# C:\Data\Academia\Nile University\Nile University Logo.jpg

# comparative study for Non-invasive detection of diabetes using Raman spectroscopy and infrared

2024

|  |  |
| --- | --- |
| Seifeldin Khaled Ibrahim  Computer science  Nile University  Giza, Egypt  s.khaled2132@nu.edu.eg | Omar Mohammed Tawfik  Computer science  Nile University  Giza, Egypt  o.mohammed2101@nu.edu.eg |
|  |  |

**Abstract --** This study compares two non-invasive techniques for detecting diabetes through spectral analysis of blood samples: Raman spectroscopy and infrared (IR). Both methods utilize spectral data from biological samples but differ in data characteristics, processing methods, and classification outcomes. The analysis covers data collection, preprocessing, augmentation techniques, model performance, and the challenges encountered. Raman data were collected from multiple body regions, including the inner arm, earlobe, thumbnail, and vein, and were augmented using SMOTE and noise addition to improve robustness. A Random Forest classifier achieved high cross-validation accuracy with the Raman dataset, yielding 97%, 95%, 91.1%, and 91.1% accuracy for the inner arm, earlobe, thumbnail, and vein, respectively. In contrast, the IR dataset faced challenges such as low feature correlation and overfitting, leading to an initial accuracy of 50%. To address this, features with low correlation were dropped, and Principal Component Analysis (PCA) was applied alongside SMOTE for resampling and Standard scaler for scaling. These adjustments improved accuracy to 70%. The results indicate that Raman spectroscopy is a more effective non-invasive method for diabetes detection compared to infrared spectroscopy.

***Keywords---*** Non-invasive detection, Diabetes, Raman spectroscopy, Infrared, Data augmentation, Random Forest, SMOTE

**Introduction**

Diabetes mellitus is a chronic condition characterized by elevated blood glucose levels due to the body's inability to produce or effectively utilize insulin. Early detection and proper management are essential to prevent serious complications. Traditional diagnostic methods rely on invasive blood sampling, which can be uncomfortable for patients and may discourage regular monitoring.

# In recent years, non-invasive techniques such as Raman spectroscopy and infrared (IR) have been investigated for diabetes detection. Raman spectroscopy utilizes vibrational energy shifts in molecules to provide a detailed biochemical fingerprint, while infrared imaging measures the absorption of infrared light by glucose in the blood. Both techniques offer promising potential for pain-free, rapid, and reliable diagnostics.

# This study compares the performance of Raman spectroscopy and infrared for non-invasive diabetes detection, with a focus on data augmentation, machine learning model performance, and overall diagnostic accuracy.

# **literature Review**

Introduction to Spectroscopy for Diabetes Detection: Non-invasive methods for detecting diabetes, particularly using Raman and Infrared spectroscopy, are gaining attention in biomedical research. These techniques can identify biochemical changes in tissues and fluids indicative of diabetes. This review focuses on two recent studies: one using Raman spectroscopy with machine learning, and another employing Attenuated Total Reflection (ATR) FTIR on saliva samples.

Raman Spectroscopy in Diabetes Screening:  
Raman spectroscopy has demonstrated promising potential for diabetes detection by analyzing molecular vibrations that change in the presence of the disease (Figure 1). In the study by Khan et al. (2023), titled *"Use of Raman Spectroscopy to Screen Diabetes Mellitus with Machine Learning Tools,"* Raman spectral data from biological samples were used to distinguish between diabetic and non-diabetic individuals. The study applied feature extraction and selection techniques to manage the high-dimensional spectral data and employed machine learning models, such as Support Vector Machines (SVM), to classify the samples. Khan et al. reported an accuracy of approximately 83% in their classification efforts. However, our study achieved higher classification accuracies with Raman spectroscopy, demonstrating the potential to further improve diabetes detection using advanced data augmentation and model tuning techniques.

Infrared Spectroscopy Using IR: focused on analyzing the biochemical composition of bodily fluids. In the study by Smith et al. (2023), *"Attenuated Total Reflection FTIR Dataset for Identification of Type 2 Diabetes Using Saliva,"* ATR-FTIR was used to analyze saliva samples. Spectral data were preprocessed to remove noise and baseline drifts, and features were extracted for potential model training. However, Smith et al.'s study was purely biomedical and did not apply any machine learning models to evaluate the performance of diabetes detection using these features. One key improvement in our work is the application of machine learning algorithms to this infrared dataset, allowing us to demonstrate better predictive performance through model training and validation, a step not taken in the previous study.

A diagram of a scatter light

Description automatically generatedComparative Insights and Gaps in the Literature: Both Raman and IR spectroscopy provide non-invasive methods for diabetes detection, but they differ in sample requirements, sensitivity, and specificity. Raman spectroscopy focuses on molecular vibrations, potentially offering higher specificity for certain biomarkers, while ATR-FTIR uses accessible saliva samples but faces challenges with overlapping spectral information. Both studies emphasize the need for optimizing data preprocessing and feature selection techniques for each spectral method.

Figure 1 shown the principle of Raman spectroscopy

Relevance to This Study: Our comparative study utilizes the datasets from Khan et al. (2023) and Smith et al. (2023) to apply machine learning models for diabetes detection. We expand upon their findings by employing advanced preprocessing techniques, data augmentation, and feature selection. Using the Raman spectroscopy dataset, which includes spectral information from various biological samples, and the IR saliva dataset, we apply and optimize machine learning models to achieve higher predictive accuracy than previously reported. This study demonstrates the effectiveness of combining these spectroscopy methods with machine learning for more accurate non-invasive diabetes diagnostics, surpassing the results from earlier works.

# **Materials and methods**

A) materials

1) Raman Spectroscopy Data: Raman spectroscopy data were collected from four anatomical locations: the inner arm, thumbnail, vein, and earlobe. These sites were selected based on their accessibility and potential to yield distinct spectral data, reflective of the biochemical changes associated with diabetes. Each sample initially contained thousands of features corresponding to Raman shifts, which required careful refinement. After preprocessing and feature selection, the dataset was reduced to 3,001 relevant features per sample, capturing the essential spectral signatures needed for classification.

The dataset was initially limited to 20 samples, necessitating data augmentation to prevent overfitting and improve model generalization. Synthetic Minority Over-sampling Technique (SMOTE) was applied to balance the classes and increase diversity. Additionally, random noise was introduced to simulate real-world measurement inconsistencies. After augmentation, the dataset expanded to 44 samples, enhancing the model’s robustness for classification tasks.

2) Infrared (IR) Spectroscopy Data: The IR spectroscopy dataset consisted of 540 samples, each with 3,740 features representing various spectral bands. Data preprocessing involved the removal of null values, irrelevant features, and standardization of the dataset. To simplify the problem, continuous glucose levels were transformed into a binary outcome, classifying samples based on whether glucose levels exceeded a threshold of 120 mg/dL.

Despite these efforts, the dataset exhibited challenges due to the high dimensionality and weak correlation between many features and the target variable. This noise and redundancy required further processing to improve model performance

B) methods

1) Data Augmentation and Preprocessing: For Raman spectroscopy, the main challenge was the limited number of samples, which risked overfitting and reduced the model's generalizability. SMOTE was applied to generate synthetic samples by interpolating between existing data points in the feature space. This technique, along with the addition of random noise to simulate measurement variability, effectively increased both the size and variability of the dataset. These augmentations resulted in cross-validation accuracies of 97% for the inner arm, 95.5% for the earlobe, and 91.1% for both the thumbnail and vein samples.

For the IR spectroscopy dataset, similar data augmentation techniques were employed, including SMOTE and noise addition. However, the high dimensionality of the IR data made it difficult to achieve significant gains. To address this, Principal Component Analysis (PCA) was applied to reduce the number of features and focus on the most informative ones. While this improved the dataset’s structure, the final classification accuracy remained at 70%, reflecting the challenges posed by the dataset's complexity and its weak feature-target correlations.

2) Model Training and Evaluation: For both Raman and IR datasets, a range of machine learning models were tested, including Random Forest, Support Vector Machine (SVM), and Logistic Regression. Ultimately, the Random Forest classifier was selected for its ability to handle high-dimensional data and capture complex feature interactions.

Hyperparameter tuning was performed using grid search combined with cross-validation to optimize model performance. The augmented datasets improved the model's ability to generalize, particularly for the Raman data, where cross-validation results demonstrated excellent predictive performance across multiple anatomical sites. Evaluation metrics such as accuracy, precision, recall, and F1-score were calculated to provide a comprehensive view of the model’s performance. While Raman data achieved exceptionally high accuracies (up to 97%), the IR dataset, even with dimensionality reduction, reached a maximum accuracy of 70%. This indicates that further feature engineering or advanced selection methods may be needed to fully leverage the potential of the IR spectroscopy data.

# **Results**

A) Raman Spectroscopy Data: The classification performance of the Raman spectroscopy data across different body parts yielded the following mean cross-validation accuracies: Inner Arm (97%), Ear Lobe (95.5%), Thumb Nail (91.1%), and Vein (91%) see Figure 2. These results demonstrate the potential of Raman spectroscopy in distinguishing relevant features for accurate classification across various sampling sites. By checking precision, recall and F1-score for each one we succeeded got the perfect score (Figure 3)

A graph of a bar chart

Description automatically generated with medium confidence Figure 2 shows the mean cross validation for each one

A graph of different colored rectangles

Description automatically generated Figure 3 shows precision, recall and F1 score

B) Infrared Data: In contrast, the classification results from the infrared data showed a decline in performance, even with data augmentation techniques. The Random Forest model exhibited lower accuracy, which can be attributed to the low correlation between the features in the dataset, indicating a limited discriminative power of the infrared features (see Figure 4).

A graph of a number of different colored bars

Description automatically generated with medium confidence

Figure 4 shows precision, recall and F1 score for each class

# **discussion**

his study presents a comparative analysis of two non-invasive spectroscopy techniques—Raman spectroscopy and infrared (IR) —for diabetes detection, focusing on their spectral characteristics, data processing approaches, and classification outcomes using machine learning.

Raman Spectroscopy: The results demonstrate that Raman spectroscopy is highly effective for diabetes detection across multiple anatomical sites. The application of machine learning, particularly the Random Forest classifier, yielded excellent cross-validation accuracies, ranging from 91.1% to 97%. These findings align with existing literature but show improved performance compared to previous studies, such as Khan et al. (2023), where accuracy reached only 83%. The augmentation techniques, including SMOTE and noise addition, played a crucial role in overcoming the challenge of small sample size and enhancing model robustness. The strong results from this study suggest that Raman spectroscopy, combined with advanced data augmentation and machine learning, holds significant promise as a non-invasive diagnostic tool for diabetes.

Infrared Spectroscopy: On the other hand, IR spectroscopy faced more significant challenges. Initial classification accuracy was low, primarily due to the high dimensionality of the data and low feature correlation with the target variable. However, by implementing dimensionality reduction through Principal Component Analysis (PCA) and applying SMOTE, we improved accuracy from an initial 50% to 70%. It is important to note that previous studies using IR spectroscopy, such as Smith et al. (2023), did not employ machine learning models and instead focused on the biomedical aspects of the spectral data. Our study extends their work by applying machine learning techniques, offering a novel approach to improve predictive accuracy. Despite this improvement, the IR dataset’s performance still lagged that of the Raman data, likely due to the complexity of the spectral information and the limitations in feature relevance for classification

Comparative Insights: The comparison highlights that Raman spectroscopy is superior to IR spectroscopy in terms of both data quality and predictive accuracy. While IR spectroscopy shows potential, especially with further refinement in feature selection and dimensionality reduction, Raman spectroscopy consistently outperforms it, suggesting that it is a more viable option for non-invasive diabetes screening. The differences between the two methods emphasize the importance of optimizing data processing and feature engineering for each technique.

**conclusion**

This comparative study establishes Raman spectroscopy as a more effective non-invasive method for diabetes detection compared to infrared (IR) spectroscopy. Raman spectroscopy's superior accuracy and reliability make it a promising candidate for integration into clinical practice as a pain-free diagnostic tool. Its ability to detect subtle biochemical changes provides a significant advantage in monitoring and diagnosing diabetes.

However, while IR spectroscopy presents potential for non-invasive diagnostics, its current limitations, including low feature correlation and high dimensionality, hinder its effectiveness. Future research should concentrate on advancing IR spectroscopy techniques to improve performance and address these challenges.

To maximize the potential of Raman spectroscopy, it is crucial to validate its effectiveness in larger and more diverse populations. Continued research in this area will help to solidify Raman spectroscopy's role in diabetes diagnostics and potentially expand its applications in other medical fields.

**References**

[1] Guevara, E., Torres-Galván, J. C., Ramírez-Elías, M. G., Luevano-Contreras, C., & González, F. J. (2018). Use of Raman spectroscopy to screen diabetes mellitus with machine learning tools. Biomedical Optics Express

[2] Lazaro, J. A., Petrie, M. S., & Stewart, J. T. (2014). Raman spectroscopy as a promising tool for noninvasive point-of-care glucose monitoring. Journal of Diabetes Science and Technology, https://doi.org/10.1177/1932296814543104

[3] Kumar, S., Sonkar, K., & Singh, R. (2022). Attenuated total reflection FTIR dataset for identification of type 2 diabetes using saliva. Computational and Structural Biotechnology Journal https://doi.org/10.1016/j.csbj.2022.08.038

[4] Khan, A. A., & Kumar, P. (2023). Infrared spectroscopy technique: An alternative technology for diabetes diagnosis. Biomedical Signal Processing and Control, 87, 105246. https://doi.org/10.1016/j.bspc.2023.105246

[5] Patil, K., & Rao, S. (2023). Use of Raman spectroscopy to diagnose diabetes with SVM. In Advances in Machine Learning for Healthcare Applications (pp. 123-134). https://doi.org/10.1007/978-3-031-28790-9\_6