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

Machine Learning Application in Spectroscopy

HDL,LDL,HGB Blood Analysis

Supervisor:

- Dr DJELLABI Brahim

Realised by:

- KHIDER Imad

2023/2024

Acknowledgement

I express my deepest gratitude to Almighty God for granting me the health, courage, and patience to start and complete this thesis.

First and foremost, my sincere thanks go to my supervisor, **Mr.Brahim DJELLABI**. His exceptional guidance, patience, and unwavering support were invaluable throughout the preparation of this work. Without his rigorous supervision, insightful recommendations, and constant availability, this thesis would not have been possible.

I am also grateful to the members of the jury for kindly attending the defense of this thesis and for their thoughtful evaluation of my work.

My heartfelt thanks extend to all the professors who have supported me throughout my university course. Your teachings and guidance have paved the way for my success.

Finally, I express my appreciation to my family and friends. Your hopeful words and continuous encouragement have been a source of strength and inspiration for me.

Thank you.

Dedication

This thesis is dedicated :

To My Father Boualem, whose unwavering support, encouragement, and love have guided me throughout my academic journey. Your belief in me has been my greatest source of strength. Thank you for always being there and for inspiring me to pursue excellence.

To My Mother, the heartbeat of our family and the one who makes every house feel like home: Thank you for being the heart and soul of our lives.

To My Brother Rayan, Thank you for your unlimited advice and support. You played a significant role in helping me arrive here.

To My Sisters, May God grant you health, happiness, courage, and above all, success. Words cannot express my gratitude for all you have done for me. I thank God for making you my sisters.

Finally, I would like to thank all my friends, colleagues, and all the students of the class. To all those who, with a word, have given me the strength to continue.

Imad khider.

Abstract

This thesis explores the integration of Near-Infrared (NIR) spectroscopy with machine learning to predict crucial blood biomarkers: HDL cholesterol, LDL cholesterol, and HGB hemoglobin levels. These biomarkers are essential for assessing cardiovascular health and overall physiological status. The study aims to enhance the precision and efficiency of blood analysis methodologies by leveraging spectroscopic insights into molecular composition and harnessing the computational capabilities of machine learning.

The investigation begins with a comprehensive review of HDL and LDL cholesterol, emphasizing their physiological significance and the challenges associated with accurate prediction using traditional methods. Similarly, the clinical implications of HGB hemoglobin, crucial for oxygen transport, are examined.

The methodology outlines a systematic approach encompassing rigorous data collection, preprocessing, and visualization techniques to ensure data quality and relevance. Initially, the XGBoost algorithm was employed to predict levels of HDL cholesterol, LDL cholesterol, and HGB hemoglobin. The achieved accuracies were **55%** for HDL cholesterol, **55%** for LDL cholesterol, and **86%** for HGB hemoglobin, suggesting limited predictive capability for cholesterol levels.

Subsequently , Recognizing the crucial significance and intricacy of predicting LDL cholesterol levels, I implemented a specialized binary classification approach using a VotingClassifier (KNN, SVC, GPC) exclusively tailored for LDL prediction. This approach yielded a notable accuracy improvement to **91%.**

Key words: Machine learning ,XGBoost , KNN ,SVC ,GPC , Near-Infrared (NIR) spectroscopy, HDL cholesterol,LDL cholesterol, HGB hemoglobin , Non-invasive techniques ,Blood biomarkers .

Contents

List of Figures

List of Tables

List of abreviations

HGB	Hemoglobin
LDL	Low-Density Lipoprotein
HDL	High-Density Lipoprotein
NIR	Near-Infrared
ML	Machine Learning
SVC	Support Vector Classifier
KNN	K-Nearest Neighbors
NMR	Nuclear Magnetic Resonance
XGBoost	Extreme Gradient Boosting
EDA	Exploratory Data Analysis
Spectroscopy Abbreviations	
UV	Ultraviolet
IR	Infrared
NIR	Near Infrared
Machine Learning Abbreviations	
NN	Neural Network
RF	Random Forest
PCA	Principal Component Analysis
GPC	Gaussian Process Classifier
SVC	support vector classifier
Statistical Abbreviations	
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
R^2	Coefficient of Determination

General introduction

High cholesterol and abnormal hemoglobin levels are critical indicators of various diseases and health conditions globally. High cholesterol, a significant risk factor for cardiovascular diseases, contributes to approximately 2.6 million deaths and 29.7 million Disability-Adjusted Life annually worldwide, according to the World Health Organization (WHO) [2]. Abnormal hemoglobin levels, affecting millions, indicate conditions such as anemia or erythrocytosis, necessitating accurate and timely diagnosis for effective management.

Traditional blood analysis methods can be labor-intensive . Advanced diagnostic techniques are needed to provide rapid, reliable, and non-invasive insights into cholesterol and hemoglobin levels. Near-Infrared (NIR) spectroscopy, coupled with machine learning (ML) algorithms, offers promising avenues for improving diagnostic accuracy and predictive capabilities.

This study explores the integration of NIR spectroscopy and ML techniques to develop models capable of predicting cholesterol and hemoglobin levels accurately. By harnessing spectroscopic data's unique capabilities and ML's analytical power, the research aims to enhance healthcare delivery, optimize treatment strategies, and improve patient outcomes.

Thesis Organization : This thesis is structured into four chapters, encompassing the following:

chapter 1 : Overview of cholesterol and hemoglobin (Hgb) analysis, including their roles, types, and health implications.

chapter 2 : Introduction to spectroscopy principles, focusing on near-infrared (NIR) spectroscopy, instrumentation, applications, and limitations. Explanation of machine learning types, supervised learning methods, and their application in analyzing spectroscopic data.

chapter 3 : Detailed steps on data collection, validation, exploratory analysis, and preprocessing techniques like normalization and feature selection.

chapter 4 : Presentation and evaluation of machine learning models for Hgb and cholesterol analysis, including results comparison and validation.

Chapter 1

Hemoglobin and Cholesterol Blood Analysis

1.1 Introduction

While traditional blood analysis remains central to medical diagnosis, it only offers a selective view of the intricate biological landscape inside the patient's body. In this chapter, we will discuss and analyse "reconceptualization" or doing things differently on how blood analysis can be done. This is by making use of new technologies and methodologies.

Among areas that are expected to yield significant dividends from this new approach is cholesterol analysis. Better knowledge on the structure and role of cholesterol in any blood specimen would give us a more detailed understanding of patients' risks for cardiovascular diseases and other health issues.As well, haemoglobin in the body can offer insight into oxygen delivery, what erythrocytes do and also reveal some haematologic pathologies. This advanced testing may help detect illness sooner,, make treatment more tailored to individual patients‘ needs and ultimately save lives.

1.2 Cholesterol: A Molecule of Duality

1.2.1 Definition of Cholesterol

Cholesterol, a waxy substance found in the blood and produced by the liver, plays a critical role in human health. It serves as a building block for healthy cell membranes, providing structure and flexibility. Cholesterol is also a precursor molecule for the synthesis of essential components.[3]

1.2.2 Roles of Cholesterol in the Body

Cell membrane constituent: Cholesterol is one of the elements that make up the cell membrane, so the cells owe it their permeability.

Production of bile salts: In order to promote good digestion, cholesterol will participate in the formation of bile salts that are essential to break down food and assimilate nutrients.

Vitamin D synthesis: In the body, cholesterol will help in the manufacture of vitamin D, which is essential to properly fix calcium on the bones.

Hormonal role: Cholesterol also has an important role at the hormonal level since it constitutes certain sex hormones such as testosterone and adrenals such as cortisone.

1.2.3 The Two Faces of Cholesterol: LDL and HDL

Cholesterol travels through the blood on proteins called “lipoproteins”. There are two types of lipoproteins that carry Cholesterol distribution in the whole body:

LDL (low-density lipoprotein) Cholesterol : often referred to as ”bad” cholesterol, constitutes the majority of the body’s cholesterol. Elevated levels of LDL cholesterol increase the risk of heart disease and stroke.

HDL (high-density lipoprotein) Cholesterol : often referred to as ”good” cholesterol, absorbs cholesterol from the bloodstream and transports it to the liver, where it is then expelled from the body. High levels of HDL cholesterol can reduce the risk of heart disease and stroke.[4]
Conversely, an excess of LDL cholesterol can accumulate on the walls of blood vessels, forming ”plaque.” This buildup can lead to serious health issues, including heart disease and stroke.

1.2.4 Normal Cholesterol Level

Normal cholesterol levels are important for maintaining cardiovascular health. Here are the recommendations of the French Agency for Health Security (Afsaps) are as follows[5]:

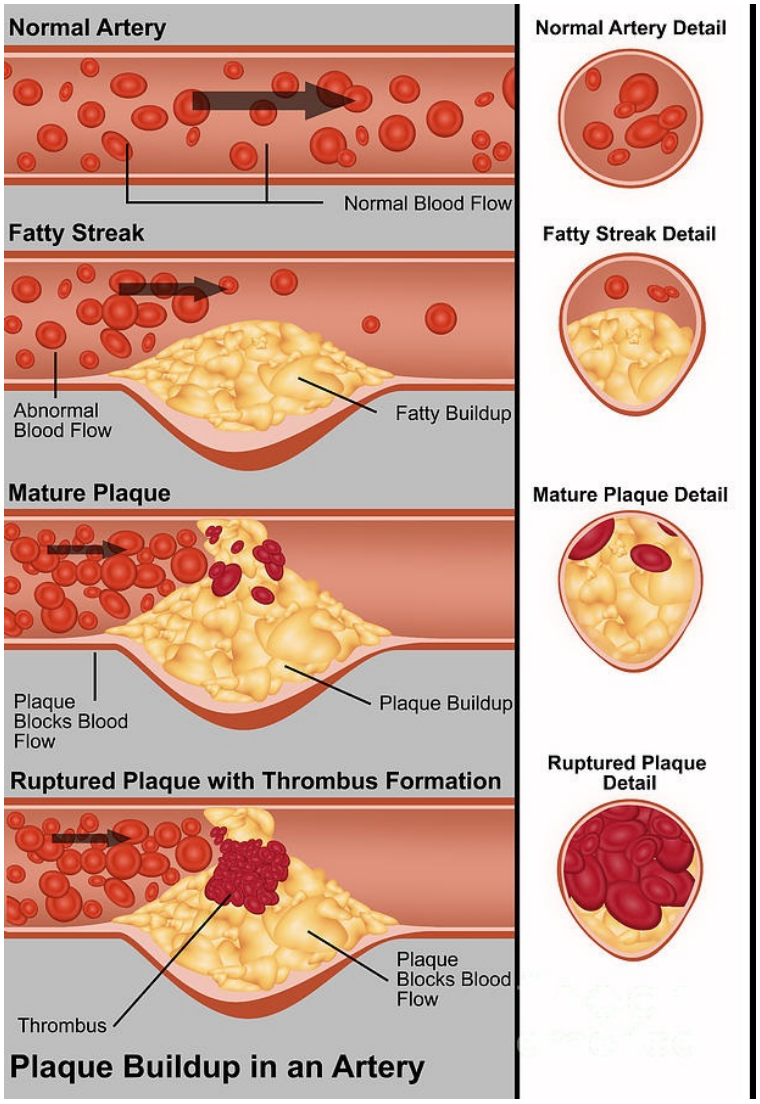
Type of Cholesterol	**Normal Values**
Total Cholestérol	< 2 g/L
HDL Cholestérol	$> 0{,}35$ g/L
LDL Cholestérol	$< 1{,}6$ g/L

Table 1.1: Types of Cholesterol and Their Normal Values

1.2.5 The Impact of high Cholesterol on the Body

Inside your blood vessels, plaque accumulates as a result of high cholesterol. The accumulation of plaque is known as atherosclerosis. Individuals who have atherosclerosis are more susceptible to a wide range of illnesses. This is due to the vital functions that your blood vessels perform throughout your body. Hence, there are consequences when an issue arises with one of your blood vessels. Your blood vessels will develop plaque if your cholesterol is high. The plaque gets larger the longer you don’t get treatment (as depicted in Figure 1.1). You experience narrowing or blockage of your blood vessels as the plaque grows. You may have prolonged blood vessel function, similar to a partially clogged drain. However, they

won't function as well as they ought to. High cholesterol raises your risk of other medical conditions, depending on which blood vessels are clogged.[6][7]



(a)

Figure 1.1: Steps of buildup plate in the artery [1]

1.2.6 Other Diseases Caused by High Cholesterol

Conditions differ according to clogged blood vessels.

Chest Pain : When the coronary arteries, which supply blood to the heart, are affected, it can result in chest pain (angina) and other signs of coronary artery disease.

Heart Attack: If plaques within the arteries rupture, a blood clot may form at the site of the rupture, blocking blood flow or breaking free and obstructing a downstream artery. This can lead to a heart attack if blood flow to a portion of the heart is interrupted.

Stroke: Similar to a heart attack, a stroke occurs when a blood clot blocks the flow of blood to a part of the brain.[6][7]

1.3 Hemoglobin (Hgb) Blood Analysis

Hemoglobin (Hgb) is a protein in red blood cells essential for transporting oxygen to tissues. Maintaining an adequate hemoglobin level is crucial for proper tissue oxygenation. Hemoglobin concentration in whole blood is measured in grams per deciliter (g/dl). The normal range for hemoglobin is 14 to 18 g/dl for males and 12 to 16 g/dl for females. [8] [9]

1.3.1 Functions of Hemoglobin in the Body

Its main function is to transport oxygen from the lungs to the organs, muscles, and all tissues through arterial circulation. After delivering oxygen to the tissues, it returns to the lungs, carrying, for example, the body's carbon dioxide, and reloads with oxygen to ensure the cellular respiration cycle.[10]

Hemoglobin comprises approximately 70% of the body's iron content and imparts the characteristic red color to red blood cells.[11]

1.3.2 Types of hemoglobin

There exist several types of hemoglobin, with the two most prevalent being:

- **Hemoglobin A (HgbA):** This is the predominant type found in healthy adults.
- **Hemoglobin F (HgbF):** Known as fetal hemoglobin, this type is present in fetuses and newborns, being gradually replaced by HgbA shortly after birth.

Abnormal forms of hemoglobin can alter both the shape of red blood cells and their capacity to transport oxygen and carbon dioxide.[10][11]

1.3.3 Low hemoglobin levels

Low hemoglobin levels indicate an insufficient number of red blood cells. This deficiency can stem from inadequate production by the body or from health conditions that decrease the red blood cell count. When hemoglobin levels are too low, muscle cells may not receive enough oxygen, leading to a lack of energy. This condition is known as **anemia**. [12] Common symptoms of anemia include: chest pain ,cold hands and feet,dizziness,fatigue,headaches.

1.3.4 Causes of Low Hemoglobin Levels

Low hemoglobin levels have many possible causes. Some of the most common causes are:

- Diet: A diet also low in iron, folate, or vitamin B12 can influence your body's ability to produce red body fluid.

- Blood loss: Meaningful bleeding can lead to depressed hemoglobin levels, but long-term ancestry loss from stomach ulcers, uterine fibroids, or heavy menstrual periods can too contribute.
- Pregnancy: Before birth your blood volume increases significantly, that can lower your hemoglobin levels.
- Genetic environments: Hereditary health environments such as G6PD deficiency, weapon cell anemia, spherocytosis, and thalassemia can demolish red blood containers.
- Cancer: Certain types of tumor, including leukemia and lymphoma, can affect your cell with hemoglobin levels.
- Chronic kidney ailment: Decreased kidney function can mean that your physique doesn't make enough erythropoietin, a hormone unavoidable for red blood cell production.
- Medication: Antiretroviral medications and chemotherapy drugs can damage your bone marrow, which may reduce red blood cell production. [12]

1.3.5 High hemoglobin levels

It's also possible to have overly high hemoglobin levels. If you have high hemoglobin levels, you might also have a high red blood cell count. This condition is called polycythemia. Primary polycythemia is a hereditary condition linked to a metamorphosis in the bone marrow containers.

Secondary polycythemia happens when another condition causes an overproduction of red ancestry cells. Cancers and courage and lung diseases are few of the conditions connected to secondary polycythemia.[8]

1.3.6 Cases of High Hemoglobin Levels

High hemoglobin levels, or polycythemia, can arise from various factors:

- High Altitude Living : The atmosphere's oxygen content decreases with altitude. In order to make up for this, the body makes more red blood cells, which improve oxygen transport and raise hemoglobin levels.
- Long-Term Smoking : Smoking causes the blood to become mixed with carbon monoxide, which attaches itself to hemoglobin and lowers its ability to carry oxygen. In order to maintain appropriate oxygen delivery, the body reacts by making more red blood cells, which raises hemoglobin levels.[12]
- Lack of fluids : Dehydration causes the blood's plasma volume to decrease, which may cause the concentration of hemoglobin to rise somewhat. The hemoglobin concentration appears higher due to the decreased plasma volume, even though the actual number of red blood cells may remain unchanged.[8]

1.3.7 Blood Analysis Techniques

A- Invasive blood analysis

Invasive blood analysis involves puncturing a vein or artery to collect a blood sample. This is normally accomplished using a needle and syringe, and the collected blood is then transferred to a vial for laboratory analysis.

Invasive blood analysis processes vary depending on the required blood volume and the specific test being performed. Here are some common methods:

Venipuncture: This approach, the most common type of invasive blood analysis, involves inserting a needle into a vein, usually in the arm, to collect a sample.[12]

Arterial puncture: This procedure, which is less common than venipuncture, targets arteries to draw blood. Arterial blood samples are frequently used to measure blood gas levels, including oxygen and carbon dioxide. Non-invasive blood analysis.

B- Non-invasive blood analysis

Non-invasive blood analysis uses techniques that do not require a skin puncture for sample collection. As technology progresses, these tactics gain traction.

Here are some popular non-invasive blood analysis methods:

Near-infrared spectroscopy (NIRS): uses light to measure the concentrations of various blood components such as cholestrol, glucose, and hemoglobin.

Pulse oximetry: This approach uses a clip-on device on a fingertip, earlobe, or toe to assess blood oxygen saturation.

Photoplethysmography (PPG): This technique uses light to assess changes in blood volume within the skin. PPG is an effective way to measure blood pressure, oxygen saturation, and heart rate.

Aspect	Invasive Blood Analysis	Non-Invasive Blood Analysis
Method of Collection	Requires puncturing a vein or artery to draw blood sample.	Blood sample collected without puncturing the skin or vessels.
Sample Collection Time	Typically quick, but may require some time for preparation.	Generally faster, as it doesn't involve the same preparation steps.
Discomfort	May cause discomfort or pain due to needle insertion.	Usually painless, as it doesn't involve needle insertion.
Risk of Complications	Slight risk of complications such as infection or bruising.	Minimal risk of complications due to non-invasive nature.
Accessibility	Requires trained healthcare professionals and specialized equipment.	Can be performed by trained personnel with appropriate devices.
Examples	Blood draws for laboratory tests, arterial blood gas analysis.	Fingerstick tests, pulse oximetry, transcutaneous blood gas monitoring.

Table 1.2: Comparison between invasive and non-invasive blood analysis

1.4 The Importance of HGB and Cholesterol Blood Analysis

The American Heart Association (AHA) recommends that those who have no risk factors have their blood cholesterol level checked now and then. Cholesterol levels affect cardiovascular health, such as heart disease and stroke, which are the leading causes of death in the world just now. So it is important to keep an eye on levels of cholesterol and make sure they are not too high.[13]

The cholesterol blood draw you may have heard frequently, or lipid panel, is an important tool to assess cardiovascular disease risk. It measures several components of cholesterol all at once -total cholesterol and its fractions LDL and HDL. In other words, by analyzing these values healthcare providers can predict a person 's chances of getting heart disease or having a stroke. Early detection through blood analysis allows for timely intervention through lifestyle changes, medications, or a combination of both, potentially preventing or delaying the onset of cardiovascular disease.

The importance of cholesterol blood analysis is crucial for assessing cardiovascular health, but it is not the only significant blood test. Hemoglobin blood analysis is equally vital for understanding overall health, particularly in relation to oxygen transport in the body. Hemoglobin, a protein in red blood cells, carries oxygen from the lungs to the rest of the body and helps remove carbon dioxide. Anemia, characterized by low hemoglobin levels, can cause symptoms like fatigue, weakness, and shortness of breath, and hemoglobin levels can indicate underlying health conditions such as iron deficiency, chronic diseases, or genetic disorders. Routine hemoglobin blood analysis is essential for diagnosing and managing various medical conditions, ensuring optimal oxygen delivery throughout the body, and maintaining overall health. Like cholesterol levels, monitoring hemoglobin levels through regular blood tests aids in early detection and timely intervention, thus preventing potential health complications.

1.4.1 Current Landscape of HGB and Cholesterol Blood Analysis

The cornerstone of blood analysis lies in established methodologies, prominently exemplified by the lipid panel test. This comprehensive assay quantifies total cholesterol, triglycerides, LDL , HDL and HGB levels in a fasting blood sample. Through enzymatic reactions , lipid panel tests offer a snapshot of an individual's lipid profile, informing clinical decision-making in cardiovascular risk assessment.

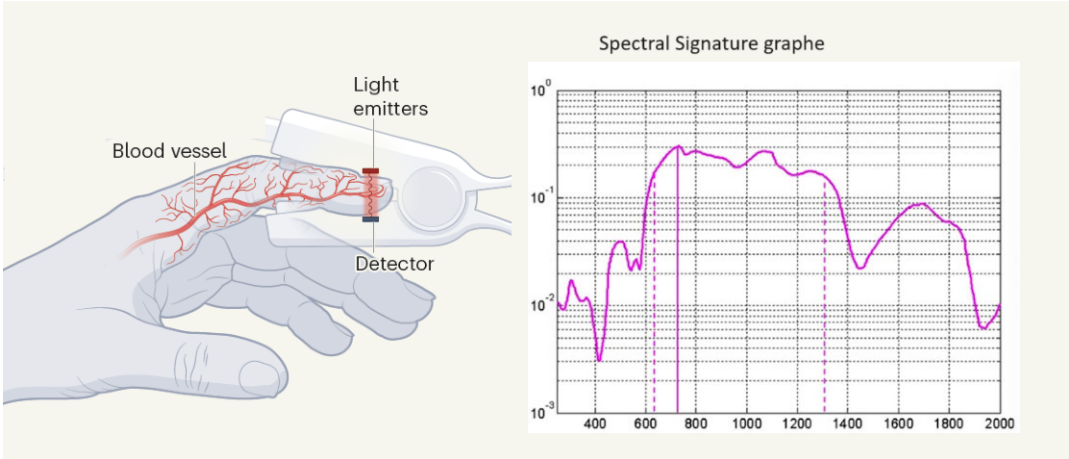
1.4.2 Limitations of Current Approaches

Current blood analysis techniques for cholesterol and hemoglobin have inherent limitations, even with their broad use. The most important of these is the requirement that patients fast before sample collection, which may delay diagnosis or therapy commencement and requires patient cooperation. Furthermore, although lipid panel tests offer useful information about major lipoprotein fractions and total cholesterol, they frequently do not provide enough detail to evaluate lipoprotein subclasses, which might make risk classification difficult. This deficit is especially important in light of the growing understanding that different LDL and HDL subtypes have different atherogenic characteristics.
Additionally, variability is introduced during sample collection and processing due to the possibility of human error, which compromises the accuracy of the results. Lipid measurements might be skewed by factors including poor venipuncture technique or sample mishandling, which can complicate clinical interpretation and jeopardize the accuracy of diagnostic results.

1.4.3 New Method of HGB and Cholesterol Blood Analysis

To determine HGB and cholesterol levels, merely have a brief scan rather than having blood drawn or fasting for a lipid panel. The fascinating combination of spectroscopy and machine learning may make this potential scenario a reality.
Spectroscopy operates by shining light on a sample, such as a single blood drop. The molecules here interact with this light, and at particular wavelengths, each molecule absorbs light. Contemplate it as a distinct fingerprint for every molecule. We may be able to determine the presence and amount of cholesterol in the sample by examining this "fingerprint."(as depicted in Figure 1.2)

Herein lies the power of machine learning. Large datasets with known HGB and cholesterol levels in blood samples are used to train powerful algorithms. These algorithms are trained to identify patterns in the spectral data that are distinct to various amounts of HGB and cholesterol. After being taught, the algorithms are able to accurately estimate the HGB and cholesterol content of a new sample by analyzing its light-matter interaction.
This approach's potential for a non-invasive test is what makes it so beautiful. In the future, picture having your skin exposed to a light beam that is analyzed by a handheld scanner that is worn on your wrist. After that, the machine learning algorithms analyze the information to give a cholesterol reading in real time. Abolition of needles and fasting, There is yet more possibility. With continued advancements in technology, cholesterol levels may be continuously monitored, giving a doctor a more complete picture



(a)

Figure 1.2: Finger Detector

of health. For those who are more likely to develop heart disease, this ongoing observation may be extremely helpful as it enables them to make proactive changes to their diet, exercise routine, and prescription regimen.

1.5 Conclusion

HGB and cholesterol blood analysis play a vital role in identifying individuals at risk for cardiovascular diseases and other health complications. By measuring different types of cholesterol (LDL, HDL) and HGB, healthcare professionals can assess a patient's overall profile and implement preventive or therapeutic strategies.

Current blood analysis methods based on HGB and cholesterol determination have been established and validated, but they may be limited in terms of being costly, accessible, and comprehensive. The emergence of a novel generation of blood analysis technology promises convenient, affordable, and potentially more informative testing methodologies.

This evolving landscape underscores the importance of continued research and innovation in cholesterol blood analysis. Early detection and management of high cholesterol and HGB remain critical for promoting cardiovascular health and overall well-being.

Chapter 2

NIR Spectroscopy

2.1 Introduction

Spectroscopy is a powerful tool that allows us to study substances without the invasiveness of procedures. For blood analysis, this will mean it is possible to obtain key health information without using needles. This chapter discusses what spectroscopy is and the types of spectroscopy. Then we describe how the wavelength affects interaction with materials. First, instrumentation, applications, and theoretical underpinning in focus with near-infrared spectroscopy will be discussed, followed by the comparison and contrast of infrared and NIR spectroscopy, and an explanation of how NIR detects bending vibrational modes.

2.2 Definition of Spectroscopy

Spectroscopy is the scientific study of how light interacts with matter [14]. Initially defined as the investigation of radiation-matter interactions relative to wavelength, spectroscopy now encompasses the analysis and measurement of spectra generated by matter interacting with or emitting electromagnetic radiation. In a spectroscopy experiment, electromagnetic radiation within a specific wavelength range is directed from a source through a sample containing compounds of interest. This interaction causes either absorption of energy from the light source during absorption or emission of light with a wavelength different from the source during emission [15].

2.3 Basic Principles of Spectroscopy

2.3.1 Electromagnetic radiation

Electromagnetic radiation is a form of energy that propagates as both electrical and magnetic fields and has the ability to transfer energy through space. The energy propagates as a wave, such that the crests and through moving in the vacuum with variable wavelengths and frequency. The distance between successive crests in a wave is called its wavelength. When Electromagnetic radiation passes through

biological material, it deposits energy in two forms: excitation and ionization.

A-Excitation: describes the deposition of enough energy to raise an electron to a higher shell without ejection of the electron.

B-Ionization: deposition of enough energy to eject one or more electrons from the atom.

2.3.2 Interactions of Electromagnetic Radiation with Matter

Spectroscopy is an analytical technique that relies on the interaction of electromagnetic radiation with matter. This radiation, encompassing a spectrum of wavelengths from radio waves to gamma rays, exhibits both wave-like and particle-like properties. As waves, it travels through space at the speed of light (c), with its wavelength (λ) and frequency (v) related by the equation $c = \lambda v$. Different regions of the electromagnetic spectrum possess distinct properties and interact with matter in unique ways.[16]

When electromagnetic radiation encounters matter, three primary outcomes are possible:

Absorption: The material absorbs specific wavelengths of radiation, corresponding to energy levels required to excite its atoms or molecules. This selective absorption forms the basis for spectroscopic analysis, as the characteristic absorption pattern provides valuable information about the material's composition and structure.[17]

Transmission: Certain wavelengths of radiation pass through the material without being absorbed. The extent of transmission depends on factors like the material's thickness and density, offering insights into these properties.[17]

Scattering: The radiation can interact with the material and deviate from its original path in various directions. This scattering phenomenon, categorized as either elastic (no change in frequency) or inelastic (frequency change), can also reveal information about the material's structure and interactions with light.[17]

By analyzing the fate of electromagnetic radiation after interacting with a sample, spectroscopy empowers scientists to unveil its chemical composition, structural features, and various other properties. This fundamental understanding forms the cornerstone of diverse spectroscopic techniques employed across various scientific disciplines.

2.3.3 Electromagnetic Spectrum

The electromagnetic spectrum refers to the entire range of electromagnetic radiation, the spectral regions of electromagnetic radiation vary in frequency, resulting in different energies. This diversity in wavelengths is advantageous, as materials react to radiation in distinct mechanisms or pathways, providing varied outcomes that can be utilized for chemical analysis, from very low frequency waves such as radio waves, to very high fre- quency waves such as gamma rays (as depicted in Figure 2.1), which are classified based on their wavelength, frequency, or energy [18].

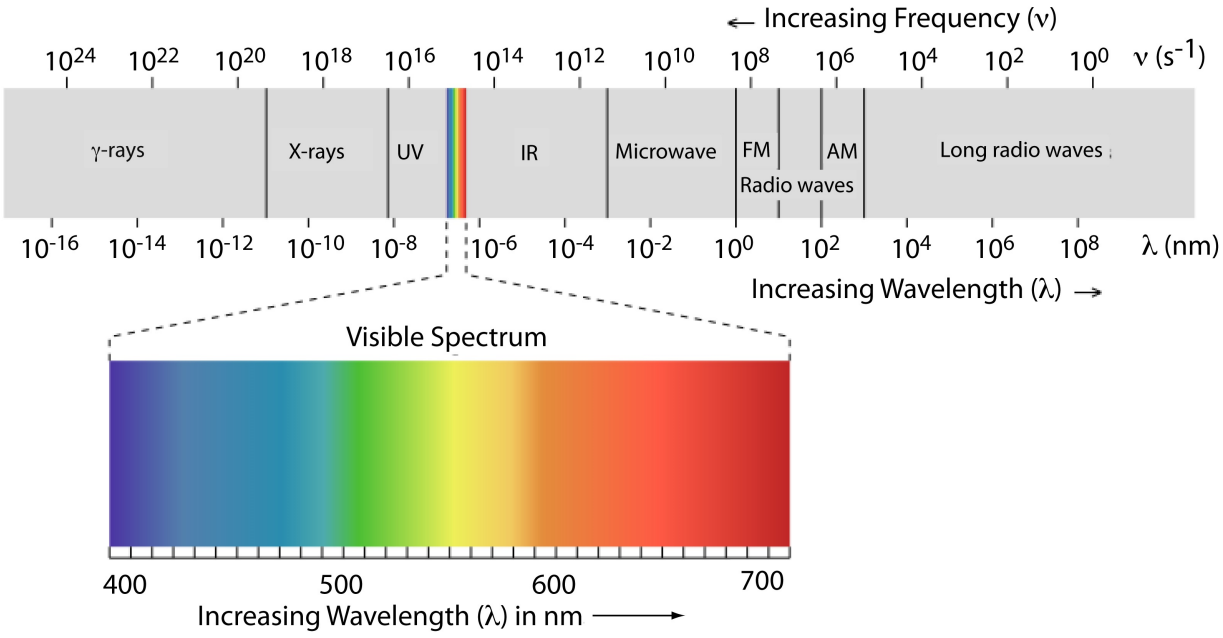
2.4 Types of Spectroscopy

Spectroscopy can be classified based on three key factors:

Type of Electromagnetic Radiation: This refers to the specific region of the electromagnetic spectrum used, such as ultraviolet (UV), visible (Vis), or X-ray (as depicted in Figure 2.1).

Sample-Radiation Interaction: This focuses on how the radiation interacts with the material being analyzed. Common interactions include absorption, emission, and reflection (absorption and emission were previously discussed).

Target Species: This identifies the type of matter under investigation, which can be atoms, molecules, or even atomic nuclei.[19]



(a)

Figure 2.1: Electromagnetic spectrum

[5]

2.4.1 Ultraviolet and Visible Spectroscopy

Ultraviolet (UV) and visible (Vis) spectroscopy are types of absorption spectroscopy in which UV-visible light from 10 to 700 nm is absorbed by the molecule. Using the Beer-Lambert Law, which states that absorbance is proportional to the concentration of the substance in solution and the path length, UV and visible spectroscopy can be used to estimate the concentration of samples,also to identify the presence of free electrons and double bonds in a molecule .Recently it used as technique for identification and quantification of organic compounds,to identify pure drug compounds.

2.4.2 X-Ray Spectroscopy

X-ray spectroscopy is a widely utilized technique for probing the atomic structure and electronic states of materials. Essentially, when an X-ray interacts with an atom, it can either elevate a core electron to an unoccupied state or expel it from the atom, resulting in the creation of a core hole. This process allows for the examination of atomic local structure and electronic configurations.[20]

2.4.3 Infrared Spectroscopy

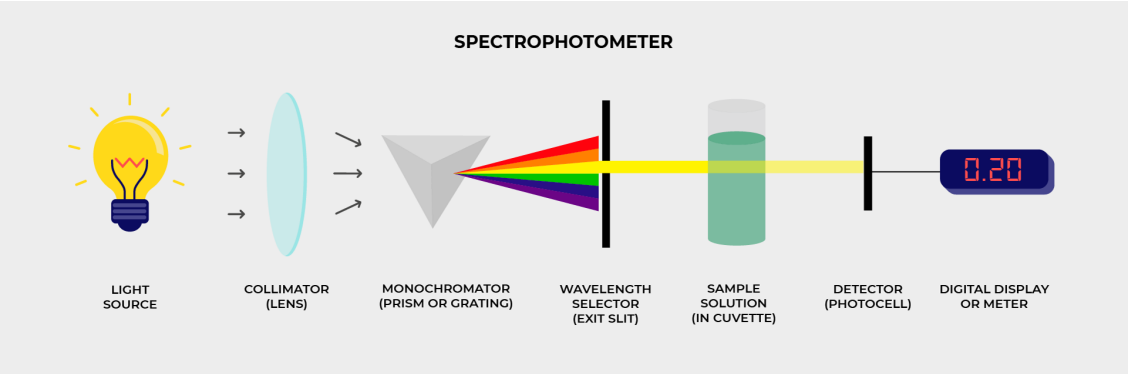
The infrared (IR) spectrum is divided into near-IR, mid-IR, and far-IR regions, each corresponding to specific ranges of wavelengths. IR spectroscopy operates on the principle that molecules exhibit vibrational motion, causing their bonds to stretch and bend upon absorption of infrared radiation. This phenomenon allows for the identification of different functional groups within molecules, as each group absorbs light at distinct wavelengths based on its structural characteristics. Consequently, a vibrational spectrum obtained through IR spectroscopy can provide valuable insights into the composition of a sample . In practice, IR spectroscopy is utilized for the qualitative and quantitative analysis of chemical species and the determination of molecular structures [21]. For our project, we concentrate on the near-IR region, which will be elaborated upon in the subsequent section.

2.4.4 Nuclear Magnetic Resonance (NMR) Spectroscopy

Nuclear magnetic resonance spectroscopy entails the measurement of magnetic fields surrounding atomic nuclei. Employing radio waves, this technique excites atomic nuclei within a sample. Sensitive radio receivers then capture the resonating nuclei. NMR spectroscopy furnishes intricate insights into the structure and reaction status of molecules, as the resonant frequency of an atomic nucleus correlates with the electronic configuration of the molecule it resides in. Consequently, it serves as a potent method for precisely determining the characteristics of monomolecular organic compounds.[20]

2.5 Near Infrared spectroscopy

Near-infrared spectroscopy (NIRS) utilizes the near-infrared region of the electromagnetic spectrum, typically ranging from 780 nm to 2500 nm. [22] This spectroscopic technique finds numerous applications in medical and physiological diagnostics, as well as in research fields such as blood sugar monitoring, pulse oximetry, functional neuroimaging, sports medicine, elite sports training, ergonomics, rehabilitation, neonatal research, brain-computer interface, urology (bladder contraction), and neurology (neurovascular coupling). Additionally, NIRS is employed in various other sectors including pharmaceuticals, food and agrochemical quality control, atmospheric chemistry, and combustion research.



(a)

Figure 2.2: SPECTROPHOTOMETER

[23]

2.5.1 Near-Infrared (NIR) Spectroscopy Instrumentation

NIR spectroscopy uses instruments that share some similarities with those for ultraviolet-visible (UV-vis) and mid-infrared (mid-IR) spectroscopy. Here's a breakdown of the key components:

Light Source: Just like a spotlight, NIR instruments use a light source to illuminate the sample. Common options include incandescent bulbs or LEDs for broad coverage. For ultimate precision, researchers might use lasers or frequency combs, even if it takes a bit longer to collect the data.

Monochromators :

- Monochromators for NIR spectroscopy include prisms, gratings, and filters.
- Prisms are manufactured using materials like Potassium bromide, Sodium chloride, or Caesium iodide.
- Filters are made from Lithium Fluoride, while diffraction gratings are composed of alkali halides.

Sample cells and sampling of substances: NIR spectroscopy can be used to characterize solid, liquid, or gas samples. Various techniques are used for preparing solid samples, while liquid samples can be held in a liquid sample cell made of alkali halides, and gas samples are sampled similarly to liquid samples

Light detectors: NIR detectors come in two flavors: single-channel (one color at a time) and multi-channel (sees many colors at once). The multi-channel ones are super cool because they allowed us to create special NIR cameras that can see the chemical makeup of different parts of an object, all at once.

Recorder: Recorders are used to record the IR spectrum.

2.5.2 Applications of NIR Spectroscopy:

Cosmic Detectives: Astronomers leverage NIR to pierce through dust clouds in space. These clouds obscure our view of distant objects in visible light, but NIR wavelengths penetrate more easily. This allows astronomers to peer deeper into the cosmos, study the atmospheres of cool stars, and even witness the birth of new stars in dust-filled stellar nurseries.[24]

Agricultural Revolution: NIR spectroscopy proves to be a valuable tool for farmers. It enables them to quickly and accurately assess the quality of their crops, measuring factors like sugar content in corn, protein levels in wheat, or moisture content in grains.[25] This information helps farmers optimize their harvest, improve yields, and ensure consistent product quality

Planetary Guardians: From airplanes and satellites, scientists utilize NIR technology for remote sensing. By analyzing the near-infrared light reflected from Earth's surface, they can monitor the health of vegetation, identify areas of drought or nutrient deficiency, and track changes in land use. This information is vital for sustainable agriculture practices, environmental monitoring, and disaster management.

Material Marvels: In the realm of material science, NIR spectroscopy shines a light on the properties of various materials. It helps researchers analyze the composition of thin films used in electronics, study the optical characteristics of nanoparticles, and even measure the thickness of coatings. This knowledge plays a crucial role in developing new materials with desired properties for diverse applications.

Medical Marvel: NIR spectroscopy finds its way into the medical field as well. Doctors can use it to measure blood flow in the brain or muscles non-invasively. This proves beneficial during surgery to monitor oxygen delivery to tissues or assess circulation problems in patients with peripheral arterial disease. NIR spectroscopy is also being explored for applications in brain-computer interfaces and neurofeedback therapy.[26]

NIR spectroscopy is a rapidly evolving technology with vast potential. As scientists continue to refine its applications, we can expect it to play an even greater role in various fields, from unraveling the mysteries of the cosmos to improving healthcare and advancing material science.

2.5.3 Theoretical Foundation of Near-Infrared (NIR) Spectroscopy

Near-infrared (NIR) spectroscopy relies on two fundamental molecular absorption processes: overtones and combinations.

Overtones: In a molecule, vibrational modes have characteristic frequencies known as fundamental frequencies. An overtone arises when a molecule undergoes excitation to higher vibrational energy levels beyond the first excited state. These transitions, occurring between non-adjacent vibrational states, result in absorption bands at multiples (integer factors) of the fundamental frequency. These overtone bands, typically designated as the first and second overtone, are significantly weaker (10-100 times) than

the fundamental band due to less probable transitions. Additionally, the strength is bond-dependent, with the first overtone appearing in the NIR region (780-2000 nm).[27]

Combinations : NIR spectra can also exhibit absorption bands due to the interaction of multiple vibrational modes. When two or more vibrational modes are excited simultaneously, the resulting energy shifts create combination bands. The frequencies of these combination bands correspond to the sum of the individual participating modes' multiples. Combination bands tend to be located in the longer wavelength region of NIR (1900-2500 nm).[28]

The intensity of both overtone and combination bands is governed by two key factors:

Dipole Moment Change: During vibration, a molecule's electrical properties, quantified by the dipole moment, undergo changes. The magnitude of this change influences the molecule's interaction with light and, consequently, the band intensity.[27]

Bond Anharmonicity: A perfectly harmonic oscillator, analogous to an ideal spring, would vibrate at a constant frequency. In reality, molecules exhibit deviations from this ideal behavior, termed anharmonicity. The lighter the atom, the greater its vibrational amplitude and anharmonicity. As a result, NIR bands are most prominent for bonds involving hydrogen (C-H, N-H, O-H, S-H), whereas bonds like C=O, C-C, and C-Cl exhibit weaker or absent NIR features.[27] [29]

2.5.4 Bending Vibrational Modes and their Detection in NIR Spectroscopy

Within the realm of molecular spectroscopy, bending vibrations represent a distinct class of vibrational modes. Unlike stretching vibrations, which involve a change in bond length, bending vibrations manifest as an alteration in the bond angles between atoms within a molecule. These bending modes can be further categorized based on their spatial orientation relative to the molecular plane.

In-plane Bending:

Confined to the plane defined by the molecule's constituent atoms, in-plane bending encompasses two primary types:

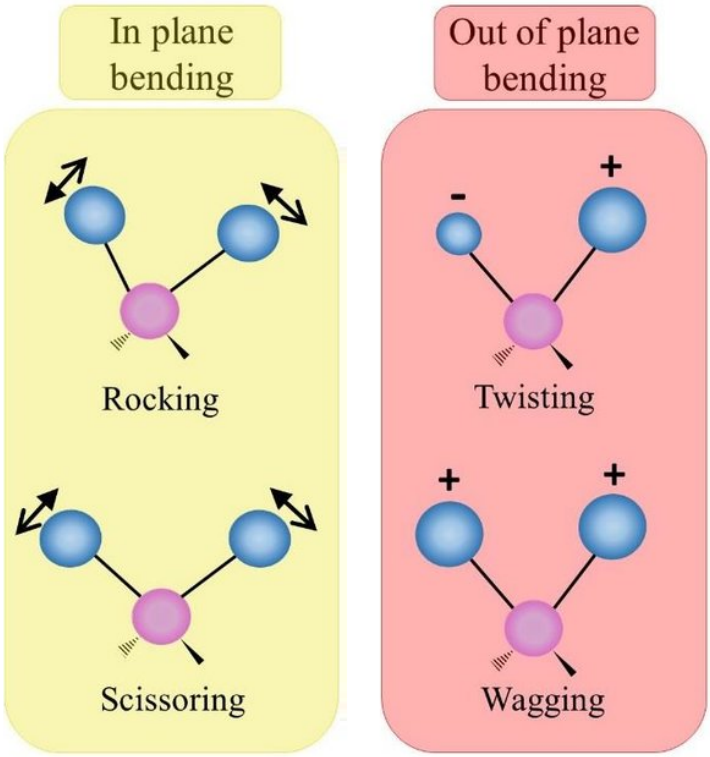
Scissoring: This mode resembles the opening and closing action of a pair of scissors, where the atoms undergo a back-and-forth shearing motion within the plane.

Rocking: Akin to a rocking chair, rocking motion characterizes this mode, with the atoms moving back and forth while remaining within the molecular plane.

Out-of-plane Bending: This category encompasses bending motions that occur perpendicular to the plane of the molecule. Two key subcategories exist:

Wagging: Here, the atoms exhibit a wagging motion, similar to a dog wagging its tail, but confined to a plane perpendicular to the molecule's main plane.

Twisting: This mode involves the twisting of one entire group of atoms on a bond relative to the group on the other side, causing the group to deviate from the molecular plane (as depicted in Figure 2.3).



(a)

Figure 2.3: vibrational-modes-of-molecules

[30]

2.5.5 Background and foundation of NIR

The two main NIR absorption processes are molecular overtones and combinations. Fundamental frequency and its overtones are known as harmonic partials

Overtone

An overtone results from molecular excitation to the second excited state, producing a series of integer multiples of the fundamental frequency. These transitions between non-contiguous vibrational states create absorption bands known as overtones (first and second overtone) at approximately multiples of the fundamental vibrational frequency. These transitions are much less probable than the fundamental ones, making the bands significantly weaker,the first overtone band is typically 10–100 times weaker than the fundamental frequency band, depending on the bond. These bands appear in the NIR range between 780 nm and 2000 nm.

Combination

Two or more vibrational modes can interact, causing simultaneous energy shifts and resulting in absorption bands known as combination bands, with frequencies equal to the sums of multiples of each interacting frequency. NIR combination bands typically appear between 1900 nm and 2500 nm. The strength of NIR bands is influenced by the change in dipole moment and the bond's anharmonicity. Due to the hydrogen atom being the lightest and exhibiting the largest vibrations and greatest deviations from harmonic behavior, the primary bands observed in the NIR region correspond to bonds containing hydrogen atoms (specifically C–H, N–H, O–H, and S–H). In contrast, bands for bonds such as C=O, C–C, and C–Cl are much weaker or even absent.

2.5.6 Advantages of NIR Spectroscopy

Non-Destructive Analysis: NIR spectroscopy allows for non-destructive analysis of samples, meaning that it does not alter or damage the sample during measurement. This is particularly useful when working with valuable or limited quantities of materials.

Rapid Analysis: One of the key advantages of NIR spectroscopy is its speed. Measurements can be performed in a matter of seconds or minutes, allowing for high-throughput analysis of large sample sets. This is especially beneficial in industries where quick decision-making is crucial, such as pharmaceutical manufacturing or food processing.

Versatility: NIR spectroscopy can be applied to a wide range of sample types, including solids, liquids, and gases. It can analyze organic and inorganic compounds, as well as complex mixtures. This versatility makes it a valuable tool in many scientific disciplines and industrial applications.

Minimal Sample Preparation: Unlike other analytical techniques that require extensive sample preparation, such as grinding or extraction, NIR spectroscopy often requires minimal or no sample preparation at all. This saves time and resources, making it an efficient choice for routine analysis.

Quantitative Analysis: In addition to qualitative analysis, NIR spectroscopy is also capable of quantitative analysis. By establishing calibration models using reference standards, it can accurately determine the concentration or content of specific components in a sample. This makes it suitable for quality control purposes in industries like pharmaceuticals or agriculture.

2.5.7 Limitations of NIR Spectroscopy

Limited Depth Penetration: One of the main limitations of NIR spectroscopy is its limited depth penetration. Near-infrared light can only interact with the outermost layers of a sample, typically up to a few millimeters. This restricts its application to surface analysis and may not provide information about deeper layers or bulk properties.

Sensitivity to Water: NIR spectroscopy is highly sensitive to water content in samples. Since water strongly absorbs near-infrared light, it can interfere with the analysis and affect the accuracy of results. Special measures, such as sample drying or using specific reference methods, may be required to overcome this limitation when working with water-containing samples.

Overlapping Absorption Bands: Another challenge in NIR spectroscopy is the presence of overlapping absorption bands from different components in a sample. This can make it difficult to resolve individual peaks or identify specific functional groups. Advanced data analysis techniques, such as chemometric modeling, are often employed to overcome this limitation and extract meaningful information.

Instrumentation Cost: While NIR spectroscopy offers many advantages, acquiring and maintaining the necessary instrumentation can be costly. Specialized NIR spectrometers, detectors, and software are required for accurate measurements and data analysis. This cost factor may limit its accessibility for smaller laboratories or companies with budget constraints.

Need for Calibration: To obtain reliable quantitative results with NIR spectroscopy, calibration is essential. This involves establishing a correlation between spectral data and reference values through a set of calibration standards. Developing robust calibration models requires expertise and time-consuming efforts. Additionally, these models may need periodic updating or validation to ensure their accuracy over time.

Despite these limitations, NIR spectroscopy remains a valuable analytical tool in various fields due to its unique advantages and versatility in material analysis.

2.5.8 Calibration and Validation Equipment

Figure 2.4 shows the equipment where the sampler and the calibrated digital circuits of the present invention are implemented, used in the calibration and validation step. The equipment is comprised of:

- **E1** - Sampler
- **E2** - Docking window of the sampler
- **E3** - Equipment lid
- **E4** - Energy source
- **E5** - Spectrophotometer 1
- **E6** - Spectrophotometer 2
- **E7** - Processing central unit
- **E9** - Forced refrigeration
- **E10** - Box
- **E11** - Power supply and data connector

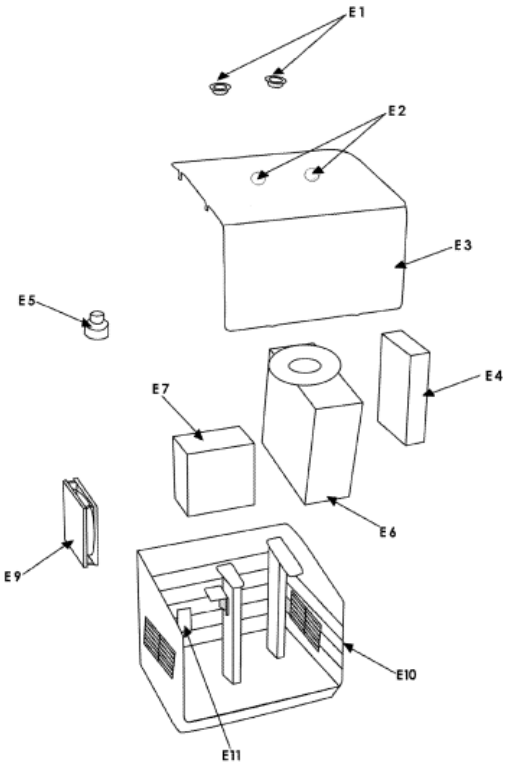


Figure 2.4: Calibration and Validation Equipment

2.5.9 Conclusion

Summary In this chapter we outlined some facts about Spectroscopy and its various types, then we discussed the NIR and elaborated on being the difference between NIR and IR, citing various chemical assignments of NIR bands, and finally, we discussed the Machine learning technique with Nir spectroscopy application. In the next chapter, we are going to focus on on Machine Learning-based Spectroscopy of the NIR.

Chapter 3

Machine learning

3.1 Introduction

Technological developments in recent years have enabled scientists to develop and refine highly effective information extraction and machine learning methods. The emergence of these new methods has made it possible to face the challenge posed by the very large number of data processed today by computer systems. Machine learning is a sub-discipline of artificial intelligence that deals with the development, analysis, and implementation of techniques to enable computers to learn and Improve with experience without being explicitly programmed. In this chapter we shall Overview This section provides an overview of the topic: it defines Machine Learning and types of machine learning algorithms with their benefits and drawbacks; and a variety of areas where the usage of machine learning is in place. Additionally, we will examine the use being made in the domain of cholestrol blood analysis and healthcare, it shows the potential to revolutionize the domain.

3.2 Machine learning

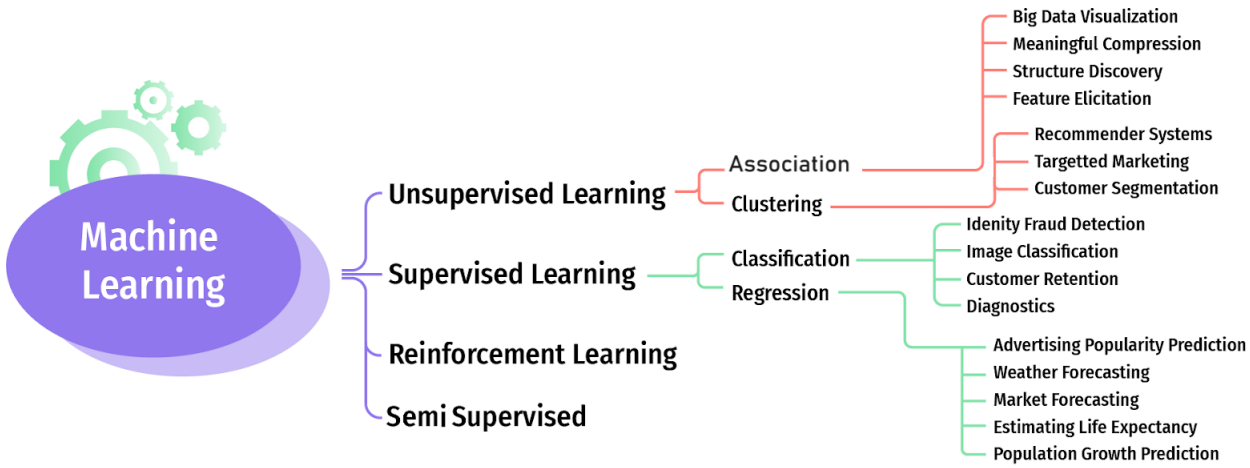
3.2.1 Definition

Machine learning is the form of AI that focuses on making systems learn or improve performance from the data being fed to them. Artificial intelligence is a broader term referring to systems or machines emulating a form of human intelligence. Machine learning and AI are often discussed together, and these terms are sometimes used interchangeably, although they do not refer to exactly the same concept. An important distinction is that, even if machine learning is based entirely on artificial intelligence, artificial intelligence is not limited to machine learning.

Today, we apply machine learning to everything. When we touch a bank, purchase something online, or use social media, machine learning algorithms work to optimize, streamline, and secure the experience. Machine learning is a fast-developing technology with its surrounding technology, and we only touch the tip of the iceberg of its capabilities.[38]

3.2.2 Types of Machine Learning

The theory of learning uses mathematical tools derived from the theory of probabilities and the theory of information to evaluate the optimality of certain methods compared to others. Figure 3.1 presents different types of ML systems, which can be classified into three main types: supervised learning, unsupervised learning and reinforcement learning. Among these types of ML, the supervised learning is the most commonly used.



(a)

Figure 3.1: Types of machine learning

1.Supervised Learning

Supervised Learning is the most popular and widely used learning technique. It corresponds to the case where the learning objective is explicitly determined through the definition of a target to be predicted. It is therefore a technique that allows learning from examples, that are accompanied by additional information about whether or not they belong to the concept [39]. Supervised Learning is generally used for regression or classification

a-Classification: is also a type of machine learning model that involves predicting a target variable that is categorical. The goal of classification is to assign new observations to predefined categories or classes based on their features or input variables. The input data for classification can be structured or unstructured, and the output can be binary or multi-class [40].

b-Regression: is a mathematical approach used in machine learning that allows data scientists to predict a continuous result (y) based on the values of one or more predictor variables (x). One of the most popular types of regression analysis is linear regression because it is so easy to apply in predicting and forecasting .[41]

2.Unsupervised learning

is almost the opposite of supervised learning. It is used to generate the classification of a model when datasets are unlabeled. Another name for unsupervised learning is knowledge discovery. The common unsupervised learning techniques include clustering, and dimensionality reduction. The algorithms of unsupervised learning discover hidden patterns or data groupings without the need for human intervention. Its ability to discover similarities and differences in information makes it the ideal solution for data analysis [39]

a-Clustering : is the method of separating or splitting a dataset into groups so that datasets belonging to the same group are more similar than datasets belonging to other groups. Briefly, the objective is to separate groups with similar characteristics and divide them into clusters [42].

b- Association: is unsupervised learning when an algorithm tries to learn without the guidance of a teacher because the data are not labeled. The association rule is a descriptive strategy, not a predictive method, that is commonly used for finding the relationships between variables in the large database. The relationship is commonly represented as a set of rules or a group of frequently occurring itemsets. Association rules mining is a technique for discovering new and interesting connections between objects in a set, such as a common pattern in transactional data or any sort of relational database [43].

3.Reinforcement Learning

Reinforcement learning is an AI approach that focuses on learning the system through its interactions with the environment.With reinforcement learning, the system adapts its parameters according to the reactions received from the environment, which then provides feedback on the decisions taken. For example, a system that models a chess player who uses the results of previous steps to improve his performance, is a system that learns with reinforcement. Current researche on reinforcement learning are highly interdisciplinary and includes researchers specializing in genetic algorithms, neural networks, psychology and control techniques [44].

4.Semi-supervised learning

Semi-supervised learning represents the intermediate ground between Supervised and Unsupervised learning algorithms. In many practical cases, marking is quite expensive because it requires qualified human experts to do so. Therefore, semi-supervised algorithms are the best candidates for model construction in cases where labels are not present in most observations but are present in a few cases. These methods exploit the idea that even if the membership of an unlabelled data group is unknown, the data contains important information about the group parameters [39].

3.2.3 Methods of Supervised Learning

Due to the large amount of documents exchanged and stored on electronic media, supervised learning has become more than necessary to facilitate the use and analysis of data. Among the most popular

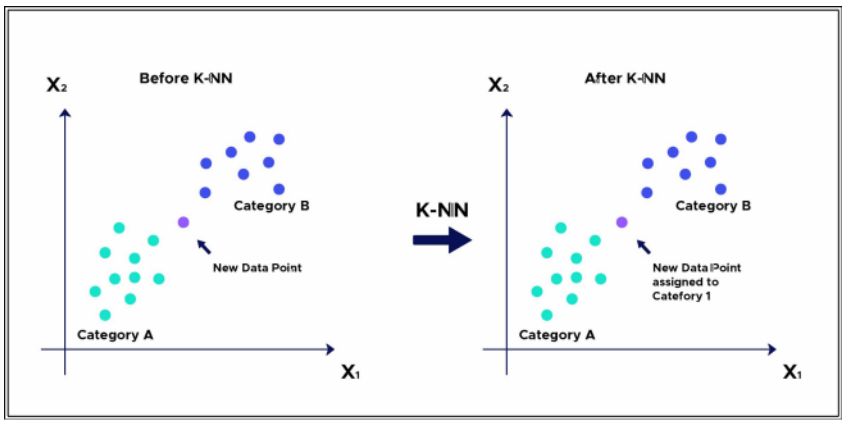
supervised learning methods are the following:

- K-nearest neighbor.
- Decision tree.
- Neural network.

K-Nearseat Neighbor

K-NN is one of the simplest and most direct classification algorithms. It is a very simple and non-parametric algorithm. Using K-NN, classification is simply based on a simple vote among the nearest observation classes. Therefore, it does not require a learning phase to generate a model as is often the case in most classification algorithms. Thus, the classification is based directly on the data of the learning base and not on a classification model as is the case for the other algorithms [45]. Figure 3.2 illustrates the data before and after applying the K-NN algorithm, where the x1 and x2 axes represent the features

- Ease of implementation of the K-NN.
- Effectiveness for classes distributed irregularly.
- Effectiveness for incomplete (or heterogeneous) data.
- K-NN method does not use a template to classify documents.



(a)

Figure 3.2: Example of K-NN

K-NN doesn't have only advantages, it also has some obstacles and disadvantages. Among these obstacles we can mention:

- Classification time: the method does not require learning which implies that all calculations are done during classification.
- Method will give poor results if the number of relevant attributes is small relative to the total number of attributes [46].
- Requires storage capacity and computing power [47].

Decision Tree

Decision Tree (DT) is a very powerful classification and decision support tool that is both descriptive and predictive, whose generated model is represented as a tree that is easy to understand and easy to use by a human user. In the decision tree, the path from the root to a sheet corresponds to a classification rule. Unlike other classification models, decision trees are extremely intuitive and provide a graphical representation to simplify the prediction process. Decision trees are now widely used in various fields, such as computer system security, data mining, medicine, etc. Their popularity is attributed to their readability, fast execution speed, and minimal number of assumptions required [48]. Decision Tree offers several advantages such as:[49]

- The ability to work on symbolic data
- The great ability and efficiency to make classification .
- The ease of learning and use.

DT doesn't have only advantages, it also has some obstacles and disadvantages. Among these obstacles we can mention:[50] [49]

- Highly sensitive to outliers and noise .
- The sensitivity to data change .
- Difficult to detect interactions between variables .
- The construction and pruning of the decision tree is too costly in terms of calculation time and storage resources .

3.2.4 Neutral Network

Neural networks are computational models inspired by the human brain. It comprises a network of artificial neurons, using weights and activation functions in processing input data to come up with the predictability of the output. The network will learn from the training data with its weights iteratively adjusted. It can capture complex patterns and relationships, which makes it quite an important component in conducting image recognition and natural language processing, among others. Neural networks have revolutionized the aspect of machine learning and have been adopted in many sectors.[51] Neutral Network is characterized by a number of benefits, which include the following:[52]

- The timeliness and effectiveness of large corpus processing .
- The possibility of combining this type of algorithms with other classification methods .
- A very low error rate compared to other classification methods .
- Neural networks do not require the use of very complex mathematical models for their functioning. Indeed, due to their learning capacity, they are mainly based on the data models to be processed .

Neutral Network doesn't have only advantages; it also has some obstacles and disadvantages. Among these obstacles we can mention:[52][53]

- The slowness of learning .
- The results obtained by the classification of neural networks are not interpretable. Indeed, the network generated by this type of algorithms is considered a black box, which means that the user has no explicit information about its internal functioning. In case of errors, it is impossible to determine the cause of the error .
- The convergence of neural network results is uncertain .
- Neural networks do not allow the integration of a priori knowledge for the processing of new data .

3.2.5 Classification types

There are mainly 3 different types of classification tasks that you may encounter in machine learning and specialized approaches. Generally, the different types of predictive models in machine learning are as follows.

Binary classification

A binary classification refers to those tasks which can give either of any two class labels as the output. where the output is restricted to two classes. Generally, one is considered as the normal state and the other is considered to be the abnormal state. [54]

Multi-Class classification

These types of classification problems refer to classification tasks that have more than two class labels. Many binary classification algorithms can also be utilized for multi-class classification .[54]

Multi-Label classification

Multi-Label Classification is the supervised learning problem where an instance may be associated with multiple labels. This is an extension of single-label classification (multi-class, or binary) where each instance is only associated with a single class label .[55]

3.2.6 Conclusion

This chapter provided an initial presentation of machine learning, including its definition and various types. We explored the different methods of supervised learning, delved into neural networks, and discussed classification types. This foundational knowledge sets the stage for understanding how machine learning can be applied to complex problems, including those in spectroscopy.

Chapter 4

Data Preprocessing and Visualization

4.1 Introduction

NIR spectroscopy classification for non-invasive blood analysis (NBA) is of paramount importance as it offers a comfortable and rapid method for analyzing blood samples without invasive procedures. It enables real-time results, minimizes risk, and allows for predictive analytics, facilitating early detection, personalized treatment planning, and cost-effective healthcare practices. After introducing the general concepts and methods of ML and spectroscopy in the previous chapters, this chapter aims to present the life cycle of ML, along with the tools and technologies that support our work. Then, we will describe the used NIR spectroscopy data. Finaly, we will propose ML models for NBA to predict the level of specific chemical compounds by utilizing spectral data obtained from NIR, followed by their description.

4.2 Flowchart of Application of ML in spectroscopy

Step 1: Sample Preparation :

Different types of samples can be analyzed by spectroscopy. Here, the concentration is on fingerprint samples.

Step 2: NIR Data Collection

Data must be collected on a series of specimens.These data are usually arranged in the form of a matrix X, having as many rows as the number of samples and as many columns as the number of measured variables. Each row of the matrix corresponds to the whole spectrum of a particular sample, whereas each column represents the absorbance of all the individuals at a particular wave number.

Step 3: Spectrum Preprocessing

Preprocessing of NIR spectra is important for:

Reducing noise and data outliers Extracting more valuable information for analysis

Step 4: Model Selection :

One of the very important steps is to determine the right algorithm among the classification algorithms, which must be adapted to the spectroscopic domain. This choice has to be based on the criterions discussed beforehand and, moreover, domain-specific requirements have to be analyzed.

Step 5: Model Evaluation

For this purpose, the chosen models will be evaluated for their performances using various techniques to make sure that they are accurate, reliable, and well generalizable for new data.

4.3 Data Collection

4.3.1 Data Source

The dataset used in this study was obtained (downloaded) from the Zindi platform , a data science community similar to Kaggle, that hosts a variety of data science competitions and provides access to datasets from various sources. The dataset was published by **bloods.ai**.
This dataset was colleted in order to predict the presence of compounds in the blood by analyzing spectroscopy readings.

4.3.2 Data Overview

In this research, each dataset collected consists of scans made using an identical scanner model. Data is collected as raw data resulting from the scanner pushing light into the target (in this case, a fingertip). The data is displayed as a function of wavelength measured in nanometers (nm) and registers the quantity of light reflecting off the target point. The application registers the results and creates an array of values to account for all wavelength data. The light that reflects back is referred to as “intensity.”

Each scan cycle pushes light as a function of wavelength, and the wavelength data ranges in intensity from 900 nm to 1700 nm. All reflected light is distinguished and categorized by wavelength (e.g., 900 nm is distinguished from those returning at 905 nm and so on). The expected intensity values are at 900, 904.71, 909.41, 914.12, and many more, resulting in an array of 170 intensities per scan.

This process is done 60 times in order to produce a reliable scan and see a holistic picture. Additionally, humidity and temperature are also taken into account, as these factors may affect scan results. Therefore, each of the 60 scans comprises 170 intensities where temperature and humidity at the time of the scan are accounted for.

4.3.3 Data validation process

Each biodata donation consists of 60 scans, and ensuring the consistency and quality of each of the 60 scans is the top priority to extract valuable results from NIR data. Any movement in the scanner or scan target may result in invalid scan data that loses its value. For the most accurate scan results, the

standard deviation of the 60 scans is calculated and, if the result wasn't consistent, the scans are excluded from the data set.

4.3.4 Data file structure

In our research we use the following 3 files :

"train.csv" : This is the dataset that we will use to train our model it contains 29 160 rows and 178 columns which are:

A-Features:

- **Absorbance**: 170 of these are labeled as *absorbance0*, *absorbance1*, and so on. This is an intensity spectrum of the target blood response to pointed light.
- **Temperature**: the temperature readings at the time of the scan.
- **Humidity**: the humidity readings at the time of the scan.
- **Donation id**: identifier of all 60 scans for each individual donation (the same donation id for the same 60 scans).
- **Id**: an identifier that refers to the number of scans for each donation (60 total) and goes from 1 to 60.
- **std**: the standard deviation is calculated for each donation in the range of 950 nm to 1350 nm.

B-Labels:

- **Hdl cholesterol human**: high-density lipoproteins which can be high, ok, or low depending on the human acceptable ranges for that compound and in numeric values.
- **Cholesterol ldl human**: low-density lipoproteins, it may be high, ok, or low level depending on the human acceptable ranges for that compound and in numeric values.
- **Hemoglobin (hgb) human**: indicates the level of Hemoglobin in blood test depending on the human acceptable ranges for that compound and in numeric values. It can be high, ok, or low.

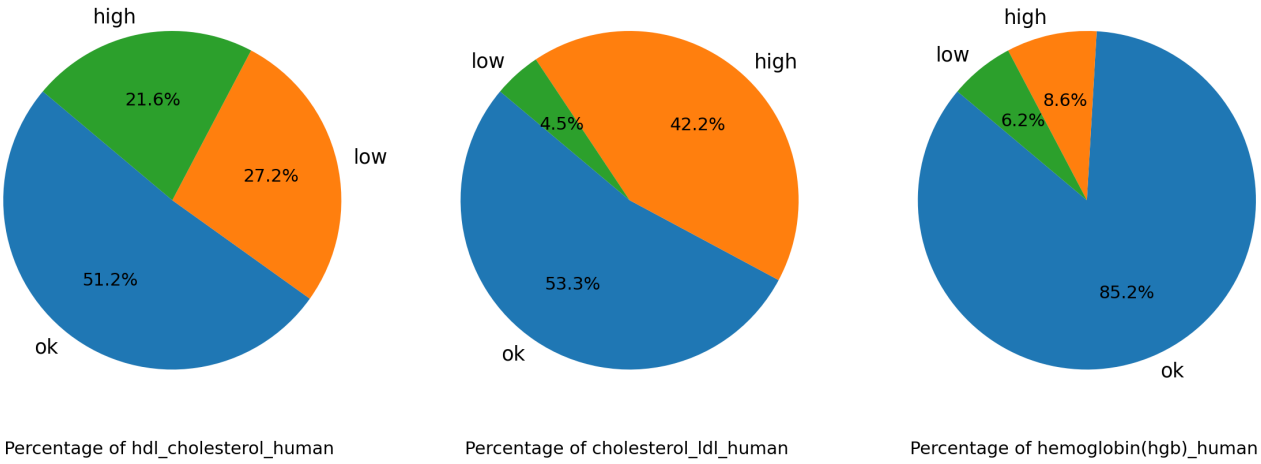
"Test.csv" : Resembles Train.csv but without the target-related columns. This is the dataset to which we will apply our model for validation.

"Zindi Contest Spectra.xlsx" :Shows The Spectral Signature that is like the "chemical fingerprint" of a compound and can be used to identify it in a sample. Glucose and cholesterol, Fat, skin, and Deoxygenated blood are the existed spectral signatures.

4.4 Exploratory Data Analysis

EDA will reveal the data set's underlying structure as well as trends, patterns, and linkages that aren't immediately apparent. It will also assist in deriving trustworthy conclusions from a large amount of data by carefully and deliberately looking at it via an analytical lens. These revelations will eventually lead to the selection of a suitable predictive model.

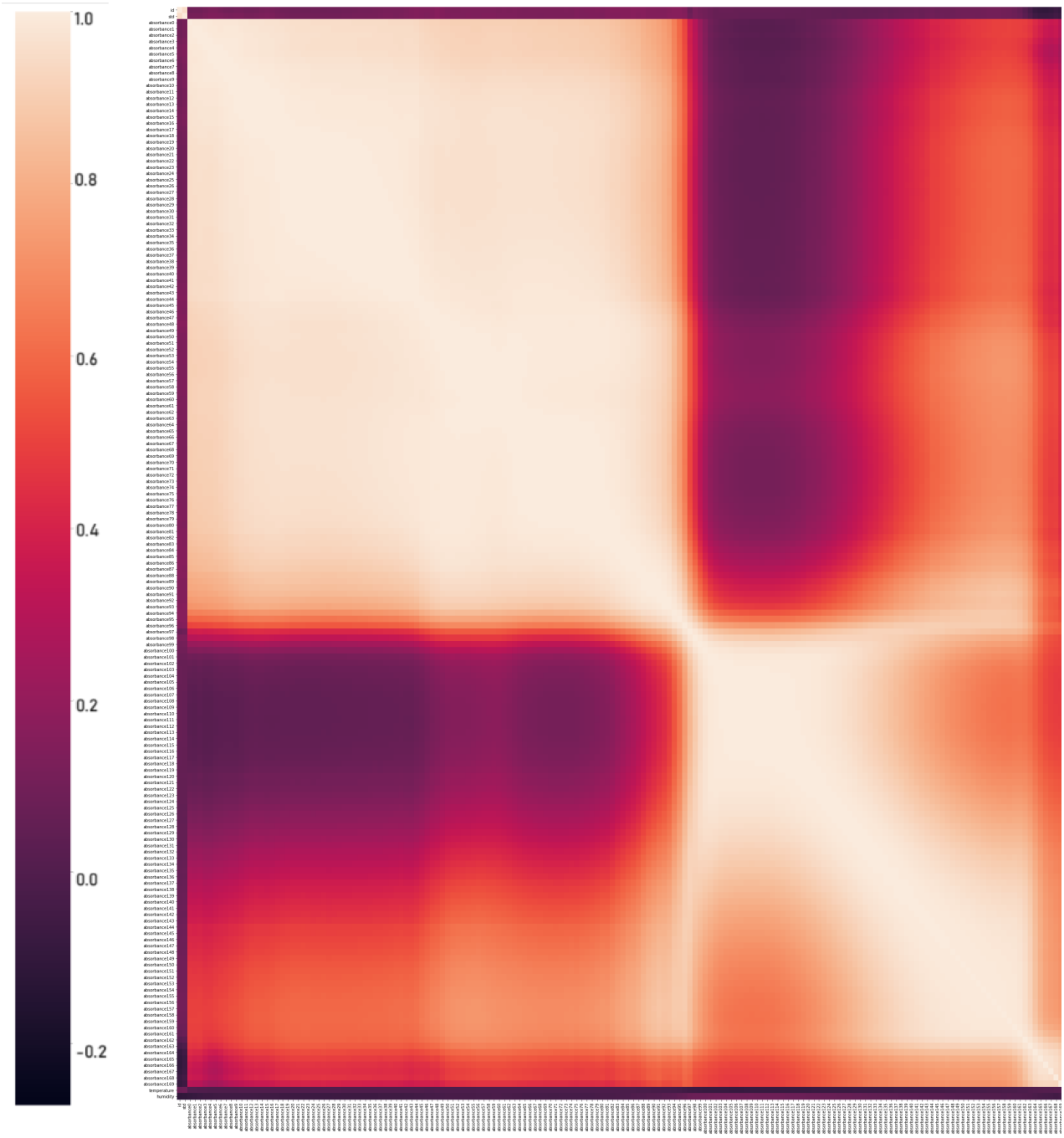
The below figures represent the distribution of the classes of each of the target variables in the dataset. It is observed that the classes are highly imbalanced for each of the target variables.



(a)

Figure 4.1: Distribution of the Classes of Each of The Target Variables.

Figure 4.2 is a correlation heatmap to check the correlation between the features. We observe that there exists multicollinearity in the data, i.e., we cannot distinguish the individual effects of independent columns. Additionally, one predictor variable in the model can be linearly predicted from others. Thus, we cannot make a direct inference as the number of features is very high.

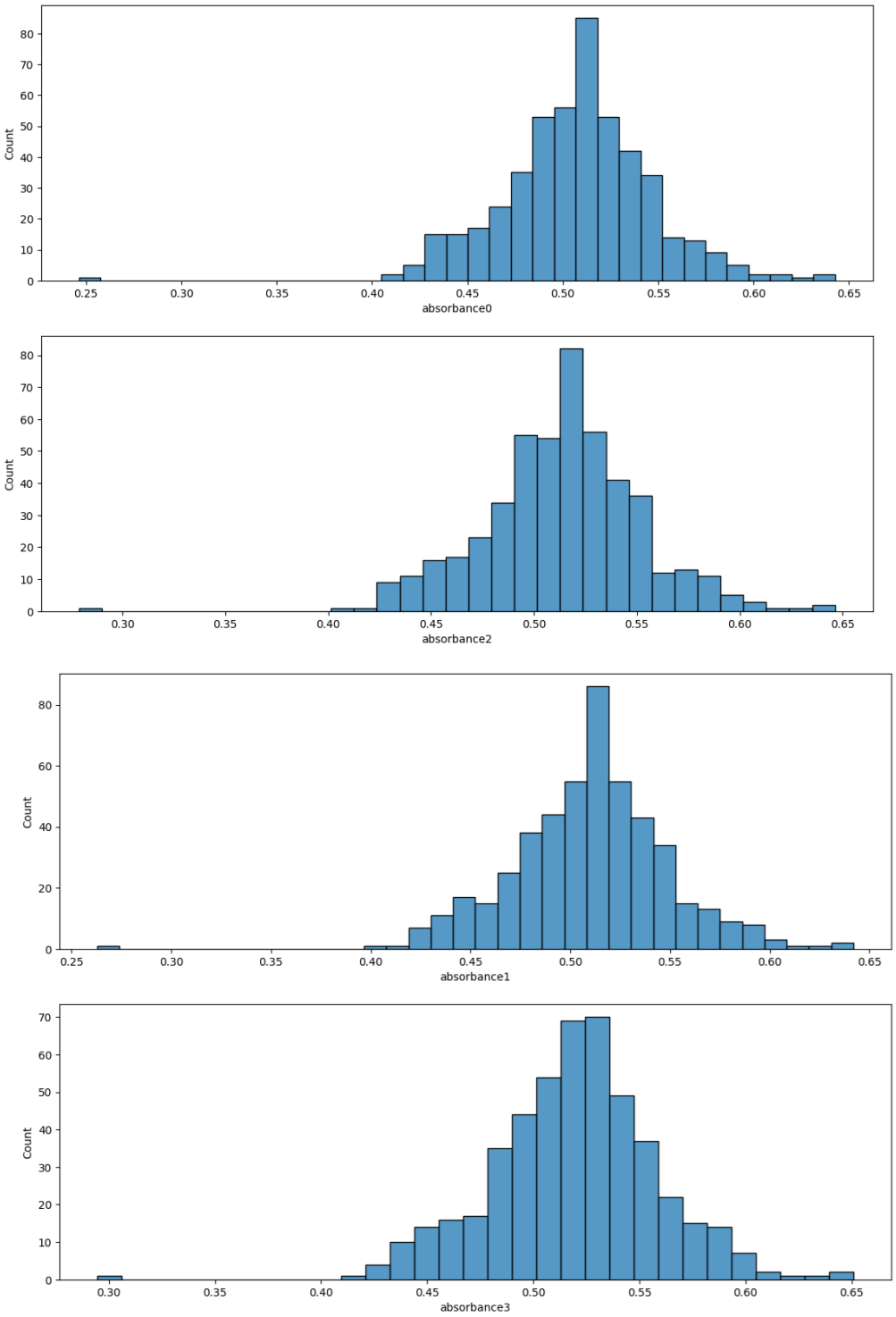


(a)

Figure 4.2: Correlation heatmap of the features

Figure 4.3 shows histograms for some of the absorbance features to check if they follow any distribution.

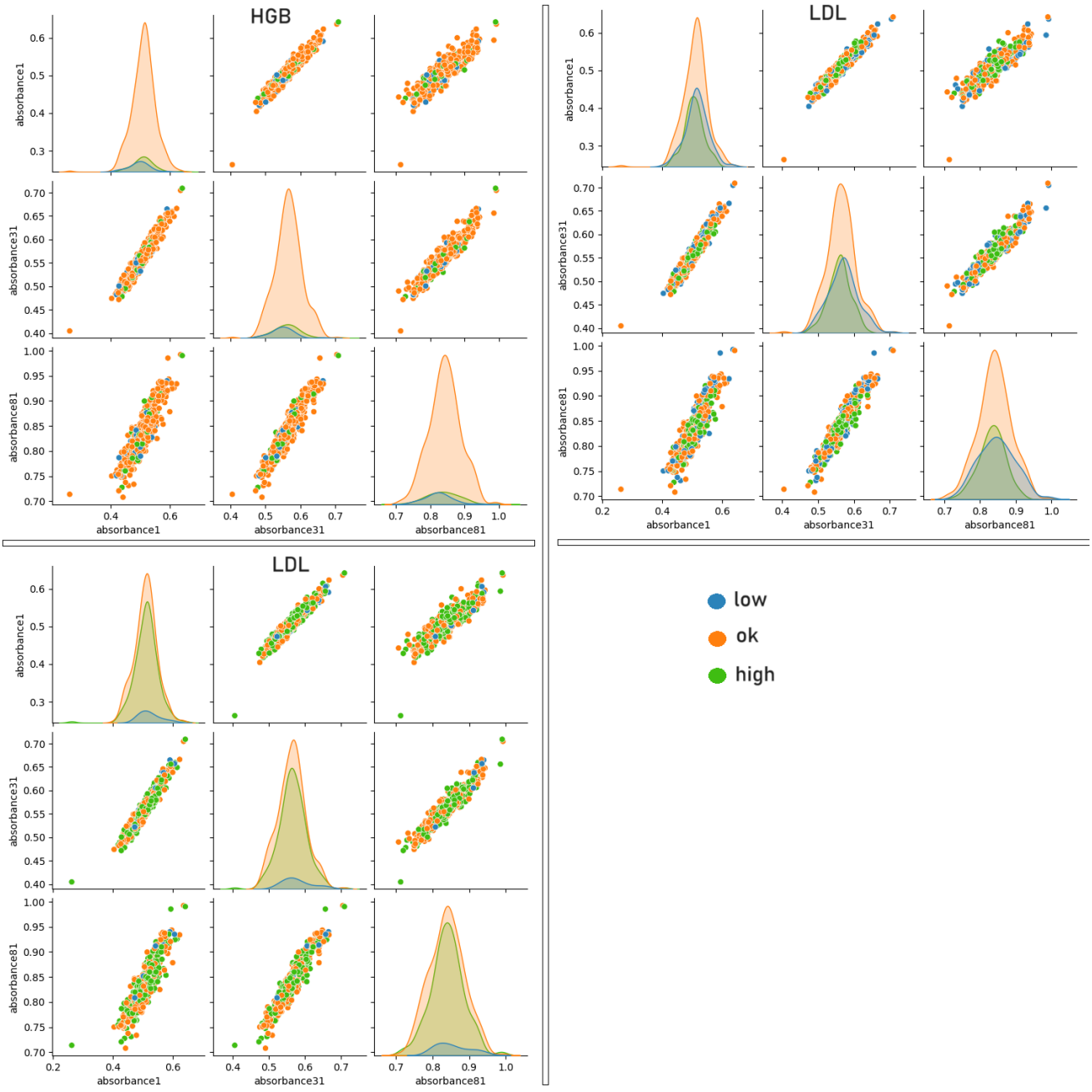
We found that the absorbance values are normally distributed as seen from the graphs below :



(a)

Figure 4.3: Distribution of Absorbance Values

Figure 4.4 visualizes the relationships between three absorbance measurements (absorbance1, absorbance31, absorbance81) and their categorization based on (HDL,LDL,HGB) levels .
Indicate that there is no structured distribution of levels (low, ok, high).and do not form well-separated clusters in the absorbance space.



(a)

Figure 4.4: Absorbance Measurements Categorized by HDL, LDL, and HGB Levels

4.5 Data Preprocessing

Data preprocessing is an essential process because it has a direct impact on the project's success rate. and the data in the real world is unclean, which minimizes the complexity of the data. I followed the below steps as part of preprocessing:

4.5.1 Drop duplicated measurements

knowing that we have 60 measurements for each sample, so to de-duplicate the readings, the most median values of absorbance intensities were aggregated across all sample donation ids.

Most median: it is calculation of the "most median" row within each group of the data. For each group it finds the row where the most individual values are closest to the median of their corresponding columns. This row essentially represents the data point in the group that is closest to the center point for each variable.

4.5.2 Remove incorrect measurements

information of Donation id =6824 contains negative values of absorbance which are outliers data, so we delete the 60 measurements of this donation id

4.5.3 Column Name Standardization

The datasets we are handling contain columns labeled from absorabance0 to absorbance169, which lack meaningful descriptors. Recognizing that each column represents absorbance values measured at specific wavelengths, we can enhance interpretability by renaming these columns to correspond with their respective wavelength values. This preprocessing step aids in clarity and understanding of the data.

absorbance3	absorbance4	absorbance5	...	absorbance164	absorbance165	absorbance166	absorbance167
0.516305	0.524040	0.530256	...	1.401995	1.393786	1.391894	1.388666
0.532937	0.538302	0.543165	...	1.298759	1.299603	1.300307	1.296176
0.531594	0.539004	0.546688	...	1.246909	1.226492	1.212731	1.206610
0.520733	0.527268	0.534909	...	1.254828	1.246104	1.249281	1.248782
0.503966	0.507523	0.513195	...	1.196684	1.196917	1.194641	...
...	...	...	...	...	...	...	...
0.510471	0.518624	0.524355	...	1.166093	1.142420	1.131919	1.119596
0.554534	0.561449	0.568376	...	1.281499	1.254623	1.240389	1.234687
0.530360	0.538104	0.545496	...	1.277367	1.265666	1.255881	1.257341
0.583375	0.590397	0.598066	...	1.240581	1.209632	1.194896	1.181866
0.538756	0.546795	0.553875	...	1.228572	1.211419	1.178484	1.159697

	900	904.733704	909.467468	914.201172	918.934937	923.66864	928.402344	933.136108	937.869812	942.603577
0	0.504811	0.506520	0.511544	0.516305	0.524040	0.530256	0.536083	0.542178	0.551150	0.567042
1	0.523239	0.523318	0.526411	0.532937	0.538302	0.543165	0.547432	0.553580	0.561792	0.575503
2	0.515089	0.517911	0.522953	0.531594	0.539004	0.546688	0.551226	0.555644	0.560248	0.573651
3	0.511924	0.513889	0.518748	0.520733	0.527268	0.534909	0.542498	0.549607	0.556574	0.567634
...	0.496975	0.497199	0.500494	0.503966	0.507523	0.513195	0.519947	0.525307	0.531738	0.541384
...	...	...	...	...	...	...	...	...	...	...
480	0.496931	0.501263	0.505982	0.510471	0.518624	0.524355	0.529103	0.533394	0.541696	0.551260
481	0.538405	0.544228	0.547718	0.554534	0.561449	0.568376	0.573394	0.577009	0.585547	0.595923
482	0.518136	0.522781	0.525928	0.530360	0.538104	0.545496	0.550113	0.556301	0.568461	0.579187
483	0.567857	0.572361	0.577658	0.583375	0.590397	0.598066	0.602257	0.605641	0.613142	0.623021
484	0.528119	0.527905	0.533717	0.538756	0.546795	0.553875	0.561601	0.564946	0.571672	0.580672

(a)

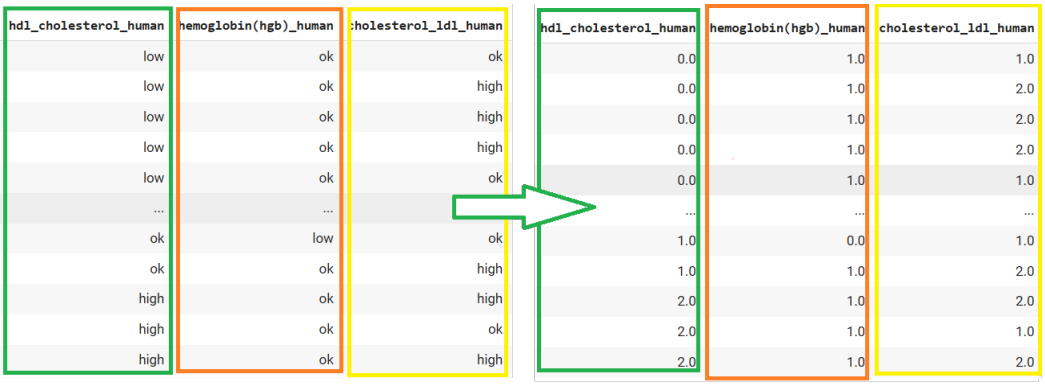
Figure 4.5: Columns before and after renaming

4.5.4 Optimized Feature Selection

In the feature selection that I basically focus on the absorbance column for wavelengths from 950 to 1350 nm. This deliberate selection is informed by the superior tissue penetration capabilities associated with this specific range within the Near Infra-Red (NIR) spectrum. for effective result applications, especially in cholesterol and hemoglobin blood analysis using NIR spectroscopy.

4.5.5 Encoding categorical data

Categorical data typically appears as "strings" or "categories" with a finite set of values, while numerical data is conveyed through numbers. In machine learning, it's common practice to convert categorical data into numerical format. (Refer to Figure 4.6)



hdl_cholesterol_human	hemoglobin(hgb)_human	cholesterol_ldl_human
low	ok	ok
low	ok	high
low	ok	high
low	ok	high
low	ok	ok
...	...	...
ok	low	ok
ok	ok	high
high	ok	high
high	ok	ok
high	ok	high

hdl_cholesterol_human	hemoglobin(hgb)_human	cholesterol_ldl_human
0.0	1.0	1.0
0.0	1.0	2.0
0.0	1.0	2.0
0.0	1.0	2.0
0.0	1.0	1.0
...	...	...
1.0	0.0	1.0
1.0	1.0	2.0
2.0	1.0	2.0
2.0	1.0	1.0
2.0	1.0	2.0

(a)

Figure 4.6: Labels before and after encoding

4.5.6 Drop unimportant columns

The two columns, "id" and "donation id," in the data offer additional details for identifying each scan but hold no significance for model development. Thus, we can safely remove these columns.

4.5.7 Normalization

The goal is to transform the data into a range that makes more suitable for the model. By ensuring that features have a similar scales, which helps in speeding up the convergence of gradient-based learning algorithms and improves the performance of the model.

4.6 Conclusion

This chapter focused on essential data preparation and visualization techniques for spectroscopy and machine learning.I explore data collection, quality checks, and structure, followed by methods to uncover patterns through EDA. Key preprocessing steps like cleaning duplicates, formatting data, and selecting features ensure the data is machine-learning-ready. Finally, I introduce two approaches to preparing spectroscopy data for building powerful models.

Chapter 5

Proposed Architecture and Evaluation

5.1 Introduction

This chapter outlines the proposed architecture and its evaluation, detailing two distinct approaches designed to address the problem. We start by describing the first proposed approach and then move on to the second approach. Each approach is rigorously evaluated based on specific criteria and methodologies, with a thorough comparative analysis of the results to assess accuracy and performance metrics. Through these evaluations, we aim to identify the strengths and weaknesses of each approach, ultimately guiding the selection of the most effective solution.

5.2 First Proposed Approach

After preprocessing data as previously described, the first proposed approach can be summarized in the following steps:

Step 1: In this step, the aim is to extract three distinct datasets from whole Data, tailored for the classification of HDL, LDL, and HGB levels. This process is crucial for optimizing classification accuracy based on specific absorbance characteristics and parameters unique to each biomarker. By segregating the data into these specialized datasets, the classification models can focus on learning the distinct patterns and relationships associated with HDL, LDL, and HGB levels separately. This approach ensures that the models are finely tuned to interpret the varying absorbance profiles and parameters relevant to each biomarker, thereby enhancing their precision in predicting or diagnosing conditions related to lipid and hemoglobin levels.

Step 2: Feature selection is crucial for identifying the most informative wavelengths for the target analytes. As I explained earlier in Chapter 2, each component in the sample, whether HDL, LDL, HGB, or other blood constituents,*" interacts with specific wavelengths of radiation, corresponding to the energy levels required to excite its atoms or molecules."* This selective absorption forms the basis for spectroscopy.

I selected a subset of absorbance columns focusing on wavelengths between 950 and 1350 nanometers, known for their effective tissue penetration. Then, I applied different numbers of Principal Component Analysis (PCA) for each target (HDL, LDL, and HGB) to choose the best absorbance variables for each. This approach helps me reduce the data's dimensionality while retaining crucial information, ensuring that the absorbance features selected are most informative for accurately predicting or analyzing HDL, LDL, and HGB levels based on their unique absorption characteristics within the chosen wavelength range.

Step 3: Split the data into training and testing sets using the train-test-split method. The training set will comprise 70% of the data, while the remaining 30% will be allocated to the testing set.

Step 4: Begin by selecting the appropriate machine learning algorithms for the task. In this context, three algorithms are considered: XGBoost, Support Vector Classifier (SVC), and K-Nearest Neighbors (KNN). Each algorithm has specific hyperparameters that need to be optimized to achieve the best performance:

- For XGBoost, consider parameters such as 'n_estimators' (number of trees), 'max_depth' (maximum depth of each tree), and 'learning_rate' (step size shrinkage).
- For SVC, focus on parameters like 'C' (regularization parameter) and 'kernel' (the type of kernel function used).
- For KNN, determine the 'n_components', which typically refers to the number of neighbors considered during classification.

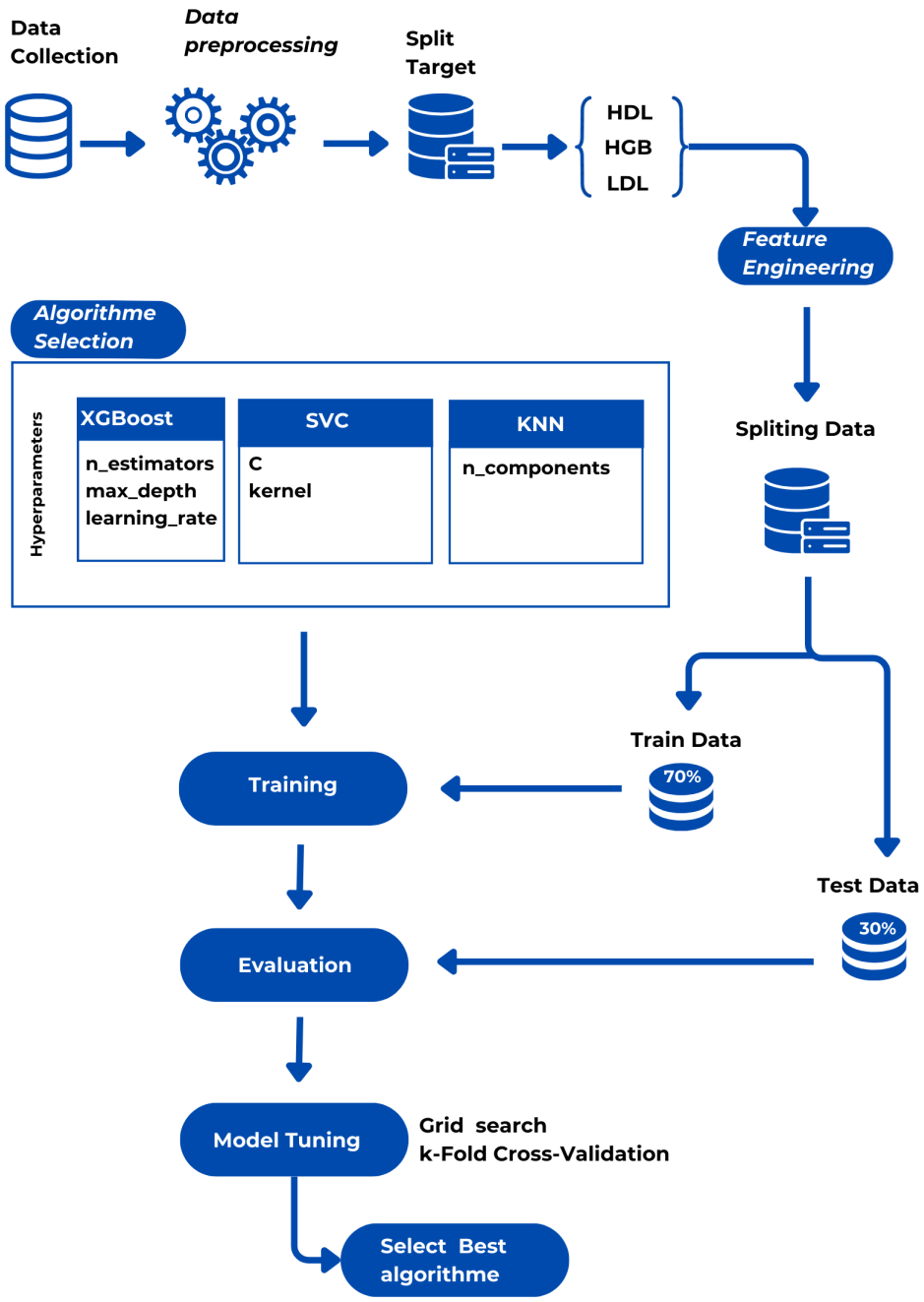
Step 5: The selected algorithms are trained on the training data. During this phase the models learn from the data by adjusting their parameters to minimize the error in predictions.

Step 6: The trained models are evaluated on the test data to assess their performance. Common metrics for evaluation might include accuracy, precision, recall, F1-score, etc.

Step 7: The models' hyperparameters are adjusted to improve performance. Grid search and k-fold cross-validation are two methods used in this step to find the best combination of hyperparameters

Based on the evaluation and tuning results, the best-performing algorithm is selected for deployment.

Step 8: The selected model is deployed into a production environment where it can be used to make predictions on new data.All these steps are illustrated in the architecture shown in figure (see figure 5.13).



(a)

Figure 5.1: Architecture of the first proposed approach

5.3 Second Proposed Approach

After preprocessing data as previously described, the second proposed approach can be summarized in the following steps:

Step 1: In this Approach, the primary focus will be on LDL (Low-Density Lipoprotein). The decision to concentrate on LDL stems from its classification challenges and its significant impact on health, making it a critical area of study due to its prevalence and serious health implications.

LDL cholesterol in a simpler way. Instead of having three labels ("low," "ok," "high"), we're suggesting just two: "ok" and "high". Because very low LDL is pretty uncommon, and some studies from Harvard Health suggest it be helpful for people who already have heart problems. This way, we can keep things clear and focus on what matters most: heart health.[31] Considering that "ok" and "low" are both useful categories, I will classify them together as a single group.

Step 2: Feature selection is crucial for identifying the most meaningful wavelengths for analyzing LDL. because it interacts with specific wavelengths of radiation that correspond to the energy levels required to excite its atoms or molecules. This selective absorption forms the basis for spectroscopy.

For the analysis, choose a subset of absorbance columns representing wavelengths between 950 and 1350 nanometers, which are known for their ability to effectively penetrate tissue.

Step 3: Split the data into training and testing sets using the train-test-split method. The training set will comprise 70% of the data, while the remaining 30% will be allocated to the testing set.

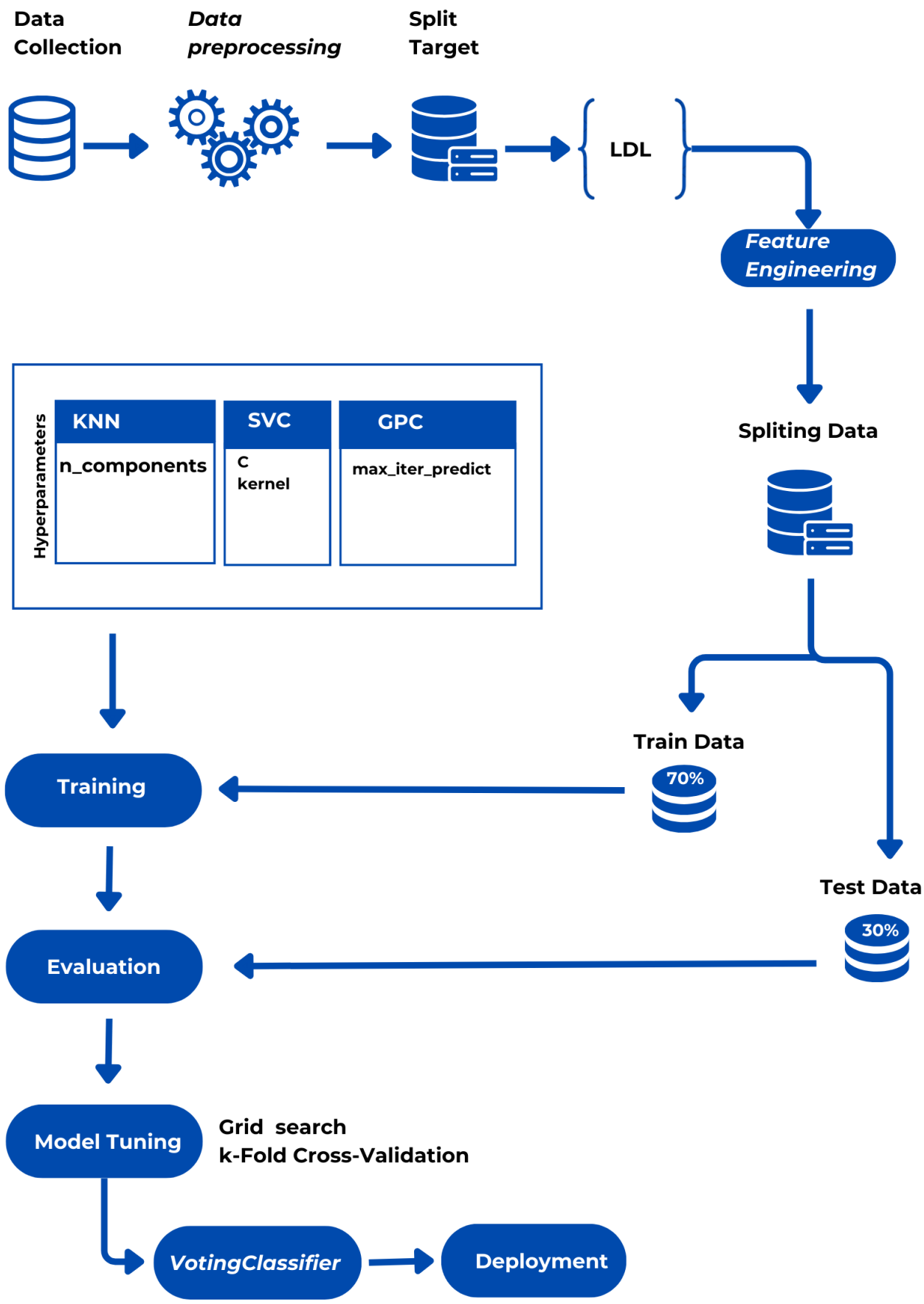
Step 4: Begin by training the appropriate machine learning algorithms for the task. In this context, three algorithms are considered: K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), and Gaussian Process Classifier (GPC). Each algorithm has specific hyperparameters that need to be optimized to achieve the best performance:

- For KNN, determine the 'n_components', which typically refers to the number of neighbors considered during classification.
- For GPC, consider parameters such as 'max_iter_predict' (maximum number of iterations).
- For SVC, focus on parameters like 'C' (regularization parameter) and 'kernel' (the type of kernel function used).

Step 5: The trained models are evaluated on the test data to assess their performance. Common metrics for evaluation might include accuracy, precision, recall, F1-score, etc.

Step 6: The models' hyperparameters are adjusted to improve performance. Grid search and k-fold cross-validation are two methods used in this step to find the best combination of hyperparameters

Step 7: Combined three algorithms using a voting classifier, which aggregates the predictions of each model to produce a final prediction. This can be done through hard voting (majority rule).
The final predictions, generated by the VotingClassifier are deployed in a production environment to enable real-time predictions on new data. All these steps are illustrated in the architecture shown in figure (see figure 5.2).



(a)

Figure 5.2: Architecture of Second proposed approach

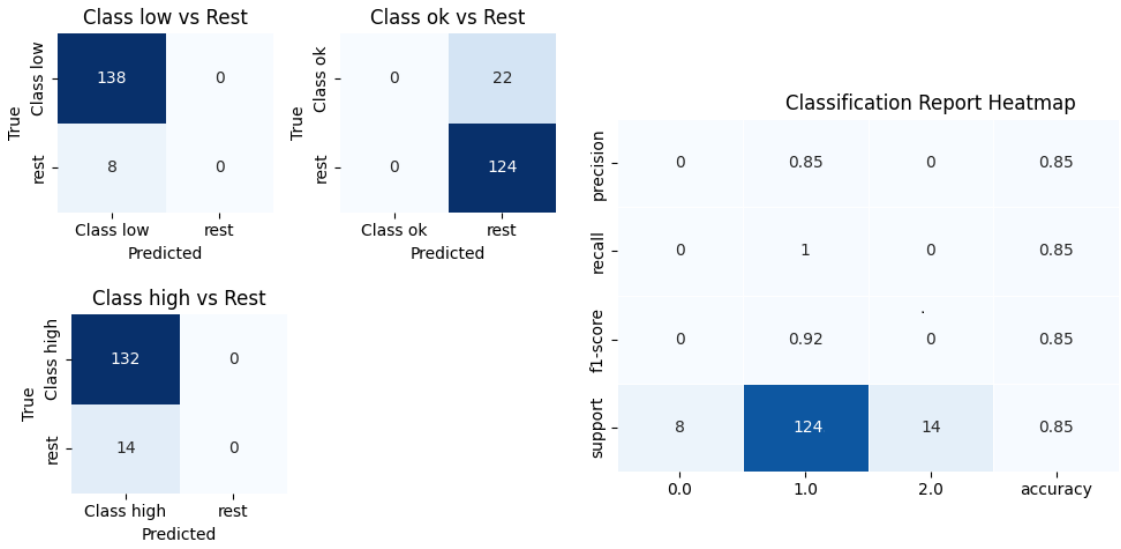
5.4 Evaluation of First Proposed Approach

In this approach, I compared three algorithms (SVC, KNN, XGBoost) to determine the best performer across three target variables(HDL, HGB, LDL), each with three labels(low, ok, high). The model evaluation was conducted using multiple machine learning techniques, including K-fold cross-validation and grid search.

5.4.1 SVC-based approach :

HGB :

The optimal hyperparameters for the SVC model include a regularization parameter C, a kernel type, and a gamma value. Specifically, a C value of 0.1, a linear kernel, and a gamma value of 0.03 yielded the best performance based on the selected metric. The model achieved a score of 0.85, reflecting its performance under these conditions, and an accuracy of 0.84.

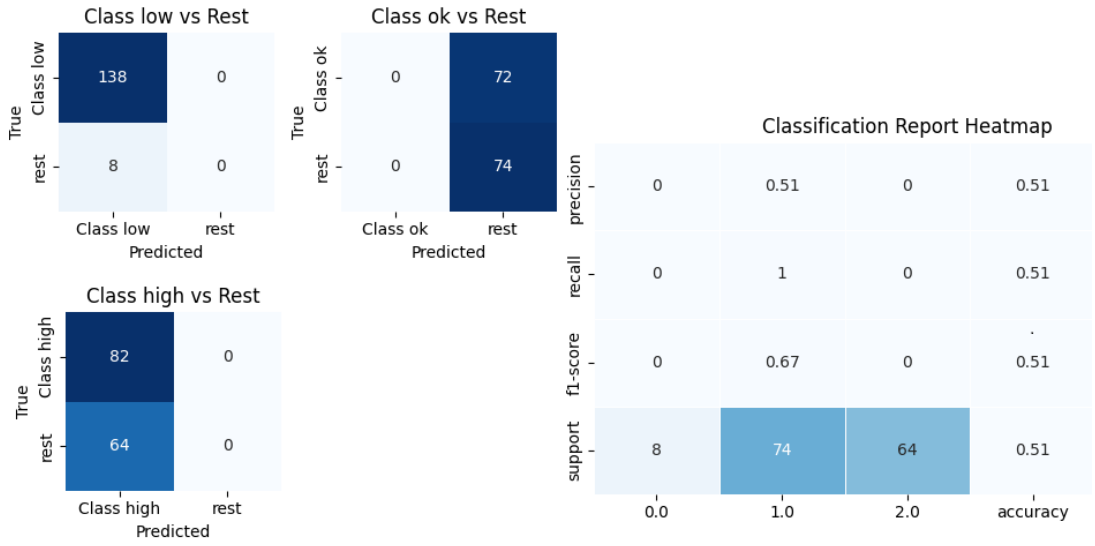


(a)

Figure 5.3: one-vs-rest confusion matrices and Classification Report HGB

LDL :

The optimal hyperparameters for the SVC model include a regularization parameter C, a kernel type, and a gamma value. Specifically, a C value of 0.1, a linear kernel, and a gamma value of 0.03 yielded the best performance based on the selected metric. The model achieved a score of 0.54, reflecting its performance under these conditions, and an accuracy of 0.51.

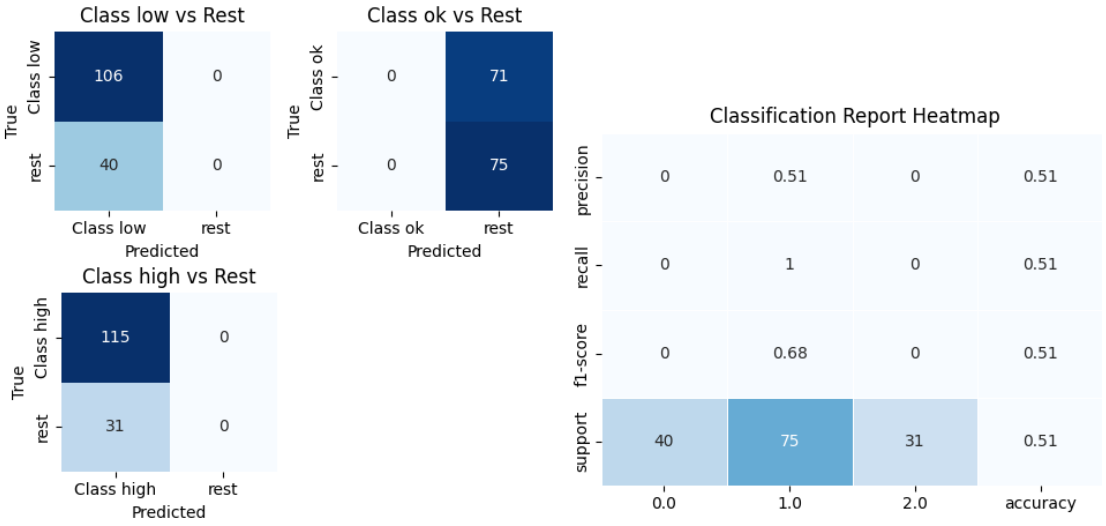


(a)

Figure 5.4: One-vs-rest confusion matrices and Classification Report LDL

HDL :

The optimal hyperparameters for the SVC model include a regularization parameter C, a kernel type, and a gamma value. Specifically, a C value of 0.1, a linear kernel, and a gamma value of 0.03 yielded the best performance based on the selected metric. The model achieved a score of 0.51, reflecting its performance under these conditions, and an accuracy of 0.51.



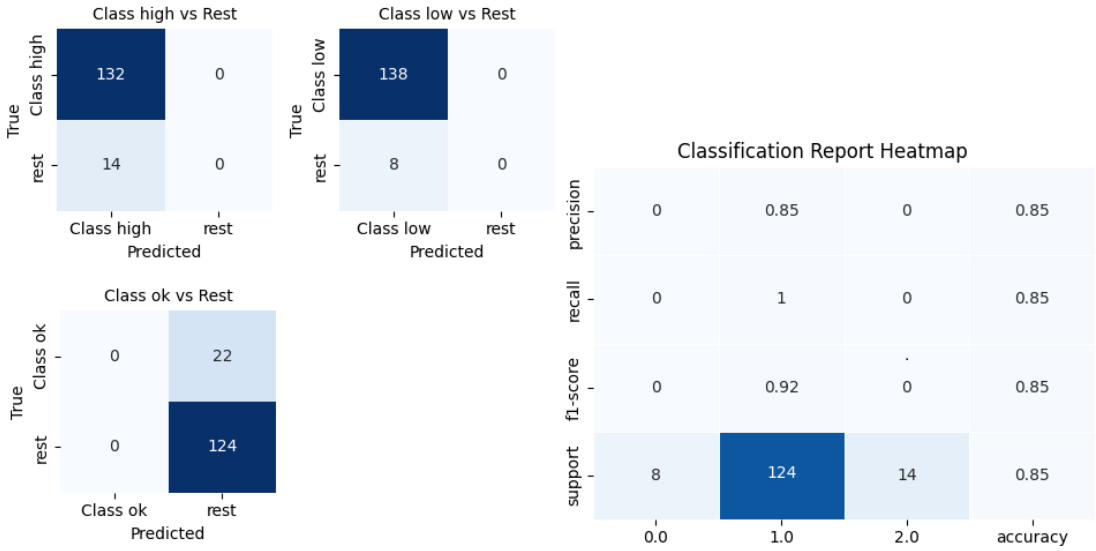
(a)

Figure 5.5: One-vs-rest confusion matrices and Classification Report HDL

5.4.2 K-Nearest Neighbors (KNN) :

HGB :

The optimal hyperparameters for the KNN model include the algorithm, the number of neighbors, and the weight function. Specifically, using the 'auto' algorithm, 10 neighbors, and 'uniform' weights yielded the best performance based on the selected metric. The model achieved a score of 0.85, indicating its performance under these conditions. Additionally, the accuracy of the model is 0.86.

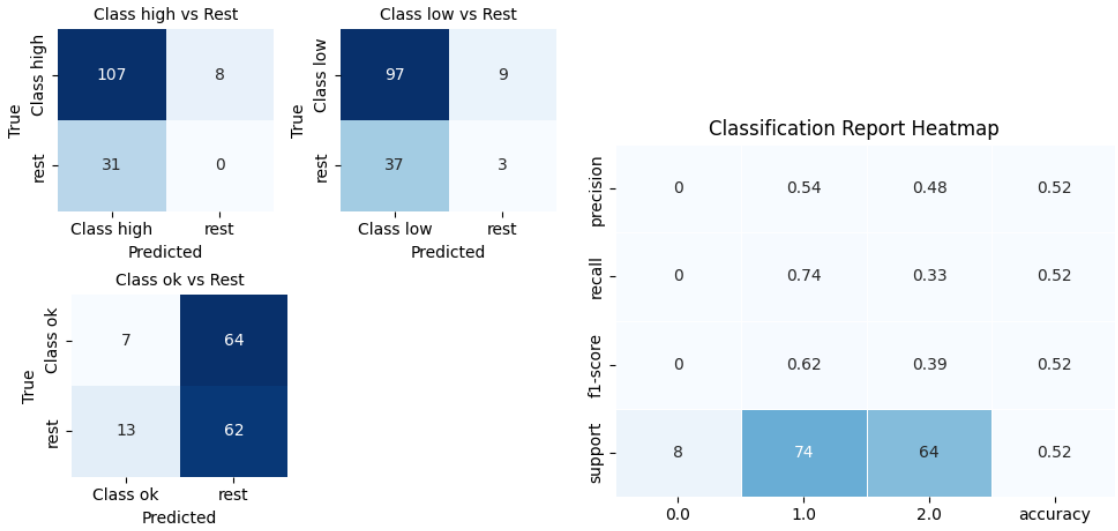


(a)

Figure 5.6: One-vs-rest confusion matrices and Classification Report HGB

LDL :

The optimal hyperparameters for the KNN model include the algorithm, the number of neighbors, and the weight function. Specifically, using the 'auto' algorithm, 14 neighbors, and 'uniform' weights yielded the best performance based on the selected metric. The model achieved a score of 0.54, indicating its performance under these conditions. Additionally, the accuracy of the model is 0.52.

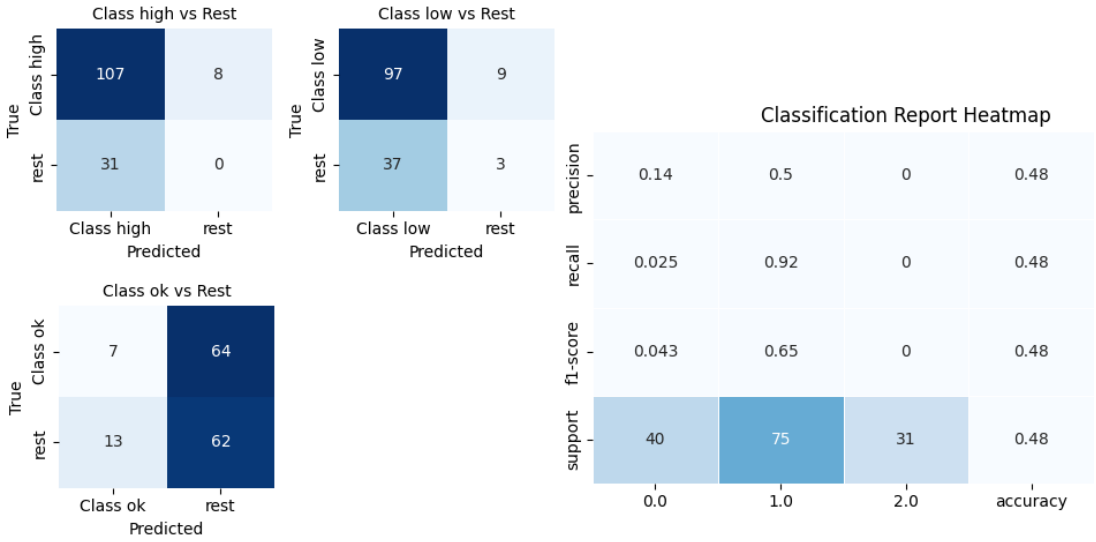


(a)

Figure 5.7: One-vs-rest confusion matrices and Classification Report LDL

HDL :

The optimal hyperparameters for the KNN model include the algorithm, the number of neighbors, and the weight function. Specifically, using the 'auto' algorithm, 14 neighbors, and 'uniform' weights yielded the best performance based on the selected metric. The model achieved a score of 0.49, indicating its performance under these conditions. Additionally, the accuracy of the model is 0.48.



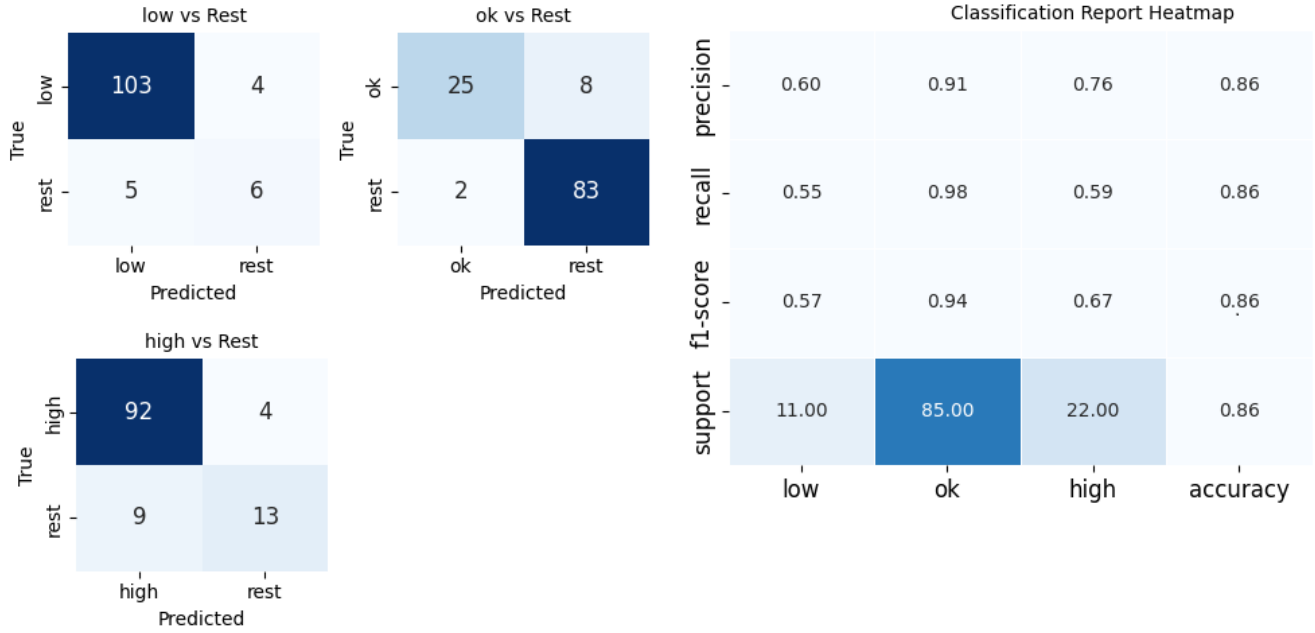
(a)

Figure 5.8: One-vs-rest confusion matrices and Classification Report HDL

5.4.3 XGBoost :

HGB :

The optimal hyperparameters for the model include a learning rate of 0.2, a maximum depth of 6, and 68 estimators, which were found to yield the best performance based on the selected metric. The model achieved a score of 0.85, indicating its performance under these conditions. Additionally, the accuracy of the model is 0.86.

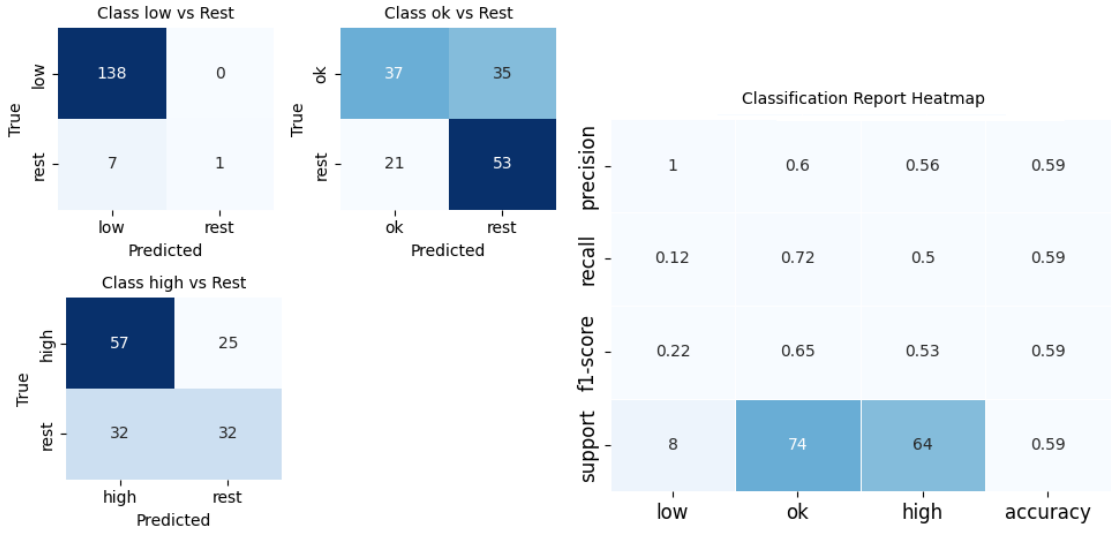


(a)

Figure 5.9: One-vs-rest confusion matrices and Classification Report HGB

LDL :

The optimal hyperparameters for the model include a learning rate of 0.2, a maximum depth of 3, and 68 estimators, which were found to yield the best performance based on the selected metric. The model achieved a score of 0.55, indicating its performance under these conditions. Additionally, the accuracy of the model is 0.59.

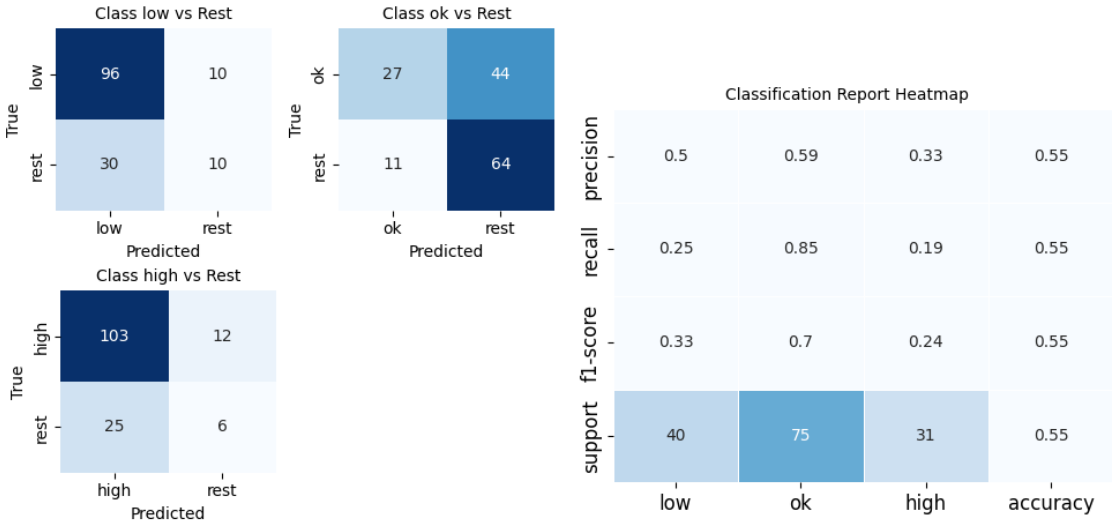


(a)

Figure 5.10: One-vs-rest confusion matrices and Classification Report LDL

HDL :

The optimal hyperparameters for the model include a learning rate of 0.2 , a maximum depth of 3, and 50 estimators, which were found to yield the best performance based on the selected metric. The model achieved a score of 0.44, indicating its performance under these conditions. Additionally, the accuracy of the model is 0.54.



(a)

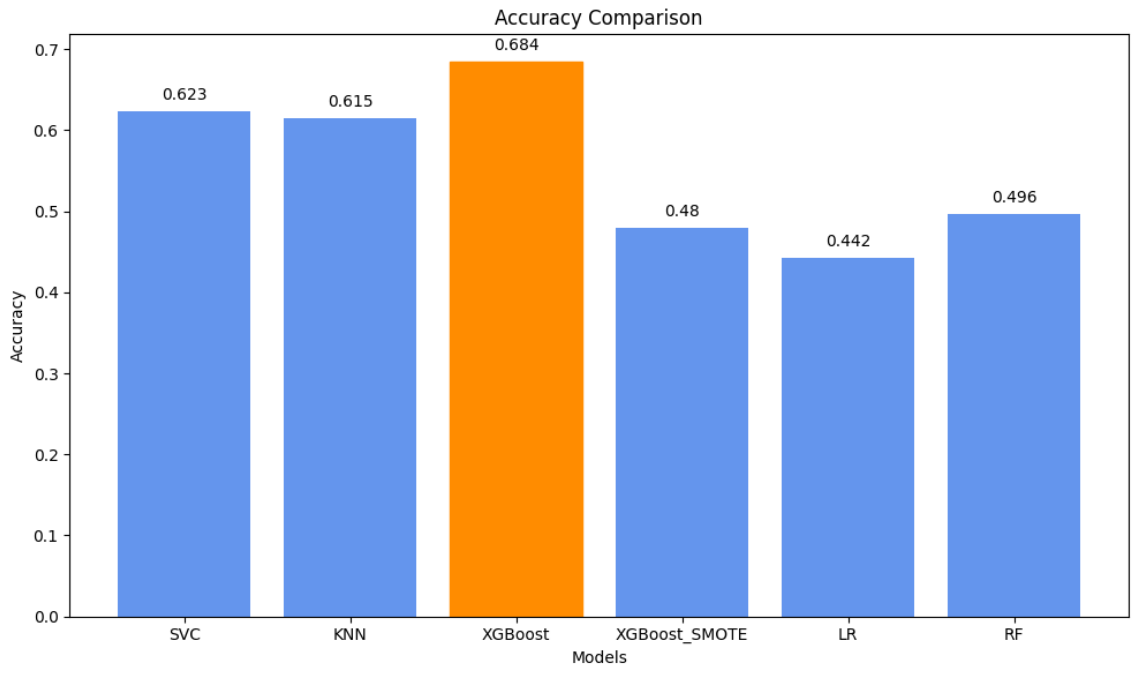
Figure 5.11: One-vs-rest confusion matrices and Classification Report HDL

5.4.4 Results Comparison

Let's compare the performance of the first proposed approach and other models on the testing data based on their accuracy. The results are presented in Table 5.1 and illustrated in Figure 5.12 .

Approach	**Accuracy**
SVC	0.623
XGBoost	0.684
KNN	0.615
XGBoost_Smote	0.48
voting classifier(SVC, Knn, XGBoost)	0.659
LR	0.442
RF	0.496

Table 5.1: Accuracy comparison .



(a)

Figure 5.12: A bar chart with the accuracy values for each model.

5.5 Evaluation of Second Proposed Approach

In this approach, I used a voting classifier combining three algorithms (SVC, KNN,GPC) to determine the best performance for predicting the target variable (LDL). Before implementing the voting classifier, it is essential to identify the optimal hyperparameters for each algorithm.

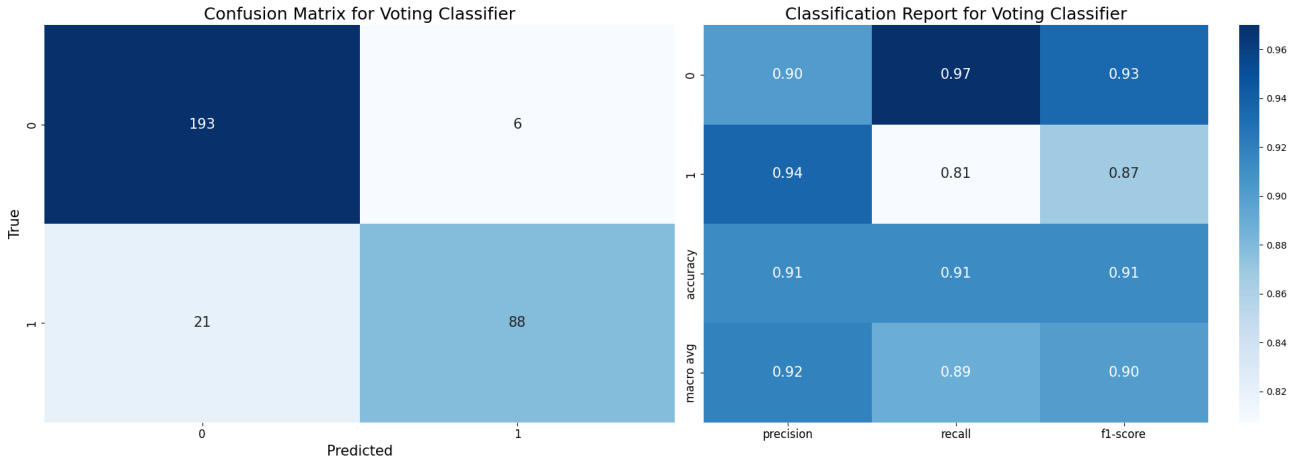
5.5.1 Hyperparameters :

Hyperparameters for the algorithms are set as follows:

- For **SVC**, 'C' is 0.2, 'gamma' is 0.01, and the kernel is 'rbf'.
- For **KNN**, the number of neighbors ('n_neighbors') is 23.
- For **GPC**, the maximum number of iterations for prediction ('max_iter_predict') is 100.

5.5.2 Evaluation :

The accuracy achieved with the second proposed approach is **0.91**.



(a)

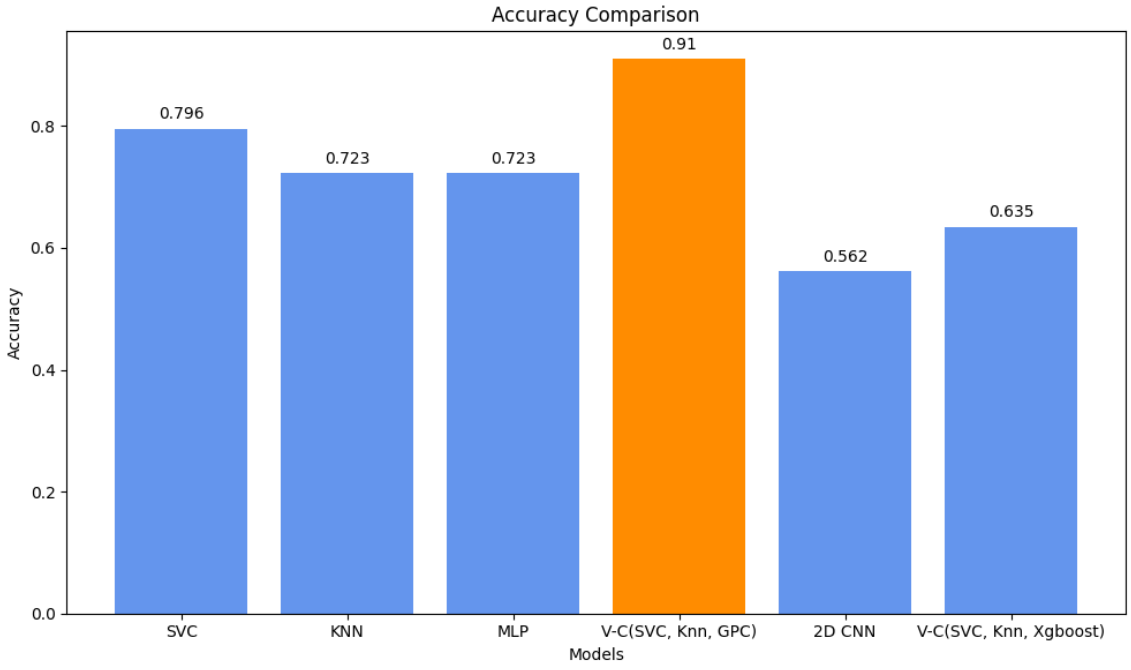
Figure 5.13: Confusion Matrix and Classification Report for Second Approach

5.5.3 Results Comparison

Let's compare the performance of the Second proposed approach and other models on the testing data based on their accuracy. The results are presented in Table 5.2 and illustrated in Figure 5.14 .

Approach	**Accuracy**
voting classifier(SVC, Knn, GPC)	0.91
knn	0.723
mlp	0.723
svc	0.796
2D CNN	0.562
voting classifier(SVC, Knn, Xgboost)	0.635

Table 5.2: Accuracy



(a)

Figure 5.14: A bar chart with the accuracy values for each model.

5.6 Conclusion

This chapter presented and evaluated two proposed approaches for machine learning models. The first approach explored individual models, including SVC, KNN, and XGBoost, revealing their respective strengths and weaknesses. The second approach utilized an ensemble method, combining the strengths of the individual models into a Voting Classifier, which demonstrated improved overall accuracy. The comprehensive evaluations and comparisons validated the effectiveness of the ensemble model, highlighting

the benefits of integrating diverse models for enhanced predictive performance.

General conclusion and perspectives

This thesis highlights the significant roles of hemoglobin (Hgb) and cholesterol in the human body, their implications on health, and the necessity of accurate blood analysis for their monitoring. Cholesterol, with its dual nature comprising LDL (bad cholesterol) and HDL (good cholesterol), plays critical roles in cellular function but poses risks of cardiovascular diseases when imbalanced. Similarly, hemoglobin is essential for oxygen transport, with deviations in its levels indicating various health conditions, from anemia to polycythemia.

Traditional blood analysis methods, while effective, present limitations in terms of accuracy, efficiency, and accessibility. Near-Infrared (NIR) Spectroscopy emerges as a promising alternative, offering non-invasive, rapid, and precise measurements. This technology, rooted in the principles of electromagnetic radiation and its interaction with matter, shows potential in improving diagnostic accuracy and patient outcomes.

The integration of Machine Learning (ML) further enhances the analytical capabilities of NIR Spectroscopy. By leveraging algorithms to process and interpret complex data patterns, ML facilitates the development of robust diagnostic tools. Techniques like Support Vector Classification (SVC), K-Nearest Neighbors (KNN), and XGBoost have demonstrated efficacy in refining blood analysis, providing more accurate and reliable results.

In summary, the convergence of NIR Spectroscopy and Machine Learning represents a significant step forward in the field of blood analysis. By addressing the limitations of conventional methods, these technologies promise to enhance diagnostic precision, improve patient care, and advance our understanding of key biomarkers like hemoglobin and cholesterol.

Despite the advancements, challenges such as data quality, instrument calibration, and the need for standardized methodologies persist. Continued research and development in these areas are essential to fully realize the benefits of NIR Spectroscopy and ML in medical diagnostics.

Bibliography

[1] Unknown. Diagram showing plaque build-up within an artery from top to bottom, 2024. Accessed: 2024-06-30.

[2] World Health Organization. Raised cholesterol, n.d. Accessed: 2024-06-17.

[3] Le cholestérol : tout savoir sur le cholestérol. Accessed: May 14, 2024.

[4] Centers for Disease Control and Prevention. Ldl and hdl cholesterol and triglycerides retrieved 14 may 2024.

[5] Fondation Recherche Cardio-Vasculaire. Le cholestérol: Le définir pour le comprendre.

[6] Cleveland Clinic. Cholesterol, high cholesterol, diseases, 2024.

[7] Mayo Clinic. High blood cholesterol: Symptoms and causes, 2022.

[8] H. H. Billett. Hemoglobin and hematocrit. In *Clinical Methods: The History, Physical, and Laboratory Examinations*. 3rd edition.

[9] Hémoglobine : Norme, taux bas, élevé chez la femme et l'homme.

[10] Verywell Health. Importance of hemoglobin, 2024. Accessed: 2024-06-10.

[11] MedlinePlus. Hemoglobin electrophoresis, 2016. Accessed: 2024-06-10.

[12] Healthline. What is hemoglobin.

[13] American Heart Association. Hdl (good), ldl (bad) cholesterol and triglycerides, February 2024. Last reviewed: February 19, 2024.

[14] Steven Chu, George Samuel Hurst, Jack D. Graybeal, and John Oliver Stoner. Spectroscopy. *Encyclopedia Britannica*, 2024.

[15] J. R. Murphy, S. R. Pasco, J. Ackerman, V. Alvarado, W. D. Rice, S. Kattel, and W. Scougale. Spectroscopic determination of ice-induced interfacial strain on single-layer graphene. *Small*, 16(42):2003892, 2020.

[16] Soufiane Boudjemaa. *Méthodes spectroscopiques d'analyse, l3 "chimie analytique"*. Publisher, 2020-2021.

[17] H. Bichsel. The interaction of radiation with matter: Datasheet from landolt-börnstein - group i elementary particles, nuclei and atoms · volume 21b1: "detectors for particles and radiation. part 1: Principles and methods" in springermaterials (https://doi.org/10.1007/978-3-642-03606-4_2). Copyright 2011 Springer-Verlag Berlin Heidelberg.

[18] Spectroscopic analysis methods, 2021. [UOMustansiriyah University].

[19] Jessica Maloney. 5 different types of spectroscopy.

[20] Types of spectroscopy. https://microbenotes.com/types-of-spectroscopy/, 2022. Accessed: March 22, 2022.

[21] Sagar Aryal. Infrared (ir) spectroscopy, 2022.

[22] IM Publications Open. An introduction to near infrared (nir) spectroscopy. www.impopen.com. Retrieved 2022-06-01.

[23] Unknown. Measuring color, 2024. Accessed: 2024-06-30.

[24] F. Quinlan, G. Ycas, S. Osterman, and S. A. Diddams. A 12.5 ghz-spaced optical frequency comb spanning ¿400 nm for near-infrared astronomical spectrograph calibration. *Review of Scientific Instruments*, 81(6):063105, June 2010.

[25] Warwick B. Dunn, Nigel J. C. Bailey, and Helen E. Johnson. Measuring the metabolome: current analytical technologies. *Analyst*, 130(5):606–625, 2005.

[26] BP Van der Sanden, A Heerschap, L Hoofd, AW Simonetti, et al. Effect of carbogen breathing on the physiological profile of human glioma xenografts. *Magn Reson Med*, 42(3):490–499, 1999.

[27] LibreTexts contributors. Combination bands, overtones, and fermi resonances.

[28] Ko Nee, Samuel A. Bryan, Tatiana G. Levitskaia, Jennifer W.-J. Kuo, and Mikael Nilsson. Combinations of nir, raman spectroscopy and physicochemical measurements for improved monitoring of solvent extraction processes using hierarchical multivariate analysis models. *Analytica Chimica Acta*, 1006:10–21, 2018.

[29] Krzysztof B. Beć. A simple guide to complex world of overtone and combination bands: Theoretical simulation and interpretation of nir spectra – summary of the workshop at nir-2021 beijing conference. *NIR News*, 32(7-8):15–18, 2021.

[30] Unknown. Application of vibrational spectroscopy in organic electronics, 2024. Accessed: 2024-06-30.

[31] Dara K. MD Lee Lewis. Ldl cholesterol: How low can you (safely) go? *Harvard Health Blog*, 2020.

[32] Anupama Sapkota. 22 types of spectroscopy with definition, principle, steps, uses. *Publication Date: August 12, 2022.*

[33]

[34] V. P. Gupta. *Molecular and Laser Spectroscopy*. ScienceDirect, 2017. Retrieved 2022-06-03.

[35] Marco Ferrari and Valentina Quaresima. A brief review on the history of human functional near-infrared spectroscopy (fnirs) development and fields of application. *NeuroImage*, (2):921–935, 2012.

[36] Master Organic Chemistry. Interpreting ir spectra: A quick guide.

[37] Unknown. Muutakin kuin klooria – mihin uv-valolaitetta tarvitaan?, 2024. Accessed: 2024-06-30.

[38] Oracle. What is machine learning?, 2023. Accessed on 15th May 2024.

[39] Khadija El Bouchefry and Rafael S. de Souza. Learning in big data: Introduction to machine learning. In *Knowledge Discovery in Big Data from Astronomy and Earth Observation*, pages 225–249. 2020.

[40] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc, 2nd edition, 2019.

[41] Gareth James, Daniela Witten, Robert Tibshirani, and Trevor Hastie. *An Introduction to Statistical Learning: with applications in R*. Springer New York, NY, 2013.

[42] Unknown. Unsupervised learning and data clustering, 2022. Accessed on 15th May 2024.

[43] Unknown. Unsupervised machine learning, 2022. Accessed on 15th May 2024.

[44] L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4, May 1996.

[45] Pascal Vincent. *Modèles à noyaux à structure locale*. PhD thesis, Université de Montréal, October 2003.

[46] S. Prabhu and N. Venkatesan. *Data Mining And Warehousing*. New Age International (P) Ltd., Publishers, September 2010.

[47] Bernard Fertil. Reconnaissance des formes: Classement d'ensembles d'objets, 2006.

[48] Julien Iguchi-Cartigny. *Scénarios d'attaques et détection d'intrusions*. 2013.

[49] F. Moutarde. Brève introduction aux arbres de décision, 2008.

[50] Gilbert Ritschard, Simon Marcellin, and Djamel A. Zighed. Arbre de décision pour données déséquilibrées : sur la complémentarité de l'intensité d'implication et de l'entropie décentrée. 2007.

[51] Neha Gupta. Artificial neural network. In *International Conference on Recent Trends in Applied Sciences with Engineering Applications*, volume 2, 2013.

[52] E-G Talbi. *Fouille de données (Data Mining) - Un tour d'horizon -*. PhD thesis, Laboratoire d'informatique fondamentale de Lille, 2022.

[53] Forest Dominic. *Application de techniques de forage de textes de nature prédictive et exploratoire à des fins de gestion et d'analyse thématique de documents textuels non structurés.* PhD thesis, Université du Québec à Montréal, 2006.

[54] Unknown. A complete guide to understand classification in machine learning, 2021. Accessed on 15th May 2024.

[55] Multi-label classification, 2021. Accessed on 15th May 2024.