AI Assisted Spectral Analysis for Diabetes Prediction

Veer Gajarlawar[1,3], Anjali Devi J S[2], Mahesh Mohan M R[3*]

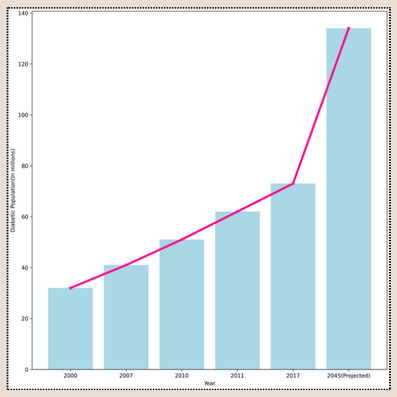
[1]Department of Chemistry, IIT Kharagpur; [2]Department of Chemistry, Kannur University;
[3] Department of Artificial Intelligence, IIT Kharagpur.

1. Introduction and Motivation

- Diabetes is approaching the scale of a global epidemic -- a major concern, particularly in India [1].
- Traditional diagnostic methods for diabetes are invasive and cause significant inconvenience, leading to improper monitoring and a large undiagnosed population.

Glucometer

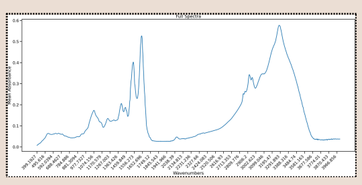
Insulin Meter



Diabetes Trend over Years

- Thus, there exists an urgent need for non-invasive methods to improve diabetes detection and monitoring.
- Fourier transform infrared (FTIR) spectroscopy, combined with Machine learning (ML), offers a non-invasive approach for analyzing saliva, providing a potential breakthrough for diabetes diagnosis [2].

Non-invasive Saliva Collection



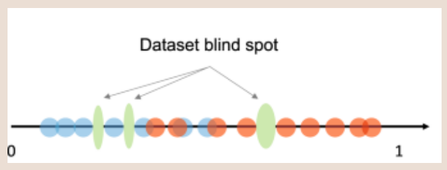
FTIR Analysis of Saliva

Low-cost setup using FTIR spectrometer

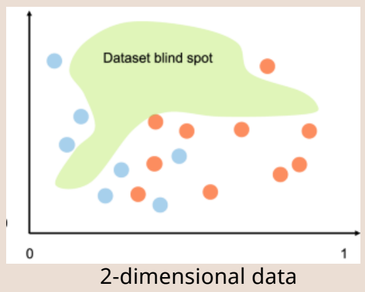
This innovative method could help in early detection, reducing the reliance on invasive tests and improving patient compliance with monitoring protocols.

2. Challenges

- A major challenge is the high dimensional data obtained by FTIR which contains absorbance at hundred or even thousands of wavenumbers.
- High dimensional data points necessitate large datasets to train an AI model (due to increase in blind spot).

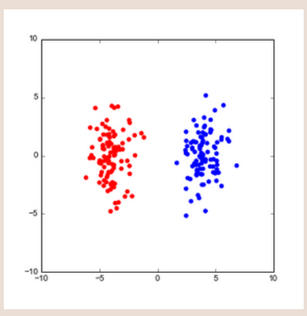


1-dimensional data

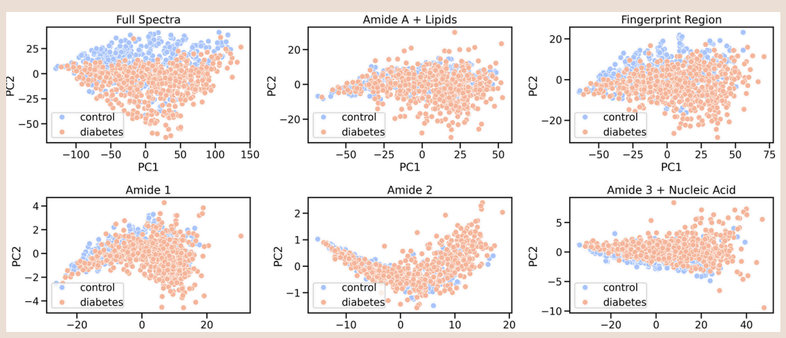


2-dimensional data

- Further it becomes harder to identify meaningful patterns due to curse of dimensionality [3]. This makes analysis difficult and less reliable.

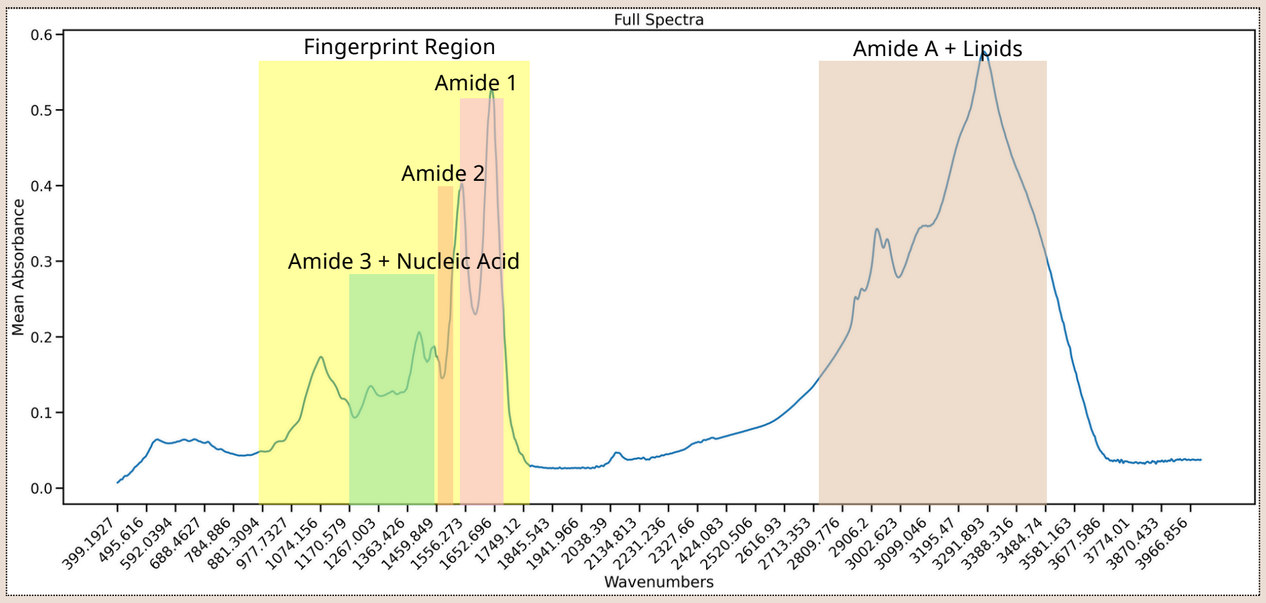


PCA Analysis for a problem with *No* Curse of Dimensionality



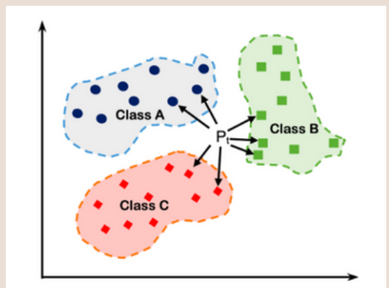
PCA Analysis of FTIR Spectra for Diabetes shows Curse of Dimensionality

3. Methodology: Divide and Conquer Approach using AI



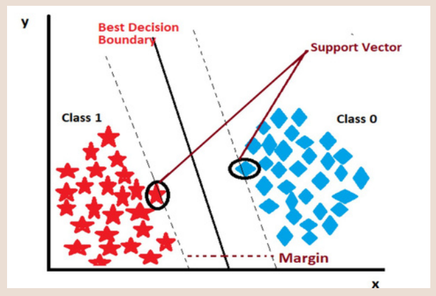
- To find the most relevant wavenumber region for diabetes diagnosis using the FTIR spectra of human saliva, we analyzed various different regions like Amide A + Lipids region (2800-3500), Amide 1 region (1600-1700), Amide 2 region (1500-1560), Amide 3 + Nucleic Acid region (1200-1500) and fingerprint region (900-1800).
- We used dimensionality reduction technique of Principal Component Analysis (PCA) to explain the variance in these regions, but it didn't prove to be very helpful.
- We used ML algorithms like Support Vector Machine (SVM) and K Nearest Neighbor (KNN) on all the possible combinations of the above-mentioned regions in order to find the region most applicable for the purpose of diagnosis.
- In order to determine the biomarkers, we used the decision tree algorithm which helped us to find the wavenumber that provided us with the best prediction results.

4. Implementation Details



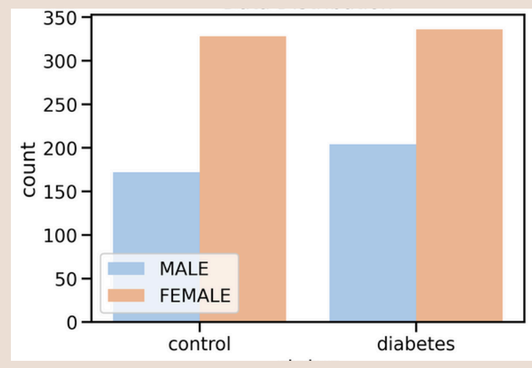
K Nearest Neighbor

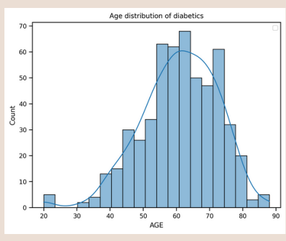
- *A supervised machine learning algorithm which is non-parametric and uses proximity (concept of nearness) to make classifications.*

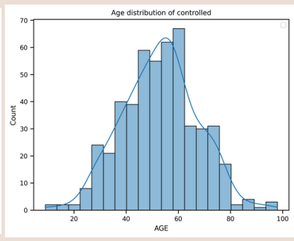


Support Vector Machine

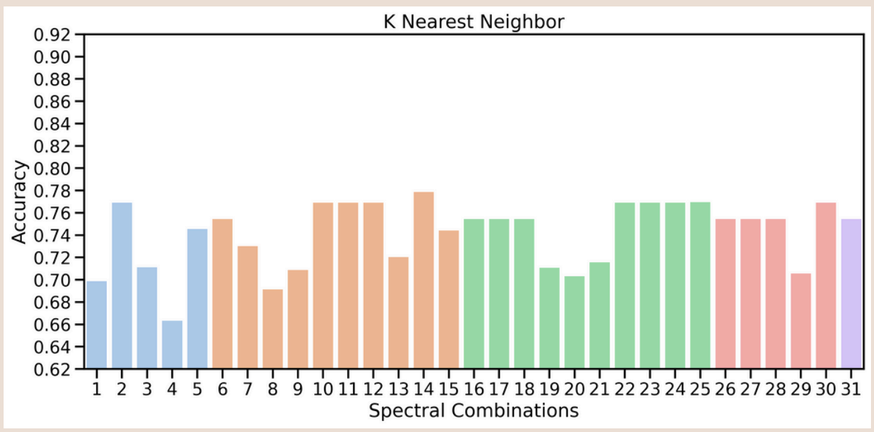
- *A supervised machine learning algorithm that classifies data by finding an optimal line or hyperplane that maximizes the distance between each class in an N-dimensional space.*



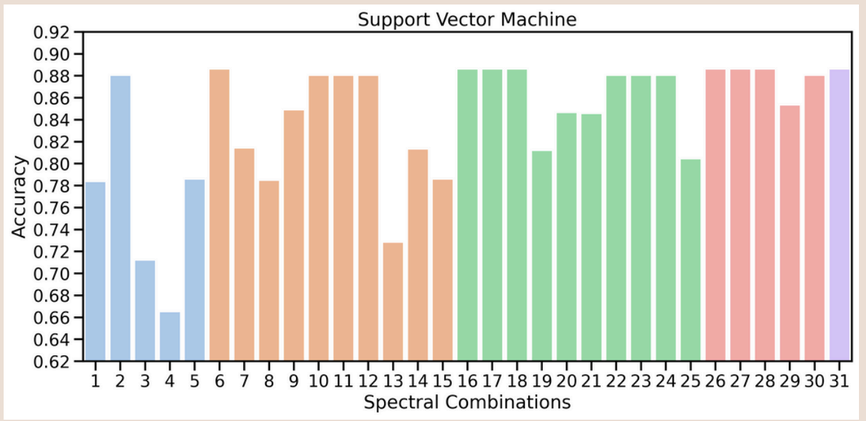




5. Results and Findings



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



KNN

- Compression: 11.15% — 14 — Amide 1 + Amide 3 + Nucleic Acid — Accuracy: 77.86%
- Compression: 12.84% — 25 — Amide 1 + Amide 2 + Amide 3 + Nucleic Acid — Accuracy: 76.95%
- Compression: 24.99% — 2 — Fingerprint Region — Accuracy: 76.9%
- Compression: 27.8% — 10 — Fingerprint + Amide 1 — Accuracy: 76.9%
- Compression: 26.67% — 11 — Fingerprint + Amide 2 — Accuracy: 76.9%
- Compression: 33.34% — 12 — Fingerprint + Amide 3 + Nucleic Acid — Accuracy: 76.9%

SVM

- 31 — Full Spectra — Accuracy: 88.56% — Compression: 0%
- 6 — Fingerprint + Amide A + Lipids — Accuracy: 88.56% — Compression: 44.44%
- 2 — Fingerprint — Accuracy: 87.99% — Compression: 24.99%
- 29 — Amide A + Lipids + Amide 1 + Amide 2 + Amide 3 + Nucleic Acid — Accuracy: 85.29% — Compression: 32.29%
- 20 — Amide A + Lipids + Amide 1 + Amide 3 + Nucleic Acid — Accuracy: 84.61% — Compression: 30.61%
- 21 — Amide A + Lipids + Amide 2 + Amide 3 + Nucleic Acid — Accuracy: 84.51% — Compression: 55.63%

6. Conclusion

- AI-assisted analysis of FTIR spectra of human saliva samples is performed to differentiate healthy and diabetic people.
- We show that the combination of fingerprint region (<1500) and Amide A + Lipid Region (2800-3500) with SVM are optimal for diabetes diagnosis.
- This leads to a 44.4% dimensionality reduction, while delivering the same performance as that of full FTIR spectrum.
- FTIR analysis of saliva samples can be developed as low cost and noninvasive alternative method for diabetes monitoring.

7. 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.

Code Available