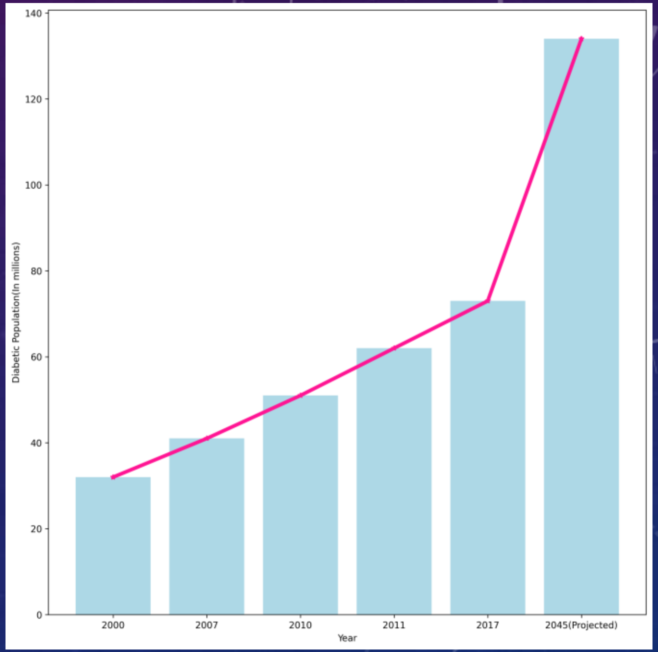
APPLICATION OF ARTIFICIAL INTELLIGENCE IN CHEMICAL SCIENCES BTP-1

VEER GAJARLAWAR | 21CY10013

GUIDE: PROF. MAHESH MOHAN MR

INTRODUCTION AND MOTIVATION
DIABETES: A GROWING EPIDEMIC IN THE MODERN WORLD

- Diabetes is a major public health problem that is approaching epidemic proportions globally. About 18 million people die every year from cardiovascular disease, for which diabetes and hypertension are major predisposing factors. India ranks second after China in the global diabetes `epidemic with 77 million people with diabetes.
- This emphasizes the need for a non-invasive technique to improve diabetes detection and monitoring. FTIR spectroscopy in combination with AI can be potential solution.

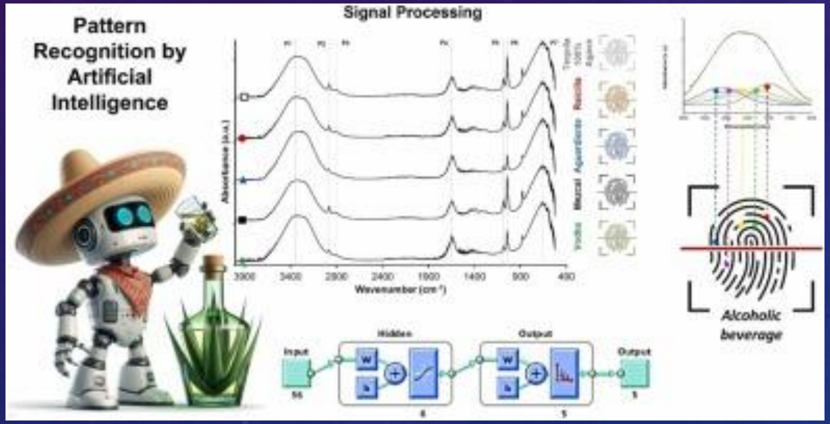


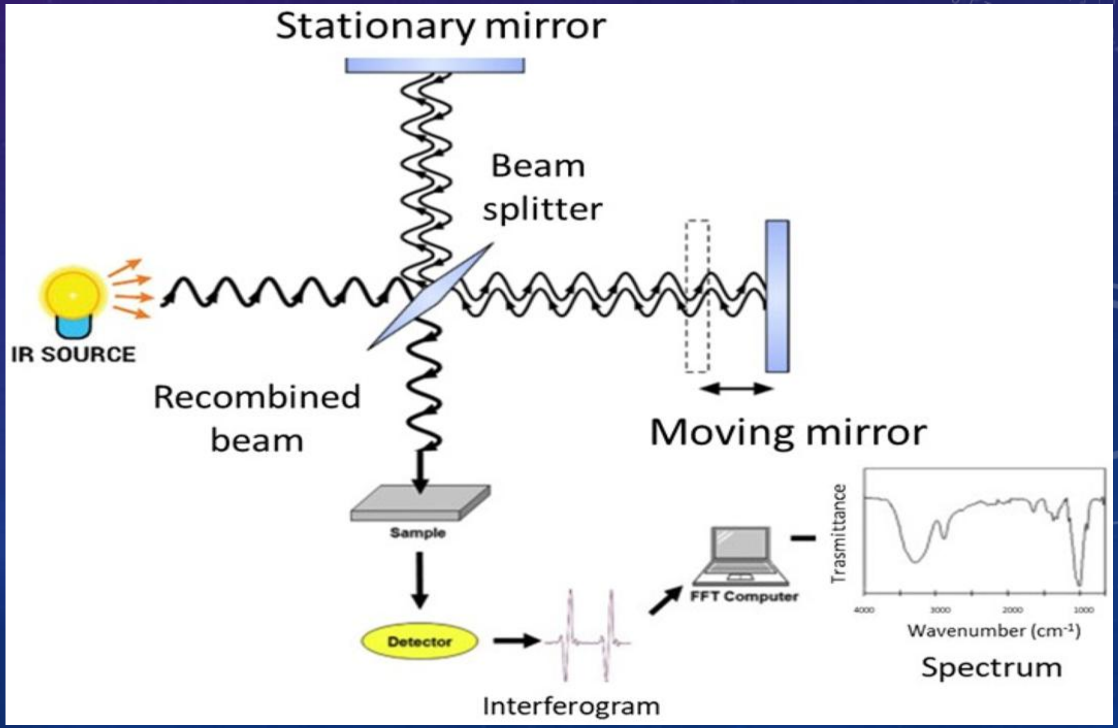
NEED FOR A NON-INVASIVE APPROACH

- Unlike traditional analytical techniques, use of FTIR spectroscopy is a non-invasive and cost effective approach.
- The consumables required for FTIR spectroscopy are minimal, as it primarily relies on the instrument's ability to analyze samples directly or with simple substrates, reducing ongoing costs.
- It provides a rapid analysis as it provides the results in minutes.
- TIR instruments are generally durable and require less frequent maintenance.

APPLICATION OF AI IN FTIR SPECTROSCOPY

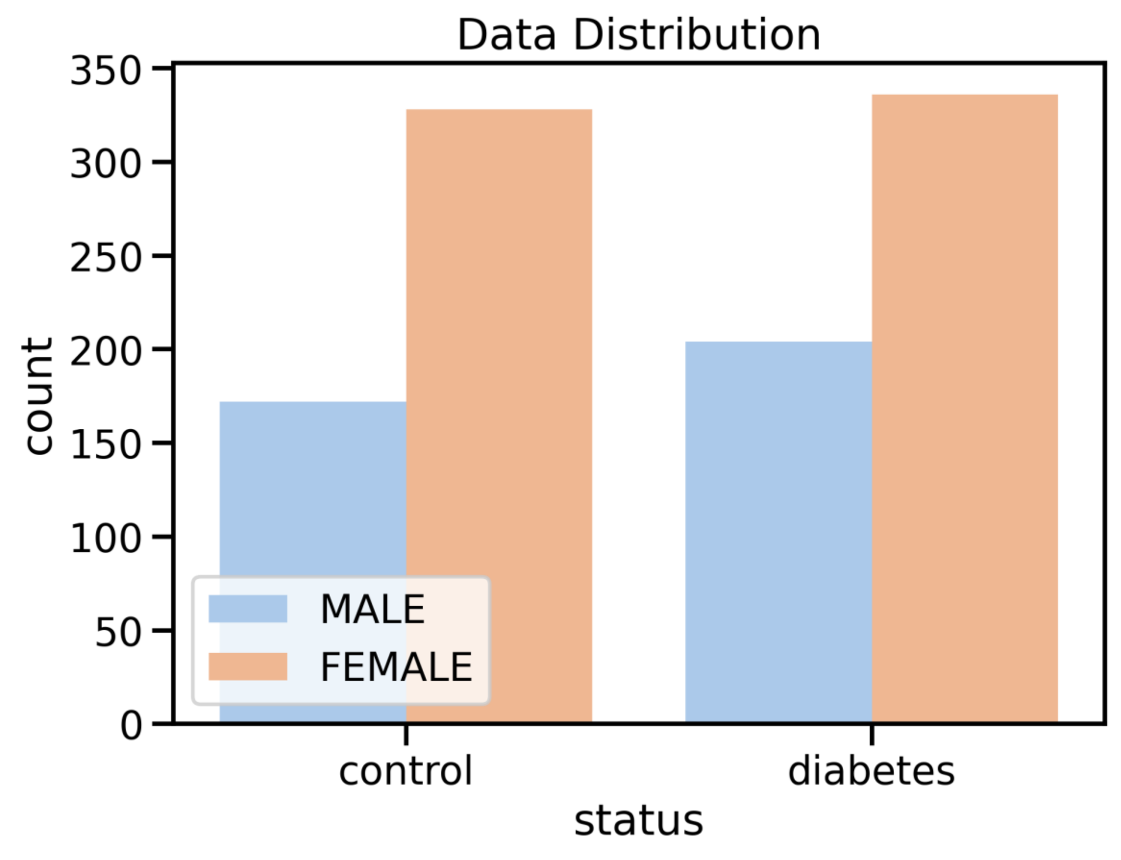
- When integrated with AI, FTIR spectroscopy has applications in a variety of domains, including medical diagnostics for disease detection, material categorization, etc.
- Some of the applications we explored were pollen grain classification, estimation of age of herbarium samples and disease detection.



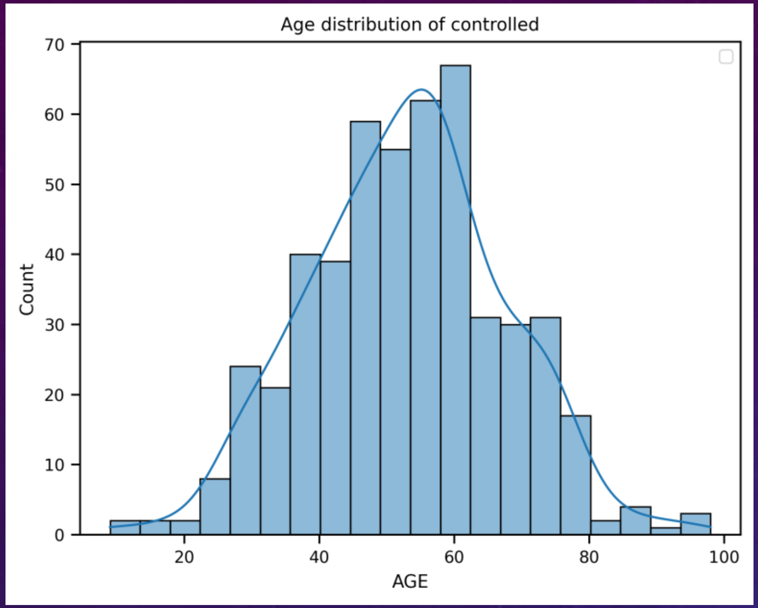


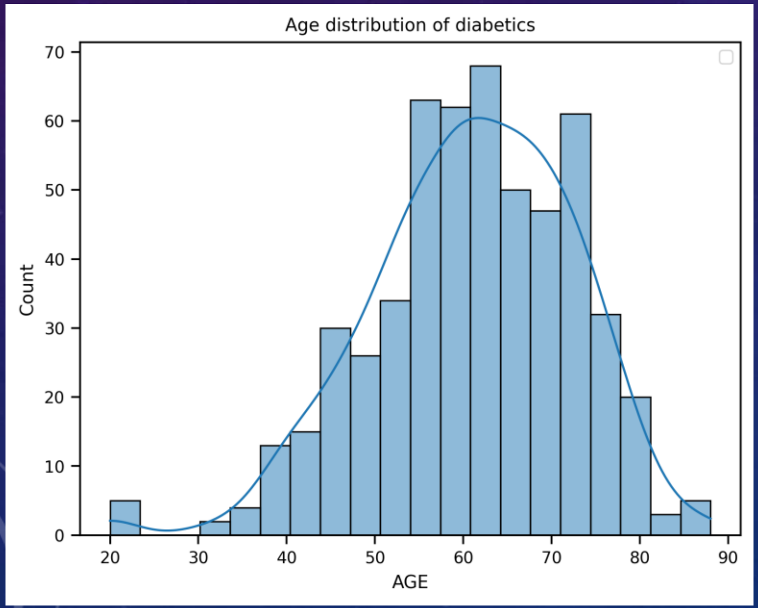
DATASET CONSIDERED FOR DIABETES DETECTION

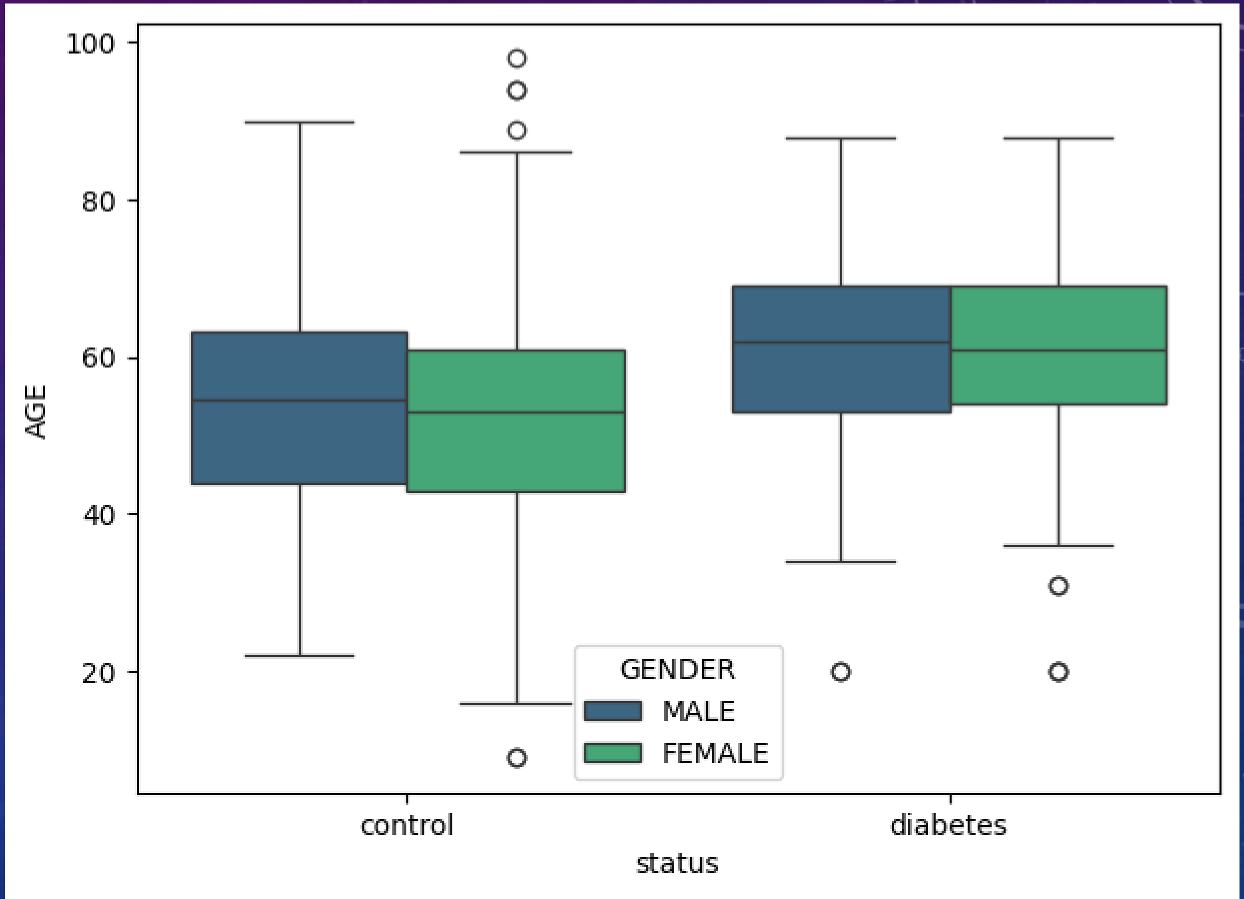
- **DATASET:**
- Total number of male diabetes spectra: 204.
- Total number of female diabetes spectra: 336.
- Total number of male control spectra: 172.
- Total number of female control spectra: 328.
- Total number of features (Wavenumbers): 3737.



DISTRIBUTION WITHIN THE DATASET

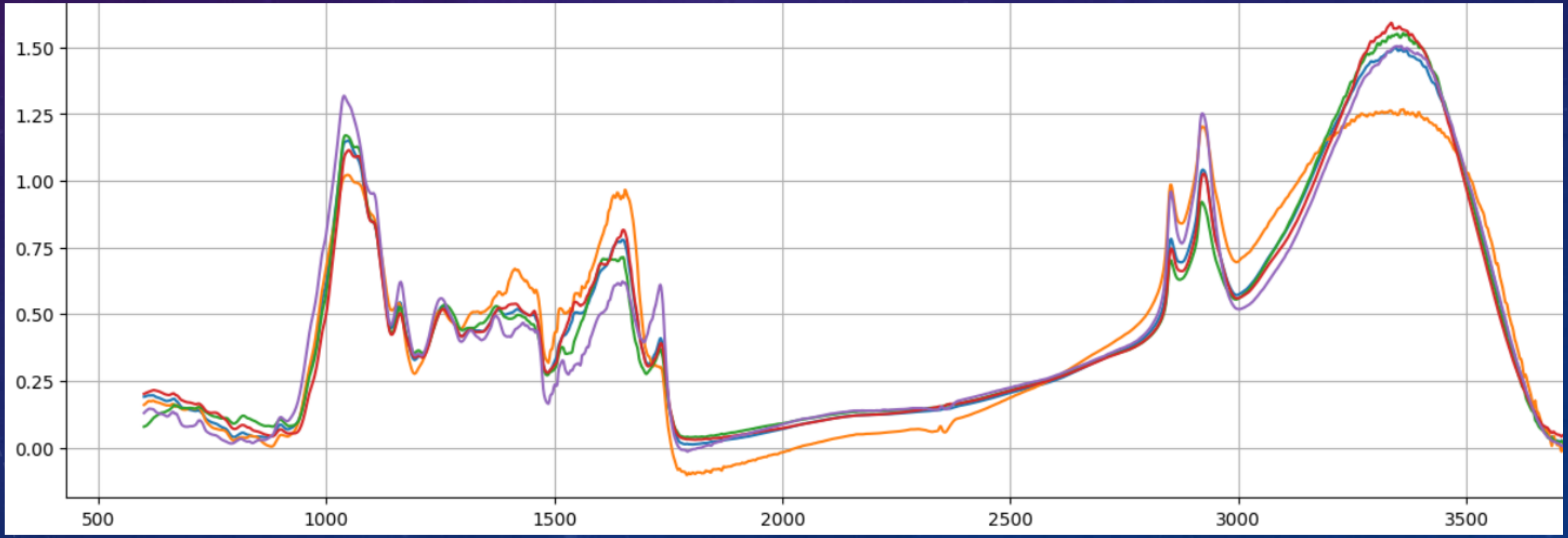






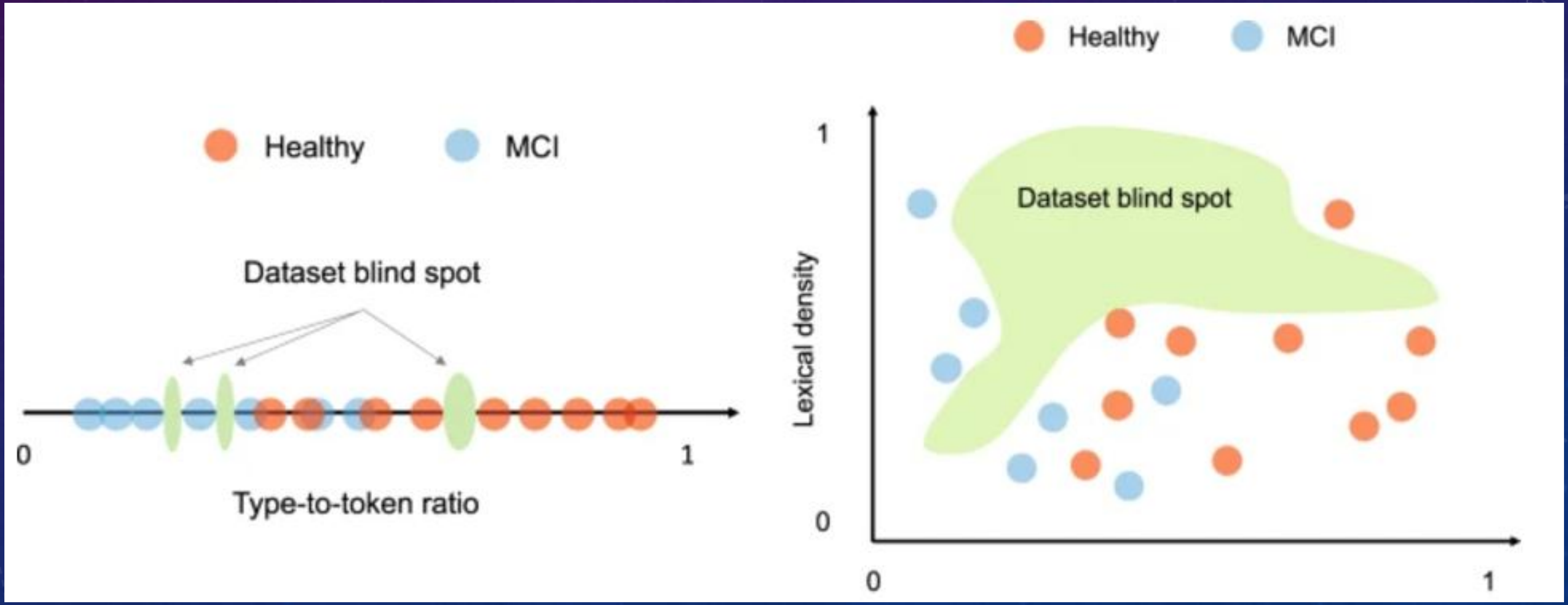
MAJOR CHALLENGES WITH FTIR SPECTROSCOPY DATA

- Interpreting complex patterns in FTIR spectral data involves understanding how the spectral features relate to the underlying chemical or biological processes.
- Each peak or feature in an FTIR spectrum corresponds to a specific bond or group of atoms vibrating in a unique way. However, in real-world samples, especially biological ones like saliva or blood, multiple compounds are present.
- This sometimes makes spectral data unreliable for analysis.
- Machine learning can help us work with such data.



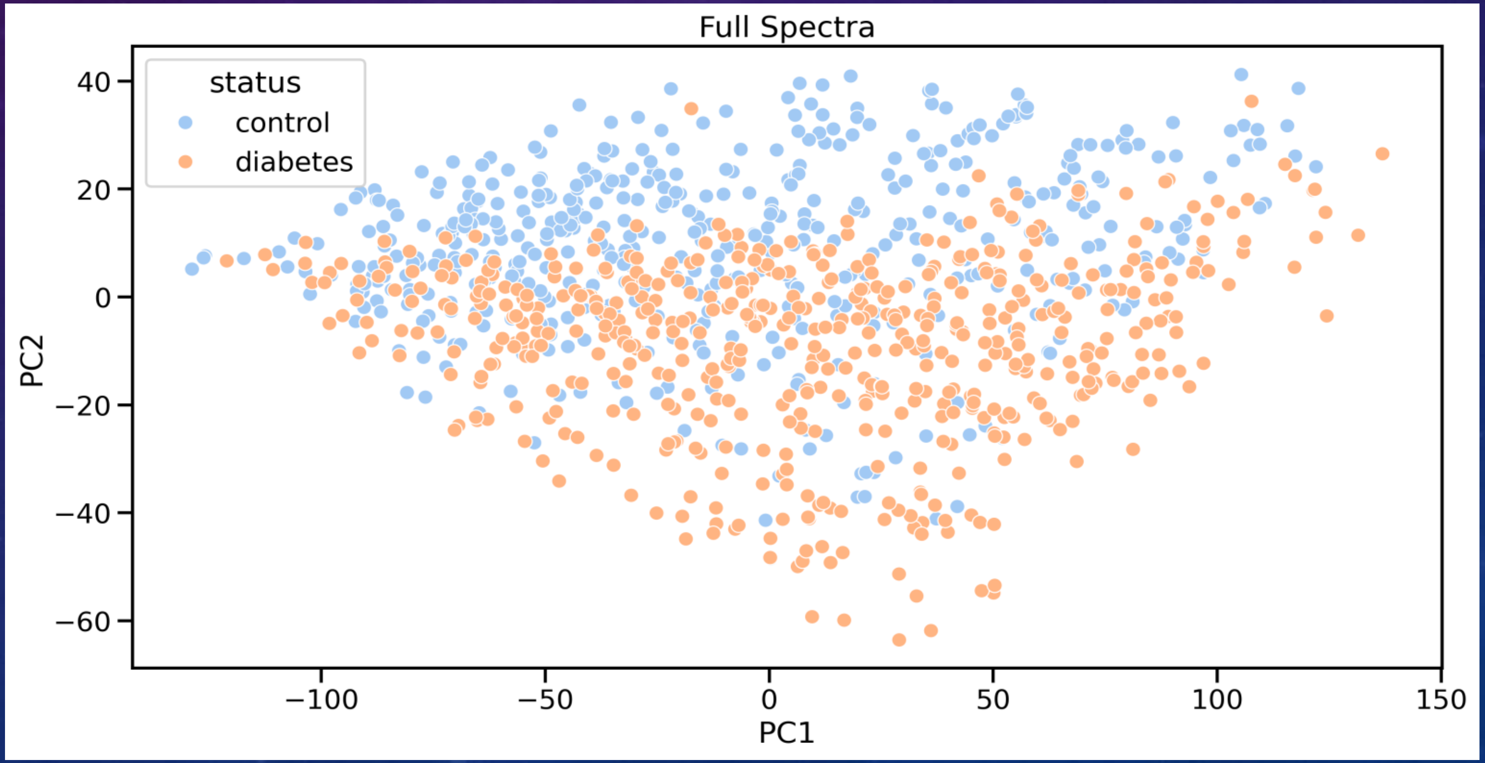
MAJOR CHALLENGES WITH FTIR SPECTROSCOPY DATA

- A major challenge is the high-dimensional nature of the spectral data obtained. FTIR spectroscopy generates data that represent absorbance at hundreds or even thousands of wavenumbers, resulting in a vast number of features or variables for analysis.
- This phenomenon is referred to as the *curse of dimensionality*.



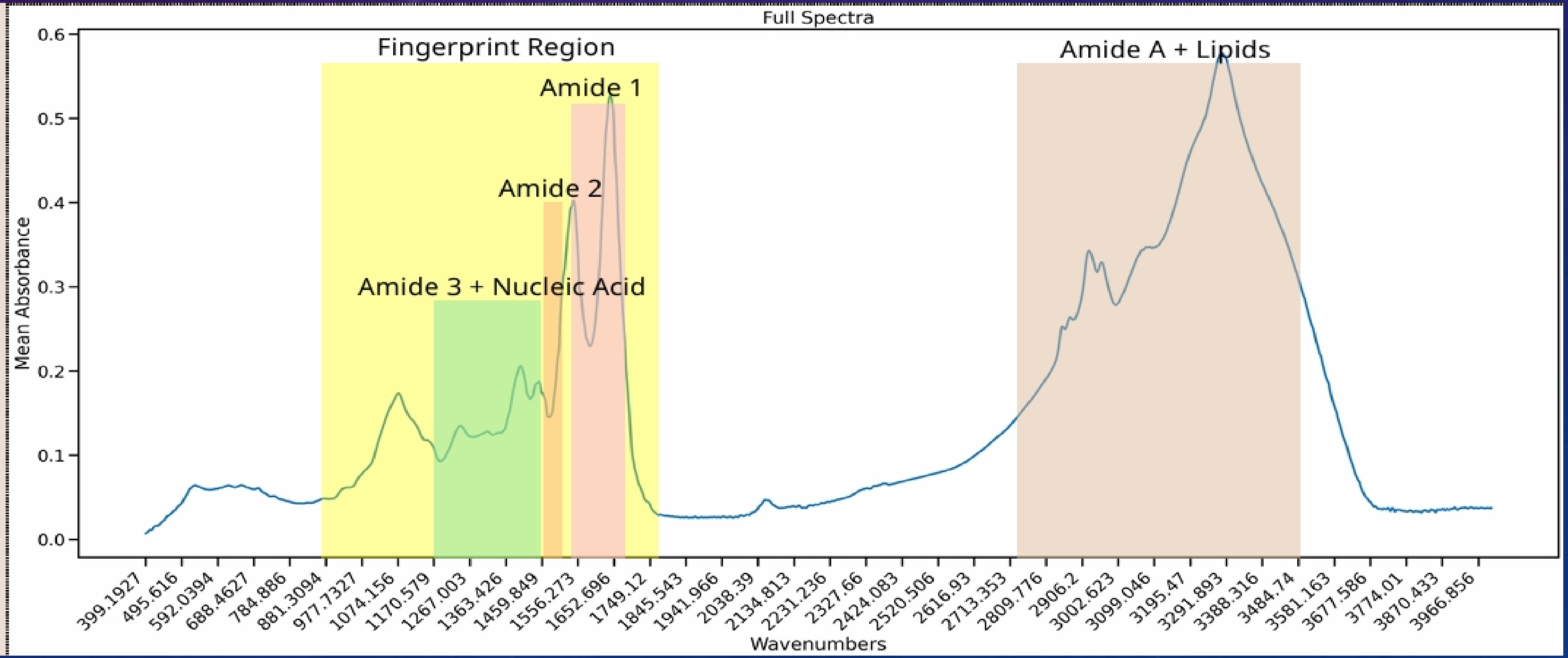
EXISTING SOLUTION

- To effectively tackle the issue of high dimensional FTIR spectral data, we can use dimensionality reduction techniques that are critical for simplifying the data while retaining as much of the underlying structure and information as possible.
- Principal Component Analysis (PCA) is a widely used dimensionality reduction technique in data science and machine learning. However, it may not always prove to be helpful.

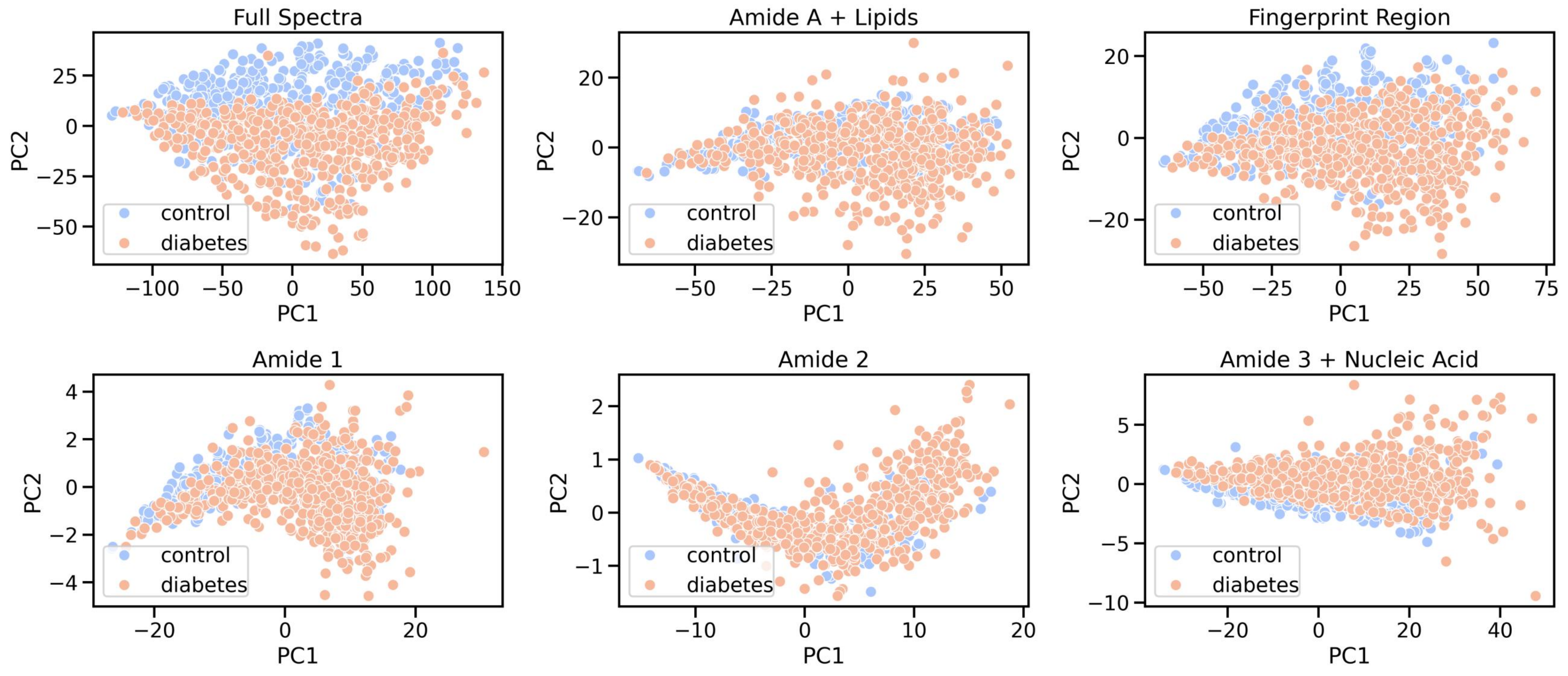


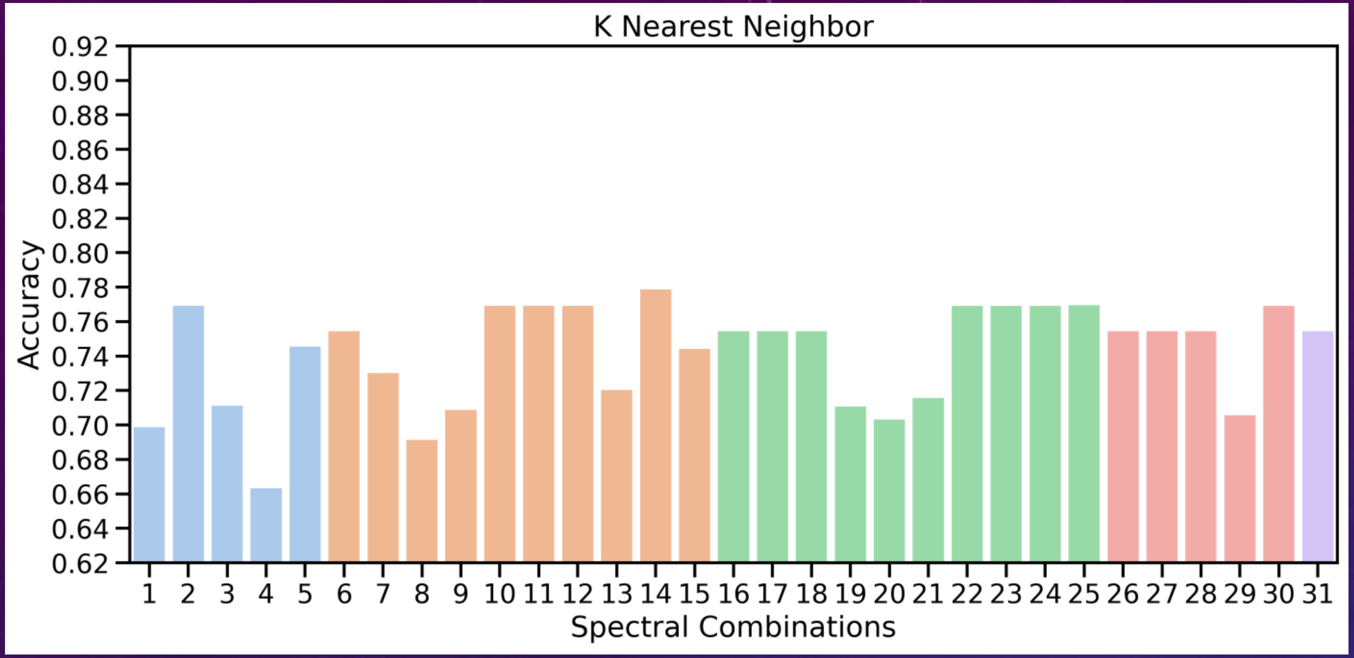
METHODOLOGY: DIVIDE AND CONQUER APPROACH USING AI

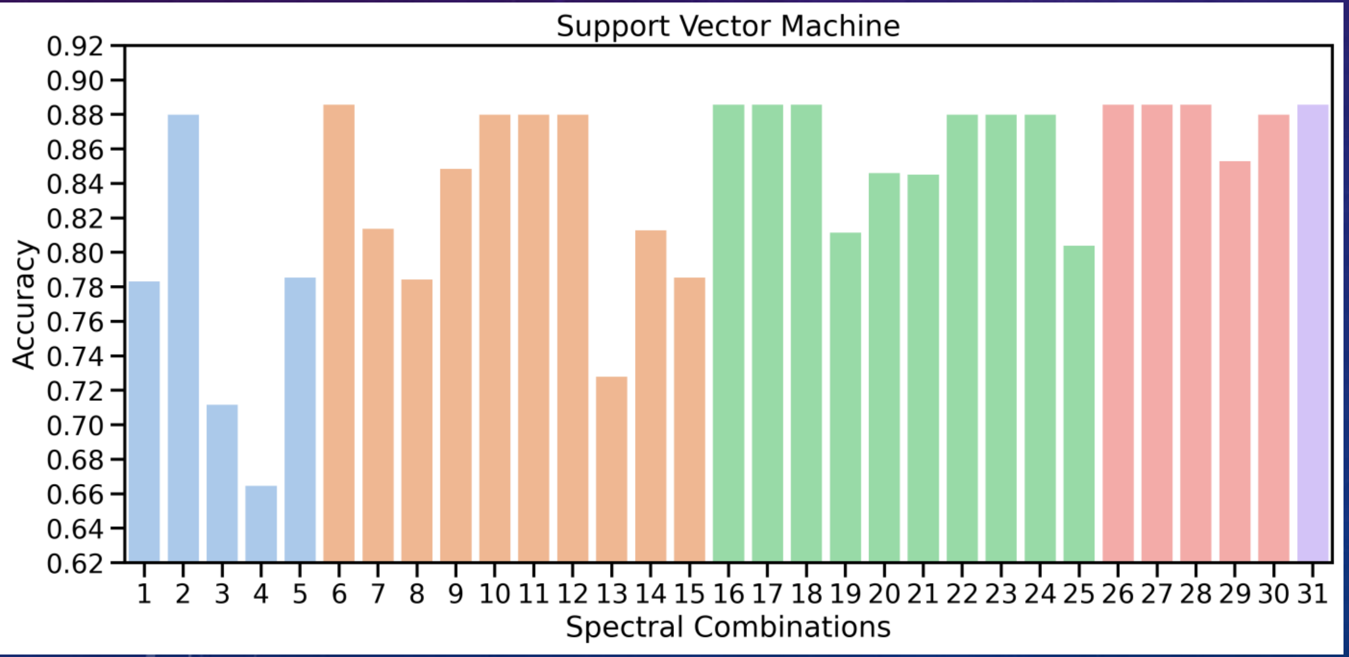
- we analyzed various different regions like Amide A + Lipids region (2800 cm^{-1}-3500 cm^{-1}), Amide 1 region (1600 cm^{-1}-1700 cm^{-1}), Amide 2 region (1500 cm^{-1}-1560 cm^{-1}), Amide 3 + Nucleic Acid region (1200 cm^{-1}-1500 cm^{-1}) and fingerprint region (900 cm^{-1}-1800 cm^{-1}).
- Performing PCA in these different regions independently, did not prove to be of much help as data points from both categories showed great amount of overlap.
- The machine learning algorithms used for the analysis were Support Vector Machines (SVM) and K Nearest Neighbour (KNN) coupled with Leave One Out Cross Validation on all possible combinations of regions.



PCA ON DIFFERENT REGIONS



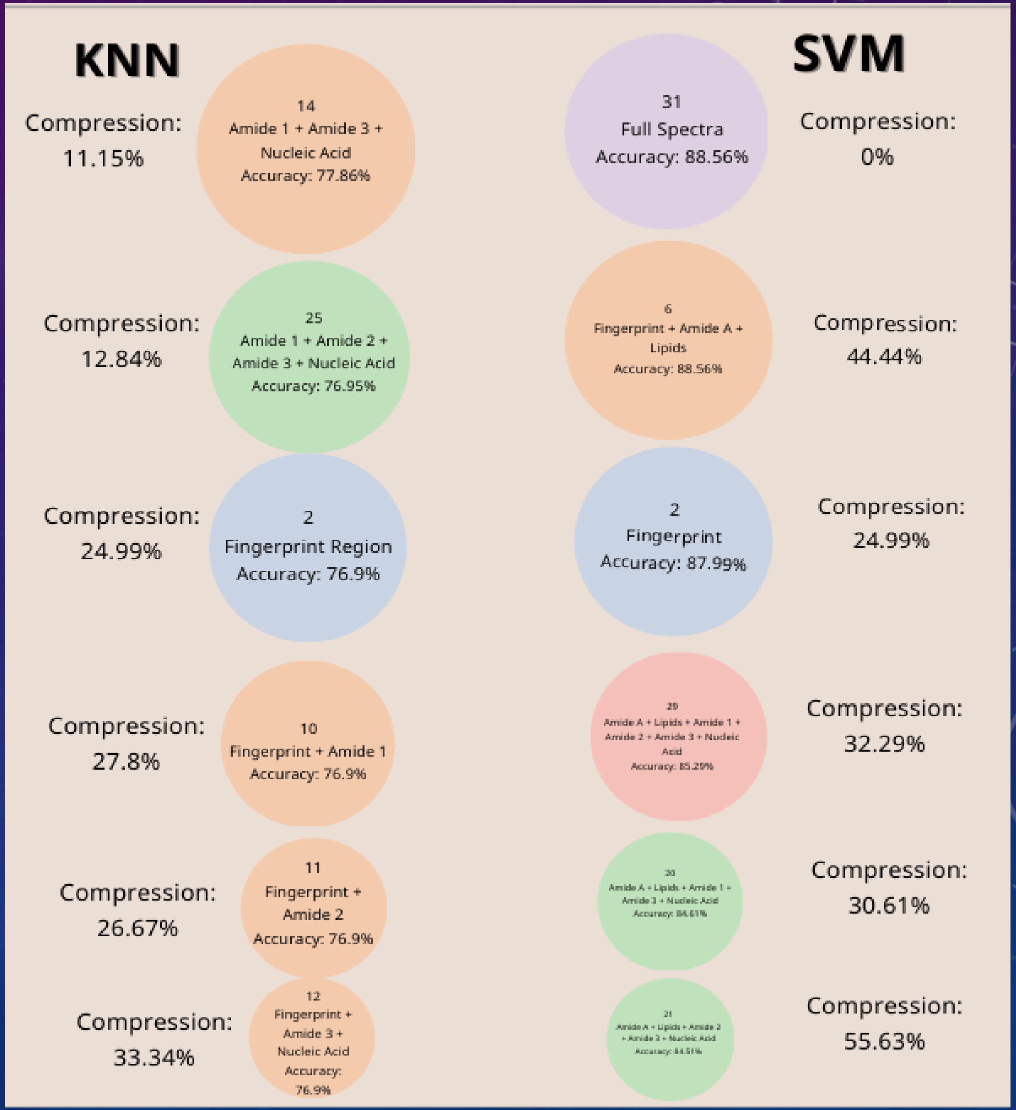




1	Amide A + Lipids
2	Fingerprint region
3	Amide 1
4	Amide 2
5	Amide 3 + Nucleic Acid
6	Amide A + Lipids + Fingerprint region
7	Amide A + Lipids + Amide 1
8	Amide A + Lipids + Amide 2
9	Amide A + Lipids + Amide 3 + Nucleic Acid
10	Fingerprint region + Amide 1
11	Fingerprint region + Amide 2
12	Fingerprint region + Amide 3 + Nucleic Acid
13	Amide 1 + Amide 2
14	Amide 1 + Amide 3 + Nucleic Acid
15	Amide 2 + Amide 3 + Nucleic Acid
16	Amide A + Lipids + Fingerprint region + Amide 1
17	Amide A + Lipids + Fingerprint region + Amide 2
18	Amide A + Lipids + Fingerprint region + Amide 3 + Nucleic Acid
19	Amide A + Lipids + Amide 1 + Amide 2
20	Amide A + Lipids + Amide 1 + Amide 3 + Nucleic Acid
21	Amide A + Lipids + Amide 2 + Amide 3 + Nucleic Acid
22	Fingerprint region + Amide 1 + Amide 2
23	Fingerprint region + Amide 1 + Amide 3 + Nucleic Acid
24	Fingerprint region + Amide 2 + Amide 3 + Nucleic Acid
25	Amide 1 + Amide 2 + Amide 3 + Nucleic Acid
26	Amide A + Lipids + Fingerprint region + Amide 1 + Amide 2
27	Amide A + Lipids + Fingerprint region + Amide 1 + Amide 3 + Nucleic Acid
28	Amide A + Lipids + Fingerprint region + Amide 2 + Amide 3 + Nucleic Acid
29	Amide A + Lipids + Amide 1 + Amide 2 + Amide 3 + Nucleic Acid
30	Fingerprint region + Amide 1 + Amide 2 + Amide 3 + Nucleic Acid
31	Amide A + Lipids + Fingerprint region + Amide 1 + Amide 2 + Amide 3 + Nucleic Acid

RESULTS AND FINDINGS

- Based on the analysis of various combinations, SVM emerged as the better per forming algorithm with the top accuracy of 88.56%, obtained using the combination of Fingerprint + Amide A + Lipids region. This result is comparable to the one achieved by considering the entire spectra while also achieving a compression of 44% in feature extraction.

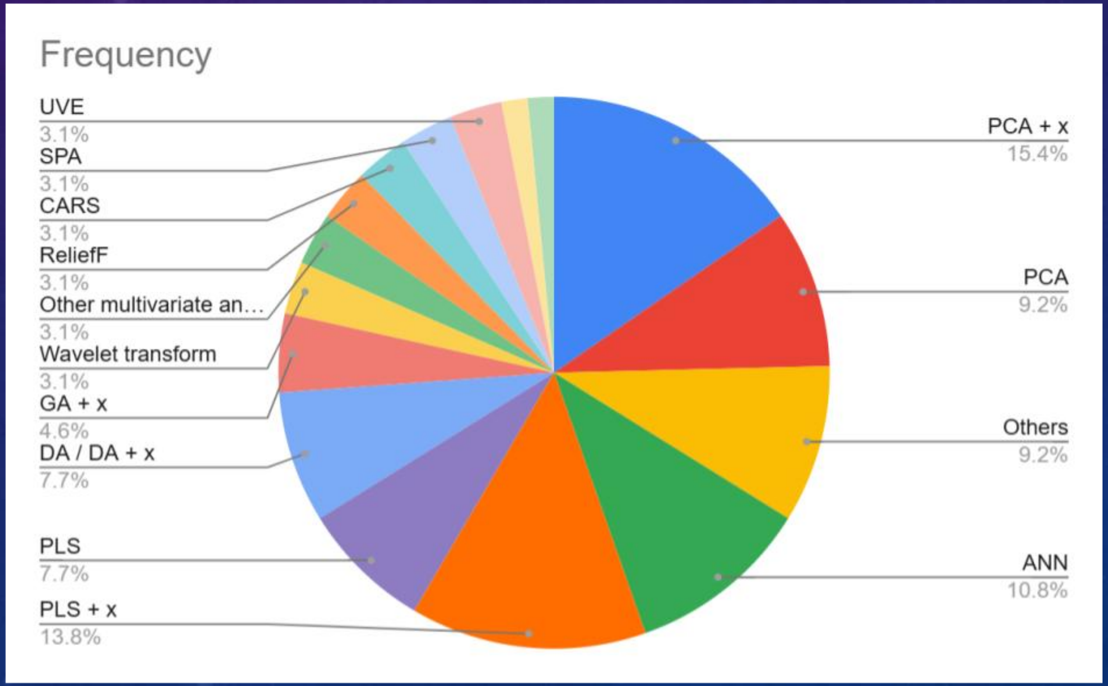


VIEW THE CODE

You can view the code used for all the analysis and predictions by scanning this barcode.

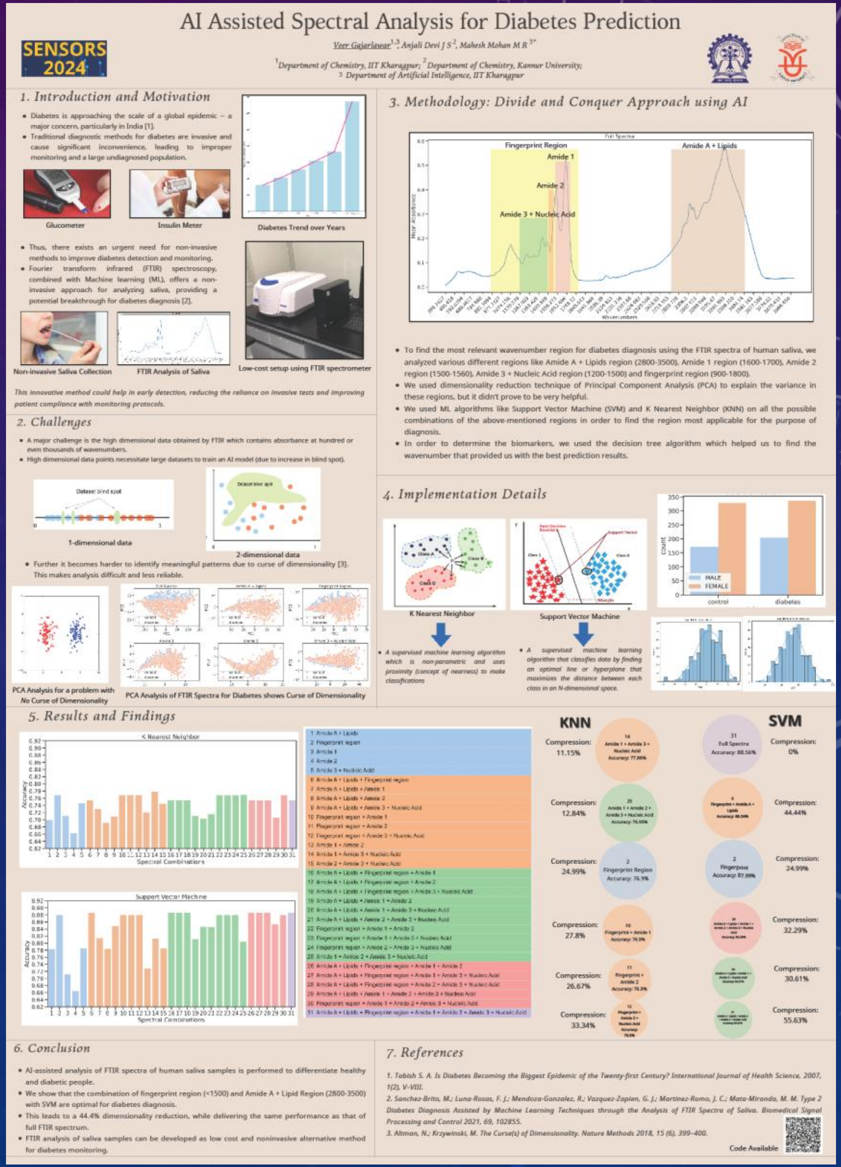
FUTURE WORKS

- As discussed early on, feature extraction is one of the key applications of AI in FTIR spectroscopy. However, the exact technique algorithm to be used is highly case sensitive and depends on factors like kind of spectral data in question and approach of the author.
- *A generalized feature extraction technique* would allow researchers to automate feature selection and dimensionality reduction without manual intervention, enabling more efficient and unbiased analysis.



ACCOMPLISHMENTS

- Our abstract titled *"AI Assisted Spectral Analysis for Diabetes Detection"* was selected for poster presentation at the Symposium on Emerging Nanotechnologies for Sensors-Organization and Recognition Systems 2024 (SENSORS 2024)



REFERENCES

1. Tabish S. A. Is Diabetes Becoming the Biggest Epidemic of the Twenty-first Century? International Journal of Health Science, 2007, 1(2), V–VIII.

2. Sanchez-Brito, M.; Luna-Rosas, F. J.; Mendoza-Gonzalez, R.; Vazquez-Zapien, G. J.; Martinez-Romo, J. C.; Mata-Miranda, M. M. Type 2 Diabetes Diagnosis Assisted by Machine Learning Techniques through the Analysis of FTIR Spectra of Saliva. Biomedical Signal Processing and Control 2021, 69, 102855.

3. Altman, N.; Krzywinski, M. The Curse(s) of Dimensionality. Nature Methods 2018, 15 (6), 399–400.

4. Baum, Zachary J., Yu, Xiang, Ayala, Philippe Y., Zhao, Yanan, Watkins, Steven P., Zhou, Qiongqiong. (2021). Artificial Intelligence in Chemistry: Current Trends and Future Directions. Journal of Chemical Information and Modelling. Received: June 1, 2021; Published: July 15, 2021.