature extraction. Finally, after all the processing, we obtain spectrum on the right in Figure 1. We can observe that, visually, this graph is very close to the graph in the original data image on the left. This means that our processing accurately captures the behaviour and variations in the original data. The regenerated spectrum image contains 945 data points instead of 3600 or more covering entire range. This shows that there is no need for 1 point for each x-value. This is because this range is sufficiently wide to contain data points which cover all the variations in the spectrum. So instead of increasing the data by interpolation which may introduce some error, we consider less number of data points.

Table 2: Wave-number Ranges of Chemical Bonds

Bond	Wave-number
C - H	2850-2950 (Alkane) 3000-3100 (Alkene) 3290-3310 (Alkyne) 3000-3040 (Aromatic)
C = C	1620-1680 (Alkene) 1400-1620 (Aromatic)
C triple bond C	2100-2260
C = O	1690-1740 (Aldehyde) 1710-1780 (Acid) 1630-1690 (Amide) 1680-1750 (Ketone) 1735-1750 (Ester)
N - H	3300-3500 (Amine) 3100-3500 (Amide)
O - H	2500-3200 (Acid) 3200-3700 (Alcohol/Phenol)
C triple bond N	2220-2260 (Nitrile)
C - N	1025-1220 (alkyl, amine, amide) 1250-1360 (aryl)
C - O	1040-1210 (alcohols/phenols) 1210-1320 (acid)
Nitro N - O bonds	1515-1560 and 1345-1385

4 RULE BASED METHOD

We refer the functional group identification process based on the standard rules used for classification as the 'Rule Based Method'. In this method, a set of deterministic rules are used to identify the functional group in a given compound. This result is reproducible and can be easily explained because the process is a white-box as against the black-box machine learning methods. This Rule Based Method is an attempt is to replicate how an average human being (spectroscopist) would classify the spectrum given a set of standard rules.

Bond ranges given in Table 2 are the base on which basic features and rules of identification depend heavily. These ranges give the exact locations in the spectra where the specific bonds may exist and from which, rules of identification of functional groups can be built; the well established rules can be found in any standard such as [4, 21–24].

Due to the peculiarities in IR spectra as mentioned in section 3.2, it is clear that the visual identification process is complex even when the rules of identification are well-defined. Hence, it is important to first get a benchmark for the Rule Based Method in order to understand what would be the accuracy given the set of standard rules.

Table 3: Results of Rule Based Method

Functional Group	Total Samples	Functional Group Priority Not Considered		Functional Group Priority Considered	
		Predictions Matched	Accuracy	Predictions Matched	Accuracy
Carboxylic Acid	100	28	0.28	28	0.28
Ester	100	0	0	0	0
Amide	100	6	0.06	6	0.06
Cyanide/Nitrile	100	85	0.85	72	0.72
Aldehyde	100	50	0.5	16	0.16
Ketone	100	68	0.68	9	0.09
Alcohol	100	96	0.96	84	0.84
Amine	100	72	0.72	0	0
Arene/Aromatic	100	76	0.76	11	0.11
Alkene	100	73	0.73	14	0.14
Alkyne	49	7	0.14	0	0
Alkane	100	98	0.98	72	0.72
Ether	100	99	0.99	0	0
Nitro	92	32	0.35	0	0
Overall Average	**1341**	**790**	**0.5891**	**312**	**0.2327**

First, the rules are applied without considering the priority order of functional groups. This is what most of the past research has focused on. In this method, it is simply checked whether the predicted functional group matches with the actual functional group without considering the priority. The results obtained are shown in column 3 and 4 under the title 'Functional Group Priority Not Considered', in Table 3. For this method, the overall accuracy is 58.91%. For eight of the functional groups, the accuracy is more than 60%. This is because, these functional groups have some distinctive patterns which can be easily identified upon observing the spectrum and hence can be included as a rule. For example, the cyanide or nitrile group can be easily confirmed from the presence of C triple bond N in the spectrum. Hence it shows high accuracy in column 4 in Table 3. On the other hand, some functional groups like carboxylic acid, ester, amide, aldehyde, ketone have many overlapping features and hence are difficult to distinguish by simple observations or basic rules. Therefore we get very low accuracy for these functional groups in column 4 in Table 3.

Furthermore, when the priority is considered in the Rule Based Approach(see column 5 and 6 under the title 'Functional Group Priority Considered' in Table 3), it is more difficult to identify the functional groups because if the higher priority group is identified incorrectly, then the remaining groups are not checked for and the compound is classified incorrectly. Hence for this method, the overall accuracy is 23.27%, which is very low.

This shows that if an amateur person in the field is given a set of rules and is asked to identify functional groups of organic molecules based on priority, the accuracy would hardly be 25%.

To improve this accuracy, more sophisticated technique is required which would be able to identify patterns and correlations between these basic features. To this end, using machine learning is a viable option.

5 MACHINE LEARNING BASED METHOD

There are two possible ways of finding features for the machine learning algorithms. First one is to use the basic human identifiable features and check whether machine is able to detect better structures and mapping. We call this the 'Intermediate approach'. The second way is to allow the machine to develop its own features from the data which the machine can use for classification. We have termed it as the 'Complete Automation Approach'.

5.1 Intermediate Approach

In this approach we try to extract features which a human would use to correctly identify the functional group of the molecule. We term them to be "handcrafted features". Here we have used bond ranges as mentioned in Table 2 to identify the transmittance levels and widths of peaks in the given ranges.

For each of the 23 bond ranges that we have considered, we identify 4 features -

- Transmittance - The peak (trough) with least transmittance value.
- Peak Width - The width of the peak with the least transmittance value.
- Number of Peaks - The count of peaks in the given bond range.
- Sum of peak widths - The sum of all the peak widths in the bond range.

So in all there are 92 features for each sample which are given as input to the machine. As the features are manually extracted and the classification is done by the machine, we consider it as an intermediate approach.

5.2 Complete Automation Approach

In the second approach, we give all data points (the complete spectrum) as features to machine learning algorithm. The transmittance levels corresponding to all the available wave-numbers are given as input features to the machine learning algorithm and we expect the algorithms to identify and extract better hidden features and patterns which would allow the machine to classify the samples correctly.

We consider a truncated range of 4000 cm^{-1} to 1000 cm^{-1} because as mentioned earlier, 1400 cm^{-1} to 500 cm^{-1} range is called fingerprint region and contains many peaks based on bending vibrations within the molecule. Therefore it may result in giving some false-positives. Since none of the bonds considered by us lie in the region 1000 cm^{-1} to 400 cm^{-1}, this part of fingerprint region should not affect the classification accuracy and hence can be neglected. As we are working with less number of classes, we can work with few features as well. We split the range uniformly such that we get data set with 250 features for each sample which acts as input to the machine learning algorithm.

5.3 Implementation

We have implemented four supervised machine learning algorithms such as Mulitlayer Perceptron (MLP), Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Random Forest Classifier (RFC), and compared the results for this problem [25].

Each sample is labelled according to the functional group to which it belongs. Each functional group is given a different class label [0, n-1] where n is the number of functional groups. Since we have 14 functional groups, we have labelled them from 0 to 13. The data is preprocessed before applying a machine learning algorithm because most of the methods work better when the features are standardized i.e mean 0 and variance 1. It helps in reducing the overall computation.

For analyzing the performance of each classifier i.e. how it behaves on unseen data, a K-Fold cross validation technique is used with a K value of 10. It is ensured that in each fold every class has an equal representation. For the K-Fold cross validation accuracy, each time the training data is split into 10 folds with 9 folds used for training the model and tested on 1 fold. This is done K times (K=10) and the average accuracy is reported.

6 TEST RESULTS AND STATISTICAL ANALYSIS

We analyze the performance of various machine learning algorithms for both intermediate approach and complete automation approach, and compare them with the rule-based approach. To that end, the 10 fold CV accuracies are indicated in Fig. 3 in which the error bars denote 95% confidence intervals for the corresponding mean values. The results indicate that the complete automaton approach performs better than intermediate approach both from statistical and practical viewpoint. This may be explained by the fact that the features for the latter approach are somewhat restrictive and may lack discrimination abilities (refer to section 3.2, where we have discussed how such handcrafted features may not provide clear distinction between functional groups). On the other hand, in the complete automation approach, we allow the ML algorithm to adaptively learn the desired mapping function f directly from the data patterns. This, however, does not rule out the possibility of using either standard or domain specific data preprocessing to make the classification process more robust and accurate.

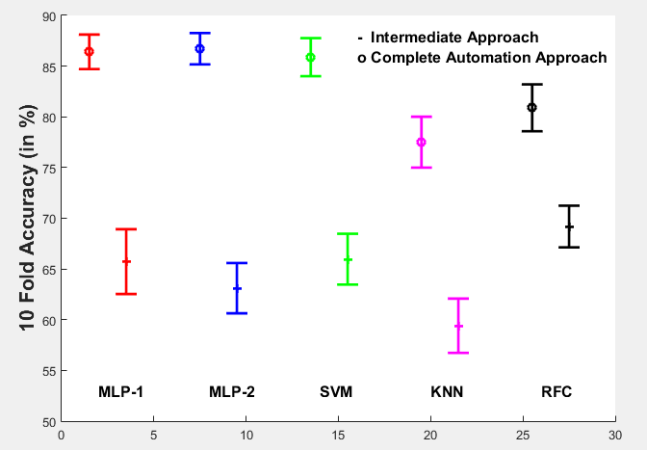

Figure 3: Comparison of the 10 Fold CV accuracies for the intermediate and complete automation methods. The error bands indicate 95% confidence intervals of the mean accuracies.

The more important result, from the view point of the targeted application, is that both the ML based approaches perform better (both practically and statistically) than the rule-based approach. This is in line with the analysis done in previous sections of the paper, and an important conclusion from the view point of applying ML to priority based functional group identification.

6.1 Further analysis using permutation tests

The results presented in the previous section indicate that the ML based approach provides promising results in comparison to the traditional and oft-utilized rule-based approach. However, ML methods requires training, and the learning process is typically a black box. This is one notable limitation of ML based approach in comparison to rule-based method which is completely transparent in terms of the rules employed (i.e. a white box model) to classify a given sample. Nevertheless, we can employ permutation test based analysis to gain some insights and obtain more descriptive information about the learning process in ML based approach. We, therefore, employed two permutation tests proposed in [26]. The first one tests the *null* hypothesis that the ML algorithm does not exploit class structure (i.e. connection between data and class labels). The second test assesses method performance in terms of using the features describing the data (or the data itself in case no features are extracted), and tests the *null* hypothesis that the ML algorithm does not use feature dependency to increase classification accuracy. For both the tests, an empirical p value can be estimated as the ratio of number of times the classification accuracy is better (or equal) than original data (i.e without any randomization) and the number of randomizations applied [26].

For the first permutation test, the class labels are randomly permuted between the samples. As a result, the underlying class structure is disturbed and the goal of the test is to assess systematically the performance of ML based method on such *unstructured* data. In this case, for each ML algorithm, the empirical p values was 0.0099 (< 0.05) (the data was permuted 100 times i.e. 100 randomizations). For each iteration the 10 fold CV accuracy obtained was between 5.5% to 9.5% (in comparison the said accuracy on the original data was nearly 80% refer Figure 3). In light of this, the corresponding *null* hypothesis can be rejected, and we conclude that all the ML algorithms exploit the class structure in the data.

As mentioned, the second permutation test was performed to test whether the ML algorithm exploits feature dependency to increase classification accuracy. Therefore, feature (data) values were permuted within each class so that the dependency between features (data), if it exists, is broken. In such case, we expect that the classifier will obtain lower classification accuracies as compared to the one on original data. We performed 100 randomizations, and obtained the 10 fold CV accuracies in each iteration.

As examples, we show in Fig. 4, the histograms of the resultant 10 fold CV accuracies for KNN and SVM. The original classification accuracy is also indicated in the respective plots. From these, we observe that KNN is exploiting data dependency because the accuracies on randomized data are typically lower than on the original data. This results in a low p value for KNN under this permutation test. On the other hand, the case of SVM is different in that when the feature structure is broken, the classification accuracy increases (similar results were obtained for the rest of the classifiers namely, MLP with 1 hidden layer, MLP with 2 hidden layers and RFC). Consequently, the p value for SVM (as well as for MLP-1, MLP-2 and RFC) is 1. Note that these results are in agreement with those published in [26] where simpler classifiers such as KNN tend to obtain lower p values as compared to the more complex classifiers, under the second permutation test.

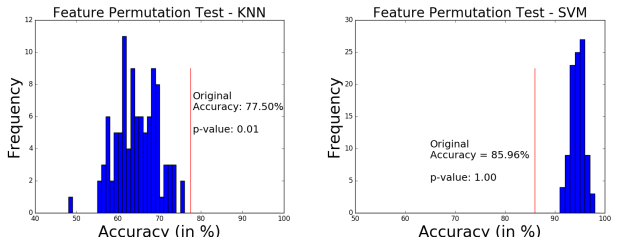

Figure 4: Results of feature permutation test. The graph on the left shows the distribution of mean accuracy for KNN. The graph on the right shows the distribution of mean accuracy for SVM.

Such high p values for the second permutation test also indicate that these classifiers (SVM, MLP-1, MLP-2 and RFC) are not fully exploiting the feature dependency to increase classification accuracies. Instead they may rely on more dominant feature/data values to classify the given test sample, and hence less affected by permuted values due to randomizations. We also note that KNN is a simpler classifier as compared to SVM, MLP and RFC, and achieves an accuracy of about 77.5% while SVM obtains relatively higher accuracy of 85.96%. This, in combination with the results of second permutation test, suggests that data preprocessing and/or domain specific feature extraction (instead of just the handcrafted features) may reduce redundant data information, and more complex classifiers such as SVM may actually exploit feature (or data) dependency to increase further the classification accuracies.

7 CONCLUSION

This paper discusses an ML based approach towards priority based identification of functional groups in organic molecules. To the best of our knowledge, this is first such work which considers the priority of the functional groups and gives a single label for each sample. The aim was to explore the possibilities of using an accurate automated process to supplement the manual rule based approach with machine based feature extraction and identification process. In the context of finding the most dominant functional group in the molecule we find that the general rule based approach is less effective (recall that it gives only 23.27% accuracy as against 58.91% when we just look whether a functional group is present or not without considering priority). Thus, we explored an ML based approach wherein we considered two cases. In the first case, features which a human would identify were used while in the second case, the ML algorithm is allowed to identify patterns directly from the given spectroscopy data. The accuracy increased to 69.18% in the first case while it improved further to 86.71% in the second case.

We also supported our analysis using confidence intervals and permutation tests in order to properly consider the statistical significance of the classifier performances. The permutation tests in particular revealed some deficiencies in complex classifiers such as SVM, and this may be remedied by using a reduced feature space by using application specific information as well as other standard dimensionality reduction techniques. Overall, the paper provides positive initial steps towards priority based function group identification of organic molecules using the ML approach.

ACKNOWLEDGMENTS

We express our sincere gratitude to National Institute of Advanced Industrial Science and Technology, Japan for open source SDBS database of IR spectra without which this work would not be possible. Author B. Chaudhury would like to acknowledge fruitful discussions on spectroscopy with Dr. K. S. Maiti of Max-Planck Institute for Quantum Optics.

REFERENCES

[1] Saul Patai. *Patai's Chemistry of Functional Groups*. Wiley, 1964-1995.
[2] John Coates. Interpretation of infrared spectra, a practical approach. *Encyclopedia of Analytical Chemistry, John Wiley Sons Ltd*, page 10815âĂŞ10837, 2000.
[3] Baker et al. Using fourier transform ir spectroscopy to analyze biological materials. *Nature America, Inc., nature protocols, VOL.9 NO.8*, pages 1771–1791, 2014.
[4] H. Favre and W. Powell. *Nomenclature of Organic Chemistry: IUPAC Recommendations and Preferred Names*. Royal Society of Chemistry, 1st edition, 2013.
[5] E. W. Robb and M. E. A Munk. neural network approach to infrared spectrum interpretation. *Mikrochim. Acta [Wien]I*, pages 131–155, 1990.
[6] M. Minsky and S. Papert. Perceptrons, mit press, cambridge, ma. 1969.
[7] R. J. Fessenden and L. GyoĬLrgyi. Identifying functional groups in ir spectra using an artificial neural network. j. *Chem*, 2:1755–1762, 1991.
[8] M. E. Munk, M. S. Madison, and E. W. Robb. Neural network models for infrared spectrum interpretation. *Mikrochim. Acta [Wien] II*, pages 505–514, 1991.

[9] M. Meyer and T. Weigelt. Interpretation of infrared spectra by artificial neural networks. *Anal. Chim. Acta*, 265:183–190, 1992.

[10] D. Ricard, C. Cachet, and D. Cabrol-Bass. Neural network approach to structure feature recognition from infrared spectra. j. *Chem*, 33:202–210, 1993.

[11] P. N. Penchev, G. N. Andreev, and K. Varmuza. Automatic classiÂőcation of infrared spectra using a set of improved expert-based features. *Anal. Chim. Acta*, 388:145–159, 1999.

[12] C. Klawun and C. L. Wilkins. Joint neural network interpretation of infrared and mass spectra. j. *Chem*, 36:249–257, 1996.

[13] C. Klawun and C. L. Wilkins. Optimization of functional group prediction from infrared spectra using neural networks. j. *Chem*, 36:69–81, 1996.

[14] Judit Ambro. Classifying organic compounds using expert system and neural networks. *Theses, Dissertations, Professional Papers*, 5104, 1991.

[15] M. Meyer, K. Meyer, and H. Hobert. Neural networks for interpretation of infrared spectra using extremely reduced spectral data. *Anal. Chim. Acta*, 282:407–415, 1993.

[16] T. Visser and H. Luinge. Recognition of visual characteristics of infrared spectra by artificial neural networks and partial least squares regression. *J.; Van der Maas, J. H*, 296:141–154, 1994.

[17] V. Tchistiakov, C. Ruckebusch, L. Duponchel, J. P. Huvenne, and P. Legrand. Neural network modelling for very small spectral data sets: reduction of the spectra and hierarchical approach. *Chemometrics and Intelligent Laboratory Systems*, 54:93–106, 2000.

[18] Chris W. Brown and Su-Chin Lo. Chemical information based on neural network processing of near-ir spectra. *Anal. Chem*, 70(14):2983–2990, 1998.

[19] Kazutoshi Tanabe et al. Identification of chemical structures from infrared spectra by using neural networks. *Appl. Spectrosc*, 55:1394–1403, 2001.

[20] K.Tanabe S.Kinugasa and T.Tamura. Sdbsweb:. http://sdbs.db.aist.go.jp(NationalInstituteofAdvancedIndustrialScienceandTechnology,Japan), accessed May-July, 2017.

[21] G. C. Bassler R. M. Silverstein and T. C. Morrill. *Spectrometric Identification of Organic Compounds*. John Wiley, 5th edition, 1991.

[22] George Socrates. *Infrared and Raman Characteristic Group Frequencies: Tables and Charts*. John Wiley and Sons, 3rd edition, 2004.

[23] Peter Larkin. *Infrared and Raman Spectroscopy*. Elsevier, 1st edition, 2011.

[24] Jr. Leroy G. Wade. *Organic Chemistry*. Pearson Education, 6th edition, 2007.

[25] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2Nd Edition)*. Wiley-Interscience, 2000.

[26] Markus Ojala and Gemma C. Garriga. Permutation tests for studying classifier performance. *J. Mach. Learn. Res.*, 11:1833–1863, August 2010.