People's Democratic Republic of Algeria
Ministry of High Education and Scientific Research
University of Akli Mohand Oulhadj Bouira
Faculty of Sciences and Applied Sciences
Computer Science department

Master Thesis

In Computer science

Speciality: *ISIL*

Topic

Raman Spectroscopy and Machine Learning for In-Vivo Glucose and HbA1C:prediction and Classification.

Supervisor:

- Dr DJELLABI Brahim

Realised by:

- BANOUH MELISSA
- DERRADJI IKRAM

2023/2024

Acknowledgement

First of all, we give thanks to Allah the Almighty, for granting us strength.

We would like to sincerely thank our research director **Dr DJELLABI Brahim**, for his attentive supervision, his enlightened advice and his valuable availability. His expertise was invaluable in refining our work and achieving our objectives.

We also express our gratitude to the members of the jury, who agreed to evaluate our dissertation with rigor and impartiality. A special thank you to all the professors and teachers. Their dedication to transmitting their knowledge and their passion for learning have been a continuous source of inspiration throughout our study.

Finally, we do not forget to show our gratitude to our families for their unconditional support, their constant encouragement and their unwavering love.

Dedication

I dedicate this thesis to my family, whose unwavering support and encouragement have been my greatest source of strength throughout this journey. Your belief in me has fueled my determination to pursue goodness. To my professors and mentors, thank you for sharing your knowledge, wisdom, and guidance. Your mentorship has shaped my intellectual growth and inspired me to push the boundaries of my capabilities.

BANOUH MELISSA.

Dedication

I dedicate this thesis to my father, whose compassionate presence and insightful advice have been a constant source of inspiration. You have always believed in me and encouraged me to pursue my dreams. To my mother, whose boundless love and unwavering support have been my pillars. Your patience, dedication, and prayers have given me the strength to persevere through challenging times. To my brothers, my lifelong companions and closest friends, for sharing with me the highs and lows of this journey. To all my family and friends, thank you for your unwavering support.

DERRADJI IKRAM.

Abstract

Raman spectroscopy is an advanced technique for analyzing spectra. This method, which does not require invasive procedures, shows great promise in blood analysis, particularly in measuring biomarkers such as HbA1c and glucose.When combined with advancements in machine learning, Raman spectroscopy enables the development of accurate predictive models using extensive spectral data sets. These models are integrated into intelligent analysis systems to automate and enhance the precision of blood diagnostics, providing a viable alternative to traditional invasive methods.Our research focuses on applying machine learning to analyze spectroscopic data for non-invasive blood analysis, aiming to overcome the limitations of conventional methods that are known for their labor-intensive nature, high cost, and occasional risks. Our goal is to explore and propose models capable of effectively quantify HbA1c and glucose levels in the blood.This approach represents a significant step forward in improving the accuracy and efficiency of non-invasive blood analysis, potentially offering safer and more accessible options compared to invasive techniques.The proposed approach, utilizing SVC-XGBoost, acheived an RMSE of 0.44% for HbA1c and 15.49 mg/dL for glucose, which is a significant improvement compared to recent literature.

Key words: Raman spectroscopy , Machine learning , Glucose , Diabetes , HbA1c , In-vivo measurements ...

Contents

List of Figures

List of Tables

List of equations

List of abreviations

EMR	Electromagnetic Radiation
NMR	Nuclear Magnetic Resonance
MS	Mass Spectrometry
IR	Infrared Spectroscopy
MIR	Mid-Infrared
FIR	Far-Infrared
ML	Machine Learning
PMTs	Photo-Multiplier Tubes
CCDs	Charge-Coupled Devices
KNN	K-Nearest Neighbor
DT	Decision Tree
SVM	Support Vector Machine
ANN	Artificial Neural Network
PCA	Principal component analysis
RF	Random Forest
SGD	Stochastic Gradient Descent
CNN	Convolutional Neural Networks
HbA1c	Glycated Hemoglobin
SOM	Self-organizing map
MAE	Mean Absolute Error
MSE	Mean Square Error
RMSE	Root Mean Squared Error
SD	Standard Deviation

General introduction

Blood tests are crucial in healthcare to diagnose many diseases such as diabetes which affects millions of people worldwide and is on the rise. The main monitoring technique is invasive, which can be expensive, uncomfortable, expensive, impractical and risky (infection, transmission of certain diseases, if not properly treated). Faced with these limitations, research has focused on finding non-invasive methods to help solve the problem.

Spectroscopy has gained popularity in the last few years due to its extensive application domain, such as identifying unknown substances in materials science, Blood tests,biology, pharmaceutics, and food science. It is a powerful analytical technique that can provide valuable information about the composition and structure of various materials. Spectroscopy has also become increasingly important in environmental monitoring and forensic analysis.

The current project goal is to find a non-invasive in-vivo blood testing method for glucose and Hb1Ac utilizing spectral data and machine learning in the hopes of achieving similar accuracy to standard monitoring devices while removing limits and invasiveness.. This approach could potentially revolutionize diabetes management by providing a more convenient and less risky way for patients to monitor their blood glucose levels. If successful, it could greatly improve the quality of life for those living with diabetes.

The structure of this thesis: is organized into four distinct chapters, each addressing a key component of the study. The chapters are as follows:

Chapter 1: Fundamentals of Spectroscopy

This chapter introduces the basic principles of spectroscopy, with a special focus RAMAN spectroscopy. It covers the theoretical background, the mechanisms behind spectroscopy, and the specific advantages and applications of RAMAN spectroscopy.

Chapter 2: ML Application on Spectroscopic Data

An overview of various machine learning algorithms , along with their applications in Raman spectroscopy.

Chapter 3:Data Preprocessing and System Architectures.

A detailed methodology for data preprocessing, including data collection, cleaning, and validation.

Chapter 4: Model Implementation and Validation.

The implementation of the proposed machine learning model, including the evaluation of the validation of results.

Chapter 1

Fundamentals of Spectroscopy

1.1 Introduction

The integration of spectroscopy with chemometrics represents a powerful approach that combines the analytical capabilities of spectroscopic techniques with advanced statistical and mathematical methods. This synergy allows researchers to extract meaningful information from complex spectral data, improve analytical performance, and solve challenging problems in various fields. In this chapter, we will review the definition of spectroscopy and its different types (MS, RAMAN, IR, etc.),then we will take a closer look at the definition and instruments of Raman spectroscopy and its advantages and disadvantages, following that we will go over qualitative Analysis and quantitative.Finally we will discuss about machine learning in raman spectroscopy.

1.2 Spectroscopy

1.2.1 Definition

"Spectroscopy is the branch of science contracts with learning about the interaction of the radiation of electromagnetic rays with substances."[4] By measuring the amount of radiation absorbed or emitted by a substance based on its wavelength or frequency, spectroscopy studies the interaction between electromagnetic radiation and matter.[5].

1.2.2 Electromagnetic radiation

EMR, or electromagnetic radiation, is a type of energy that can be found in a variety of forms, including microwaves, radio waves, visible light, infrared, ultraviolet X-rays, and gamma rays, as well as sunlight, which is a small part of the electromagnetic spectrum [4], It is defined as waves and particles. Waves explain refraction and absorption, whereas particles explain absorption and emission. The precise nature of electromagnetic radiation is unknown, although dual models offer a reasonable description.[6]

1.2.3 Interaction phenomena

By interacting with matter, light produces a variety of phenomena that can be examined to learn more about the biological system under study. The phenomena are :

Absorption

Matter absorbs energy from electromagnetic waves in specific regions of the spectrum; for example, leaves absorb green light, causing it to be transmitted and reflected preferentially.[7]

Emission

Light can be released from a heated body (light bulb) or after absorbing one wavelength, the longer wavelength of the released light results in less energy in the released light.[7]

Scattering

Matter scatters electromagnetic waves, changing their propagation direction. The energy difference between the scattered and incident light determines whether this occurs elastically (Rayleigh scattering) or inelastically (Raman scattering). [7]

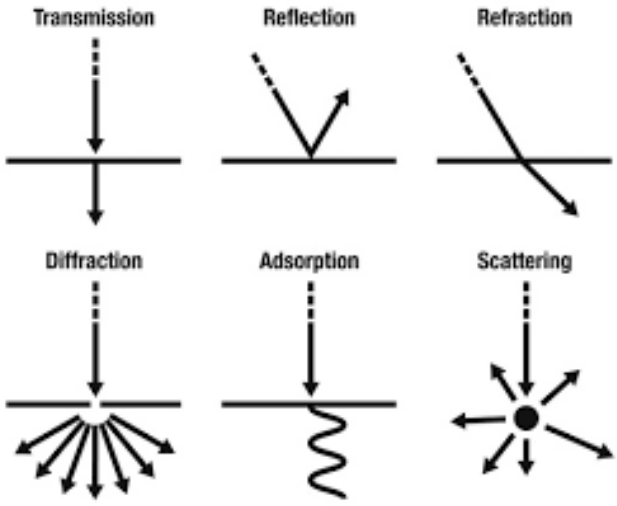


Figure 1.1: interaction of electromagnetic radiation (light) [1]

1.2.4 Types of Spectroscopy

Nuclear magnetic resonance (NMR) spectroscopy

Nuclear magnetic resonance (NMR) spectroscopy is a powerful technique that uses magnetic properties to record energy absorption between quantized levels, revealing all nuclei in the molecule under research. [8]

Mass spectrometry (MS)

Mass spectrometry is an analytical tool used to determine the mass-to-charge ratio of molecules in a sample, calculate molecular weight, identify unknown compounds, quantify known compounds, and discover the structure and chemical properties of molecules. [8]

Infrared (IR) spectroscopy

Infrared Spectroscopy (IRS) is a technique for analyzing the frequencies of bond vibrations in molecules. It is typically used to identify functional groups in samples. Covalently bound molecules absorb electromagnetic radiation in the infrared spectrum, which lies between visible light and microwaves. This radiation, mostly thermal energy, causes larger vibrations in molecules, making it suitable for nondestructive or nondestructive determination and quantitative compound analysis.[4]

IR can be divided into three major bands:

- **Near-Infrared** (NIR, 0.78_ 3.0 um) : The first region (NIRS) allows the study of overtones and harmonic or combination vibrations.[9]
- **Mid-Infrared** (MIR, 3.0 _50.0 um): The MIRS region is to study the fundamental vibrations and the rotation-vibration structure of small molecules.[9]
- **Far-Infrared** (FIR, 50.0_1000.0 um): the FIRS region is for the low heavy atom vibrations (metal-ligand or the lattice vibrations).[9]

Raman spectroscopy

Raman spectroscopy is a method of analysis that utilizes the scattering of light to determine the energy modes of vibration in a given sample. Raman spectroscopy provides valuable chemical and structural information, as well as enabling the identification of substances through their unique Raman 'fingerprint'. When molecules scatter light, Raman scattering occurs, leading to the polarization of the electronic cloud of the molecules and the creation of a transient complex known as the virtual state of the molecule. [8]

1.3 Raman Spectroscopy

1.3.1 History

Raman spectroscopy was discovered by Sir Chandrasekhara Venkata Raman in India in 1928 and is named after him. It was initially uncommonly used during most of the 20th century, with infrared spectroscopy being more popular at the time. The technique faced instrumental difficulties in its early stages, but advancements in technology eventually made it more accessible. The first commercialized Raman spectrometer was introduced in the 1960s as a result of special government projects. Raman spectroscopy gained popularity in the 1970s, with improvements in sensitivity and the development of laser technology. It became widely used in various fields, including chemistry, materials science, biology, and medicine.[10]

1.3.2 Instrumentation for Raman Spectroscopy

Raman spectroscopy requires specific instrumentation to perform accurate measurements and analysis.A Raman spectrometer typically consists of a laser source, a sample holder,

a spectrometer, and a detector.

The laser source

Provides the excitation light, which interacts with the sample and generates Raman scattering. Commonly used lasers include solid-state lasers, diode lasers, and gas lasers.

The sample holder

Holds the sample in place and allows for precise positioning during measurements. It can accommodate various sample types, including solids, liquids, and gases.

The spectrometer

Disperses the Raman scattered light into its different wavelengths, allowing for the identification and analysis of the Raman spectrum. It consists of a diffraction grating or a prism and a detector.

The detector

Captures the dispersed Raman signal and converts it into an electrical signal for further analysis. Commonly used detectors include charge-coupled devices (CCDs) and photomultiplier tubes (PMTs).

Additional accessories

Such as temperature and pressure cells, optical fibers for remote analysis, and microprobes for high spatial resolution, can be incorporated into the instrumentation setup for specific experimental requirements. [10]

1.3.3 Advantages and limits of Raman Spectroscopy

Advantages of Raman spectroscopy

- Applicable to all states: Solids, liquids, and gases can be analyzed.
- Requires only a small amount of sample due to the narrow laser source bandwidth. [11]

- Raman spectroscopy offers noninvasive analysis of biological samples, providing valuable diagnostic information.
- It allows for the identification and characterization of various compounds, without the need for sample preparation.
- Raman spectroscopy can be used to study the molecular composition and structure of tissues, aiding in the diagnosis of diseases.
- Raman spectroscopy is a versatile technique that can be applied to various fields.
- Raman is advantageous for studying living tissues (in vivo) due to its reduced water interference. [12]

Limits of Raman spectroscopy

- Large and complex datasets, interferences from instrumentation noise, and sample properties can mask the true features of samples, making Raman spectroscopy challenging.
- Preprocessing steps such as cosmic ray removal, smoothing, and baseline correction are often required for Raman-based regression procedures.
- Distinguishing similar peaks: Complex samples, have many molecules causing overlapping peaks in the spectra.
- Separating weak signals from background: Weak signals from certain molecules might be masked by the background noise and strong signals from other components. [13]

1.3.4 Chemometrics

Chemometrics is a discipline that utilizes statistical and mathematical techniques to examine chemical data. It plays a significant role in the realm of Raman spectroscopy, which is used in various fields to analyze complex spectral data. Chemometrics models are developed by merging data from different sources to enhance discrimination and prediction

capabilities in Raman spectroscopy. These models can be formed via supervised or unsupervised learning and are utilized to forecast sample properties or parameters based on Raman spectra. [14]

1.3.5 Quantitative analysis

Quantitative analysis in Raman spectroscopy entails the determination of the concentration or quantity of particular elements in a given sample. This can be accomplished by making a comparison between the Raman spectra of the sample and reference spectra that contain known concentrations. [15] Quantitative analysis is the process of determining the quantity or concentration of a particular component in a sample. It involves measuring the physical or chemical properties and using mathematical calculations to ascertain the amount of the substance being analyzed. This type of analysis is commonly employed in fields like chemistry, biology, and environmental science to measure substances of interest. [14]

1.3.6 Qualitative analysis

Qualitative analysis in Raman spectroscopy focuses on the identification of the chemical composition or molecular structure of a sample. It entails a comparison between the Raman spectra of the sample and reference spectra that contain known compounds. The aim is to determine the presence of specific functional groups or molecular vibrations. [15] Qualitative analysis, on the other hand, involves identifying the presence or absence of specific components in a sample. Its main focus is on determining the identity or characteristics of the substance being analyzed rather than its quantity. Qualitative analysis techniques can include visual observation, chemical tests, and spectroscopic methods like Raman spectroscopy. These methods aid in identifying the functional groups, chemical bonds, or molecular structures present in a sample. [14]

1.3.7 Machine learning in Raman spectroscopy

Despite being a highly effective method for identifying chemical materials, Raman spectroscopy has drawbacks because of the equipment's limits and complicated data. Deep learning shows promise in assisting researchers in overcoming obstacles, modeling intricate

connections, and extracting information from Raman-based chemical analysis. [16]

1.4 Conclusion

In conclusion, this chapter has provided an in-depth exploration of spectroscopy, electromagnetic radiation, and the fundamental interaction phenomena that underpin these techniques. Various types of spectroscopy, including Raman spectroscopy, were examined, with detailed discussions on its principles and applications. The advantages of Raman spectroscopy, such as its non-destructive nature, were highlighted, alongside its limitations. Furthermore, we explored chemometrics and both quantitative and qualitative analyses. Finally, the chapter discussed the application of machine learning techniques in Raman spectroscopy, illustrating their potential to overcome the challenges inherent in spectroscopic analysis. This comprehensive exploration provides a foundational understanding of spectroscopic techniques.

Chapter 2

ML Application on Spectroscopic Data

2.1 Introduction

Machine learning (ML) is an essential technology because of its ability to analyze large data sets, discern patterns, and make predictions autonomously. Integrating ML with spectroscopy addresses the limitations of traditional chemometrics by providing more efficient handling of high-dimensional spectral data, developing adaptable models less dependent on assumptions, and optimizing analytical tools for improved performance. This synergy not only improves precision and efficiency, but also promotes innovation, opening new avenues for scientific exploration and technological advancement. In this chapter, we will first examine the definition of ML and its types, then we will take look at the tools and descriptions that are commonly used in Machine Learning in Raman spectroscopy .At last we will mention the distinct applications of machine learning in raman spectroscopy.

2.2 Machine learning

2.2.1 Definition

Machine learning (ML) represents a groundbreaking shift in computing. It empowers systems to learn and adapt through dedicated training data, enabling them to automate the creation of analytical models and independently tackle associated tasks. This eliminates the need for explicit programming. ML strives to uncover meaningful relationships and patterns within provided examples and observations. This process has led to the develop-

ment of intelligent systems with remarkable cognitive capabilities, mirroring the human mind. Such advancements have significantly automated aspects of our lives. [17]

2.2.2 Deep learning

Deep learning, a subset of machine learning that utilizes artificial neural networks (ANNs), has recently gained significant traction in chemical research due to its ability to create powerful models that can both explore and predict from large, raw datasets. [17]

2.2.3 Types of machine learning

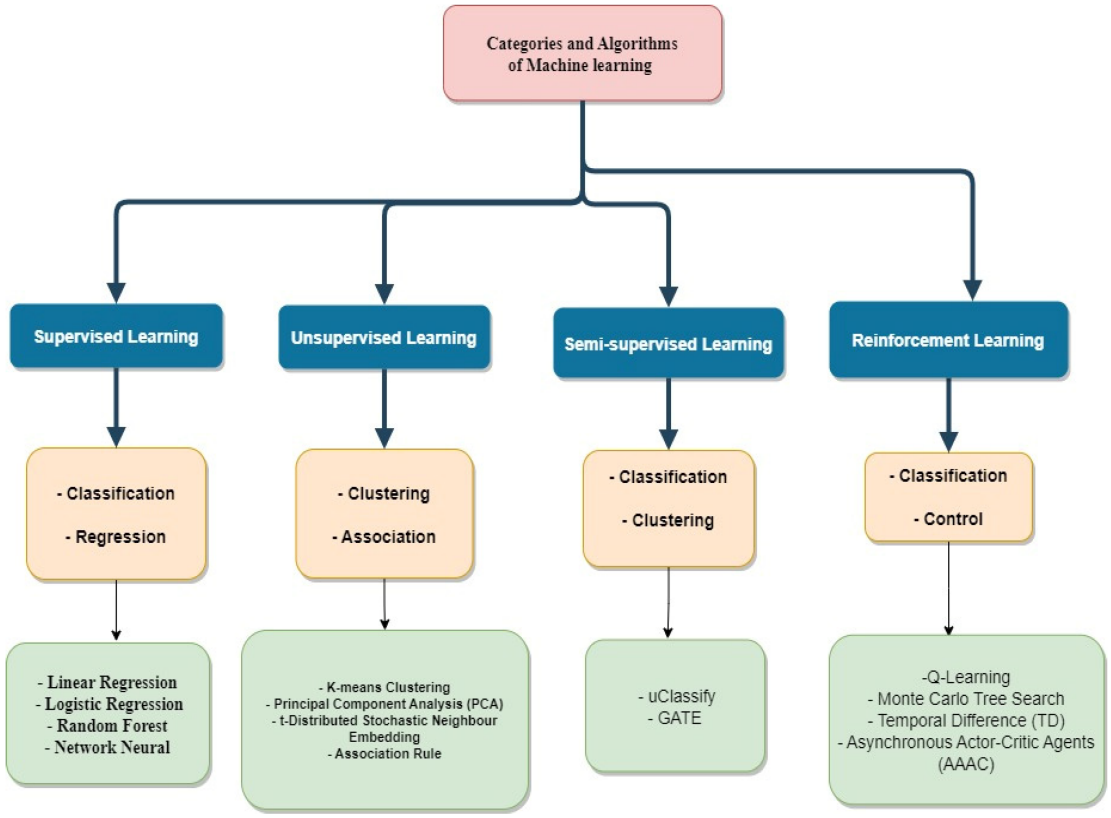


Figure 2.1: machine learning types [2]

1. **Supervised machine learning**

   A machine learning technique that involves the use of algorithms to model the relationships and dependencies between the target prediction output and the input features. The main objective is to make predictions for new data based on the

relationships learned from previous datasets. This type of learning, known as supervised learning, is driven by specific tasks such as regression and classification. One commonly employed method in supervised deep learning is the convolution neural network. The purpose of supervised learning algorithms is to provide answers to questions like "Based on the Raman fingerprint of this newly collected sample, which class in my database does it most likely belong to?" or "What is the level of purity of this substance?"[13]

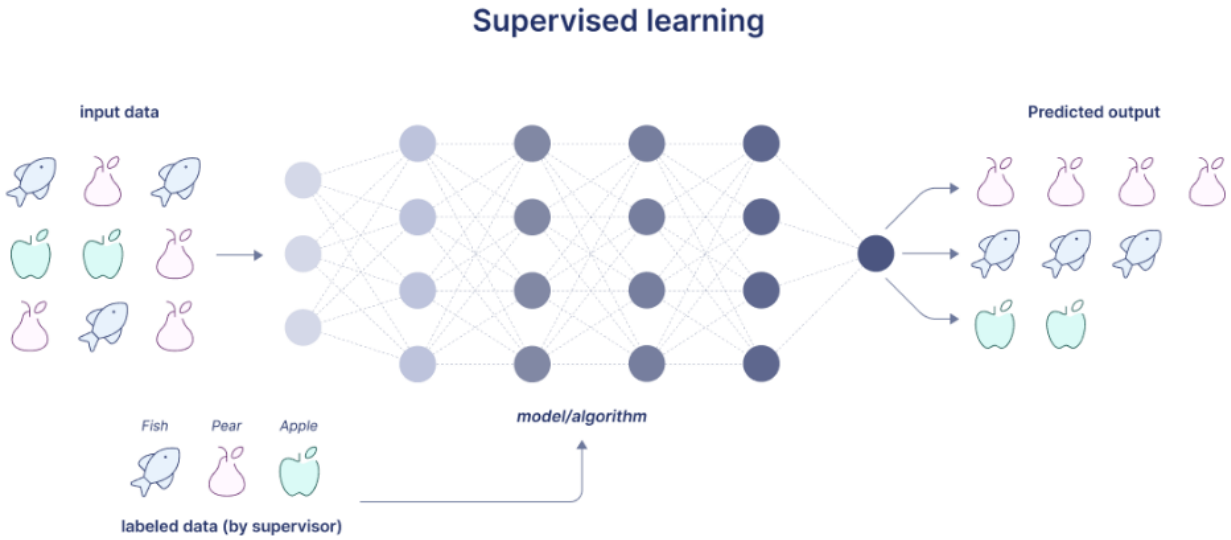


Figure 2.2: Supervised machine learning [3]

There are two type of algorithms :

1.1 Classification Analysis: Classification is a supervised machine learning algorithm that predicts class labels based on supplied exemple. It transforms input information into output variables to anticipate the goal, label, or categories. Spam detection is a common categorization challenge among email service providers.

- **Binary categorization** : Is a way of categorizing tasks using two labels, such as "true and false" or "yes and no." It enables for one class to represent the normal state and another for the aberrant condition. For example, "cancer not detected" is a normal result in a medical test.
- **Multiclass classification** : A technique for categorizing activities with several class labels that does not rely on the concept of normal and abnormal results is called multiclass classification. As an illustration, consider categorizing the

many kinds of network attacks seen in the NSL-KDD dataset into four class labels: DoS, U2R, R2L, and probing assaults.

- **Multi-label** : Multi-label classification is a machine learning approach that associates an example with many classes or labels. It is an extension of multiclass classification. It uses hierarchically constructed classes, which allow any example to belong to several classes at each level. This method use sophisticated algorithms to forecast mutually non-exclusive classes or labels, as opposed to typical tasks in which class labels are mutually exclusive.

[18]

1.2 Regression Analysis: Regression enables continuous quantity prediction. Applied in a variety of sectors, including financial forecasting, cost estimating, trend analysis, marketing, and medication response modeling. The most common regression algorithms are linear, polynomial, lasso, and ridge regression. [18]

2. **Unsupervised machine learning**

A technique in machine learning where models are not trained using a dataset. Instead, the models themselves discover hidden patterns and insights from the given data without labels. This process is similar to how the human brain learns new things. Unsupervised learning allows users to perform more complex processing tasks compared to supervised learning and is known as a data-driven approach. Some tasks that can be achieved with unsupervised machine learning include dimensionality reduction, clustering, and association. Examples of unsupervised deep learning algorithms include autoencoders, sum-product networks, recurrent neural networks, and Boltzmann machines. Unsupervised learning algorithms focus on answering questions like "How similar are these samples to each other based on their Raman fingerprints?". [13]

There are two types of algorithms :

2.1 Cluster Analysis: An unsupervised machine learning method called cluster analysis is used to find and organize similar data points in big datasets. It arranges items into categories based on how similar they are to one other compared to other

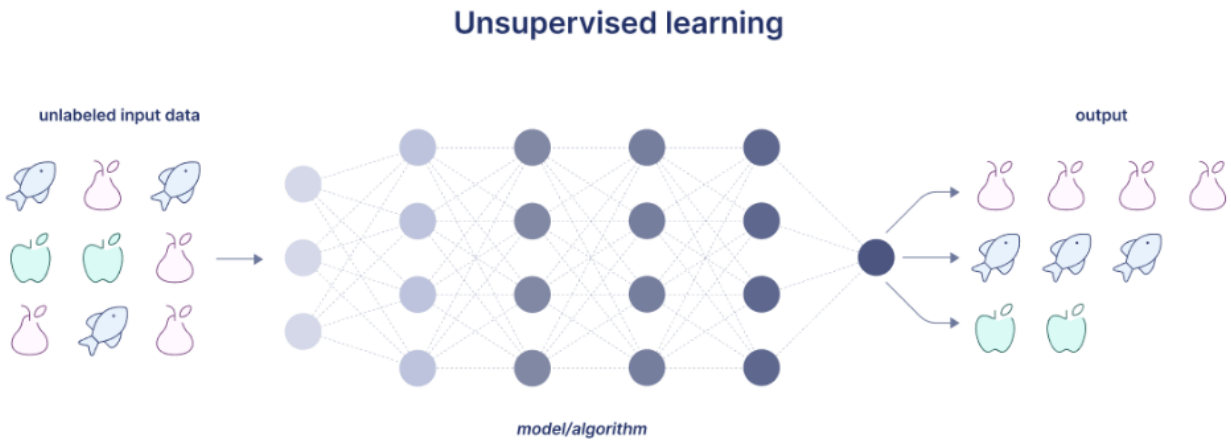


Figure 2.3: Unsupervised machine learning [3]

groupings. Finding trends or patterns in data, such customer groupings based on behavior, is frequently accomplished through the use of clustering. Applications for it include e-commerce, user modeling, health analytics, and cybersecurity. [18]

2.2 Association Rule Learning: Association rule learning is a rule-based machine learning technique that identifies associations between variables in huge datasets. It is utilized for a variety of applications, including IoT services, medical diagnostics, online usage statistics, and cybersecurity.The data mining literature has offered a variety of methodologies, including logic-dependent, frequent pattern-based, and tree-based approaches. [18]

3. **Semi supervised machine learning**

Semi-supervised learning may be regarded as a combination of the aforementioned supervised and unsupervised approaches, since it works with both labeled and unlabeled data . Thus, it lies in between learning "without supervision" and learning "with supervision." In the actual world, labeled data may be scarce in various scenarios, but unlabeled data is abundant, making semi-supervised learning helpful . The ultimate aim of a semi-supervised learning model is to offer a better prediction output than that obtained by the model utilizing only labeled input. Machine translation, fraud detection, data labeling, and text classification are all examples of semi-supervised learning applications. [18]

4. **Reinforcement machine learning**

   Instead of providing input and output pairs in a reinforcement learning system, we describe the system's current state, specify a goal, provide a list of allowable actions and their environmental constraints for their outcomes, and allow the ML model to experience the process of achieving the goal on its own using the trial and error principle to maximize a reward. [17]

2.3 ML-based models used in Raman Spectroscopy

Machine learning techniques have been applied more and more to improve data interpretation and analysis in Raman spectroscopy.Multivariate analysis techniques like as support vector machines (SVM), partial least squares regression (PLSR), and principal component analysis (PCA) are often used for problems including regression, classification, and dimensionality reduction. These techniques make it possible to extract useful data from intricate spectral datasets, which helps with the identification and measurement of analytes in Raman spectra.For a more specification, view the table on the next page2.1.

Tools	**Descriptions**
K-Nearest Neighbor	KNN classifier is to classify unlabeled observations by assigning them to the class of the most similar labeled examples.[19]
Decision Tree	DT A decision tree is a type of tree structure that resembles a flowchart, with each leaf node representing the result, the branch representing a decision rule, and the internal node representing a features (or attribute).[20]
Random Forest (RF)	The idea behind random forests is to combine a number of randomly selected decision trees. The goal is to pool a set of predictors (not necessarily optimal).[21]
Support Vector Machine	SVM Find the hyperplane that distinguishes the different categories with maximum margins and separate the dataset into different categories by selecting appropriate support vectors. [22]
Artificial Neural Network	ANN mathematical model that simulates the brain's neuronal activity as a set of connected input/output units, where each connection has a weight associated with it.[22]
Stochastic Gradient Descent (SGD)	Is a simple but highly effective method for fitting linear classifiers and regressors to convex loss functions [23] .
XGBoost	Is a lightweight, efficient, and versatile distributed gradient boosting library that utilizes the Gradient Boosting framework for machine learning algorithms, including parallel tree boosting..[24]
Principal component analysis	PCA is a statistical technique that is useful for compression and extract useful information from multivariate data sets ,the objective is to reduce the dimensionality of a data set.[25]
Isomap	Isometric mapping is an additional spectral theory-based distance-preserving nonlinear dimensionality reduction method. [26]

Table 2.1: Tools and descriptions that are commonly used in Machine Learning in Raman spectroscopy.

2.4 Discriminant analysis in Raman Spectroscopy

Discriminant analysis in Raman spectroscopy is a technique used to differentiate between groups based on their Raman spectra.[27]

2.4.1 Multivariate analysis

Multivariate analysis is a statistical technique used to analyze data with multiple variables, identifying trends, outliers, and classifying data. In this case, unsupervised principal component analysis (PCA) and supervised partial least square discriminant analysis (PLS-DA) were used to analyze Raman spectroscopy data of waste cooking oil samples, identifying a chemical fingerprint characteristic of each sample.[28]

2.4.2 Principal component analysis (PCA)

Principle component analysis is a powerful statistical approach for reducing a case-by-variable data table to its core properties, known as principle components. A small number of linear combinations of the original variables, known as principal components, can account for the majority of the variation in all the variables. Using just a few key elements, the technique approximates the original data table in the process. The definition, geometry, and interpretation of the method's numerical and graphical findings are all covered in detail in this primer.[29]

2.5 Developing a Machine Learning Model for Raman Spectroscopy

The several processes required to develop a machine learning model for Raman spectroscopy will be covered , with an emphasis on important procedures such data collection and preprocessing, model selection and training, and model evaluation. Refer to the figure 2.4

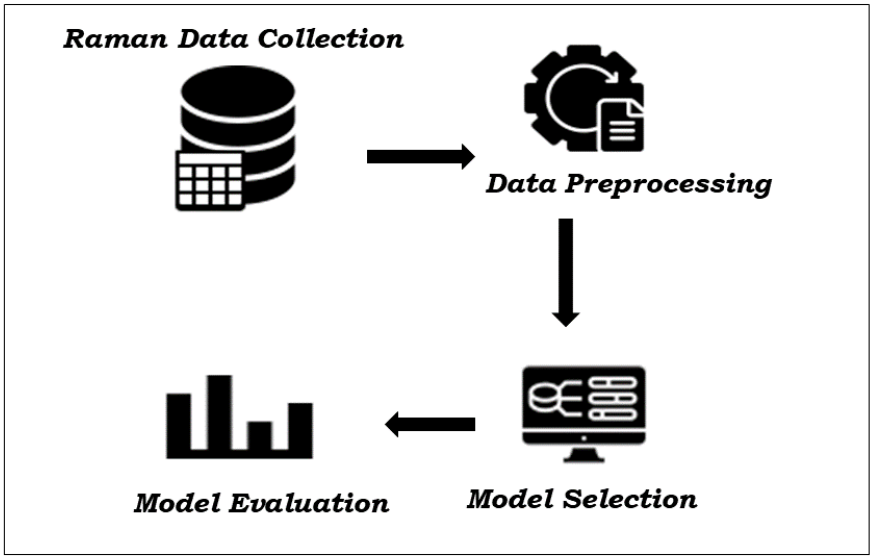


Figure 2.4: Steps for Developing a Machine Learning Model for Raman Spectroscopy

2.5.1 Data Collection

The dataset offers Raman spectra measurements of 46 volunteers, alongside their HbA1c and glucose levels. While the dataset doesn't detail the exact collection method, it does indicate the in vivo nature of the study, meaning the Raman spectra were directly obtained from the living subjects. In vivo Raman spectroscopy typically involves a specialized probe directing a laser onto a target body area and collecting the resulting scattered light. This scattered light provides information on the molecular makeup of the sampled tissue.

2.5.2 Data Preprocessing

Data preprocessing is an important phase in machine learning that includes operations such as cleaning, scaling, and encoding data. Preprocessing improves the caliber and dependability of machine learning results by addressing irregularities and converting data into a format that is readable.

Checking missing values and outliers

Outliers and missing values in data can have a major negative influence on a study's statistical power and dependability, which can result in severe bias and decreased efficiency. Handling outliers and missing values has a big influence on the outcomes of data analysis.[30]

Naming columns

Column naming in a DataFrame aims to provide clear and meaningful labels for the data included inside them. This clarity makes it easier to use and analyze data and helps to grasp the structure and substance of the dataset.

Normalisation

Normalization is a data preprocessing method that rescales numerical values within a defined range. It aim is to reduce disparities in the ranges of values while maintaining a consistent scale for all aspects .[31]

Encoding categorical data

Encoding techniques are required to convert these category variables into numerical values since machine learning algorithms can only handle numerical inputs.[32]

Oversample

Oversampling is a resampling approach that can be used to balance a dataset by increasing the number of minority class instances or samples, as well as creating new instances or repeating some of them. Borderline is one oversampling technique example.-SMOTE.[33]

Smoothing

Data smoothing is a statistical strategy that removes outliers from datasets to highlight trends. It is accomplished by removing statistical noise from datasets using algorithms. Data smoothing is a useful tool for predicting patterns, such share price trends.[34]

Feature Extraction and Selection

Feature extraction retains the original variables processes them into a smaller set to retain more information, and feature selection removes input variables that do not significantly impact model performance.[35]

2.5.3 Model Selection

From the algorithms suggested for spectroscopy in the studied field, we choose the most suitable one for exemple ANN,SVM...etc.

2.5.4 Model Evaluation

TO understanding its performance, detecting issues making informed decisions, and improving its effectiveness.

2.6 Applications of machine learning in Raman spectroscopy

2.6.1 Materials science

Raman spectroscopy is a powerful tool for studying material structures and composition. Combining computer science innovations with materials synthesis and characterization could save costs and time. AI algorithms like ML and deep learning can efficiently identify materials and understand their behaviors[36].Refer to the table 2.2

Applications	**approach**	**Result**	**Years**
Identify the number of graphene layer	PCA	Accuracy $< 90\%$	2018
Recognize minerals	CNN	Accuracy is 98.43%	2022
Classify the plastics	SVM ,PCA ,ANN	Accuracy for SVM ,PCA $> 95\%$ and for ANN close 100%	2019
Distinguish phases of matter	SVM, PCA	Accuracy 98.7% and 99.7%	2019

Table 2.2: Applications of machine learning in Raman spectroscopy for Materials science.

2.6.2 Food science

The following are some examples of how Raman spectroscopy is used in the examination of food, with an emphasis on its effectiveness in identifying food adulteration, unapproved additions, antibiotics, medications, pesticide and fungicide residues, and heavy metals,[37]Refer to the table 2.3.

Applications	**approach**	**Result**	**Years**
Detection of edible oils type andadulteration	PCA, CNN, RF...	All ML algorithms were used and acheive 100 % accuracy	2022
Identification of intact beef, venison andlamb	PCA, PLS-DA, SVM	Accuracy: 80 % (PLS-DA) , 92 % (SVM) and 100 %(SVM and PLS-D)	2021
Detection of fruit distillates	DT,DA,SVM,KNN, Ensemble classifiers	Accuracy is 95.5 %	2020

Table 2.3: Applications of machine learning in Raman spectroscopy for Food science .

2.6.3 Pathogens in biomedicine

Spectroscopic raman have been finding growing uses in the field of biomedicine, notably in the field of illness diagnosis and monitoring, despite the quick introduction of various molecular biology-based approaches. Here are some examples:[38]Refer to the table 2.4.

Applications	**approach**	**Result**	**Years**
Detection of bacteria	CNN	Accuracy $\approx$ 86 %	2021
Analysis of Raman spectra of humanand avian viruses	CNN	Accuracy 99 %	2022
Identification of Marine Pathogens	RNN (LSTM)	Accuracy >94 %	2021

Table 2.4: Applications of machine learning in Raman spectroscopy for Pathogens in biomedicine .

2.6.4 Healthcare

Machine learning (ML) and Raman spectroscopy are revolutionizing healthcare by providing precise, non-invasive diagnostic tools. ML algorithms can detect molecular signatures associated with health conditions like cancer and neurological disorders. By analyzing biological samples such as blood, saliva, or urine and it can differentiate between healthy and diseased states, enabling early disease detection and personalized treatment. This technology also holds promise in drug development and quality control, improving patient outcomes and medical research,here are some examples refer to the table 2.5.

Applications	**Approach**	**Result**	**Years**
Alzheimer's disease (AD) diagnosisbased on saliva analysis	ANN	Accuracy is 99 %	2019
Saliva-based detection of COVID-19 infection	MILES	Auc t Max=0.80	2022
Diagnosis of lung cancer	STFT based CNN	Accuracy is 96.5 %	2021
Identify blood species	RNN	Accuracy is 97.7 %	2021
Identification of kidney tumor tissue	SVM	Accuracy is 92.89 %	2021
Use of Raman spectroscopy to screen diabetes mellitus with machine learning tools	ANN SVM	Accuracy: 88.9–90.9% 76.0–82.5%	2018
Classification of cerebral infarction andcerebral ischemia	PCA, PLS, MRMR, SVM,KNN, PNN, DT	Accuracy >85 %	2022
Quantifcation of glycated hemoglobin and glucose in vivo using Raman spectroscopy and artifcial neural networks	FFNN	RMSE :0.69% for HbA1c and 30.12 mg/dL for glucose.	2022
Blood glucose concentration estimation by Raman spectroscopy based on particle swarm optimized SVR	PSO-SVR	R-Squared is 0.8041 RMSE is 1,8580	2023

Table 2.5: Applications of machine learning in Raman spectroscopy for Healthcare .

2.7 Conclusion

To conclude , we have explored the foundational concepts of ML and its various types. We have delved into the tools and methodologies commonly employed in Machine Learning, particularly within the context of Raman spectroscopy. By understanding these tools and techniques, we gain insight into how they facilitate the analysis and interpretation of spectral data. Additionally, we have highlighted the distinct applications of Machine Learning in Raman spectroscopy, showcasing how ML can enhance the precision and efficiency of spectroscopic analysis. Through these discussions, it becomes evident that the integration of Machine Learning with Raman spectroscopy holds significant potential for advancing both research and practical applications in this field.

Chapter 3

Data Preprocessing and System Architectures

3.1 Introduction

Preprocessing spectral data is an important step for spectral data in machine learning because it helps enhance the quality and usability of the data , thus improving the performance of machine learning models. In this chapter, we will explore the process of machine learning implementation and our data structure and specificity. Additionally, we will examine the architecture of our proposed models, laying the foundation for effective spectral data analysis and modeling.

3.2 Data overview

3.2.1 Acquisition information

The data was taken from 46 participants who were fasting and had their blood drawn, and tests for glycated hemoglobin and glucose were performed. The first test used afinidad de boronato, while the second used glucosa oxidasa. These results are regarded as gold standards.[39]

In vivo measurements

The Raman measurements equipment consisted of a 60 mW power Raman probe InPhotonics® RIP-RPS-785 and a Raman spectrometer QE65000 from Oceans Optics® with a resolution of 0.14–7.7 nm FWHM. The observations were made in Tuxtla Gutierrez, Chiapas, Mexico, where the temperature was 19.76 ±1.02 °C and the humidity was 63.09

± 5.25 %. Ten measurements were made for each participant and body area using 30 seconds of integration time; however, some volunteers moved while the spectra were being acquired. The American National Standard for the safe use of lasers (ANSI Z136.1-2007) was followed in the calculation of the laser power and integration time.[39]

3.2.2 Data file structure

The files listed below are used in our study:HbA1c and glucose in vivo Raman spectra[40]

The file Volunteers information

Presents details on the participants, including their id, age, weight, height, diabetes, and HbA1c and glucose...etc.Consult the figure 3.1.

- Features:
    1. ID:An identifier of Volunteers and ranges from 1 to 46.
    2. Glucose (mg/dl):Glucose level of patient varied from 56 to 400 mg/dL.
    3. HbA1c (%):Glycated hemoglobin test estimates the blood glucose average for the previous 2 to 3 months, the values varied from 5.2 to 14% .
    4. Age :Count the age of volunteers
    5. Gender:Refers to the gender of volunteers
    6. Others :Height ,Weight,Humidity,Enviremental tamperatur(°C)...etc
- Labels:
    1. Diabetes:Represent if the Volunteers has type 2 diabetes or not (Binary labels).

```
1 info_data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 46 entries, 0 to 45
Data columns (total 17 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Acquisition date                46 non-null     datetime64[ns]
 1   ID                              46 non-null     object
 2   Glucose (mg/dL)                 46 non-null     int64
 3   HbA1c (%)                       46 non-null     float64
 4   Gender                          46 non-null     object
 5   Age                             46 non-null     int64
 6   Diabetes                        46 non-null     object
 7   Height (cm)                     46 non-null     int64
 8   Weight (kg)                     46 non-null     float64
 9   Wrist width (cm)                46 non-null     int64
 10  Forearm width (cm)              46 non-null     int64
 11  Long forearm (cm)               46 non-null     int64
 12  Forearm temperature (°C)        46 non-null     float64
 13  Wrist temperature (°C)          46 non-null     float64
 14  Finger temperature  (°C)        46 non-null     float64
 15  Humidity  (%)                   46 non-null     float64
 16  Environmental temperature (°C)  46 non-null     float64
dtypes: datetime64[ns](1), float64(7), int64(6), object(3)
memory usage: 6.2+ KB
```

Figure 3.1: Information about data "Volunteers information"

The folder Filtered Raman spectra

Holds the input and output for glucose and HbA1c, or the values of glucose and HbA1c for each volunteer. This folder also contains the filtered Raman spectra for each volunteer and body area (forearm, wrist, index finger).

There are Raman spectra in each tab for the forearm, wrist, and index finger, covering a spectral range of 200 to 1800 cm-1 (788 features). It is evident that just 414 measurements each volunteer obtaining 9 measurements will be used for the wrist and forearm, whereas the index finger's Raman spectra tab has 3 measurements.[39]

- Input and Output HbA1c: A file for each body part(forearm, wrist, and index finger).Refer to the figure 3.2.

    1. features:

       - ID: An identifier of 46 Volunteers.

       - Spectral data :788 features from F0 to F787 refer to the Raman spectroscopy measurements of HbA1c molecules in the bloodstream at specific Raman shift values.

    2. labels:

       - HbA1c: The values varied from 5.2 to 14% .

```
1 h_wrist.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 414 entries, 0 to 413
Columns: 790 entries, ID to F787
dtypes: float64(789), object(1)
memory usage: 2.5+ MB
```

```
1 h_forearm.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 414 entries, 0 to 413
Columns: 790 entries, ID to F787
dtypes: float64(789), object(1)
memory usage: 2.5+ MB
```

```
1 h_finger.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 138 entries, 0 to 137
Columns: 790 entries, ID to F787
dtypes: float64(789), object(1)
memory usage: 851.8+ KB
```

Figure 3.2: Information about data "Input and Output HbA1c"

- Input and Output Glucose:A file for each body part(forearm, wrist, and index finger).Consult the figure 3.3.

  1. features:

     - ID: An identifier of 46 Volunteers.

     - Spectral data :Features from F0 to F787 refer to the Raman spectroscopy measurements of Glucose molecules in the bloodstream at specific Raman shift values.

  2. labels:

     - Glucose: The values varied from 56 to 400 mg/dL.

```
1 g_forearm.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 414 entries, 0 to 413
Columns: 790 entries, ID to F787
dtypes: float64(788), int64(1), object(1)
memory usage: 2.5+ MB
```

```
1 g_wrist.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 414 entries, 0 to 413
Columns: 790 entries, ID to F787
dtypes: float64(788), int64(1), object(1)
memory usage: 2.5+ MB
```

```
1 g_finger.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 138 entries, 0 to 137
Columns: 790 entries, ID to F787
dtypes: float64(788), int64(1), object(1)
memory usage: 851.8+ KB
```

Figure 3.3: Information about data "Input and Output Glucose"

3.3 System architectures

3.3.1 Data Exploration

In this phase, we explore the dataset HbA1c and glucose in vivo Raman spectra to comprehend the underlying problem and explore the relationships between its columns. Our primary focus lies in detecting any inaccuracies, missing values, or outliers.

3.3.2 Diabete classification

We utilized three datasets:

1. "Volunteers Information": We used ID,Gender,Diabetes,age.

2. "Input and Output HbA1c" and "Input and Output Glucose" :The complete set of features comprising the dataset.
   For binary classification (yes/no).Take a look at the figure 3.4.

Step 1:

We merged the three datasets for each body part (forearm, wrist, and index finger), resulting in datasets each with 1592 columns.

Step 2: Preprocessing

1. We checked for any missing values or outliers in the datasets to ensure data completeness. Upon thorough examination, it becomes evident that the dataset is devoid of any missing values or outliers.

2. We appropriately named the columns to make the datasets more understandable and easier to manipulate.

3. Through correlation analysis, we identify certain columns that appear to be less significant. These columns were removed.

4. We encoded categorical columns (such as ID, gender, and diabetes status) to numerical values .This step is essential for algorithms that require numerical input.

5. We normalized the data using StandardScaler to ensure that all features contribute equally to the analysis.

6. In addressing the issue of unbalanced data, we've opted to employ resampling techniques as a solution. Given the limited number of volunteers in our dataset (only 46), oversampling emerged as the more suitable approach. Specifically, we've chosen the Synthetic Minority Over-sampling Technique (SMOTE) to augment the minority class.

7. the features In our dataset represent spectral data values.However,due to factors such as instrument noise and fluctuations, spectral data can be prone to noise and variability, which can complicate the task of building accurate machine learning models. To address this challenge, we employed the Whittaker smoother technique.

8. During our data exploration phase, a notable observation was the large number of features within the dataset, posing a risk of dimensionality. To address this risk, we've chosen feature selection/extraction as our strategy. In our approach, we've employed Self-Organizing Maps (SOM) and Relief techniques.

Step 3:Train-Test Split

We split the data into training and testing sets for each body part to evaluate the models' performance. This step ensures that we have a separate dataset to test the models' generalizability.While keeping data with same "ID" (same volunteer) in one group to have a more reliable evaluation.We divided the data as follows: 80% for the train set and 20% for the test set.

Step 4:Models Application

We applied the top three models to the datasets:

- Xgboost, SVC, SGDClassifier

Step 5:Model Evaluation

We evaluated each model on the testing data using various metrics (such as accuracy,Specificity). ,Sensitivity) to measure their performance comprehensively.

Step 6:Model Validation

We validated the performance of the models to ensure their reliability and effectiveness in predicting outcomes. This validation step helps in confirming that the models can be trusted for real-world applications.

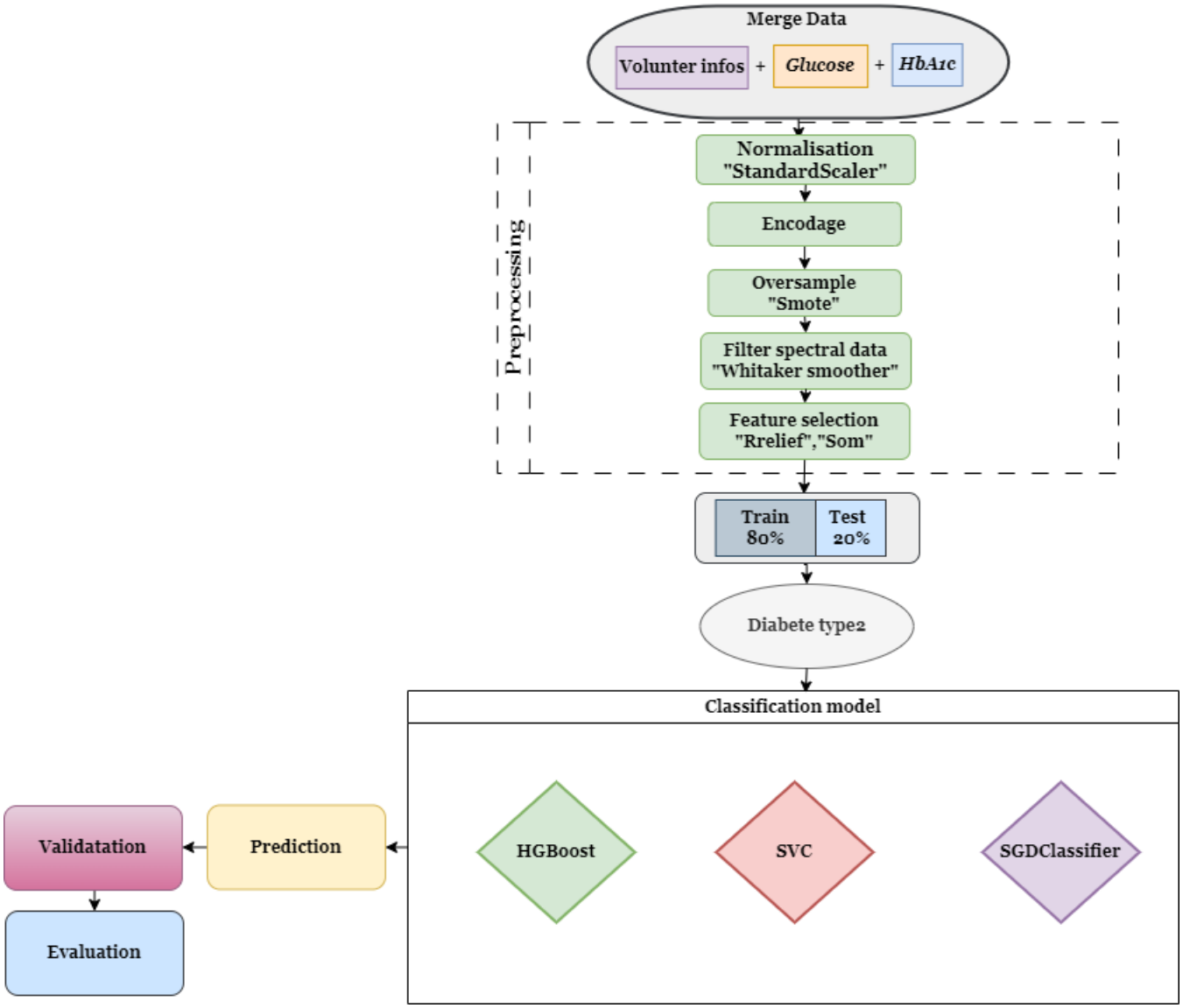


Figure 3.4: "Diabete classification architecture"

3.3.3 Regression

The same steps were used for both "Input and Output HbA1c", and "Input and Output Glucose"individualy.Refer to the figure 3.5.

Step 1:Preprocessing

Those preprocessing phases (Checking missing values,naming columns,Filtring spectral data Whittaker,Features selection/extraction) were used as in the previous architecture ,in addition to :

1. In our implementation, we applied the Z-score outlier detection technique to the dataset. However, despite its application, we observed no significant improvement in the model's performance. Given the limited dataset and the absence of notable enhancements We concluded that, due to the absence of notable enhancements and the limited nature of our dataset, it was prudent to discard this technique from our data preprocessing approach.

Step 2:Train-Test Split

We divided the data as follows: 80% for the train set and 20% for the test set.

Step 3:Models Application

We applied four models to the datasets:

- **XGboost:** To refine our XGBoost model's performance, through grid search, we tested key parameters such as learning rate, maximum tree depth (max_depth), and the number of estimators (n_estimators).
- **SVR:** To refine SVR's performance, we examined the key parameters such as regularization (C), epsilon, kernel coefficient (gamma), choice of kernel function, and polynomial degree.
- **KNN:** To optimize the performance of the KNN algorithm, we employed grid search, a hyperparameter tuning technique. We explored various combinations of hyperparameters such as the number of neighbors (n_neighbors), the weighting scheme (weights), and the distance metric (algorithm) .

- **Gaussian process regressor:** To optimize the performance of this regression model, we employed grid search methodology. we systematically explored a range of hyperparameters, including the regularization parameter (alpha), choice of kernel function (kernel), the number of restarts for optimization (n_restarts_optimizer), and whether to normalize the target variable (normalize_y).

Step 4:Model Evaluation

We assessed the performance of each model on the test dataset using multiple evaluation metrics, including RMSE and SD.

Step 5:Model Validation

We conducted thorough validation of the models to ensure their reliability and effectiveness in predicting outcomes.

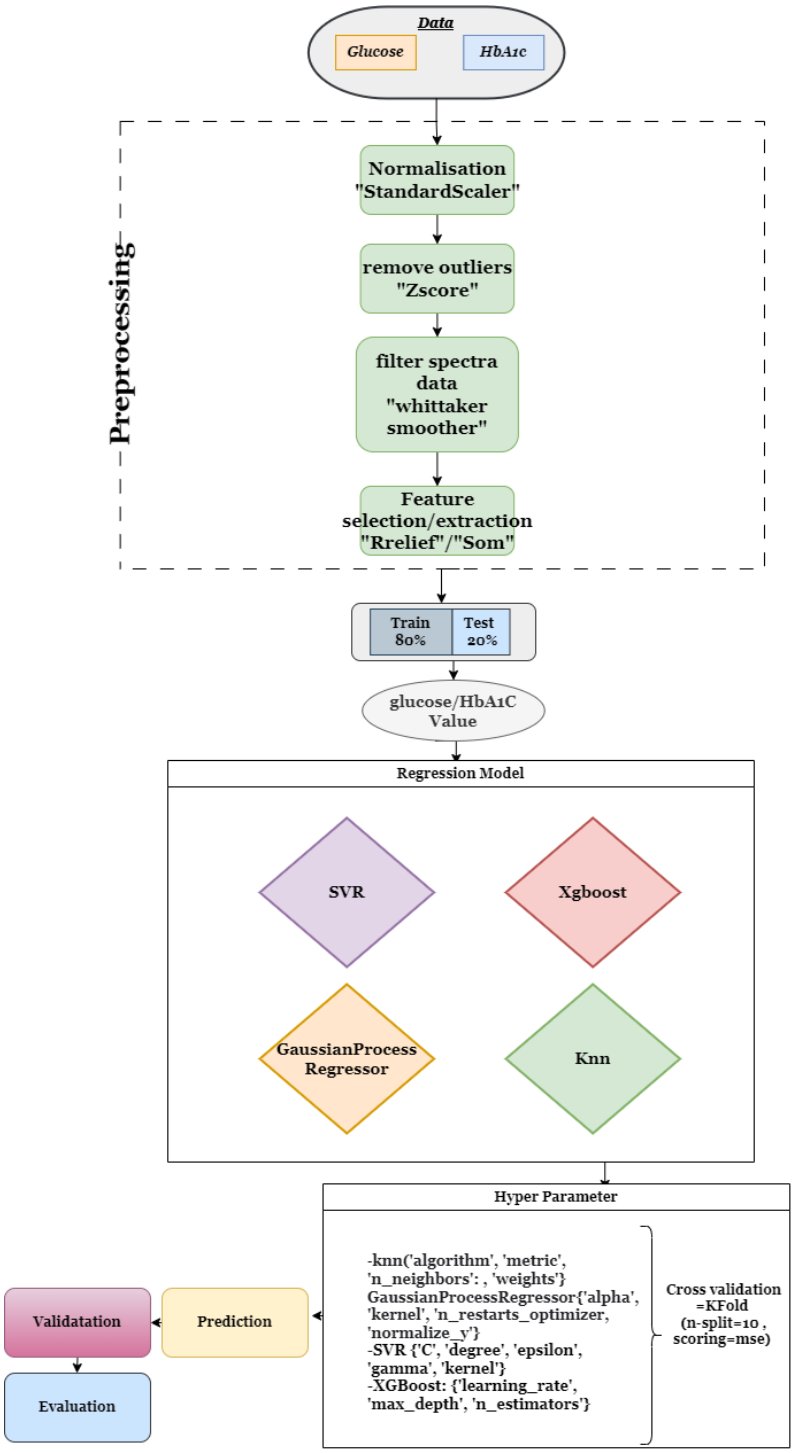


Figure 3.5: "Regression architecture "

3.3.4 HbA1c/Glucose Classification

Similar procedures were applied for glucose and HbA1c data independently for multiclass classification(ok/high/low).Refer to the figure 3.6.

The same process were utilised as Diabete classification plus :

In the models application :SVC ,KNN,SGDClassifier were used.

In the validation steps :Cross validation (GroupKFold) was used.

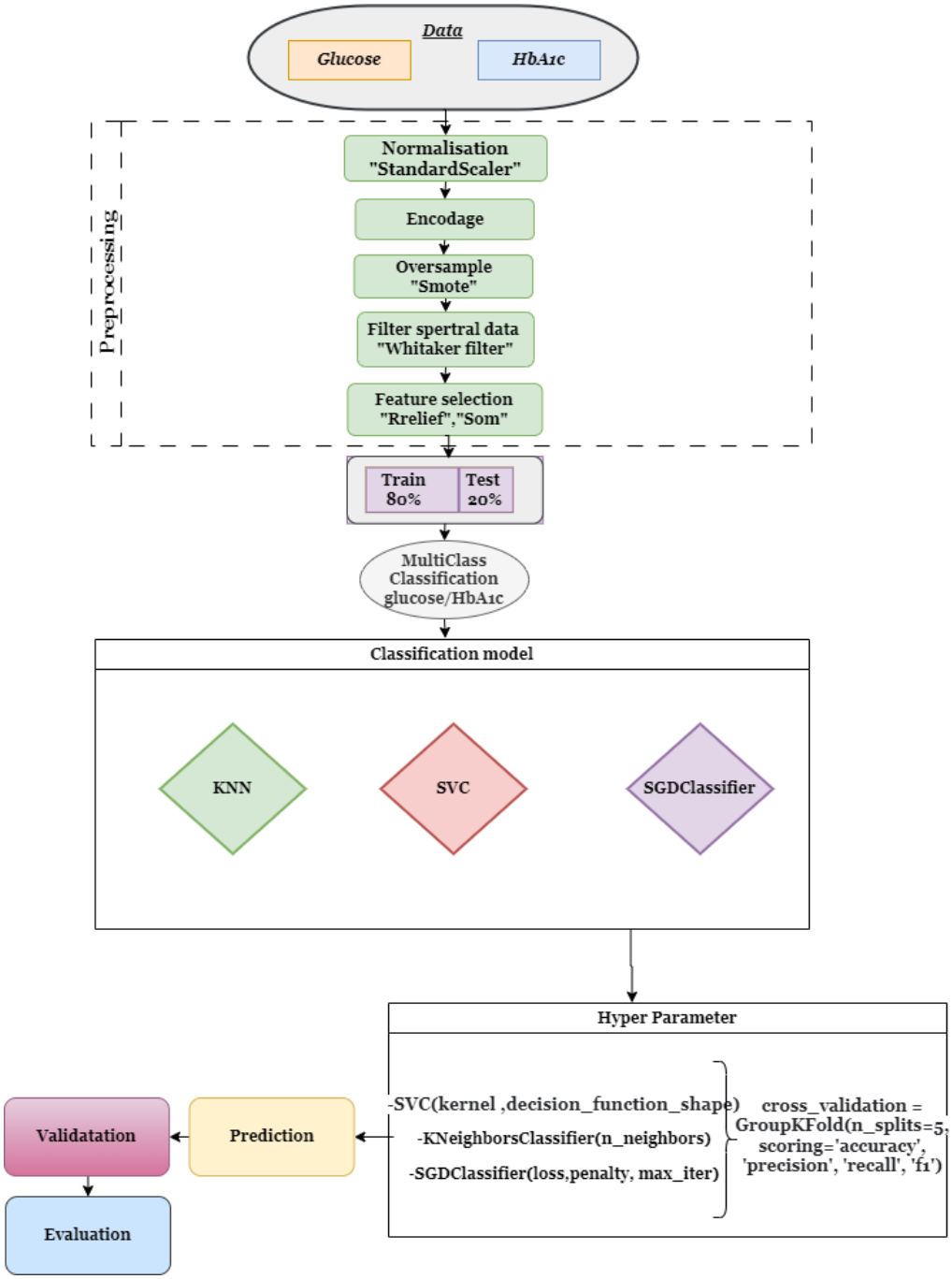


Figure 3.6: "Architecture of HbA1c/Glucose Classification"

3.3.5 Combined classification and regression model

The dataset utilized for classification consisted of all spectral data from "Input and Output HbA1c" and "Input and Output Glucose", serving as the input for Support Vector Classifier (SVC). The input for XGBoost for the quantification included all spectral data from "Input and Output HbA1c" and "Input and Output Glucose", augmented by the predicted class (high, low, ok) of glucose and HbA1c from SVC.

- To boost the previous model performance, we tested introducing a new class column. This column categorized the target glucose/hbA1c values into specific classes(ok,high,low), aiming to provide additional structure and context for the model.

- This approach made an improvement in our results. To allow the addition of this column, we first implemented a classification model(HbA1c/Glucose Classification)its architecture was explained previously.

- The choosed model from the models applied in HbA1c/Glucose Classification is SVC,The output from the Support Vector Classification (SVC) model was integrated as an additional feature into our initial dataset.

- Then we rescaled the data using StandardScaler to adjust the values of the data to a standard range. The updated data was passed into the XGBoost regression model to generate predictions.

- After that ,we evaluated and validated model using the evaluation metrics(RMSE,SD) plus the Clarke Error Grid for assessing the accuracy in predicting glucose levels,and Bland-Altman plot for HbA1c.
  Consult the figure 3.7.

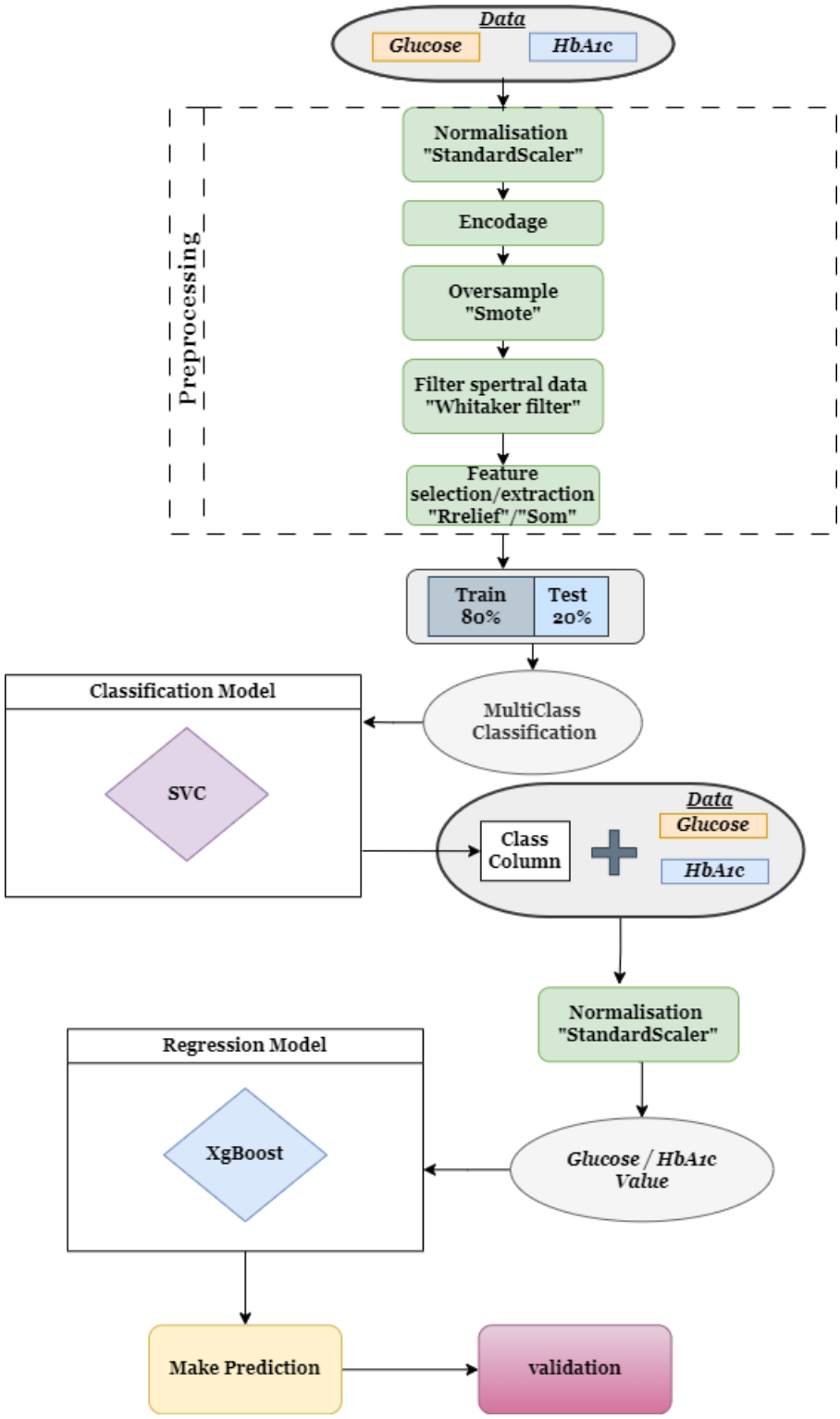


Figure 3.7: "Architecture Combined classification and regression model"

3.4 Conclusion

This chapter has provided an in-depth understanding of the key stages of machine learning development, from data collection and preparation to modeling and performance evaluation. Than, we have offered a detailed overview of the data used in this study .Finally, we presented the proposed architectures for the curren work.

Chapter 4

Model Implementation and Validation

4.1 Introduction

The performance of machine learning models is crucial for their effectiveness in solving real-world problems. In this chapter, we will emphasize the significance of model performance evaluation and optimization techniques. We'll delve into the implementation tools and libraries utilized in our project, providing insights into their roles and functionalities. Additionally, we'll conduct a detailed examination of the results obtained from each architecture, elucidating their strengths, weaknesses,and overall performance, thereby offering valuable insights into the efficacy of our models.

4.2 Development tools

4.2.1 Definition language Python

Python is an object-oriented, interpreted, high-level language with dynamic semantics. It is particularly appealing for Rapid Application Development as well as for usage as a scripting or glue language to join existing components together because of its high-level built-in data structures, dynamic typing, and dynamic binding. Python's easy-to-learn syntax prioritizes readability, which lowers software maintenance costs. Python promotes code reuse and software modularity by supporting modules and packages. For free on all major systems, both the Python interpreter and the large standard library are accessible in source or binary format. They may be shared without restriction.[41]

4.2.2 Python library

Pandas

Pandas is a Python module that offers expressive, quick, and flexible data structures that make dealing with "relational" or "labeled" data simple and straightforward. It seeks to serve as the essential high-level building block for using Python to undertake useful, real-world data analysis. Its overarching objective is to become the most potent and adaptable open source data analysis/manipulation tool accessible in any language. [42]

NumPy

NumPy is Python's essential scientific computing library. It is a Python library that includes a multidimensional array object, various derived objects (such as masked arrays and matrices), and a collection of routines for performing fast array operations such as mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation, and much more.[43]

Matplotlib

Matplotlib is a feature-rich Python visualization toolkit for static, animated, and interactive graphics. Matplotlib allows for visualizing both simple and complex data.[44]

Scikit-learn

Scikit-learn is a Python package that combines a variety of cutting-edge machine learning methods for medium-scale supervised and unsupervised applications. This package aims to make machine learning accessible to non-specialists by adopting a general-purpose high-level language. The focus is on simplicity of use, performance, documentation, and API consistency. [45]

1. **Pipeline:** Pipeline is a tool for preprocessing data using a list of transformers and, if desired, concludes with a final predictor for predictive modeling. It puts together steps with adjustable parameters and cross-validation capabilities. Pipelines make

modeling simpler than merely writing code by condensing logic into a single function call, streamlining routine tasks.[46]

2. **Grid search:** Grid search is the easiest method for modifying hyperparameters. In essence, we create a discrete grid within the hyperparameter domain. Next, we experiment with every possible combination of values in this grid, utilizing cross-validation to compute a few performance metrics. The ideal set of hyperparameter values is the grid point that maximizes the average value in cross-validation.Grid search finds the optimum location in the field since it is a thorough method that covers all possible combinations. Its major drawback is its extreme slowness.[47]

3. **StandardScaler:** Is a scikit-learn pretreatment approach that removes the mean and scales to unit variance in order to standardize features.[48]

4. **SMOTE:** Interpolates several existing data points from the minority class to create fresh samples of the positive class.[49]

TensorFlow

TensorFlow is an open-source toolkit that allows developers and academics to design Deep Learning templates and conduct complicated machine learning tasks. It is a toolset for tackling highly complicated mathematical issues in an easy and straightforward manner.[50]

4.3 Development platform

4.3.1 Collaboratory

Google Colaboratory, often known as Colab, is a cloud-based platform for machine learning research and instruction that is built on top of Jupyter Notebooks. It offers a runtime that is ready for deep learning and unfettered access to a powerful GPU.[51]With this platform, you can train machine learning models on the cloud directly. As a result, we don't need to install anything on our computer other than a browser. [52]

4.3.2 JupyterLab

JupyterLab is an extremely feature-rich and extensible notebook writing and editing program. Its main objective is to provide tools (and standards) for interactive computing using computational notebooks.JupyterLab is the brother program of Jupyter Notebook and Jupyter Desktop, two additional notebook writing programs that are part of the Project Jupyter family. When compared to Jupyter Notebook, JupyterLab provides a more sophisticated, feature-rich, and customized experience.[53]

4.4 Classification Evaluation Metrics

4.4.1 Confusion Matrix

The confusion matrix is a cross-tabulation that documents the frequency of occurrences between two raters, as well as the anticipated and true/actual classifications. The rows show the actual categorization, while the columns represent the model prediction for consistency's sake across the whole publication.[54]

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

- **True Positives (TP)**: Number of cases where the model correctly predicted the positive class.
  - *Example*: A medical test correctly identifies a sick person as sick.
- **True Negatives (TN)**: Number of cases where the model correctly predicted the negative class.
  - *Example*: A medical test correctly identifies a healthy person as healthy.
- **False Positives (FP)**: Number of cases where the model incorrectly predicted the positive class for an instance that is actually negative.
  - *Example*: A medical test incorrectly identifies a healthy person as sick.
- **False Negatives (FN)**: Number of cases where the model incorrectly predicted the negative class for an instance that is actually positive.

– *Example*: A medical test incorrectly identifies a sick person as healthy.

4.4.2 Classification Report

Accuracy

Calculates the proportion of accurate predictions to all cases examined. [55]

$$Accuracy = \frac{CorrectPredictions}{TotalPredictions} = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.1}$$

Precision

The number of accurately predicted positive patterns from the total number of anticipated patterns in a positive class.[55]

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

F1-score

The F1 score is the harmonic mean of memory and accuracy. It offers a clear indication of the model's performance and is used to evaluate test accuracy. The objective is to acquire the F1 score as close to 1 as is practical. The score goes from 0 to 1.[56]

$$F1 - score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.3}$$

4.4.3 Sensitivity(Recall)

Is employed to calculate the percentage of positive patterns that receive accurate classification. .[55]

$$Sensitivity(recall) = \frac{TP}{TP + FN} \tag{4.4}$$

4.4.4 Specificity

Used to calculate the percentage of negative patterns that are appropriately categorized.[55]

$$Specificity = \frac{TN}{TN + FP} \tag{4.5}$$

4.5 Regression Evaluation Metrics

4.5.1 Mean Absolute Error (MAE)

The mean absolute error (MAE) describes the difference between the original and predicted values and is calculated as the dataset's total alteration mean. [57]

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4.6}$$

4.5.2 Mean Square Error (MSE)

Also called the Mean Squared Deviation the average squared error between the predicted and actual values. It takes positive or zero values .[58]

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.7}$$

4.5.3 Root Mean Squared Error (RMSE)

Also called the square root of the mean of the squares of all the mistakes is the deviation . Stated differently, the RMSE can be defined as the standard deviation of the errors. Once more, RMSE indicates the proximity of the line of best fit to the data set.[58]

$$RMSE = \sqrt{\text{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.8}$$

4.5.4 Standard Deviation (SD)

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{4.9}$$

4.6 Validation Technique

4.6.1 Cross-validation

Cross-validation is a statistical approach for assessing and comparing learning algorithms that divides data into two sections: one for learning or training a model and another for validating the model. To ensure that every data point has an equal chance of being

verified against, the training and validation sets must cross over in subsequent rounds of traditional cross-validation. K-fold cross-validation is the most fundamental type of cross-validation. Special instances of k-fold cross-validation include other types of cross-validation.[59]

4.6.2 Bland-Altman

The Bland-Altman analysis is frequently utilized in studies examining the concordance between two methods of the same medical measurement[60]. Bland and Altman introduced the Bland-Altman plot as a tool to assess agreement between two quantitative measurements. They devised a method to quantify this agreement by establishing limits of agreement, calculated using the mean and standard deviation (s) of the measurement differences. To validate assumptions like normality, they employed graphical techniques. The resulting XY scatter plot displays the difference (A-B) on the Y-axis and the average of the measurements ((A+B)/2) on the X-axis. [61].

4.6.3 Clarke Error Grid

The Clarke error grid method is used to determine the clinical relevance of variations between the glucose measurement methodology being tested and the venous blood glucose reference data. [62]

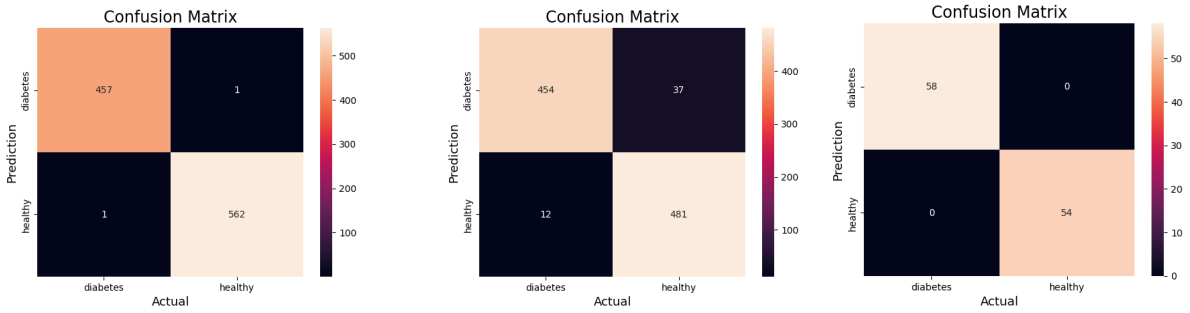
- Zone A indicates accurate predictions with no risk to patient care.
- Zone B suggests slight deviations in treatment decisions.
- Zones C, D, and E represent increasing levels of clinical risk, indicating potential dangers in treatment decisions.

4.7 Models implementation and evaluation

4.7.1 Implementation and evaluation of 'Diabete Classification'

Evaluation of Model

- SVM: Let's examine the outcomes of the diabetes binary classification by reviewing the confusion matrices shown in image 4.1

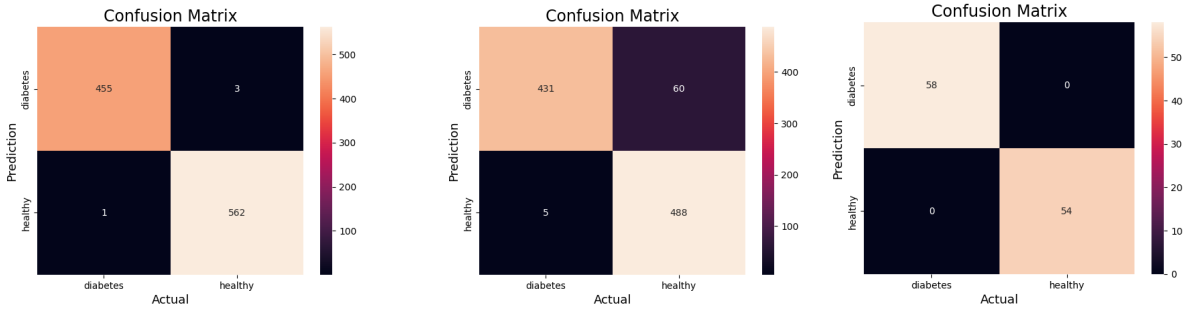


(a) Forearm (b) Wrist (c) Index finger

Figure 4.1: Confusion matrix for SVM

– The analysis of confusion matrices reveals the classification performance across different body parts for diabetes and healthy classes. In the forearm, the model correctly classified 457 diabetes cases (TP) and 562 healthy cases (TN), with only 1 diabetes case misclassified as healthy (FN) and 1 healthy case misclassified as diabetes (FP). Moving to the wrist, all 491 diabetes cases were correctly identified (TP), but 37 were misclassified as diabetes(FP), while 493 healthy cases were correctly identified (TN) with 12 misclassified as healthy (FN). Notably, the finger consistently showed perfect classification for both diabetes (58 TP) and healthy (54 TN) cases, indicating high accuracy across both classes.

- SGDClassifier:Let's analyze the results SGDClassifier's classification by examining the confusion matrices depicted in the image 4.2



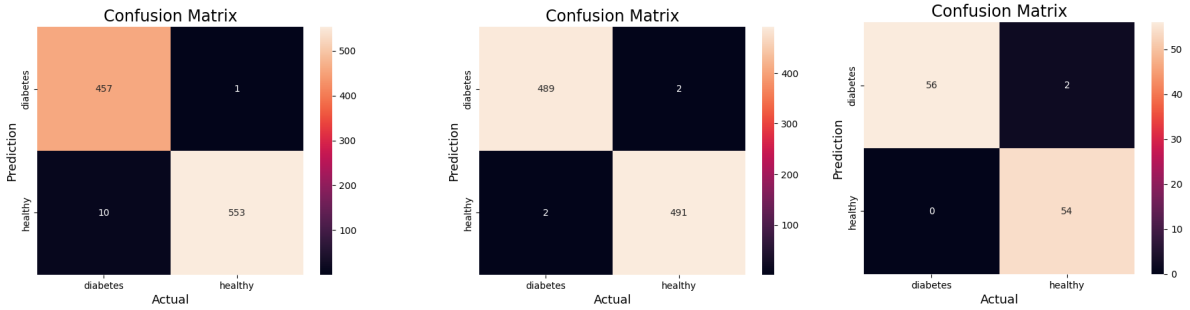
(a) Forearm (b) Wrist (c) Index finger

Figure 4.2: Confusion matrix for SGDClassifier

– Using SGDClassifier, the confusion matrix analysis reveals the classification

performance across different body parts. In the forearm, the model correctly classified 455 diabetes cases (TP) out of 458, with 3 misclassified as diabetes (FP). For healthy cases, it correctly identified 562 (TN) out of 563, with 1 misclassified as healthy (FN). Moving to the wrist, the model correctly classified 431 diabetes cases (TP) out of 491, but missed 60 (FP). It correctly identified 488 healthy cases (TN) out of 493, with 5 misclassified as healthy (FN). In the finger, the model achieved perfect classification with all 58 diabetes cases and all 54 healthy cases correctly identified.

- HGBoost:Let's examine the outcomes of the HGBoost classification as depicted in the figure. 4.3



(a) Forearm (b) Wrist (c) Index finger

Figure 4.3: Confusion matrix for HGBoost

  - In the forearm, the model correctly classified 487 diabetes cases (TP) , with 1 cases misclassified as diabetes (FP). It also correctly classified 553 healthy cases (TN) , with 10 cases misclassified as healthy (FN). Moving to the wrist, the model correctly classified 489 diabetes cases (TP) but missed 2 (FP), and it correctly classified 491 healthy cases (TN) but missed 2 (FN). In the finger, the model correctly classified 56 diabetes cases (TP) but missed 2 (FP), and all 54 healthy cases were correctly classified.

Result of model

- The analysis reveals distinct performances across different regions: SVC excelled in the forearm with an accuracy of 0.94, specificity of 0.95, and sensitivity of 0.93. Meanwhile, HGBoost showed superior results in the wrist, achieving an accuracy of

0.95, specificity of 0.97, and sensitivity of 0.94. In the index finger, both SVM and HGBoost achieved high accuracy rates of 0.96, with SVM demonstrating specificity and sensitivity of 0.94 and 0.97 respectively, and HGBoost showing specificity of 0.93 and sensitivity of 0.98. As shown in figure 4.4

Diabete Classification accuracy| specificity | sensitivity			
region\model	SVM	SGDClassifier	HGBoost
Forearm	0.94 | 0.95| 0.93	0.91| 0.88 | 0.95	0.93| 0.95| 0.92
Wrist	0.91 | 0.95 | 0.86	0.89 | 0.88 | 0.92	0.95|0.97|0.94
Index finger	0.96 | 0.94 | 0.97	0.94 | 0.91 | 0.97	0.96 | 0.93| 0.98

Figure 4.4: Diabete classification result

4.7.2 Implementation and evaluation of 'Regression'

After performing cross-validation, we calculated the RMSE and SD for each body region for both Glucose and HbA1C using diffrent algorithms as presented in the figure below4.5

HbA1c RMSE | SD				
region\model	KNN	Gaussian	SVR	XGBoost
Forearm	2.01 | 0.27	8.97 | 0.72	1.72 | 0.36	1.25| 0.45
Wrist	2.01| 0.69	8.95 | 0.72	1.31 | 0.22	0.98| 0.40
Index finger	1.22 | 0.47	0.31 | 0.09	1.18 | 0.29	0.98 | 0.57

(a) HbA1c

Glucose RMSE | SD				
region\model	KNN	Gaussian	SVR	XGBoost
Forearm	72.09 | 17.42	162.83 | 21.93	69.05 | 25.67	46.66 | 17.62
Wrist	74.95 | 12.86	158.67 | 23.58	70.64 | 25.51	33.73 | 13.74
Index finger	52.50 | 25.57	38.60 | 10.52	73.83| 22.43	38.86 | 14.84

(b) Glucose

Figure 4.5: Regression result

- In the forearm region, XGBoost and SVR emerged as the top performers for accurately quantifying hbA1c and glucose levels, showcasing their robust predictive capabilities. Moving to the wrist region, these models continued to excel with minimal error, reaffirming their effectiveness in precise measurements of hbA1c and glu-

cose. Meanwhile, in the index finger region, Gaussian and XGBoost demonstrated superior performance in quantifying hbA1c and glucose levels.

Xgboost with added column

- To improve the performance of regression model we tryed adding a new feature (class of glucose/HbA1c) the result improved as illustrated in figure 4.6.

XGboost with added column RMSE|SD		
region	**HbA1c**	**Glucose**
Forearm	0.33 | 0.08	16.60 | 9.58
Wrist	0.35 | 0.048	13.07| 7.97
Index finger	0.56 | 0.33	24.93| 15.14

Figure 4.6: Xgboost with added column result

4.7.3 Implementation and evaluation of 'HbA1c/Glucose Classification'

HbA1c accuracy| Precision|Recall			
region\model	**knn**	**svm**	**SGDClassifier**
Forearm	0.99|0.99|0.99	0.99|0.99|0.99	0.99|0.99|0.99
Wrist	0.98|0.98|0.98	0.99|0.99|0.99	0.99|0.99|0.99
Index finger	0.97|0.97|0.97	0.98|0.98|0.98	0.98|0.98|0.98

(a) HbA1c

Glucose accuracy| Precision|Recall			
region\model	**knn**	**svm**	**SGDClassifier**
Forearm	0.87| 0.87 | 0.91	0.84| 0.88 | 0.92	0.83 | 0.83 | 0.89
Wrist	0.83| 0.827|0.83	0.92| 0.81 | 0.89	0.94 | 0.81 | 0.89
Index finger	0.92| 0.88 |0.79	0.96 | 0.82 | 0.82	0.85| 0.85 | 0.79

(b) Glucose

Figure 4.7: 'HbA1c/Glucose Classification'result

- This model was implemented to get the class feature mentioned previously.We picked the best performing model SVM in both Glucose and HbA1c data.

4.7.4 Implementation and evaluation of 'Combined classification and regression model'

- Combined model regression classification result :

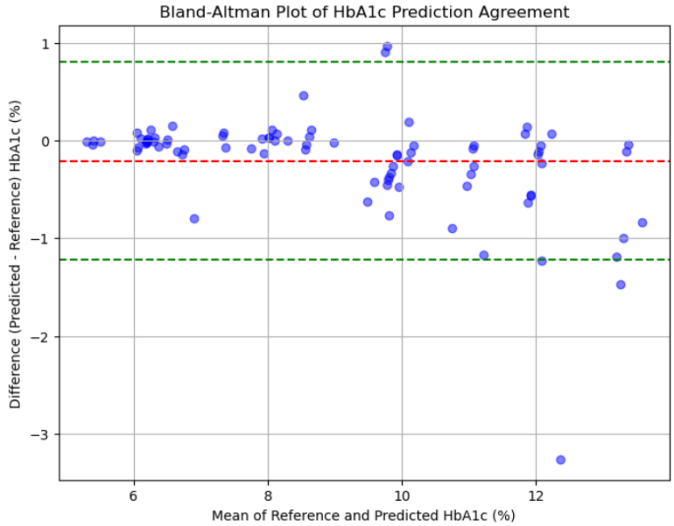
RMSE | SD		
region	**HbA1c**	**Glucose**
Forearm	0.55 | 0.51	21.28| 20.76
Wrist	0.44 | 0.41	15.49| 15.43
Index finger	1.08 | 0.97	18.62 | 18.54

Figure 4.8: Combined model regression classification result

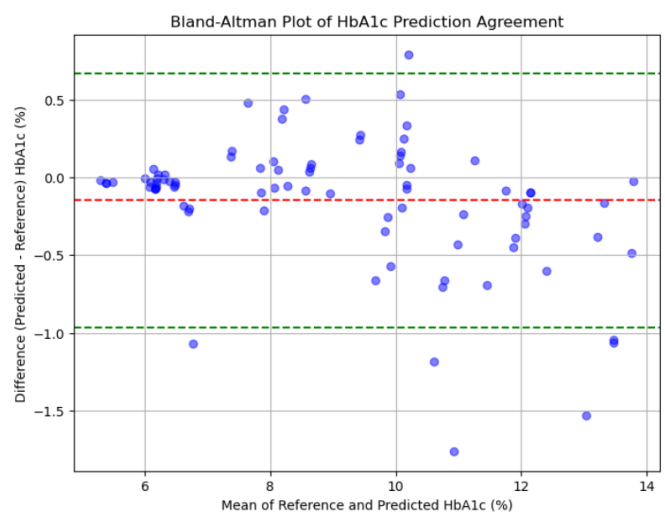
- We employed a combination of classification and regression techniques to analyze glucose and HbA1c levels across various body regions. The most effective model utilized SVC for classification and Xgboost for regression. Notably, the optimal results were achieved in the wrist area for both glucose and HbA1c predictions.Consult the figure3.7.

4.7.5 Evaluation HbA1c

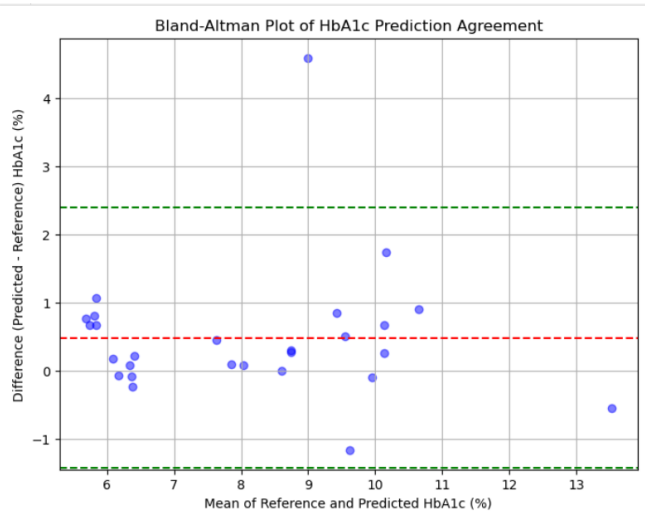
- We utilized the Bland-Altman plot to evaluate the agreement between predicted and reference values of HbA1c(true value) in a quantification model. This plot compares the differences between predicted and reference HbA1c values against their mean. Each data point on the scatter plot represents a pair of these values, visualizing both the spread and any potential bias between predictions and actual measurements. The red dashed line across the plot indicates the average difference between predicted and reference values. green dashed lines would depict the 95% limits of agreement around the mean difference, reflecting the range within which most differences lie. This plot aids in identifying any systematic bias between the methods (illustrated by deviations from zero on the mean difference line) and evaluating the precision of agreement (indicated by the width of the limits of agreement).This graphical analysis assess the accuracy and reliability of the HbA1c quantification model.Refer to the figure4.9.



(a) Forearm



(b) Wrist



(c) Index finger

Figure 4.9: Bland altman evaluation of HbA1C

4.7.6 Evaluation Glucose

- The forearm results show that the majority of measurements (73 out of the total) fall within Zone A, indicating clinically acceptable accuracy. There are very few errors (2 in Zone B and 1 in Zone D), and no measurements in the higher risk Zones C and E. This suggests that overall, the device performs well within acceptable limits for clinical use, with only minor errors observed in a small number of cases.Consult the figure 4.10.

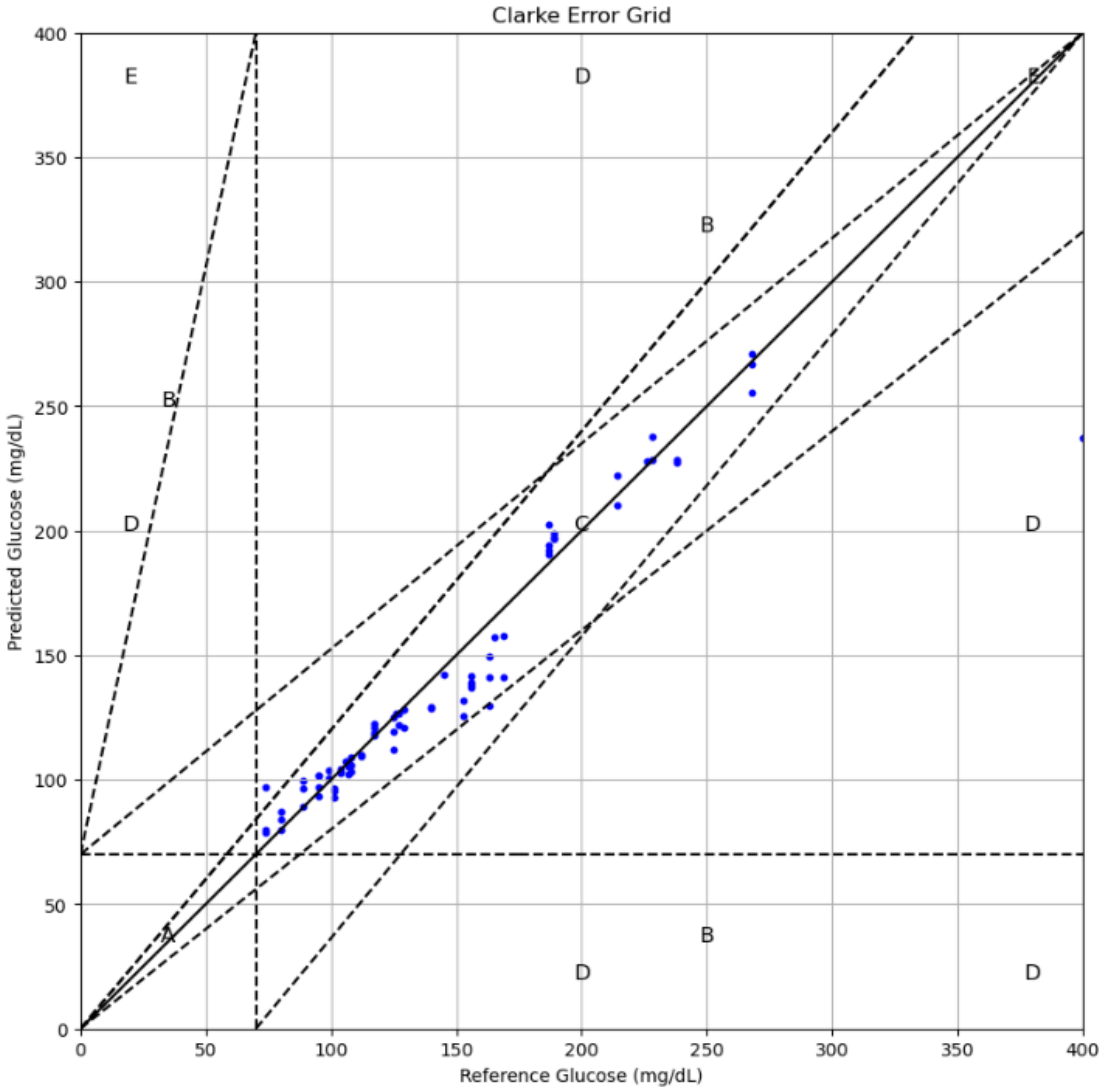


Figure 4.10: Clarck error grid forearm

- The wrist results indicate that the model generally performs well, with 72 measurements falling within Zone A, demonstrating clinically acceptable accuracy. No measurements were found in Zones B, C, or E, indicating the absence of minor, moderate, or critical errors. However, 4 measurements fell into Zone D, suggesting that in some cases, the device's readings significantly deviated from true glucose levels, which could impact treatment decisions. Therefore, while the overall performance is satisfactory, further investigation and potential error mitigation in Zone D are recommended.Refer to the figure 4.11.

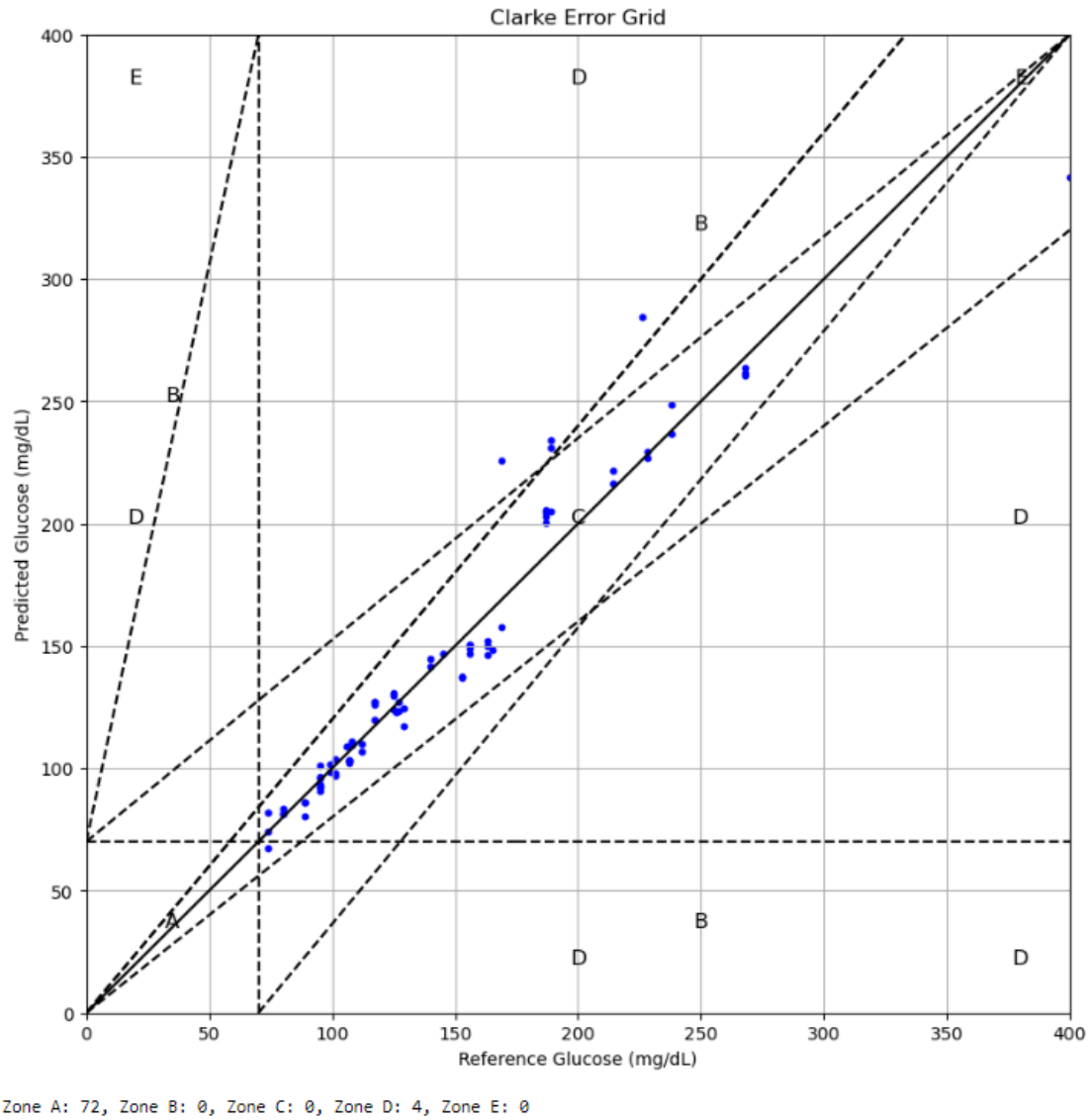
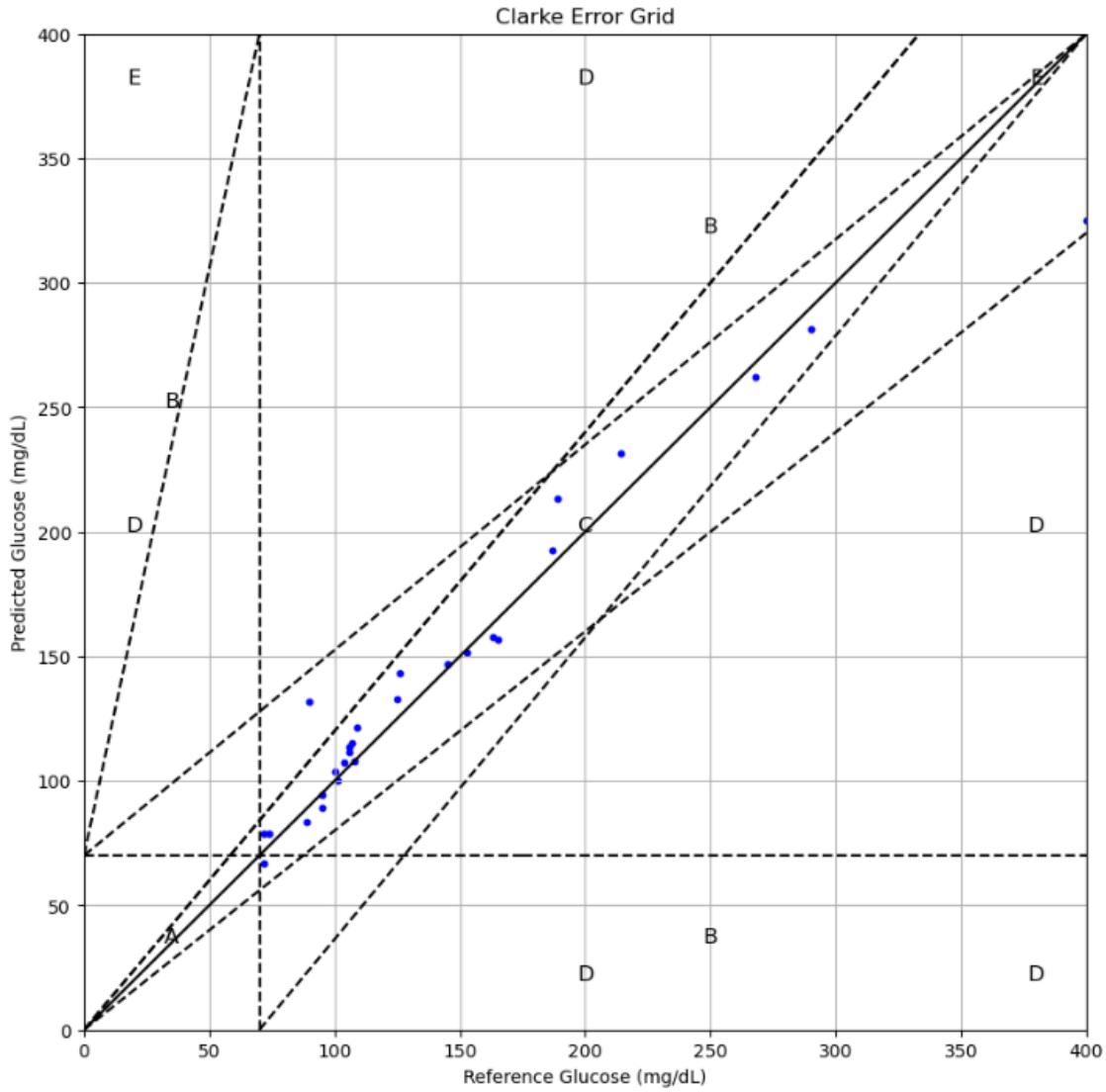


Figure 4.11: Clarck error grid wrist

- The Clarke Error Grid analysis in the index finger the glucose measurements shows that the majority of readings (26 out of 27) fall within Zone A, indicating clinically acceptable accuracy. One measurement is in Zone B, suggesting a minor error that is unlikely to significantly affect treatment decisions. Importantly, there are no measurements in Zones C, D, or E, indicating no errors that could lead to benign, clinically significant, or critical treatment decisions opposite to what is needed. Overall, these results demonstrate that the device performs well within acceptable limits, with only a minor deviation observed in a single instance.As illustrated in the figure 4.12



Zone A: 26, Zone B: 1, Zone C: 0, Zone D: 0, Zone E: 0

Figure 4.12: Clarck error grid finger

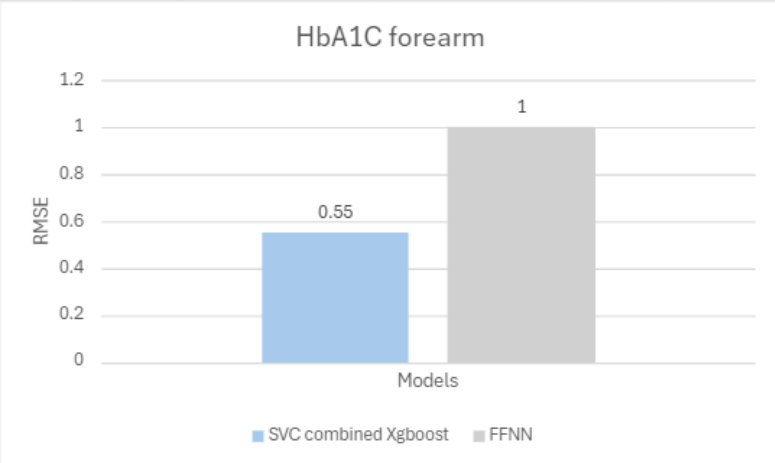
4.8 Related work

In "Quantifcation of glycated hemoglobin and glucose in vivo using Raman spectroscopy and artifcial neural networks" article, two approaches were implemented, and both of them used the same preprocessing technique and deep learning algorithm:Zernike polynomial filttering combined with genetic algorithms and Whitaker filtter to reduce fuorescence and shot noise, followed by RReliefF and SOM for feature selection and extraction,first combined with FFNN algorithm for regression achieved an error in the predictive model of 0.69% for HbA1c and 30.12 mg/dL for glucose and the second for the classification the Patients were classifed into three categories:healthy,prediabetes, and T2D. The FFNN obtained a accuracy of 96.01% .[63]

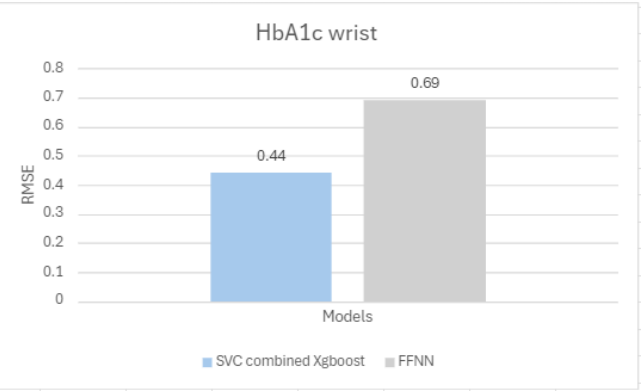
4.9 Our proposed models vs related work Comparison

4.9.1 Combined classification regression comparison

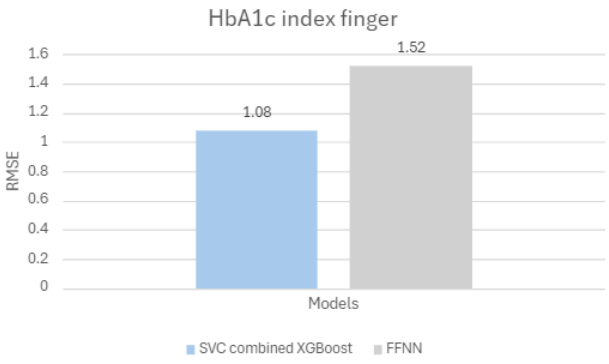
1. A clustered column chart illustrating a comparative analysis between the RMSE results of related work and the proposed method for HbA1c quantification across each body region, Refer to the figure 4.13.



(a) Forearm



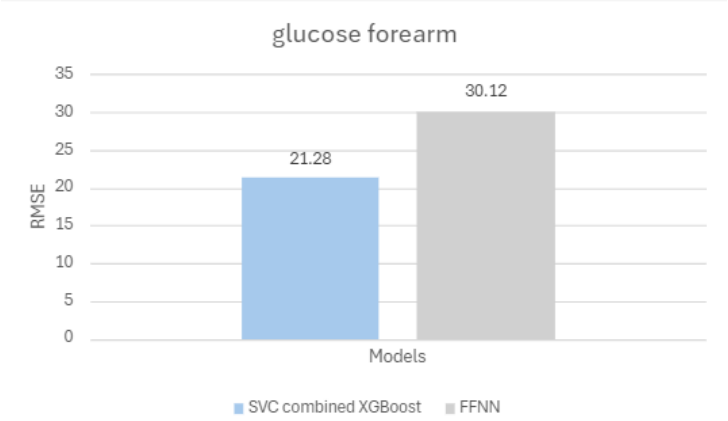
(b) Wrist



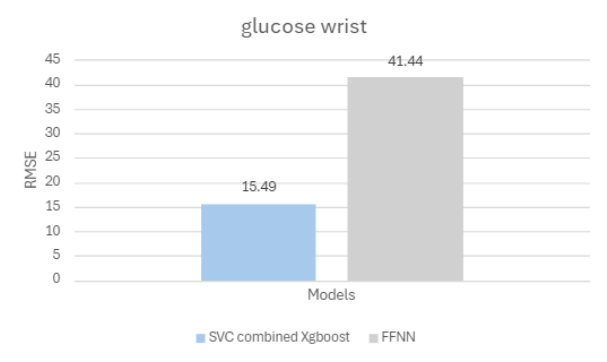
(c) Index finger

Figure 4.13: Clusterd column HbA1c

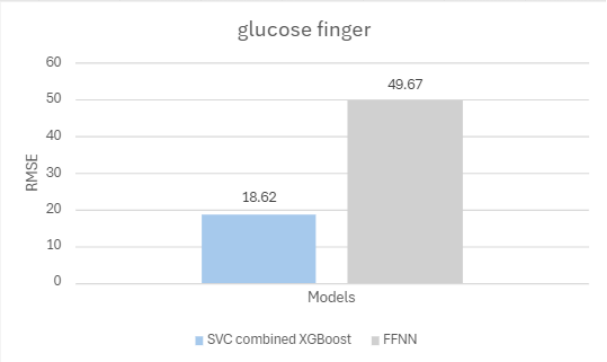
2. A clustered column chart comparing RMSE results between existing methods and our proposed approach for glucose quantification across various body regions, As illustrated in figure 4.14.



(a) Forearm



(b) Wrist



(c) Index finger

Figure 4.14: Clusterd column glucose

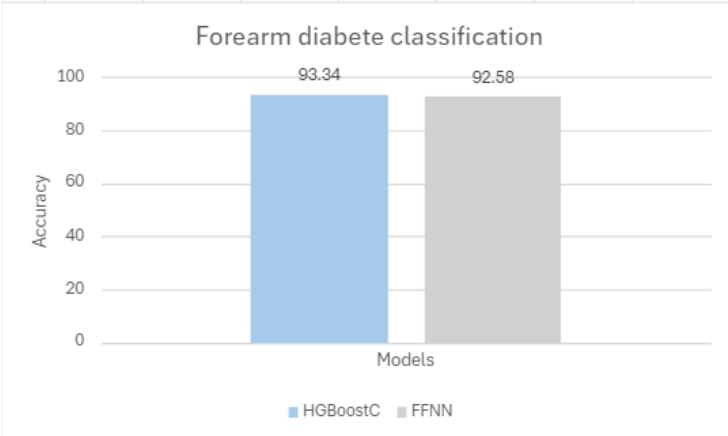
- In a comparison , the article previously discussed using the same dataset reported predictive errors of 0.69% for HbA1c in the wrist region and 30.12 mg/dL for glucose in the forearm region using FFNN.

  In contrast, our approach utilizing a combined classification and regression model (SVC-XGBoost) achieved significantly lower errors of 0.44% for HbA1c and 15.49 mg/dL for glucose in the wrist region.

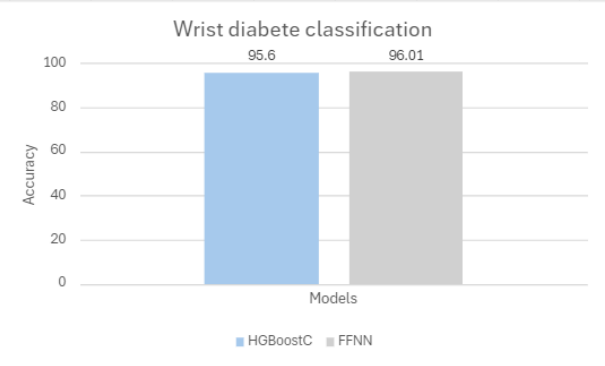
  This improvement underscores the efficacy of integrating multiple machine learning techniques into a unified predictive framework, highlighting advancements in accuracy and performance over previous methodologies. Consult the figure 4.13 and 4.14.

4.9.2 Diabetes classification comparison

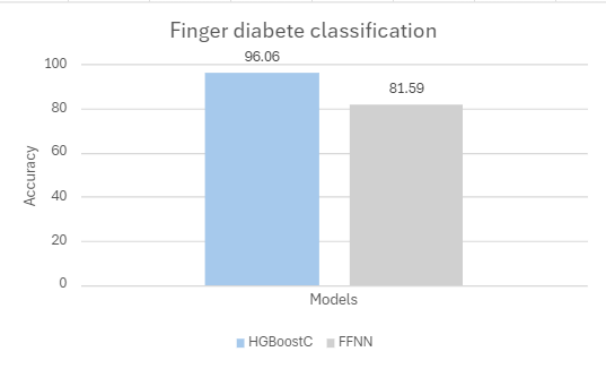
1. A clustered column chart comparing accuracy results between existing methods and our proposed model across various body regions, As illustrated in figure 4.15.



(a) Forearm



(b) Wrist



(c) Index finger

Figure 4.15: Clusterd column diabete classification

- In the evaluation of classification performance across different body parts, both FFNN and HGBoost exhibited similar levels of accuracy for wrist and forearm classifications. However, a significant improvement was observed in the accuracy for the index finger classification when using HGBoostC . Specifically, the accuracy for the index finger increased from 81.59% with FFNN to an impressive 96.06% with HGBoostC.Consult the figure4.15.

4.10 Conclusion

In this chapter, we dive into the essential components required for implementing various architectures, including development tools, platforms, and libraries. We explored the significance of selecting appropriate evaluation metrics and validation techniques to ensure the robustness and effectiveness of these architectures. Through detailed implementation and evaluation of each architecture provided a hands-on understanding of their applicability and performance.Finally, we conducting a comprehensive comparison of the proposed models with the existent model.

General Conclusion

Spectroscopy integrated with machine learning has emerged as a potent tool across diverse fields, promising innovative solutions to longstanding challenges. In the realm of healthcare, traditional methods for quantifying glucose and HbA1c levels often entail invasive blood analyses, presenting inherent drawbacks such as discomfort, inconvenience, and potential health risks. Recognizing this limitation, the fusion of spectroscopy with machine learning offers a transformative approach. The central objective of our work lies in implementing these integrated models, aiming to pioneer a non-invasive alternative that addresses the shortcomings of conventional blood analysis techniques. Through the seamless integration of spectroscopy and ML, we aspire to contribute to the advancement of healthcare diagnostics, facilitating improved patient care and outcomes.

To address the shortcomings of conventional blood analysis techniques, we implemented a hybrid approach combining classification and regression models. Specifically, we utilized support vector classification (SVC) for categorizing glucose and HbA1c levels, alongside XGBoost for regression tasks. This integrated model achieved remarkable accuracy, with errors of only 0.44% for HbA1c and 15.49 mg/dL for glucose levels.

However,Our dataset is limited in size, This highlights the importance of ongoing efforts to expand and diversify our data collection practices.
The most promising outcomes were achieved in the wrist area. This model has the potential to be integrated into a watch, where it could display whether the Glucose/HbA1c levels are high, low, or within the normal range, along with the specific numerical value. This innovation could greatly enhance monitoring and management of glucose levels for individuals needing regular health assessments in a non invasive way.

Bibliography

[1] https://weeklysciencequiz.blogspot.com/2011/09/when-light-meets-matter.html. Accessed on February 12, 2024.

[2] Mohammad Mustafa Taye. Understanding of machine learning with deep learning: Architectures, workflow, applications and future directions. *Computers*, 12(5), 2023.

[3] Accessed on june, 2024.

[4] Nabeel Othman. Ir spectroscopy in qualitative and quantitative analysis. In Marwa El-Azazy, Khalid Al-Saad, and Ahmed S. El-Shafie, editors, *Infrared Spectroscopy*, chapter 4. IntechOpen, Rijeka, 2022.

[5] Donald L. Pavia, Gary M. Lampman, George S. Kriz, and James R. Vybiral. *Introduction to Spectroscopy (4th Edition)*. Cengage Learning, 2014.

[6] David Harvey. https://resources.saylor.org/wwwresources/archived/site/wp-content/uploads/2012/07/Chapter1011.pdf. Accessed on February, 2024.

[7] Andreas Barth. https://www.su.se/polopoly$_fs$/1.521101.1602178917!/$menu/standard/file/Intro$

[8] https://sist.sathyabama.ac.in/sist$_c$$oursematerial/uploads/SCY1612.pdf.AccessedonFebruary$, 2024.

[9] Theophile Theophanides. Introduction to infrared spectroscopy. In Theophile Theophanides, editor, *Infrared Spectroscopy*, chapter 0. IntechOpen, Rijeka, 2012.

[10] *Raman spectroscopy as a powerful analytical tool: probing the structure of matter*. Centres Científics i Tecnològics. Universitat de Barcelona, 2012.

[11] Cecilia Carlota Barrera-Ortega, America Rosalba Vazquez Olmos, Roberto Isaac Sato Berrú, and Pineda Dominguez Karla Itzel. Application of raman spectroscopy for dental enamel surface characterization. In Marwa El-Azazy, Khalid Al-Saad, and Ahmed S. El-Shafie, editors, *Infrared Spectroscopy*, chapter 10. IntechOpen, Rijeka, 2022.

[12] Jianhua Zhao, Harvey Lui, David I. McLean, and Haishan Zeng. Real-time raman spectroscopy for noninvasive in vivo skin analysis and diagnosis. In Domenico Campolo, editor, *New Developments in Biomedical Engineering*, chapter 24. IntechOpen, Rijeka, 2010.

[13] M.H. Wathsala N. Jinadasa, Amila C. Kahawalage, Maths Halstensen, Nils-Olav Skeie, and Klaus-Joachim Jens. Deep learning approach for raman spectroscopy. In Chandra Shakher Pathak and Samir Kumar, editors, *Recent Developments in Atomic Force Microscopy and Raman Spectroscopy for Materials Characterization*, chapter 5. IntechOpen, Rijeka, 2021.

[14] Chemometrics for raman spectroscopy harmonization. *Applied Spectroscopy*, 76(9):1021–1041, 2022.

[15] Ramanspy: An open-source python package for integrative raman spectroscopy data analysis. *bioRxiv*, 2023.

[16] Recent progresses in machine learning assisted raman spectroscopy. *Advanced Optical Materials*, 11, 2023.

[17] Machine learning and deep learning. *Electronic Markets*, 31(3):685–695, 2021.

[18] Iqbal H Sarker. Machine learning: Algorithms, Real-World applications and research directions. *SN Computer Science*, 2(3):160, March 2021.

[19] Zhongheng Zhang. Introduction to machine learning: k-nearest neighbors. *Ann. Transl. Med.*, 4(11):218, June 2016.

[20] https://www.datacamp.com/tutorial/decision-tree-classification-python. Accessed on march 04, 2024.

[21] Robin Genuer and Jean-Michel Poggi. *Random Forests*, pages 33–55. 09 2020.

[22] Yaping Qi, Dan Hu, Yucheng Jiang, Zhenping Wu, Ming Zheng, Esther Xinyi Chen, Yong Liang, Mohammad A. Sadi, Kang Zhang, and Yong P. Chen. Recent progresses in machine learning assisted raman spectroscopy. *Advanced Optical Materials*, 11(14):2203104, 2023.

[23] https://scikit-learn.org/stable/modules/sgd.html. Accessed on june, 2024.

[24] https://xgboost.readthedocs.io/en/stable/. Accessed on march 14, 2024.

[25] Deeman Mahmood. Principal component analysis (pca), 03 2018.

[26] B.K. Tripathy, S Anveshrithaa, and Shrusti Ghela. *Isomap*, pages 53–65. 07 2021.

[27] Gunganist Kongklad, Ratchapak Chitaree, Tana Taechalertpaisarn, Nathinee Panvisavas, and Noppadon Nuntawong. Discriminant analysis pca-lda assisted surface-enhanced raman spectroscopy for direct identification of malaria-infected red blood cells. *Methods and Protocols*, 5(3), 2022.

[28] Waste cooking oil classification through raman spectroscopy and multivariate analysis, 2023.

[29] https://doi.org/10.1038/s43586-022-00184-w. Accessed on june, 2024.

[30] Sang Kwak and Jong Kim. Statistical data preparation: Management of missing values and outliers. *Korean Journal of Anesthesiology*, 70:407, 08 2017.

[31] Elisha Blessing and Hubert Klaus. Normalization and standardization: Methods to preprocess data to have consistent scales and distributions. 2237:10, 12 2023.

[32] Kedar Potdar, Taher Pardawala, and Chinmay Pai. A comparative study of categorical variable encoding techniques for neural network classifiers. *International Journal of Computer Applications*, 175:7–9, 10 2017.

[33] Roweida Mohammed, Jumanah Rawashdeh, and Malak Abdullah. Machine learning with oversampling and undersampling techniques: Overview study and experimental results. pages 243–248, 04 2020.

[34] https://corporatefinanceinstitute.com/resources/business-intelligence/data-smoothing/. Accessed on june, 2024.

[35] https://www2.informatik.uni-hamburg.de/WTM/ps/dalicann.pdf. Accessed on june, 2024.

[36] Yaping Qi, Dan Hu, Yucheng Jiang, Zhenping Wu, Ming Zheng, Esther Xinyi Chen, Yong Liang, Mohammad A. Sadi, Kang Zhang, and Yong P. Chen. Recent progresses in machine learning assisted raman spectroscopy. *Advanced Optical Materials*, 11(14):2203104, 2023.

[37] https://jhhhm.halal.ac.ir/article$_1$92174$_e$63d3fe4bf94aa8b78d0a6d15836c061.$pdf.Accessedonjune$, 2024.

[38] Ranjit Sahu and Shaul Mordechai. Spectroscopic techniques in medicine: The future of diagnostics. *Applied Spectroscopy Reviews*, 51:484–499, 07 2016.

[39] https://osf.io/v32d4/. Accessed on june, 2024.

[40] https://osf.io/v32d4/. Accessed on june, 2024.

[41] https://www.python.org/doc/essays/blurb. Accessed on may, 2024.

[42] https://pandas.pydata.org/docs/getting$_s$$tarted/overview.html.Accessedonmay$, 2024.

[43] https://numpy.org/doc/stable/. Accessed on may, 2024.

[44] https://matplotlib.org/. Accessed on may, 2024.

[45] https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf. Accessed on may, 2024.

[46] https://scikitlearn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html: :text=Pipeline Accessed on may, 2024.

[47] https://www.yourdatateacher.com/2021/05/19/hyperparameter-tuning-grid-search-and-random-search/. Accessed on may, 2024.

[48] https://www.geeksforgeeks.org/what-is-standardscaler/. Accessed on june, 2024.

[49] Peter Gnip, Liberios Vokorokos, and Peter Drotár. Selective oversampling approach for strongly imbalanced data. *PeerJ Comput. Sci.*, 7(e604):e604, June 2021.

[50] https://blent.ai/blog/a/tensorflow-deep-learning-python. Accessed on may, 2024.

[51] Tiago Pessoa, Raul Medeiros, Thiago Nepomuceno, Gui-Bin Bian, V.H.C. Albuquerque, and Pedro Pedrosa Filho. Performance analysis of google colaboratory as a tool for accelerating deep learning applications. *IEEE Access*, PP:1–1, 10 2018.

[52] https://ledatascientist.com/google-colab-le-guide-ultime/. Accessed on may, 2024.

[53] https://jupyterlab.readthedocs.io/en/latest/. Accessed on may, 2024.

[54] Margherita Grandini, Enrico Bagli, and Giorgio Visani. Metrics for multi-class classification: an overview. *ArXiv*, abs/2008.05756, 2020.

[55] Mohammad Hossin and Sulaiman M.N. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining Knowledge Management Process*, 5:01–11, 03 2015.

[56] Ismail Muraina, Olayemi Adesanya, and Solomon Abam. Data analytics evaluation metrics essentials: Measuring model performance in classification and regression. 08 2023.

[57] Anand Singh Rajawat, Omair Mohammed, Rabindra Nath Shaw, and Ankush Ghosh. Chapter six - renewable energy system for industrial internet of things model using fusion-ai. In Rabindra Nath Shaw, Ankush Ghosh, Saad Mekhilef, and Valentina Emilia Balas, editors, *Applications of AI and IOT in Renewable Energy*, pages 107–128. Academic Press, 2022.

[58] Vagelis Plevris, German Solorzano, Nikolaos Bakas, and Mohamed Ben Seghier. Investigation of performance metrics in regression analysis and machine learning-based prediction models. 06 2022.

[59] Payam Refaeilzadeh, Lei Tang, and Huan Liu. *Cross-Validation*, pages 532–538. Springer US, Boston, MA, 2009.

[60] Nurettin Özgür Doğan. Bland-Altman analysis: A paradigm to understand correlation and agreement. *Turk. J. Emerg. Med.*, 18(4):139–141, December 2018.

[61] Davide Giavarina. Understanding bland altman analysis. *Biochem. Med. (Zagreb)*, 25(2):141–151, June 2015.

[62] https://www.mathworks.com/matlabcentral/fileexchange/20545-clarke-error-grid-analysis. Accessed on june, 2024.

[63] Naara González-Viveros, Jorge Castro-Ramos, Pilar Gómez-Gil, Hector Humberto Cerecedo-Núñez, Francisco Gutiérrez-Delgado, Enrique Torres-Rasgado, Ricardo Pérez-Fuentes, and Jose L Flores-Guerrero. Quantification of glycated hemoglobin and glucose in vivo using Raman spectroscopy and artificial neural networks. *Lasers in Medical Science*, 37(9):3537–3549, 2022.

[64] S. Lakshmi Reddy, Tamio Endo, and G. Siva Reddy. Electronic (absorption) spectra of 3d transition metal complexes. In Muhammad Akhyar Farrukh, editor, *Advanced Aspects of Spectroscopy*, chapter 1. IntechOpen, Rijeka, 2012.

[65] OP-TEC. $Basics_of_Spectroscopy_2008_CORD.pdf, 2008. Accessed on February 12, 2024.$

[66] https://research-repository.griffith.edu.au/bitstream/handle/10072/34561/62679$_1$.*pdf*, 2010. *Accessedon*

[67] https://medium.com/@sasirekharameshkumar/deep-learning-basics-part-10-feed-forward-neural-networks-ffnn-93a708f84a31: :text=A Accessed on june 7, 2024.

[68] Ryan J. Urbanowicz, Melissa Meeker, William La Cava, Randal S. Olson, and Jason H. Moore. Relief-based feature selection: Introduction and review. *Journal of Biomedical Informatics*, 85:189–203, 2018.

[69] Tineke Vankeirsbilck, Ann Vercauteren, W. Baeyens, G Weken, Francis Verpoort, Geert Vergote, and J.P Remon. Applications of raman spectroscopy in pharmaceutical analysis. *TrAC Trends in Analytical Chemistry*, 21:869–877, 12 2002.

[70] P. Hildebrandt. Raman spectroscopy, biochemical applications. In John C. Lindon, George E. Tranter, and David W. Koppenaal, editors, *Encyclopedia of Spectroscopy and Spectrometry (Third Edition)*, pages 906–914. Academic Press, Oxford, third edition edition, 2017.

[71] Danting Yang and Yibin Ying. Applications of raman spectroscopy in agricultural products and food analysis: A review. *Applied Spectroscopy Reviews*, 46(7):539–560, 2011.

[72] Raman spectroscopic techniques for meat analysis: A review. *Teoriâ i Praktika Pererabotki Mâsa*, 7(2):97–111, 2022.

[73] René Staritzbichler, Pascal Hunold, Irina Estrela-Lopis, Peter Werner Hildebrand, Berend Isermann, and Thorsten Kaiser. Raman spectroscopy on blood serum samples of patients with end-stage liver disease. *PLOS ONE*, 16(9):1–18, 09 2021.

[74] A. C.-T. Ko, L.-P. Choo-Smith, R. Zhu, M. Hewko, C. Dong, B. Cleghorn, and M. G. Sowa. Application of NIR Raman spectroscopy for detecting and characterizing early dental caries. In Anita Mahadevan-Jansen and Wolfgang H. Petrich, editors, *Biomedical Vibrational Spectroscopy III: Advances in Research and Industry*, volume 6093, page 60930L. International Society for Optics and Photonics, SPIE, 2006.

[75] Gregory W Auner, S Kiran Koya, Changhe Huang, Brandy Broadbent, Micaela Trexler, Zachary Auner, Angela Elias, Katlyn Curtin Mehne, and Michelle A Brusatori. Applications of raman spectroscopy in cancer diagnosis. *Cancer Metastasis Rev.*, 37(4):691–717, December 2018.

[76] Haoyue Liang, Ruxue Shi, Haoyu Wang, and Yuan Zhou. Advances in the application of raman spectroscopy in haematological tumours. *Front. Bioeng. Biotechnol.*, 10, January 2023.

[77] https://www.dss.uniroma1.it/en/system/files/pubblicazioni/Fordellone.pdf. Accessed on march 04, 2024.

[78] https://databasecamp.de/en/ml/semi-supervised-learning-en. Accessed on june, 2024.

[79] Himel Mondal and Shaikat Mondal. Clarke error grid analysis on graph paper and microsoft excel. *J. Diabetes Sci. Technol.*, 14(2):499, March 2020.