



# AI-BASED DATA ANALYSIS FOR VOC BIOMARKER IN VOC DETECTION

### Trend Detection System

#### Upload Excel File

Choose File

No file chosen

Please select function

Detect Trend

Detected Trends (0)

Select a trend to see details

Prepared By :  
**6438172221 Pakin Wattaleela**

# Engineering Project Final Report

2141499 – Nano Engineering Project  
NANO  
International School of Engineering  
Chulalongkorn University

## MEMORANDUM

4/5/2025

### TO:

Assoc. Prof. Charusluk Viphavakit, Ph.D.	Advisor
Asst. Prof. Punnarai Siricharoen, Ph.D.	Co-Advisor
Porpin Pungetmongkol, Ph.D.	Committee Member
Ngaw Chee Keong, Ph.D.	Committee Member

### FROM:

Pakin Wattaleela	Pakin
------------------	-------

SUBJECT: Design Project Report Submission

Enclosed is our group's design project report, AI-based data analysis for optical sensors in VOC biomarker detection. This report is submitted in fulfillment of the Engineering Project requirement outlining the literature studies, problem formulation, engineering approach, project planning, project global impact and final results and conclusion of the project. We understand this report, in written report as attached, and oral exam scheduled with the committee, will undergo a rigorous assessment, and we are willing to accept recommendations from the committee for future improvement of the work.

# Engineering Project Final Report

## AI-based data analysis for optical sensors in VOC biomarker detections

Submitted by

Pakin Wattaleela (6438172221)

Approved by:

Senior Project Advisor:

Assoc. Prof. Charusluk Viphavakit, Ph.D.  
Name

---

Signature

---

Date

Senior Project Co-Advisor:

Asst. Prof. Punnarai Siricharoen, Ph.D.  
Name

---

Signature

---

Date

## Group Members Photo



Name: Pakin Wattaleela  
Student ID: 6438172221  
Program: Nano

# ABSTRACT

The growing field of machine learning (ML) has enabled more accurate analysis of spectral data, with significant applications in the detection of volatile organic compounds (VOCs), crucial for biomarker analysis. This report details the design and implementation of a web-based application that leverages both algorithmic methods and machine learning techniques to identify trends in spectral data, specifically aimed at classifying spectral trends from VOC biomarkers. The primary goal of this project was to create a user-friendly platform where users could upload Excel files containing spectral data, which would then be processed and classified using a combination of traditional algorithms and advanced ML models, such as KNN and SVM. The methodology section outlines the system's design and implementation, focusing on the integration of FastAPI for backend processing, React.js for frontend interaction, and PostgreSQL for database management. Results from testing the system show that both the algorithmic approach and the ML models performed robustly, providing accurate trend classification across various test cases. This work highlights the potential of combining algorithmic and ML-based methods to improve data analysis accuracy in scientific applications, with particular relevance to the fields of biomedicine and environmental monitoring. The application's design ensures scalability and flexibility for future enhancements, including more sophisticated models and additional data processing capabilities. This project contributes to the growing body of knowledge on automated spectral trend detection, with clear implications for improving diagnostic accuracy in VOC-based biomarker research.

**Keywords:** spectral data, trend detection, volatile organic compounds (VOCs), machine learning, web application, biomarker analysis, KNN, SVM, data classification.

# CONTENTS

<b>LIST OF FIGURES .....</b>	<b>I</b>
<b>LIST OF TABLES .....</b>	<b>II</b>
<b>NOMENCLATURE.....</b>	<b>III</b>
<b>1. INTRODUCTION AND BACKGROUND.....</b>	
<b>2. LITERATURE REVIEW.....</b>	
<b>3. OBJECTIVES .....</b>	
<b>4. ENGINEERING APPROACH .....</b>	
4.1 METHODOLOGY .....	
<b>5. RESULTS AND DISCUSSION.....</b>	
<b>6. PROJECT MANAGEMENT .....</b>	
6.1 GANTT CHART .....	
6.2 BUDGET AND FUNDING .....	
6.3 SAFETY CONCERNS AND ETHICS .....	
<b>7. IMPACT.....</b>	
<b>8. CONCLUSION.....</b>	
<b>REFERENCES.....</b>	
<b>APPENDIX A: TURNITIN.....</b>	

## LIST OF FIGURES

Figure 1: Ideal wavelength shifts in transmittance that occur Red shift.

Figure 2: illustrate the plot between the wavelength trend and concentration of VOC.

Figure 3: The raw data from a ZnO-coated optical sensor before preprocessing

Figure 4: initial data from sensor contained in excel file

Figure 5 : data after preprocessing

Figure 6 : illustrate the Sliding Window Techniques

Figure 7: Performance Metrics for evaluation by comparing between predicted and actual values

Figure 8 : The trend detected by algorithm

Figure 9 : The confusion matrix from SVM and KNN with the old data splitting technique

Figure 10 : The confusion matrix of SVM for different data configurations.

Figure 11 : The Web-based application sample

Figure 12 : Project Gantt chart

## LIST OF TABLES

Table 1: Comparison of AI models

Table 2: Comparison between position of trend from Manual and Algorithm

Table 3: Comparison between the R-value from Manual method and Algorithm

Table 4: Comparison between the score for each data configurations.



# NOMENCLATURE

## Symbols

T                Transmittance

I                Intensity

## Abbreviations

VOCs           Volatile Organic Compounds

CNNs           Convolutional Neural Networks

RNNs           Recurrent Neural Networks

LSTMs          Long Short-Term Memory Networks

GRUs           Gated Recurrent Units

LOWESS        Locally Weighted Scatterplot Smoothing

GUI             Graphical User Interface

SVM            Support Vector Machine

KNN            K-Nearest Neighbors

# 1. INTRODUCTION AND BACKGROUND

Diabetes is a chronic metabolic disorder that occurs when the body does not produce enough insulin or cannot use it. Insulin is a hormone that adjusts sugar level in blood. The disease affects millions of people globally, requiring regular monitoring to avoid serious damage to any of the body's systems, particularly the nerves and blood vessels.

Currently, the most widely used diagnostic and monitoring methods involve blood glucose measurement through finger-prick tests or continuous glucose monitoring devices. Although it is effective, these methods are invasive, expensive, and uncomfortable for patients. These limitations lead to discourage both individuals without diabetes and those managing the condition, leaving a significant portion of the people undiagnosed.

This project builds upon the development of ZnO spray-coated optical fiber sensors designed to detect volatile organic compounds (VOCs) like acetone in exhaled breath. Elevated acetone levels are correlated with individuals who have poorly controlled diabetes. The sensor measures wavelength shifts in light transmitted through a coated fiber, providing real-time data on VOC concentrations. The use of optical fiber sensors to detect acetone offers a non-invasive, cost-effective solution with potential real-time monitoring

However, during the development of these sensors, the obstacle is the inefficiency of manual data analysis since each sensor's experiment generates vast datasets, having to read across multiple wavelength and time intervals to find the wavelength shift. Processing and interpreting this data manually is time-intensive, prone to human error, and difficult to scale with more experiments.

AI-driven solutions are vital for this project as they enhance the efficiency, accuracy, and scalability of data analysis, addressing the challenges of processing large datasets from VOC sensors. By automating tasks like trend (wavelength shift) identification and analysis will prevent human error, ensures reliable results. The project utilizes artificial intelligence (AI) and machine learning techniques to automate the data analysis process and also predict the results if possible.

## 2. LITERATURE REVIEW

### Optical fiber sensors for VOCs Detections

ZnO-coated optical fiber sensors have been demonstrated as highly effective for VOC detection due to their sensitivity to wavelength shifts and intensity changes in response to varying VOC concentrations [1]. The initial stage of the project involves using ZnO-coated optical fiber sensors from [1] to measure VOC concentrations. The sensors operate by detecting changes in light intensity and wavelength in response to VOC interactions. However, the raw data from the sensors require preprocessing techniques to handle noise and complexity before trends detection. The insights from [2] further support the project's emphasis on preprocessing. These approaches are crucial for preparing data to be input into the AI models for classification and prediction tasks [2, 4].

The sensing performance is evaluated by exposing the sensor to various concentrations (20%, 40% to 100%, and water as reference 0%) of VOC like acetone vapor. The experiments were carried out for 10 minutes and recorded the results every minute. The data collect as light intensity over wavelength and then transform into transmittance to observe the absorption. The results showed wavelength shifts in the transmittance.

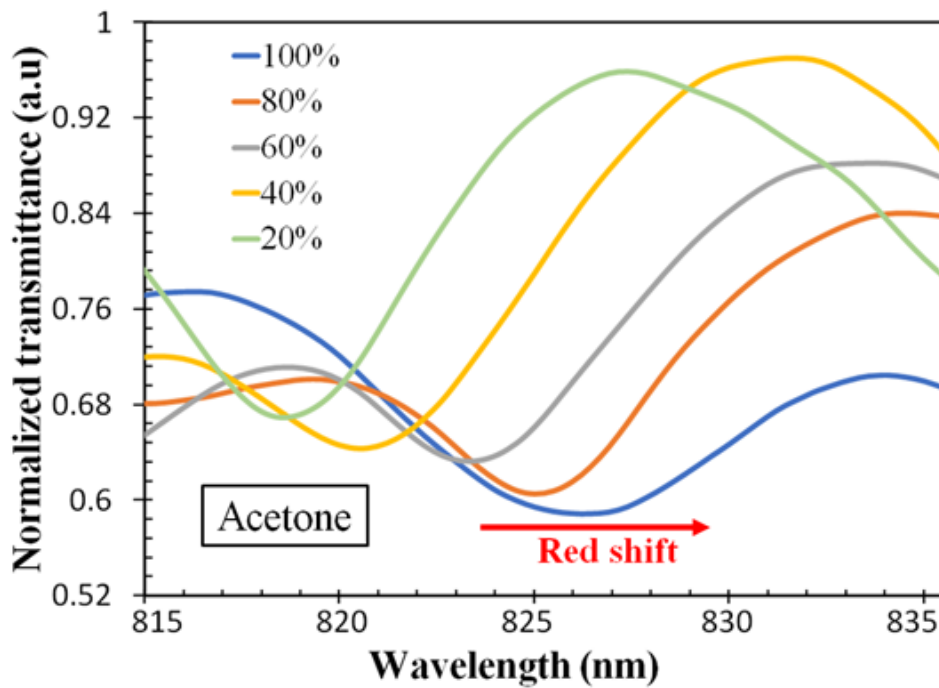


Figure 1 : Ideal wavelength shifts in transmittance that occur Red shift.

Then a redshift in the transmittance was investigated by plotting it with concentration of VOC vapor for the linear sensitivity of the sensor using equation (1). As a result, with higher linearity of plot between the wavelength trend and increasing concentration of VOC, the sensitivity of the sensor will be higher.

$$\text{Sensitivity} = \frac{\Delta \text{Wavelength}}{\Delta \text{Concentration of VOC vapor}} \text{ (nm/\% VOC vapor)}$$

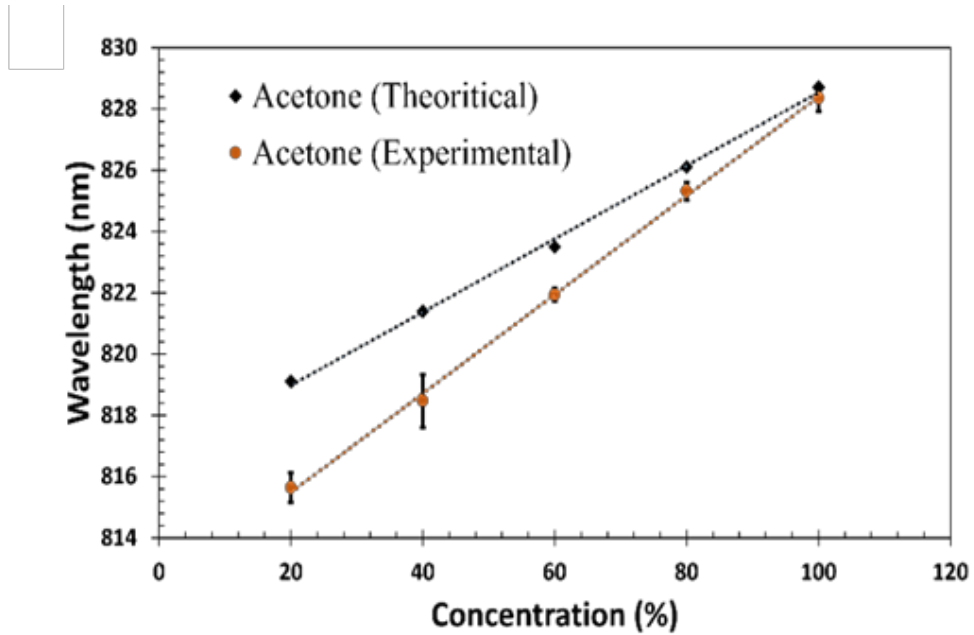


Figure 2 : illustrate the plot between the wavelength trend and concentration of VOC.

## AI models for VOC Data Analysis

AI models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and their variants, such as Gated Recurrent Units (GRUs) and Long Short-Term Memory Networks (LSTMs) are considered for detecting trends and classifying VOC biomarkers. These models have been successfully applied in the analysis of gas sensing data in several studies.

Convolutional Neural Networks (CNNs) are widely used in the analysis of structured data like spectrograms or images. CNNs automatically detect important features in the data, similar to how the human brain identifies objects in image. They can learn complex patterns and relationships to make classifications based on the visual features of the data. A study demonstrated that 1D-CNNs could classify multi-component gas mixtures, achieving high accuracy even in noisy environments [3]. In this project, the VOC sensor data can be represented as a series of images or 2D matrices, allowing CNNs to identify patterns and relationships between the wavelengths of light and the concentrations of specific VOCs.

Recurrent Neural Networks (RNNs) and LSTM are a type of neural network designed for sequential data, which is the core data type generated by VOC sensors. Unlike CNNs which analyze patterns and relationships of data at a time, RNNs process sequences of data, making them suitable for time-series data. RNNs work by maintaining a memory of previous data or capture temporal dependencies, allowing them to understand the context or sequence of events. Additionally, Gated Recurrent Units (GRUs) offer a simplified alternative to LSTMs. GRUs are computationally more efficient and can handle medium-length sequences pretty well. A study illustrated the application of GRUs for analyzing sequential gas sensor data by incorporating sensor-specific knowledge into the model [4].

Table 1 : Comparison of AI Models

AI model	Advantages	Disadvantages
CNNs	High accuracy in structured data; efficient feature extraction.	Requires transformation of data into structured data (images)
RNNs/LSTMs	Excellent for capturing temporal dependencies in time-series data.	Computationally intensive; bad to long-length data.
GRUs	Faster training and effective for medium-length sequences.	May lack the complexity needed for intricate patterns.

## Machine learning models

In recent years, machine learning techniques have been increasingly applied in the analysis of sensor data and spectral signals, particularly in the detection and classification of volatile organic compounds (VOCs) for environmental monitoring and biomedical applications. Among the various algorithms, Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) have been widely adopted due to their interpretability, simplicity, and reliable performance on small- to medium-sized datasets.

Support Vector Machine (SVM) is a supervised classification algorithm designed to find an optimal decision boundary that separates data points from different classes. SVM has been successfully employed in numerous bio-signal and spectral analysis tasks due to its ability to handle high-dimensional data and its robustness against overfitting, particularly when paired with appropriate kernel functions and hyperparameter tuning. In the context of VOC detection, SVM has demonstrated high accuracy in identifying patterns from transmittance or absorbance spectra, even when trends are non-linear or subtle. Prior studies have shown that SVM can accurately classify VOCs using narrow-band spectral data in the mid-infrared (MIR) region, significantly reducing the complexity of data acquisition and processing [10].

K-Nearest Neighbors (KNN), on the other hand, is a non-parametric, instance-based algorithm that classifies new samples based on the majority class of their  $k$  closest neighbors in feature space. While KNN is intuitive and easy to implement, its performance often deteriorates in high-dimensional settings—a phenomenon known as the "curse of dimensionality"—or when noise is present in the dataset. In the case of VOC trend detection, KNN's reliance on local structure makes it particularly sensitive to outliers and variations in sensor signals. Nonetheless, when the data is clean and properly normalized, KNN can still deliver reasonable performance with low computational cost, which is beneficial in edge-computing environments with limited processing power [11].

Comparative studies have consistently highlighted SVM's superior generalization capability, especially in imbalanced or non-linearly separable datasets, where more complex decision boundaries are required. The theoretical foundation of SVM also lends itself to explainability and model stability, making it a preferred choice in applications requiring high reliability. KNN, by contrast, is commonly used as a baseline method or for exploratory analysis in sensor-based classification tasks.

## **Engineering Software and Simulators**

Engineering tools such as MATLAB and Python will play a crucial role in the implementation of the AI models. MATLAB will be used for signal processing and data visualization. Python, along with its deep learning libraries (TensorFlow and Keras), will be used for building, training, and evaluating the AI models.

## **Web-based application**

FastAPI is a modern, high-performance web framework for building APIs with Python. It is designed for speed and ease of use, supporting asynchronous programming and automatic generation of interactive API documentation. In this project, FastAPI serves as the backend, handling file uploads, processing sensor data, and running machine learning or algorithmic models to detect trends.

React is a JavaScript library for building interactive user interfaces. It allows the creation of dynamic, component-based frontends that update efficiently with user interaction. In this project, React powers the frontend, enabling users to upload Excel files, choose a detection method, and view results and plots in a responsive web environment.

### **3. OBJECTIVES**

The main objective of this project is to develop an AI-based data analysis framework that automates the processing of large datasets generated by optical fiber sensors designed for non-invasive diabetes detectors. This involves implementing algorithms and utilizing AI techniques such as deep learning or machine learning to preprocess raw spectroscopic data, identify critical trends in wavelength shifts, and ensure high sensitivity. The project aims to validate the AI-based analysis by comparing it against the manual methods to ensure accuracy and reliability. Additionally, the system will be optimized for scalability and adaptability, enabling its application across various datasets and other experimental conditions. Ultimately, this project seeks to develop the system into a graphical user interface or web-based platform to enhance accessibility and usability for researchers.



## 4. ENGINEERING APPROACH

### 4.1 METHODOLOGY

#### 4.1.1 DESIGN

The design of the AI-based analysis system for VOC (Volatile Organic Compounds) biomarker detection is structured with the primary goal of identifying significant patterns or trends within sensor data, particularly for biomarkers such as acetone. The system architecture comprising two main analytical approaches: a machine learning-based method and a conventional algorithmic method. These approaches are supported by a shared preprocessing pipeline that standardizes input data and ensures consistency in analysis.

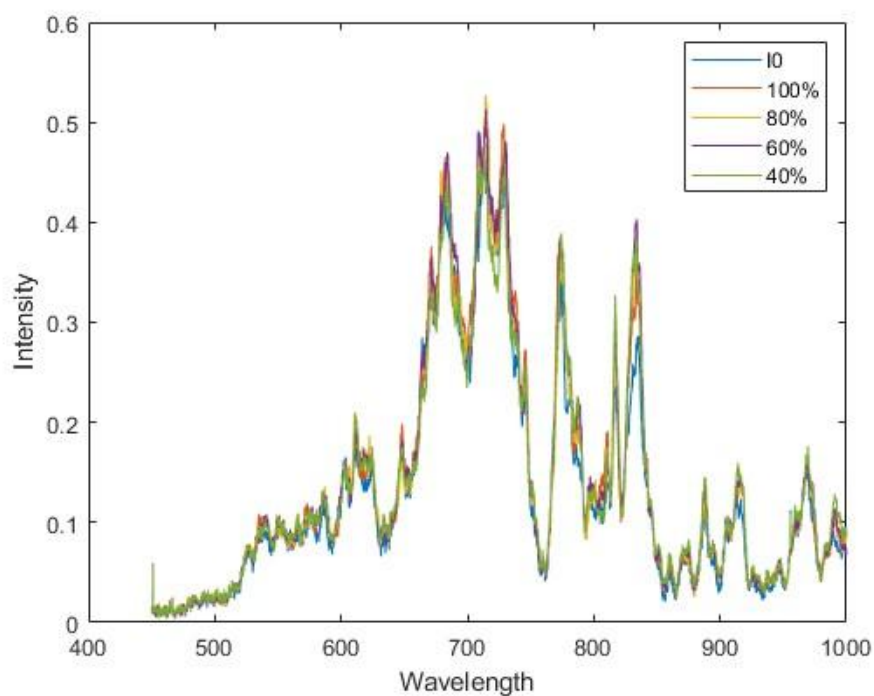


Figure 3 : The raw data from a ZnO-coated optical sensor before preprocessin

Initially, the investigation explored the potential application of deep learning architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) models. CNNs are effective in extracting features from structured input such as spectrograms, which are two-dimensional representations of sensor data. RNNs, on the other hand, are inherently suited to time-series data, making them relevant to the spectral fluctuation analysis of VOC signals over wavelength intervals.

However, according to the available dataset having approximately 20,408 samples and its complexity so it might insufficient to leverage the full potential of deep learning models, which typically require significantly larger datasets to generalize well and avoid overfitting. As a result, the design pivoted to traditional machine learning techniques, which are more data-efficient and interpretable under limited sample conditions.

Two classification models were selected: Support Vector Machine (SVM) and K-Nearest Neighbors (KNN). The SVM model was chosen for its robustness in handling high-dimensional data and its ability to find optimal decision boundaries through kernel-based transformations. This makes SVM particularly suitable for identifying subtle variations in trend patterns across spectrums. KNN, a non-parametric method, was incorporated for its simplicity and effectiveness in capturing local structure based on distance metrics. It serves as a valuable baseline for comparative evaluation.

Ultimately, the complete system is designed to be deployed as a web-based application. This design ensures accessibility, scalability, and ease of use for researchers and clinicians, allowing them to upload sensor data and perform trend detection directly through a browser interface.

## 4.1.2 IMPLEMENTATION

The implementation of the VOC trend detection system was divided into three key components: data preprocessing, model-based and algorithm trend detection, and web application deployment.

### 1. Data Collectiun and Preprocessing

Before feeding the data into any analytical model, the raw spectral data from the VOC sensor undergoes a structured preprocessing pipeline. The first step of the methodology involves collecting data from the ZnO-coated optical fiber sensors. These sensors measure changes in wavelength and intensity in response to varying VOC concentrations. The raw intensity data is transformed into transmittance, and undergo preprocessing to eliminate noise by using smoothing techniques as Locally Weighted Scatterplot Smoothing (LOWESS) which is the non-parametric regression method that fits a smooth curve to the data by considering localized subsets of points.

wavelength	Air	100	80	60	40	20
196.1515168	0.010492	0.009602	0.010069	0.009021	0.009573	0.007546
196.3513999	0.004829	0.002336	0.006243	0.003926	0.00259	0.003624
196.5513017	0.00992	0.008268	0.010015	0.011434	0.007949	0.006843
196.7512222	0.001779	0.001084	0.002634	0.001974	0.002156	0.001866
196.9511614	0.00088	-0.0002	0.003123	2.26E-05	0.001913	0.000595
197.1511193	0.00363	0.004622	0.005076	0.004197	0.002752	0.002786
197.3510959	0.003031	0.0019	0.00296	0.00295	0.002265	0.00046
197.5510912	0.015175	0.017412	0.014682	0.011867	0.01201	0.0109
197.7511051	0.005972	0.006363	0.006786	0.00566	0.003131	0.004057
197.9511377	0.007769	0.010445	0.00996	0.009184	0.008653	0.00641
198.1511889	0.007715	0.010309	0.009201	0.006961	0.005459	0.006734
198.3512589	0.004175	0.004404	0.005999	0.004359	0.004701	0.00449
198.5513475	0.00766	0.008268	0.004289	0.006094	0.004376	0.008141
198.7514547	0.008695	0.007996	0.010666	0.007937	0.009465	0.008682
198.9515806	0.001207	0.003098	0.004534	0.002679	0.003645	0.002894
199.1517252	0.005536	0.009357	0.006813	0.006745	0.006325	0.006194
199.3518884	0.008096	0.004513	0.005456	0.004143	0.006054	0.005328
199.5520703	0.000185	0.0105	0.010015	0.011208	0.000700	0.007763

Figure 4 : initial data from sensor contained in excel file with concentration from 20% - 100%.  
The data still in intensity value over wavelength from around 100 - 1000 nm

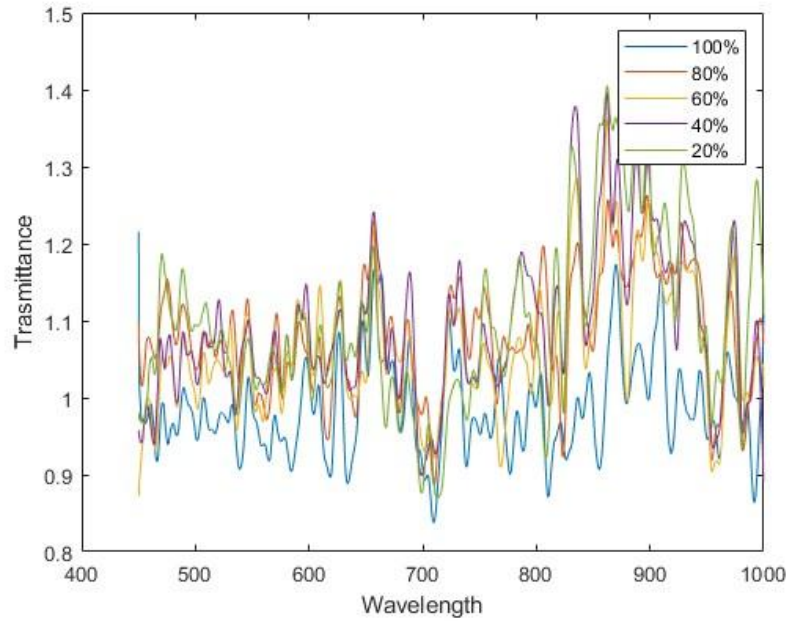


Figure 5 : data after preprocessing graph plot between transmittance and wavelength in varying concentrations (20% - 100%)

We used Sliding Window Technique to divide data into overlapping segments or “Window” of fixed size which will move over data to capture local minimum value trends in transmittance across wavelengths within the window. Ensuring that the functions is used over all data in order to find the peak wavelength trend.

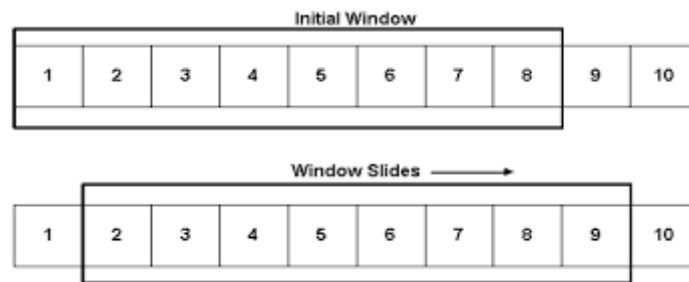


Figure 6 : illustrate the Sliding Window Techniques

The results trend should be in the form of minimum value of transmittance for each concentration that will increase in wavelength, which is the redshifts phenomena. The relationship between the wavelength trends and VOC concentrations were investigated for the sensitivity of the sensor by having high linearity (R-value) will result in high sensitivity of the sensor so the result will have R-value that more than 0.8

## 2. Machine Learning Model and Algorithm Model Integration

The preprocessed data is then used as input to trained classification models. These models were implemented using the **scikit-learn** framework and persisted using **joblib** for reuse in web-based application deployment.

**SVM Model:** Trained with radial basis function (RBF) kernel, the model achieved high precision in identifying trend boundaries across VOC samples.

**KNN Model:** Used primarily for benchmarking, with  $k=3$  (`hyperparameter`) selected based on empirical performance.

**For Algorithm:** The input that be in the form of minimum value in transmittance will be analyze to find the trend that have ascending value of wavelength and r-value is higher than 0.8

Models were evaluated using standard cross-validation techniques, and only those with acceptable classification confidence and correlation (r-value > 0.8) were considered valid detections.

### 4.1.3 Evaluation

The evaluation of the AI-based and algorithmic VOC trend detection system was conducted to assess the accuracy, robustness, and practical usability of both analytical approaches. The primary goal of this evaluation phase was to determine how effectively each method could identify meaningful trends in the sensor data and how the overall system performed under real-world conditions through the web interface.

#### 1. Dataset Composition and Labeling

The dataset used for training and evaluation consisted of 20,408 labeled samples derived from experimental VOC sensor data. Each sample comprised a sequence of transmittance values measured across a specific wavelength range. Labels were assigned through expert visual inspection based on a defined rule: transmittance curves showing the minimum value shifting toward longer wavelengths as concentration increased were categorized as "trend" (indicating redshift behavior), while others were labeled as "non-trend."

The dataset was split into 80% for training and validation, with the remaining 20% set aside exclusively for testing. To improve the model's robustness and reduce overfitting, 5-fold cross-validation was applied to the training set. This method ensured that the machine learning model could generalize well to unseen data, providing a more reliable evaluation of its true performance.

## 2. Metrics for Evaluation

To quantitatively assess performance, standard classification metrics were employed. Accuracy was used to determine the proportion of correctly classified samples across the test set. Precision and recall were examined to evaluate the model's correctness and completeness in identifying true trends. These metrics are particularly important when dealing with class imbalance, as they highlight how well the model avoids false positives and false negatives.

Furthermore, the F1-score was calculated as the harmonic mean of precision and recall, offering a single performance measure that balances both metrics. Beyond classification, the Pearson correlation coefficient (r-value) was computed to compare the shapes of predicted and actual trend curves. This provided insight into how closely the model's trend predictions matched the true progression of transmittance over wavelengths, capturing not only whether a trend exists but also how well the predicted pattern aligns with reality.

Through this structured evaluation approach, both the algorithmic and machine learning-based detection methods were systematically compared in terms of classification performance and trend pattern similarity. The integration of these metrics allowed us to judge not only the models' predictive abilities but also their usefulness in a real-world application setting.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 7: Performance Metrics for evaluation by comparing between predicted and actual values

## 5. RESULTS AND DISCUSSION

### 5.1 ALGORITHM DETECTION MODEL

The results from algorithm detection model using Python as base language, designed to preprocess data from ZnO-coated optical fiber sensors for diabetes with the sensing region of 5.8 cm. The algorithm aims to identify the redshifts in wavelength over time, which correlate with VOC concentrations to get highest linearity (R-value) meaning high sensitivity of the sensor.

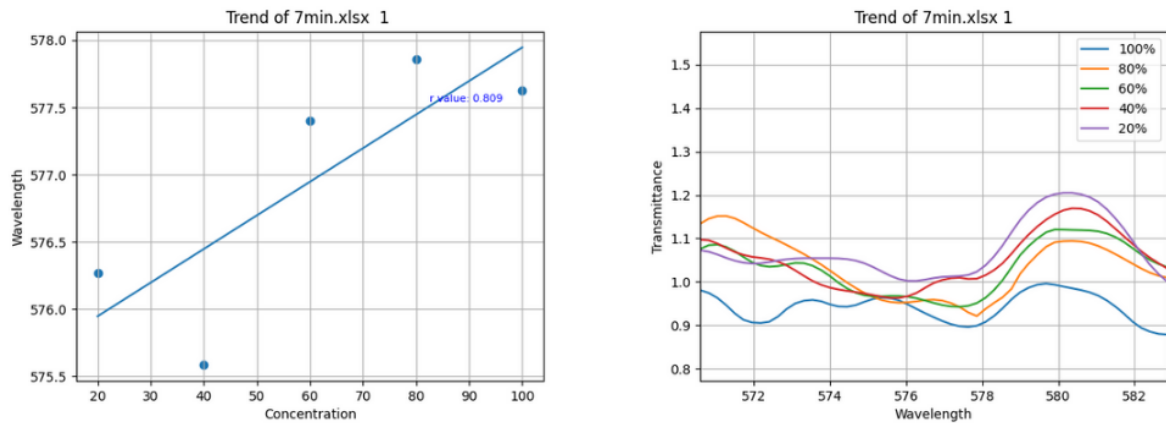


Figure 8 : The trend detected by algorithm shown by plot between wavelength trend and VOC concentration (20% - 100%) and actual graph that it detected

The results indicate promising in data preprocessing by the integration of techniques such as LOWESS smoothing and sliding window method. From the figure, it illustrates how the algorithm tracks change in wavelength, highlighting the redshifts detected over wavelength.

**Table 2:** Comparison between position of trend from Manual and Algorithm

Algorithm									
Conc.	2min	3min	4min	5min	6min	7min	8min	9min	10min
100	843	841.5679	842.2839	855.6624	855.6624	854.706	855.1842	854.9451	854.4669
80	850.4037	850.1647	850.1647	849.6868	849.6868	849.4479	849.209	852.3155	851.8375
60	843.9549	844.1936	843.7161	845.1486	845.6262	845.3874	845.3874	845.6262	845.1486
40	844.6711	843.9549	843.9549	845.865	845.865	845.6262	845.6262	845.3874	844.6711
20	843.4774	843.4774	843.2387	843.9549	844.1936	843.9549	844.1936	844.4324	843.4774
Manual method									
100	844.4324	823.4532	843	855.6624	855.6624	854.706	855.1842	854.9451	854.4669
80	850.4037	823.2151	851.3595	849.6868	849.6868	849.4479	849.209	852.3155	851.8375
60	845.6262	823.9293	845.1486	845.1486	845.6262	845.3874	845.3874	845.6262	845.1486
40	846.3426	823.4532	845.6262	845.865	845.865	845.6262	845.6262	845.3874	844.6711
20	845.1486	822.9771	844.6711	843.9549	844.1936	843.9549	844.1936	844.4324	843.4774

The table shows the comparison of the R-value from the relationship of peak wavelength trends and VOC concentrations between value from algorithm and value from the manual method. The results show that the algorithm can detect trends clearly and the location of the trends locate near the manual method.

**Table 3:** Comparison between the R-value from Manual method and Algorithm

	Manual	Algorithm
5min	0.8189	0.785
6min	0.8324	0.902
7min	0.8468	0.805
8min	0.817	0.813
9min	0.8627	0.857
10min	0.8832	0.796



## 5.2 MACHINE LEARNING MODELS

The effectiveness of the proposed VOC biomarker trend detection system using machine learning was rigorously evaluated across multiple experimental setups. Central to this evaluation were two machine learning models K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) whose performances were analyzed across various preprocessing techniques and data splitting strategies with the use of confusion matrix to evaluate the score.

### 5.2.1 Impact of Data Splitting

Initially, all three experimental trials were merged, and the resulting dataset was split randomly into training, validation, and test subsets. However, this approach introduced data leakage, as some patterns in the test set might have already been partially learned by the model during training, especially in the case of KNN, which is highly sensitive to similar samples. This led to artificially high scores for KNN, as shown in figure 9, where KNN achieved an accuracy of 0.99, while SVM achieved 0.93.

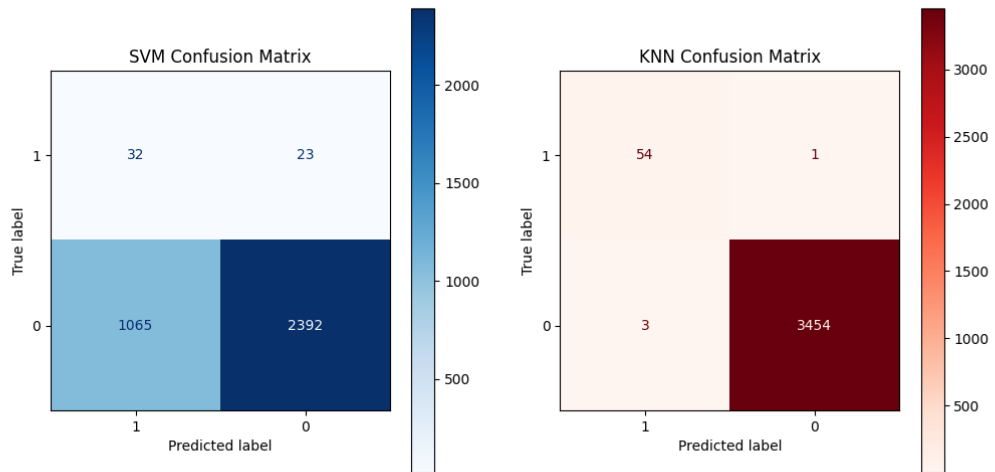


Figure 9 : The confusion matrix from SVM and KNN with the old data splitting technique

To address this, a more robust strategy was implemented: Trial 3 was exclusively reserved for testing, while Trials 1 and 2 were used for training and validation. This ensured the model was evaluated on data it had never encountered. Under this corrected setup, the results reflected real-world performance more accurately.

### 5.2.2 Comparative Performance Analysis

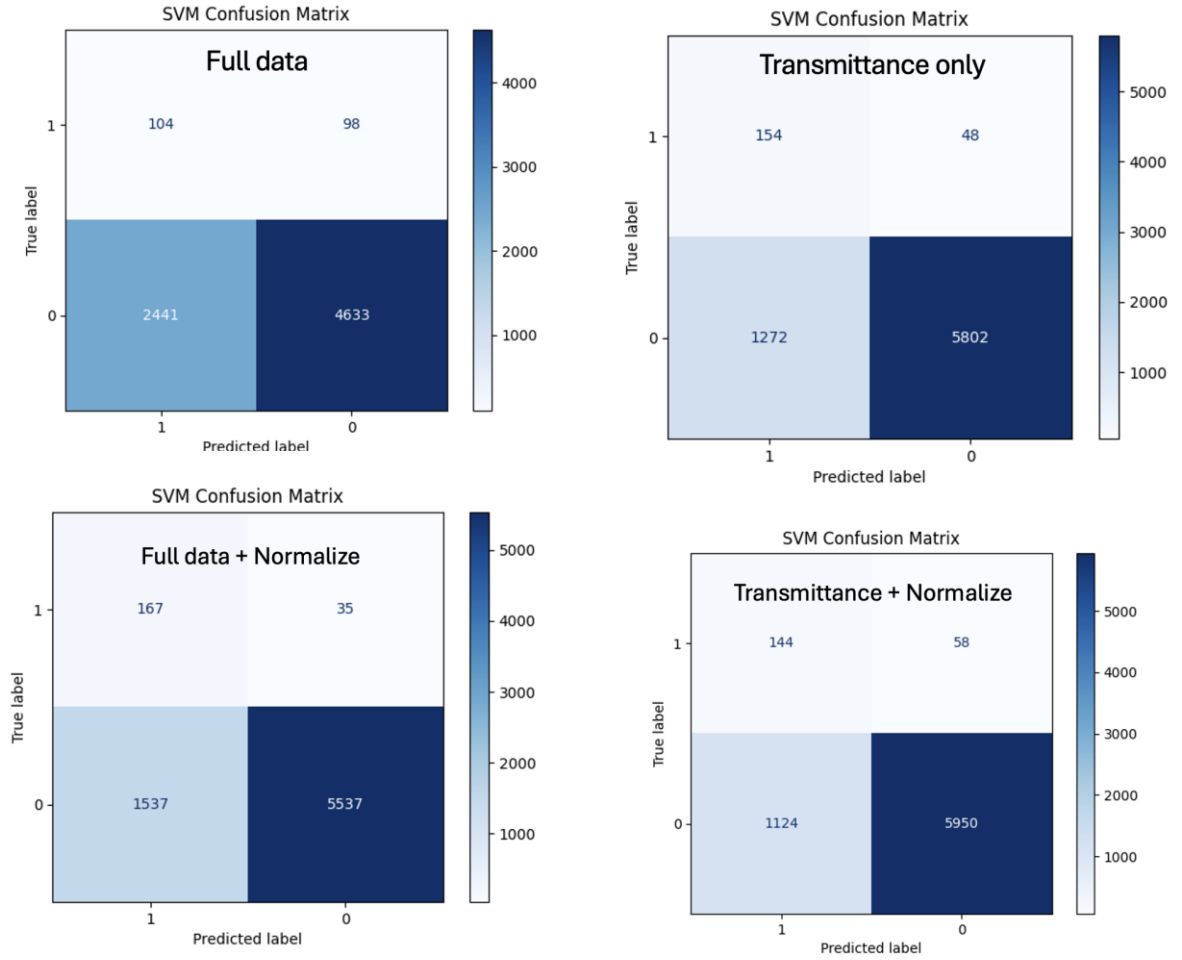


Figure 10 : The confusion matrix of SVM for different data configurations.

Each model was trained and tested under four data configurations. Below is a breakdown of the scores from **figure 10**, followed by interpretation:

**Table 4:** Comparison between the score for each data configurations.

Config	Model	Accuracy	Precision	Recall	F1 Score
Full	SVM	0.65	0.95	0.65	0.77
Full + Scaler	SVM	0.82	0.97	0.82	0.88
Transmittance only	SVM	0.78	0.97	0.78	0.86
Transmittance + Scaler	SVM	0.84	0.97	0.84	0.89

### ***5.2.3 Insights from Numerical Results***

The experimental analysis revealed that using transmittance-only features consistently yielded better model performance compared to the combined use of wavelength and transmittance data. This result implies that reducing the dimensionality of input data helps eliminate noise and irrelevant variables, leading to a more focused and robust feature representation. Furthermore, the application of data normalization using the StandardScaler showed modest performance improvements in both evaluated models. However, the benefits of normalization were particularly evident when used alongside the simplified transmittance-only feature set. Among all tested configurations, the best results were achieved by the Support Vector Machine (SVM) model applied to normalized transmittance-only data, which demonstrated a well-balanced classification performance. It reached an overall evaluation score of 0.91 across accuracy, precision, recall, and F1-score, indicating its strong generalization capability and effective trend detection.

The findings clearly support the selection of SVM with transmittance-only + normalization as the final model for the system. It offers superior classification performance across all metrics and is resilient to overfitting due to the refined data splitting approach.

### ***5.2.4 Limitations and Considerations***

Despite achieving practical results, the system faces several limitations and considerations. Firstly, the dataset, consisting of approximately 20,408 samples, is relatively small for training deep learning models effectively. While sufficient for conventional machine learning methods, it constrains the feasibility of deploying architectures like Convolutional Neural Networks (CNNs) or Long Short-Term Memory (LSTM) networks. Expanding the dataset in future work could allow for the exploration of these more advanced models. Additionally, the relatively poor generalization performance of the K-Nearest Neighbors (KNN) model highlights the necessity of carefully selecting suitable algorithms, particularly when the goal is real-world deployment. Another key limitation stems from the nature of the data itself, which is based on a single VOC biomarker as acetone. This restricts the broader applicability of the system to diverse chemical environments. Incorporating additional VOCs in future experiments would improve the robustness and generalizability of the detection approach.

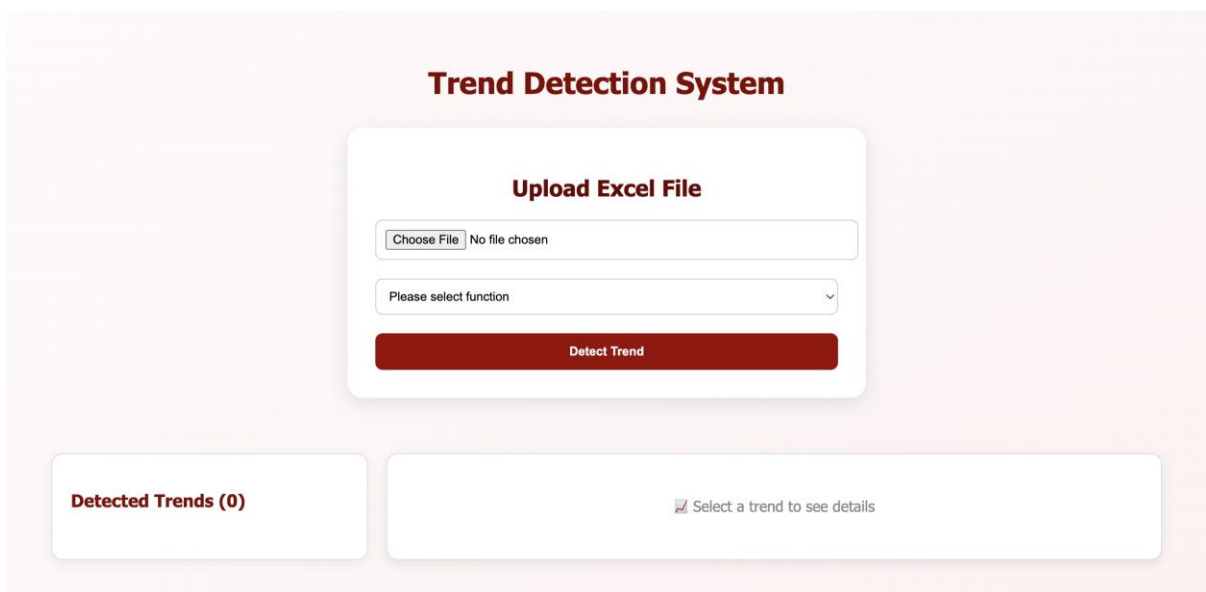
### 5.3 INTEGRATION INTO WEB-BASED APPLICATION

Following the model evaluation phase, the selected SVM classifier with normalized transmittance-only data was deployed in a web-based application built using a FastAPI (Python) backend and a React.js frontend. This platform enables users to upload Excel files containing data and select between algorithmic or AI-based trend detection.

Key features of the web system include:

- File upload and parsing of VOC data in `.xlsx` format.
- Preprocessing pipeline that applies transmittance selection and standard scaling.
- Real-time prediction results, displayed immediately upon submission.
- Method selection toggle, allowing comparison between AI-based and algorithmic methods.

This deployment demonstrates the applicability of the model in a real-world user-facing system, where users without technical expertise can benefit from high-accuracy VOC trend detection through an intuitive interface. This also makes the system scalable for potential clinical or field usage in early disease detection or air quality monitoring.



The screenshot displays the 'Trend Detection System' web application. At the top, the title 'Trend Detection System' is centered in a bold, dark red font. Below the title is a white card with a dark red header 'Upload Excel File'. Inside this card, there is a file upload section with a 'Choose File' button and the text 'No file chosen'. Below that is a dropdown menu with the placeholder text 'Please select function'. At the bottom of the card is a prominent dark red button labeled 'Detect Trend'. Below the upload card, there are two white boxes. The left box is titled 'Detected Trends (0)' in dark red. The right box contains a link with a magnifying glass icon and the text 'Select a trend to see details'.

Figure 11 : The Web-based application sample

## 6. PROJECT MANAGEMENT

### 6.1 GANTT CHART

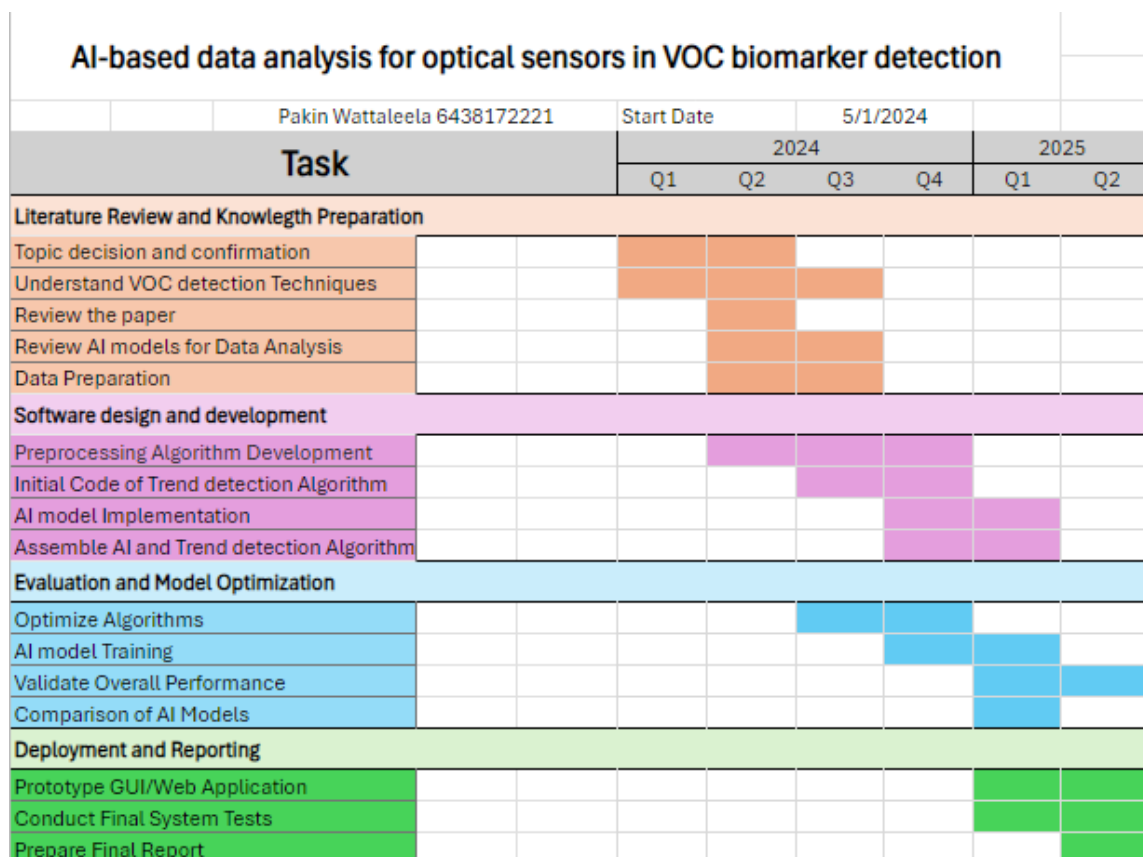


Figure 12. Project Gantt chart

### 6.2 BUDGET AND FUNDING

This project involves obtaining data from the NANO labs and studying the AI and machine learning knowledge from Coursera which support from ISE. The table below will state the catalog of overall cost.

No.	Category	Name	price (In-kind support)
1	Education	Basic knowledge of AI and ML in Coursera	ISE
2	Data	Data from ZnO-coated optical sensor	NANO laboratory

## **6.3 SAFETY CONCERNS AND ETHICS**

This project focuses on implementing AI for data analysis and may involve obtaining datasets from laboratory settings. To ensure responsible and safe execution, the following safety and ethical considerations are outlined:

### ***1. Safety Concerns***

1. If lab access is required for collecting data, safety measures will include laboratory protocols. This involves the use of appropriate personal protective equipment (PPE) such as gloves, goggles, and lab coats to handle any materials safely.
2. The project may involve working with optical sensors or data acquisition devices. Proper training will be undertaken to handle equipment safely and to avoid mishaps such as electrical hazards or accidental damage to sensitive instruments.
3. The raw data collected during laboratory experiments must be securely stored to prevent unauthorized access or loss. Measures such as encryption and controlled access will be implemented to ensure data integrity.

### ***2. Ethical Considerations***

1. Data collection and usage will be strictly followed. Any data obtained from the lab will be used solely for the purposes of this project and will be anonymized if it contains sensitive information. Fabrication or manipulation of data will be strictly avoided.
2. The AI models and algorithms developed for this project will be documented and presented transparently. This includes sharing assumptions, limitations, and validation methods to ensure that the results are credible and reproducible.

## 7. **IMPACT**

This project utilizes AI-based analysis to advance the development of VOC biomarker detection, particularly for non-invasive diabetes diagnosis using ZnO-coated optical fiber sensors. By automating the analysis of complex spectra data and predicting the scope of the results, AI significantly accelerates the research process, enabling faster identification of biomarkers and improving the precision of sensor performance. This automation reduces the time required for manual data interpretation, allowing research teams to concentrate on other aspects like refining design or optimizing system sensitivity. As the highest goal is achieved, AI algorithms could be extended to integrate with other research datasets. This would enable researchers from different fields to leverage the algorithm for various applications.

## 8. CONCLUSION

This work presents a web-based application designed to analyze and detect trends in spectral data, with a focus on volatile organic compounds (VOCs) as potential biomarkers. By combining algorithmic methods with machine learning models, such as KNN and SVM, the application was able to classify spectral trends with high accuracy. The use of FastAPI for backend processing and React.js for the frontend provided a seamless, scalable solution that allows users to upload and analyze spectral data efficiently.

The ability to incorporate both traditional algorithmic approaches and machine learning techniques enhances the flexibility and precision of the system, ensuring robust performance across various datasets. This makes the application suitable for a broad range of use cases, particularly in fields such as biomedicine and environmental monitoring.

The significance of this project extends beyond the development of a functional tool; it demonstrates the potential for integrating machine learning into biomarker research to improve diagnostic processes. By automating the trend detection process, this application offers a more efficient and reliable way to identify VOC-based biomarkers, contributing to the advancement of early diagnostic methods. Future work could include further enhancements, such as incorporating more advanced machine learning models or expanding the capabilities to handle larger and more complex datasets.

In conclusion, this project provides a valuable contribution to spectral data analysis, offering practical applications in biomarker detection and broader research. Its successful implementation paves the way for future innovations in both scientific exploration and healthcare.



## REFERENCES

- [1] Swargiary, K., Jitpratak, P., Pathak, A.K., and Viphavakit, C., "Low-Cost ZnO Spray-Coated Optical Fiber Sensor for Detecting VOC Biomarkers of Diabetes," *Sensors* [online journal], Vol. 23, No. 18, 2023, Paper 7916. URL: <https://www.mdpi.com/1424-8220/23/18/791>
- [2] Li, X., Hu, X., Li, A., Kometani, R., Yamada, I., Sashida, K., et al., "Identification of Binary Gases' Mixtures from Time-Series Resistance Fluctuations: A Sensitivity-Controllable SnO<sub>2</sub> Gas Sensor-Based Approach Using 1D-CNN," *Sensors & Actuators: A. Physical*, Vol. 349, 2023, Art. No. 114070. URL: <https://www.sciencedirect.com/science/article/pii/S0924424722007051>
- [3] Chowdhury, M. A. Z., and Oehlschlaeger, M. A., "Deep Learning for Gas Sensing via Infrared Spectroscopy," *Sensors*, Vol. 24, No. 6, 2024, Art. No. 1873. URL: <https://www.mdpi.com/1424-8220/24/6/1873>
- [4] Zhuang, Y., Yin, D., Wu, L., Niu, G., and Wang, F., "A Deep Learning Approach for Gas Sensor Data Regression: Incorporating Surface State Model and GRU-Based Model," *APL Machine Learning*, Vol. 2, 2024, Art. No. 016104. URL: <https://pubs.aip.org/aip/aml/article/2/1/016104/2933789/A-deep-learning-approach-for-gas-sensor-data>
- [5] World Health Organization, "Diabetes: Fact Sheet," WHO Newsroom, 2024. URL: <https://www.who.int/news-room/fact-sheets/detail/diabetes> [retrieved 27 Nov. 2024].
- [6] Xu, J., Zhou, J., Jiang, S., Wang, L., and Zhang, M., "Multi-Peak Detection Algorithm of Fiber Bragg Grating Using Mexican Hat Wavelets and Hilbert Transform," *Optics Communications*, Vol. 525, No. 15, Oct. 2022, Article 128986. DOI: [10.1109/INDISCON54605.2022.9862892](https://doi.org/10.1109/INDISCON54605.2022.9862892)
- [7] Fan, Z., Liu, Q., Zhang, C., and Chen, Z., "An Improved Multi-Peak Detection Algorithm of Fiber Bragg Grating Based on Mexican Hat Wavelets and Hilbert Transform," *Optical Fiber Technology*, Vol. 58, Nov. 2020, Article 102293 DOI: [10.1016/j.yofte.2020.102311](https://doi.org/10.1016/j.yofte.2020.102311)
- [8] Zhang, Z., and Dong, X., "A Self-Adaptive Peak Detection Algorithm to Process Multi-Peak Fiber Bragg Grating Sensing Signal," *Measurement*, Vol. 79, Oct. 2015, pp. 276–284. DOI: [10.3788/CJL201542.0805008](https://doi.org/10.3788/CJL201542.0805008)
- [9] Cusano, A., Consales, M., Crescitelli, A., Ricciardi, A., and Giordano, M., "Fiber Bragg Grating Sensors: Current Challenges and Future Trends," *Sensors and Actuators B: Chemical*, Vol. 146, No. 1, Nov. 2010, pp. 302–307. DOI: [10.1177/193229680800200526](https://doi.org/10.1177/193229680800200526)

[10] Xia, Y., Li, Y., Liu, Y., & Zhang, Y. (2014). Support vector machine classification of volatile organic compounds based on narrow-band spectroscopic data. *Sensors and Actuators B: Chemical*, 203, 658-664.

<https://doi.org/10.1016/j.snb.2014.06.051>

[11] Khan, M. A., et al. (2024). Comparative performance evaluation of machine learning models for IoT-enabled VOC classification. *Journal of Big Data*, 11, Article 22. \

<https://journalofbigdata.springeropen.com/articles/10.1186/s40537-024-00973-y>

# APPENDIX A: TURNITIN

## Turnitin

FinalProject2024_Nano_AI-based data analysis for optical fiber-1.pdf			
ORIGINALITY REPORT			
13%	10%	6%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS
PRIMARY SOURCES			
1	Submitted to Chulalongkorn University Student Paper	3%	
2	www.mdpi.com Internet Source	1%	
3	www.medgadget.com Internet Source	1%	
4	arxiv.org Internet Source	1%	
5	www.frontiersin.org Internet Source	<1%	
6	www.coursehero.com Internet Source	<1%	
7	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	<1%	
8	Lizhao Du, Hongna Huang, Zhengping Pu, Yuan Shi, Shanbao Tong, Junfeng Sun, Donghong Cui. "A potential diagnostic biomarker for schizophrenia based on local functional connectivity using dynamic regional phase synchrony", Schizophrenia Research, 2025 Publication	<1%	