Detection of Formalin In Milk Using Machine Learning

A Dissertation for

Course Code and Course Title: ELE-625 & Project

Credits: 16

Submitted in Partial Fulfillment of Masters Degree

M. Sc. In Electronics

By

Ms. Delinka Genoveva Rosa

Roll Number: 22P0360005

ABC ID:- 887083941881

Under the supervision of

Prof. Jivan S. Parab

School of Physical and Applied Sciences
Electronics

Goa University

Date:

Examined by :

Seal of the School

DECLARATION BY STUDENT

I hereby declare that the data presented in this Dissertation report entitled, **Detection of Formalin In Milk Using Machine Learning** is based on the results of investigations carried out by me in the M.Sc Electronics at the School of Physical and Applied Sciences, Goa University under the Supervision of Prof. Jivan Parab and the same has not been submitted elsewhere for the award of a degree or diploma by me. Further, I understand that Goa University or its authorities will be not responsible for the correctness of observations / experimental or other findings given the dissertation.

I hereby authorize the University authorities to upload this dissertation on the dissertation repository or anywhere else as the UGC regulations demand and make it available to any one as needed.

Name of Students: Delinka Genoveva Rosa

Signature: 

Seat no: 22P0360005

Date:

Place: Goa University

COMPLETION CERTIFICATE

This is to certify that the dissertation report "Detection of Formalin In Milk Using Spectroscopy" is a bonafide work carried out by Ms. Delinka Genoveva Rosa under my supervision/Membership in partial fulfillment of the requirements for the award of the degree of M.Sc in Electronics Discipline at School Of Physical and Applied sciences, Goa University.

Prof. Jivan S. Parab

Electronics Disciplines

School of Physical and Applied Sciences

Date:

Prof. Ramesh V. Pai

School of Physical and Applied Sciences

School Stamp

Date:

Place: Goa University

PREFACE

This dissertation represents the culmination of several months of research and hard work. It is with great pleasure that I present it to the academic community.

Throughout this journey, I have been fortunate to receive support and guidance from many individuals and the institution, to whom I owe immense gratitude.

The dissertation is organized as follows: Chapter 1 provides an introduction to the topic, including its significance and relevance. Chapter 2 reviews the existing literature on Food adulteration and it's detection using various methods, synthesizing key findings and identifying gaps in the literature. Chapter 3 outlines the methodology employed in this study, detailing the research design, data collection methods, and analytical approach. Chapter 4 explains the machine learning algorithms used on the datasets. Finally, Chapter 5 presents the results with empirical findings and analysis with various ML algorithms.

ACKNOWLEDGEMENT

I would like to express my sincerest thanks to everyone who has contributed to this project First and foremost, I would like to thank all the students who have voluntarily participated in this study and have provided me with the milk from their villages giving me their time to help me collect the data.

I am grateful to Prof. Jivan S. Parab my guide for his constant support, assistance, and guidance at every stage, which has been instrumental in the success of this project.

I express my deep gratitude towards all the teaching faculty Prof. Rajendra . S. Gad, Dr. Narayan Vetrekar, Dr. Marlon Sequeira, Dr. Aniketh Gaonkar and Dr. Sandeep Gawali.

I would also like to thank Mr. M.G. Lanjewar (Technical Officer-I) for his assistance and provision for necessary help throughout the completion of the project.

A sincere thank you to Mrs. Ashwini, Sir. Lopes, Miss Krishna, Sir Vishant, Miss Marissa and Sir Ramchandra for their help and support.

Their knowledge and suggestions have been crucial in helping me carry out,

And execute the project.

I would like to express my gratitude to my family and close friends for their unfailing support and inspiration during this Project.

TABLE OF CONTENTS

TITLE OF FIGURES

TITLE OF TABLES

ABBREVIATIONS USED

SVR	Support Vector Regression
KNR	K Nearest Regression
PLSR	Partial Least Square Regression
PCA	Principal Component Analysis
SG	Savitzky Golay
RFR	Random Forest Regression
DTR	Decision Tree Regression
ML	Machine Learning
NIR	Near Infrared Spectroscopy
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
MSE	Mean squared Error

"DETECTION OF FORMALIN IN MILK USING MACHINE LEARNING"

Delinka Genoveva Rosa
Electronics Discipline , SPAS
Goa University

Abstract

Food adulteration, commonly referred to as 'food fraud,' is a serious issue worldwide that can have dire consequences on the health and well-being of consumers.Common types of adulterants added to food include water, starch, urea, detergents, and chemical preservatives. Adulteration can lead to severe health impacts such as digestive issues, toxicity, and long-term diseases like cancer. To enhance the shelf life of milk during long-distance transportation, formalin is often employed as an adulterant in milk. However, formalin is a highly toxic substance that can lead to severe damage to the liver and kidneys if consumed. In this research A non-destructive system based on machine learning (ML) is developed to detect formalin adulteration in milk using near-infrared (NIR) spectroscopy. A comprehensive database was produced by combining variable quantities of formalin in two different types of milk (ranging from 0% to 50%) and recording spectra using a Jasco V770 spectrophotometer from 400nm to 1400nm. The spectral data obtained was pre-processed using the Savitzky Golay filter (SG- Filter) and Principial Component Analysis (PCA). To predict the formalin concentration in milk, five different ML models were used for regression analysis. Among the regression models, KNN outperformed for buffalo Milk, jersey cow milk and for the combination of both the milk having excellent values for R^2, RMSE, MAE as follows. For Buffalo milk an R^2

of 0.999, RMSE of 0.28 ml (% v/v), MAE of 0.08 ml (% v/v) and average RMSE 1.08. For Jersey Cow Milk an R^2 value of 0.999, RMSE of 0.13 ml (% v/v), MAE of 0.03 ml (% v/v) and average RMSE 0.48. For the combination of both the milks an R^2 value of 0.999, RMSE of 0.5 ml (% v/v), MAE of 0.18 ml (% v/v) and average RMSE 1.06.

C H A P T E R 1

Introduction

Food is a fundamental source of nourishment for all living beings, serving as a means of sustenance and growth. The recommended daily intake of dairy-based food products for adults is 2-3 portions, as per the food guide pyramid. The quality of milk is essential in ensuring the production of high-quality dairy products and promoting consumer health. However, milk adulteration is a significant concern in developing countries, as it is the most frequently adulterated food commodity. The dairy industry suffers major economic losses due to this practice, while consumers are exposed to health risks. Overpopulation, rapid urbanization, and scattered settlements are some of the contributing factors to the growing demand for milk production. In an attempt to meet this demand, milk dealers often resort to adulteration. "Food fraud, or Food Adulteration" which involves the deliberate compromise of food quality and safety, is a serious public health threat. This unethical practice is a growing concern worldwide as it not only defrauds consumers but also poses severe health implications. Adulteration is often carried out to increase profits or due to a lack of proper detection technology and confusion about appropriate drug administration practices. Therefore, it is crucial to raise awareness about this issue and take appropriate measures to prevent and deter food adulteration. Milk is considered the "ideal food" due to the abundance of nutrients required by both infants and adults, including protein, fat, carbohydrate, vitamins, and minerals (Moore et al.,2014; Azad and Ahmed, 2016). Cow milk includes 3.7% fat, 4.8% lactose, 12.7% total solids, 3.4% protein, 0.7% ash, 86% water, casein 2.8%, it varies from breed to breed and species to species. Today, India is the largest producer of milk in the world, contributing 23% of global milk production (Milk Production in India, 2022). From 2014-15 to 2022-23,

increased by 58%, reaching 230.58 million tonnes in 2022-23. (Invest India, n.d., 2023).

1.1 Background

Milk is undoubtedly a ubiquitous food in the human diet. This is the first food of mammals and, as such, provides all the energy and nutrients needed for proper growth and development. The nutritional richness of milk is indisputable; it is a good source of protein with high biological value and polyvalent roles in immune function as well as nutrient transport and absorption, and of course, it contains essential vitamins and minerals (Pereira, 2014). However, the quality of milk can vary significantly depending on various factors such as the health of the animal, the method of milking, and storage conditions, among others.

1.1.1 Types of Adulterants in Milk

Various substances are used as adulterants in milk, each with its own implications.

Water

The most common method of adulterating milk is by adding water to increase its volume, which can lead to a reduction in the nutritional value of the product (Francis et al., 2020). However, this practice poses a serious health risk to consumers since contaminated water can contaminate milk and cause various diseases such as diarrhoea, typhoid, rotavirus, and hepatitis A and E (Bhuiyan and Noor, 2020). Consumers are particularly concerned about the safety of the water used in milk

production, as it may contain harmful contaminants such as pesticides and heavy metals that can pose a risk to their health.

Urea

Milk is often adulterated with urea to improve its brightness, fluidity, and non-protein nitrogen content, as well as to balance the levels of solids not fat in the correct proportion. Urea is also used to generate artificial milk. However, even a small amount of urea can cause health problems such as vomiting, nausea, gastritis, ulcers, and even cancer. Urea is particularly harmful to the heart, liver, and kidneys, as the kidneys have to work harder to eliminate it from the body. The presence of ammonia in milk can result in impairment, loss of acquired speech, and visual impairment. Inconsistent cow nutrition can also cause the urea content of milk to increase, leading to productivity issues in dairy cows. Urea is also used for heat consistency, highlighting the importance of detecting urea in milk (Kandpal et al., 2016).

Formalin

Formalin is a type of disinfectant that is widely use to enhance the storage life of liquid milk, during transportation. However, applying any form of preservative to milk is illegal. Formalin is primarily used to conserve biological specimens, and its use can help save money on refrigeration and electricity. Formalin can induce carcinogen agents. It can also induce gut corrosion, which can lead to ulcers and intestinal inflammatory illnesses, all of which can lead to renal failure (Chemical Fact Sheets: Formaldehyde, 2022).

Detergents

Cleansers are often used to dissolve and disperse water in oil, resulting in a foaming mixture with a milk-like white color. The cosmetic nature of milk can be improved by detergents, which are primarily used to make it thicker and more viscous. However, detergents can cause gastrointestinal problems and are hazardous to human health. Detergents such as dioxane, sodium lauryl sulfate, and phosphates are commonly used, but they can have harmful effects. Dioxane is a carcinogen, while sodium lauryl sulfate can cause conjunctivitis, liver damage, cytotoxicity, endocrine disruption, mutation, and cancer. Phosphates can cause symptoms such as nausea, diarrhea, and skin irritation (Hemanth Singuluri and Sukumaran, 2014).

Hydrogen Peroxide

Hydrogen peroxide is often used to keep milk fresh for a longer period of time, but it can harm the cells in the gastrointestinal tract, leading to cancer, ulcers, and intestinal inflammation. It disrupts the antioxidants in the body, which can accelerate the aging process. Hydrogen peroxide is an oxidizing and bleaching agent that is colorless and odorless. It is commonly used in deodorants, water and sewage treatment, and the production of other compounds. Like formalin, hydrogen peroxide can extend the storage life of milk and inhibit bacterial growth. However, milk contaminated with hydrogen peroxide has been linked to an increase in heart rate and the development of cardiac arrhythmia (Lindmark-Månsson and Åkesson, 2000).

Synthetic Milk

Synthetic milk is not real milk, but rather a heavily adulterated product that is designed to increase profits by enhancing the quantity of liquid. It typically contains a

combination of liquid cleanser or soap, caustic soda, vegetable fats, sodium, and ammonia, among other ingredients. Although it looks and tastes like milk, synthetic milk lacks the nutrients found in real milk. The technology for producing synthetic milk was developed by milkmen in Kurukshetra around 15 years ago and has since spread to other countries. According to some estimates, up to 1.10 crore liters of synthetic milk are produced and sold each day in various states across the country. Synthetic milk can cause significant harm to the human body, including eye inflammation, liver and kidney complications. It is especially dangerous for pregnant women and anyone with high blood pressure. Urea and sodium carbonate, two common ingredients in synthetic milk, are extremely toxic to the heart, liver, and kidneys, and can turn the human body into a breeding ground for disease (Francis et al., 2020)..

Chlorine

In order to counteract the viscosity of diluted milk, chlorine is often utilized as an additive subsequent to its dilution (Reddy et al., 2017). However, oxygenated milk can pose potential health risks, such as the development of cardiac conditions and blockage of arteries (van der et al., 2021). The incl usion of chloride in milk can disrupt the alkaline ratio and pH of the blood, leading to potential health implications.

Melamine

Melamine is utilized as an artificial means to increase the protein concentration of milk powder. However, under severe circumstances, its intake can lead to renal failure and even fatality (Cheng et al., 2010). Melamine is a compound consisting of cyanamide and 1,3,5-triazine, which typically occurs in the form of crystal shards in nitrogen. It is commonly employed in the production of amino polymers and plastic

materials, textiles, nitrogenous pesticides, and other products that are only slightly miscible in water. Despite its industrial applications, melamine has the potential to cause harm to one's health. Clinical trials have revealed that melamine ingestion alone can lead to urinary blockages and the formation of kidney stones. This can cause a disruption in the effective functioning of kidneys, leading to renal failure.

Whey

Whey, a cost-effective by-product of cheese production, is added to liquid milk to increase volume as well as protein content (de Carvalho et al., 2015).

Oil

Milk is composed of various components, with fat being a major constituent, typically accounting for 3-5% (m/m) of cow's milk. Triacylglycerols are the primary type of fat found in milk, comprising around 97-98% of its fat content, and are responsible for the characteristic flavor and texture of milk products. Fat is a target of adulteration due to its value in the production of milk derivatives and its ability to compensate for fraudulent dilution. Vegetable oils, such as soybean, sunflower, groundnut, coconut, palm, and peanut oil, as well as animal fat, are the primary adulterants used for this purpose (Rani et al., 2015).

Some more adulterants that are added in milk are listed below:

To decrease microbial growth and increase the shelf life of the product. Such substances include hypochlorite, salicylic acid, and even potassium dichromate. However, these substances are toxic to humans, and hence it is crucial to monitor their use for quality control purposes. Ammonium Sulphate is a chemical fertilizer, which is added to milk to raise the density of watered milk and increases the lactometer reading by maintaining the density of milk. The serious health risk of these adulterants are gastrointestinal complaints, liver and kidney damage. Starch is adulterated in milk to increase the solid content. Caustic soda is added to the blended mixture of chemical and natural milk to neutralise the effect of increased acidity, thereby preventing it from turning sour during transport. Benzoic acid and Salicylic acid are added to milk to increase the shelf life of milk. Maltodextrin is a common additive used in milk. It increases the volume of milk and milk products. Consumption of milk adulterated with Maltodextrin may cause Allergy and diarrhea. Melamine is added into the milk to increase the protein count falsely in milk and dairy products. Melamine is described as being harmful if swallowed, inhaled or absorbed through the skin. Many food colorants are added to improve the appearance of milk and have hazardous effects on health.Synthetic milk has bitter after taste, gives a soapy feeling on rubbing between the fingers and turns yellowish on heating Synthetic milk is made by adding while colour water. chalk powder is added to the milk to increase its apparent volume, or to make it appear whiter. (Nascimento et al.,2017; Raturi et al., 2022). The table 1.1 provides a list of adulterants added to milk, along with the purpose of the adulteration, their impact on health, and the absorption maxima in nanometers.

Table 1.1 List of adulterants with milk along with their Spectral Signatures

Adulterants in Milk	***Used***	***Effects on Health***	***Absorption Maxima (nm)***
Water	to increase the quantity	water-borne illnesses.	Broad band around 1450 nm
Urea	show higher protein content in milk	harmful to the kidney and gastrointestinal system.	Peaks at 232 and 280 nm
Formalin	Preservation	Damage effect on the liver and kidneys.	Broad band around 1100 nm
sodium carbonate	emulsify and dissolve the oil in water giving a frothy solution, which is the desired characteristic of milk	irritation to your eyes, skin, mouth, and lungs.	Peaks at 220 and 260 nm
Starch	increase the solid content.	blockage in the intestines and stomach pain.	Broad band around 1020 nm
ammonium sulphate	raises the density of watered milk and increase the lactometer	gastrointestinal complaints, and liver and kidney damage.	Peaks around 205 and 230 nm
melamine	increase the protein	swallowed, inhaled or absorbed through the skin irritating to skin and eye mucous membranes.	Peaks at 232 and 280 nm

1.1.2 Detection Methods

Detecting formalin adulteration in milk is a critical task for regulatory authorities and consumers alike. In most cases of contamination, different types of analysis methods

are used to give the analyst flexibility in selecting the appropriate method. This enables the analyst to accurately detect and identify the type of adulteration or contamination present in the food product. It is essential to use these testing methods to ensure that food products are safe and free from any harmful contaminants, thereby ensuring the health and safety of consumers. Various techniques, especially in the fields of electronics and spectroscopy, are employed to detect adulterants in milk. Here's an overview of some of these techniques:

Electronics-Based Techniques

Electronics-based techniques use various sensors and devices to detect changes in electrical properties, chemical composition, or sensor responses that indicate adulteration. Here are the most prominent methods:

- **Electronic Nose (E-nose) and Electronic Tongue (E-tongue)**

These devices are designed to mimic human olfactory and gustatory senses, respectively, and use arrays of sensors to detect adulteration in milk. E-nose detects volatile organic compounds (VOCs) by analyzing their unique patterns. A trained electronic nose can identify the presence of foreign substances in milk based on changes in these patterns. E-tongue detects non-volatile compounds in a similar fashion, using sensor arrays to recognize alterations in the chemical composition of milk. It is particularly useful in detecting flavor adulterants, like sweeteners or bittering agents (Cristian Olguín et al., 2014; Zhang, L., et al., 2014)

- **Electrical Conductivity**

This technique measures the electrical conductivity of milk to detect the addition of substances like water, salt, or urea. Since each compound has a distinct conductivity profile, any deviation from normal levels can indicate adulteration.

- **Potentiometric Sensors**

These sensors measure changes in voltage due to specific ion concentrations. They are useful in detecting substances like ammonium compounds and formaldehyde, which are often used in milk adulteration.

Spectroscopy-Based Techniques

Spectroscopy-based techniques analyze the interaction between light and matter to detect adulterants. These methods are widely used due to their precision and sensitivity.

- **Near-Infrared Spectroscopy (NIRS)**

NIRS involves the absorption of near-infrared light by milk samples. Each compound absorbs light at specific wavelengths, creating unique spectral signatures. This technique is effective in detecting adulterants like water, urea, and melamine. It is non-destructive and allows for rapid analysis, making it ideal for large-scale testing. (Pierna et al., 2012; Salgó and Gergely, 2012; Albanell et al., 2012)

- **Fourier Transform Infrared Spectroscopy (FTIR)**

FTIR uses infrared light to obtain molecular fingerprints of milk samples. It is particularly useful for identifying complex adulterants and can be combined with

chemometric analysis to increase accuracy. FTIR is capable of detecting a wide range of adulterants, including detergents and other chemicals used to dilute milk (Jawaid et al., 2013).

- **Raman Spectroscopy**

Raman spectroscopy analyzes the scattering of monochromatic light as it interacts with molecular vibrations. This technique can detect adulterants based on their unique Raman spectra. It is sensitive to changes in composition and structure, enabling real-time analysis. Raman spectroscopy is especially useful in detecting protein-based adulterants and melamine (Khan Mohammad Khan et al., 2015)

- **Ultraviolet-Visible (UV-Vis) Spectroscopy**

UV-Vis spectroscopy measures the absorption of ultraviolet and visible light by milk samples. It is helpful for detecting adulterants that affect the optical properties of milk, such as dyes or colorants (Agharkar and Mane, 2021).

Chemical and Biochemical Techniques

These methods use chemical reactions or biochemical processes to detect adulterants in milk.

- **pH and Acidity Tests**

These tests measure the pH level or acidity of milk to detect substances like detergents or certain chemicals. A significant deviation from normal pH levels indicates adulteration.

- **Chromatography**

Chromatography techniques, such as gas chromatography (GC) and high-performance liquid chromatography (HPLC), separate and analyze compounds in milk. They are effective in detecting a wide range of adulterants, including melamine, antibiotics, and other chemicals (Filazi et al., 2012; Tittlemier, S. A. , 2010; Jablonski et al., 2014).

- **Enzyme-Based Tests**

Enzyme-based tests use specific enzymes to detect certain adulterants. For example, the urease test detects urea in milk by measuring enzyme activity in the presence of urea.

- **Immunoassays**

Immunoassays use antibodies to detect specific proteins or antigens in milk. They are useful in detecting adulterants like melamine or other protein-based substances (Matabaro et al., 2017).

- **Polymerase chain reaction (PCR)**

Polymerase chain reaction (PCR) has also been used for the specific and sensitive detection of milk and other food adulterants. PCR is a detection method that can be used for both qualitative and quantitative detection of milk adulterants including milk from other sources. The different variants of PCR such as multiplex PCR, Real-Time PCR and restriction fragment length polymorphism (RFLP) etc. are used for the detection of microbial and exogenous milk from different sources in raw and processed milk. The addition of exogenous proteins in milk has been detected specifically using PCR. The use of PCR as a regular milk adulterant detection method is still not in practice because of some pitfalls. The high level of substances such as

fats and proteins are inhibitory to PCR, and inability to detect non-DNA based milk adulterants limits PCR use in milk adulterant detection method (Di Domenico et al.) (Ewida and El-Magiud) (Hazra et al.) (Yang et al.).

Choosing the best technique depends on several factors, including the type of adulterant, required sensitivity, cost, and analysis speed. Electrical conductivity and pH tests are generally more affordable, making them ideal for initial screening. However, they may not detect all types of adulterants. Spectroscopy-based techniques, such as NIRS, FTIR, and Raman spectroscopy, offer high sensitivity and specificity, making them suitable for comprehensive analysis. They are especially useful when detecting a wide range of adulterants. (Azad and Ahmed, 2016; Yadav et al., 2022). Compared to other techniques, NIR spectroscopy has several advantages in detecting formalin in milk. It is non-destructive, enabling repeated measurements without affecting the sample's quality. NIR spectroscopy is highly sensitive and accurate, capable of detecting even trace amounts of formalin. It can also analyze multiple components at once, providing comprehensive milk quality assessment beyond formalin detection. It is easy to use and requires minimal sample preparation, making it cost-effective and suitable for routine screening. The technique can be applied to various milk matrices and sample forms. NIR spectroscopy is a rapid technique, providing results within minutes, enabling timely decision-making and intervention.

1.1.3 Advanced techniques Implemented On Adulterants

- **Water**

Milk adulteration with water is a common fraudulent practice aimed at increasing the volume of milk, thereby reducing its quality and nutritional value. Research papers

have proposed several techniques to detect water adulteration in milk. For example, the measurement of milk's electrical conductivity is a simple yet effective method, as added water changes the ionic concentration, leading to altered conductivity (Montalvo et al., 2010). Similarly, the freezing point of milk is a critical parameter; added water raises the freezing point, and this anomaly can be detected through cryoscopy or other freezing point determination methods (Garcia et al., 2012). The analysis of specific gravity or density is another common approach, as the addition of water reduces these values (Patil et al., 2015). Additionally, spectroscopy-based techniques, such as near-infrared spectroscopy (NIRS), can detect changes in the milk's composition due to added water, providing a non-destructive and rapid method of analysis (Santos et al., 2011). Using these techniques, researchers and regulatory agencies aim to identify and combat milk adulteration to ensure food safety and maintain consumer trust.

- **Melamine**

Various advanced techniques are employed for the quantitative detection of melamine in milk and milk products. The Surface Enhanced Raman Spectroscopy (SERS) method has been utilized to identify melamine, with a portable sensor based on this technology allowing for instant detection (Zhang, Zou, Qi, Liu, Zhu, & Zhao, 2010; Kim, Barcelo, Williams, & Li, 2012). SB-ATR FTIR (Single Bounce Attenuated Total Reflectance - Fourier Transform Infrared Spectroscopy) is another technique used to quantify melamine in both liquid and powdered milk (Jawaid et al., 2013). Mass spectrometry methods such as LC-MS/MS, APCI-MS (Atmospheric Pressure Chemical Ionization Mass Spectroscopy), and EESI-MS (Extractive Electrospray Ionization Mass Spectrometry) have also been employed to detect melamine in

various milk products (Yang et al., 2009; Zhu et al., 2009). Another technique used for melamine detection is High-Performance Liquid Chromatography (HPLC), which has been employed to quantify melamine in milk and other dairy products (Gopalakrishnan Venkatasami, 2010; Ruicheng Wei et al., 2009). Raman spectroscopy has also been used to immediately detect melamine in dried milk powder without extracting it, focusing on the Raman band at 676 cm-1 (Okazaki et al., 2009). Portable screening systems based on Laser Raman Spectroscopy have been designed to quantify melamine (Cheng et al., 2010). Gold nanoparticles offer an innovative approach to melamine detection. When these nanoparticles, which are grafted with melamine and cyanuric acid derivatives, bind to melamine, they change color from red to blue, providing an instant on-site detection method (Ai, Liu, & Lu, 2009). Additionally, the use of oxidized polycrystalline gold electrodes has been reported to detect melamine, along with traditional approaches like Gas Chromatography-Mass Spectrometry (GC-MS) (Tsai, Thiagarajan, & Chen, 2010). Recent advancements in melamine detection techniques are discussed in Liu et al., 2012.

- **Urea**

Urea is naturally found in milk, constituting a significant portion of the non-protein nitrogen content. However, urea is often used to adulterate milk, either through deliberate addition or by combining unspecified synthetic milk with natural milk. According to the Food Safety and Standards Authority of India (FSSAI) Act 2006 and the Prevention of Food Adulteration (PFA) rules of 1955, the maximum permissible limit for urea in milk is 70 mg/100 mL (Sharma et al., 2012). Various techniques have been developed to detect urea in milk. Near-infrared Raman spectroscopy allows for

the quantification of urea without pre-processing (Khan et al., 2014), while Liquid Chromatography (LC) is another method to identify urea as an adulterant (Dai et al., 2013). Gas Chromatography/Isotope Dilution Mass Spectrometry (GC/IDMS) has also been used for quantifying urea (Xinhua Dai et al., 2010), and High-Performance Liquid Chromatography (HPLC) has been suggested for detecting urea by converting it into a derivative containing a chromophore (Czauderna & Kowalczyk, 2009). A combination of the Kjeldahl method and spectrophotometry has been proposed to detect melamine, urea, and ammonium sulfate adulteration (Virginia de Lourdes et al., 2013). In terms of infrared technology, an optical waveguide sensor that detects ammonia at a characteristic wavelength of 1530 nm has been developed, aiding in the detection of multiple analytes including ammonia (Bamiedakis et al., 2013). Other technologies, such as a Field Effect Transistor (FET) with a graphene channel and ionic liquid (IL) gate, can detect ammonia and carbon dioxide at specific thresholds (Inaba et al., 2013) Two research papers discuss the classification of biosensors for detecting urea, based on enzymatic and non-enzymatic approaches and transduction signal systems (Farzaneh Shalileh et al., 2023) (None Jyoti et al., 2022) . Another paper focuses on the detection of urea as a milk adulterant using a fiber optic sensor, while another presents a point-of-use sensor for urea quantification in milk with minimal sample reprocessing and seamless readout (Ruchira Nandeshwar et al., 2023) .

- **Formalin**

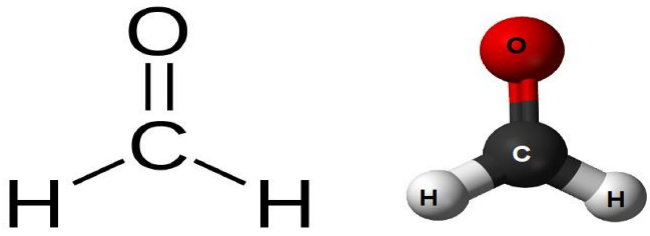

Fig 1.1 Formalin Structure

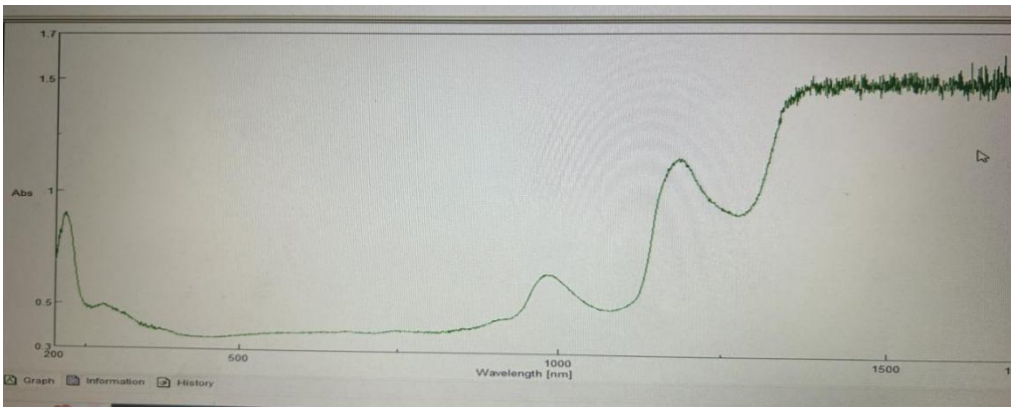

Fig 1.2 Recorded Formalin Spectra on Jasco

Formalin (CH_2O fig 1.1), a solution of formaldehyde in water , is sometimes used as an adulterant in milk due to its preservative properties. However, its use in food products poses significant health risks.The spectroscopic data obtained for formalin appears as depicted in figure 1.2. According to research, several techniques are used to detect formalin in milk. For instance, Fourier Transform Infrared Spectroscopy (FTIR) can identify specific functional groups that are characteristic of formalin, allowing for its detection in adulterated milk (Alhendi et al., 2014). Chromatography techniques, such as High-Performance Liquid Chromatography (HPLC), offer another method to separate and quantify formaldehyde in milk (Kaminski et al., 1973). Additionally, colorimetric tests based on the Nash reagent have been employed, where the reagent reacts with formaldehyde to produce a color change, providing a simple yet effective method for detection (Mathaweesansurn and Detsri, 2022). Some

studies also suggest the use of Gas Chromatography-Mass Spectrometry (GC-MS) for more accurate quantification of formaldehyde in milk samples (Vaz et al., 2022). Overall, these methods provide a variety of approaches to detect formalin adulteration in milk, catering to different analytical needs and contexts.

- **Other Compounds**

Various advanced spectroscopy techniques have been employed to detect and quantify adulterants in milk. Near-infrared (NIR) spectroscopy, in the 1100-2500 nm range, has been used to identify whey in cow's milk (Kasemsumran et al., 2007). Santos et al. (2013a) conducted a comparative study between NIR and medium-infrared (MIR) spectroscopy and found that MIR outperformed NIR in detecting a variety of adulterants, including tap water, whey, hydrogen peroxide, and synthetic urine. In another study, the presence of synthetic urine in UHT milk was identified in all tested samples through a chemometric approach (Souza et al., 2011). Moreover, synthetic urine concentrations as low as 0.78 mg/L could be detected using infrared microspectroscopy and chemometric analysis (Santos et al., 2013b). Additionally, changes in sodium and calcium concentrations, measured with flame atomic absorption spectroscopy, can indicate the presence of synthetic urine (Santos et al., 2012).

In other research, Matrix-assisted Laser Desorption/Ionization Time of Flight Mass Spectroscopy (MALDI-QTOF MS) has been used to detect vegetable oil in milk (Saraiva et al., 2012), and Raman chemical imaging has allowed for the identification of various adulterants, including ammonium sulfate, dicyandiamide, and other contaminants in powdered milk (Qin et al., 2013). The adulteration of milk fat is also

a common issue, with various detection methods available. Techniques such as fluorescence spectroscopy (Ntakatsane et al., 2013), derivative spectroscopy (Jirankalgikar & De, 2014), and Raman spectroscopy (Uysal et al., 2013) have proven effective in detecting these types of adulteration.

In recent years, Fourier-transform infrared (FTIR) spectroscopy and near-infrared (NIR) spectroscopy have gained attention for their potential in detecting chemical contaminants in food products. These techniques provide high sensitivity and specificity, making them suitable for detecting formalin in milk at trace levels. By leveraging the distinctive spectral signatures of formalin and milk components, spectroscopic methods offer a rapid and reliable alternative to traditional detection methods.

Analysis and experimental methods used for milk composition measurement are time-consuming, costly, and require manpower and are not automated [22]. NIR spectroscopy is a rapid and less complex method that can be used for real-time analysis of milk samples in a laboratory or in-line production environment. It is a cost-effective and efficient technique that can provide accurate and reliable results for milk analysis. In milk analysis, NIR spectroscopy can be used to describe and determine the various components present, such as protein, fat, and lactose content of the milk. This information is essential and important for ensuring the quality and safety of dairy products and can help detect milk adulteration. NIR spectroscopy can be suitable for specific adulterant detection when applied to well- characterized samples with known adulterants. However, its effectiveness depends on the complexity of the sample matrix and the availability of a comprehensive spectral database (Sneha et al.)

1.2 Aim

This dissertation aims to develop an effective, and non-invasive technique to identify formalin in milk using Near Infrared (NIR)spectroscopy. Milk typically has a short shelf life of up to 48 hours when stored at a temperature of less than 7°C. However, the addition of preservatives can help to extend its shelf life. By adding preservatives, milk can be stored for longer periods without spoiling. Formaldehyde, also known as formalin, is the oldest and cheapest chemical used as a preservative. Even a small amount of it can significantly increase the shelf life of milk. However, formalin is extremely harmful to the human body. It is highly carcinogenic and can cause vomiting, abdominal pain, dizziness, and in extreme cases, even death. Formalin is also nephrotoxic, which means it is highly toxic to the kidneys. When consumed, it reacts with macromolecules such as Deoxyribonucleic acid, Ribonucleic acid, and proteins, forming reversible adducts or irreversible cross-links. Although the use of formalin in food products has been banned worldwide, some people continue to use it for the same purpose. Foods that are commonly adulterated using formalin include noodles, salted fish, tofu, and even chicken and beer. Even today, formalin is illegally used as a preservative in some foods, which exposes consumers to its consumption and its dangerous consequences. Acute exposure to formaldehyde can irritate the eyes, nose, throat, and skin. Long-term exposure, on the other hand, has been associated with certain types of cancers, such as sinonasal cancer, and can also trigger asthma.

Adulteration of milk is a malpractice in which dealers either incorporate cheap substances or subtract valuable components from milk to increase its volume and thus profit margin. Excessively documented adulterants used to adulterate milk are diluent (water and ice) thickening agents (starch, glucose, urea, flour, salt, and chlorine, etc.),

preservatives (sodium carbonate, sodium bicarbonate and formalin, etc.), reconstituting agents (seed oils, cane sugar and animal fats and milk powder), cosmetic agents (Detergent/soap and bleaching powder, etc.) melamine and others. In a research paper by Singuluri (2014) found that Sucrose and skim milk powder were present in 22% and 80% of the milk samples respectively. Urea, neutralizers, and salt were present in 60%, 26%, and 82% of the milk samples respectively. Formalin, detergents, and hydrogen peroxide were present in 32%, 44%, and 32% of the milk samples obtained.

1.3 Objective

Development of a milk sample database containing formalin adulterant.

Analysis of milk and adulterated samples using NIR spectroscopy

Developing best machine learning algorithm for detecting formalin with good accuracy and low RMSE

Test and validate the ML models on a diverse range of milk samples cow and buffalo

1.2 Hypotheses

Formalin (formaldehyde solution) is an illegal adulterant sometimes used to preserve milk, extending its shelf life. Formalin, a 37% aqueous solution of formaldehyde gas, is a highly toxic substance extensively used as an antiseptic, disinfectant, and preservative. Milk lasts for 48 hours when it is stored at a temperature less than 7 °C. Still, its shelf life can be extended further by adding preservatives like formalin which ensures long-distance transportation without refrigeration, resulting in significant cost savings for suppliers. Despite its effectiveness in preserving milk, formalin poses a

severe health hazard to consumers as it can cause liver and kidney damage. Ingesting formalin can result in a range of adverse effects, including diarrhea, vomiting, abdominal pain, and even blindness. Given the critical need for rapid and reliable detection methods, this study hypothesizes that Near-Infrared (NIR) Spectroscopy can accurately and efficiently detect formalin in milk. NIR Spectroscopy operates in a wavelength range that allows for detailed molecular analysis, and its sensitivity to changes in chemical composition suggests it could effectively identify formalin adulteration in milk. The hypothesis assumes that formalin will produce unique spectral patterns in the NIR range that distinguish it from typical milk components, enabling accurate detection even at low concentrations. By analyzing these spectral patterns, the study aims to confirm whether NIR Spectroscopy is a viable method for detecting formalin in milk samples.

The primary research question guiding this study is: "Can Near-Infrared (NIR) Spectroscopy reliably detect the presence of formalin in milk, and if so, at what concentration levels?" This question seeks to address two critical aspects: the accuracy and sensitivity of NIR Spectroscopy in identifying formalin adulteration. Additionally, the study will explore whether certain reprocessing or chemometric techniques can enhance the detection accuracy of NIR Spectroscopy. A secondary research question could focus on practical applications: "Is NIR Spectroscopy a feasible technique for real-time, on-site formalin detection in dairy production environments?" This question aims to assess whether the method can be used effectively in field settings, contributing to food safety and quality control. By addressing these questions, the study intends to provide valuable insights into the capability of NIR Spectroscopy for detecting formalin in milk and its potential application in ensuring dairy product safety.

1.3 Scope of Detection of Formalin in Milk Using NIR Spectroscopy

Formalin detection in milk is a crucial aspect of food safety and public health. Near-Infrared (NIR) Spectroscopy offers a promising approach for this detection due to its non-destructive nature, rapid analysis, and high sensitivity to chemical composition changes. The scope of using NIR Spectroscopy for formalin detection in milk encompasses several key areas. First, it includes the development and validation of spectral models to identify and quantify formalin in milk samples. This involves analyzing the unique spectral fingerprints of formalin in the NIR range, typically between 400 and 1400 nm, where distinct absorption peaks may be indicative of its presence.

Additionally, the scope extends to the evaluation of NIR Spectroscopy's sensitivity to various formalin concentrations. This involves exploring detection thresholds to determine the minimum concentration of formalin that can be reliably identified. The approach should also account for potential spectral interference from other milk components, such as fats, proteins, and lactose, which could affect the accuracy of detection. Therefore, chemometric techniques and data reprocessing methods, such as baseline correction and noise reduction, become integral to the successful application of NIR Spectroscopy in this context.

The practical scope of using NIR Spectroscopy for formalin detection in milk includes its application in both laboratory and real-time settings. In a laboratory context, the focus is on achieving high precision and accuracy through controlled experiments and calibration with known formalin concentrations. This establishes the method's reliability for quality control and compliance with food safety standards. The real-

time application involves adapting the NIR Spectroscopy setup for use in dairy processing environments, allowing for rapid, on-site detection of formalin during milk production and distribution.

The scope also covers the integration of NIR Spectroscopy into existing quality assurance workflows, potentially enabling automated detection systems that can continuously monitor milk for formalin adulteration. This approach could help dairy producers and regulatory agencies identify contamination early, reducing the risk of formalin-tainted milk reaching consumers. Additionally, the portability of some NIR Spectroscopy instruments broadens the scope to field applications, where inspectors can quickly assess milk samples for formalin content at various points in the supply chain. Ultimately, the scope encompasses not only the technical aspects of formalin detection using NIR Spectroscopy but also the practical considerations for ensuring food safety and regulatory compliance in the dairy industry.

C H A P T E R 2

Literature Review

In recent years, spectroscopic methods have emerged as a promising tool for detecting adulterants in food products. Using machine learning techniques, it is possible to analyze the spectral data of food products to detect adulterants. While these techniques have been used to detect formalin in milk, only a small number of researchers have focused on detecting formalin as an adulterant in milk.

Bezuayehu Gutema Asefa (2022) proposed a method for detecting water adulteration in milk using digital image analysis combined with machine learning techniques. The support vector machine-based class prediction model outperformed the other machine learning tools in classifying adulterated milk samples based on the quantity of added water, with 94% total accuracy and 97% precision (Bezuayehu et al., 2022). .

Lucas de Souza Ribeiro et al.(2016) developed a hardware architecture based on diffuse reflectance spectroscopy in the near-infrared. To improve the signal-to-noise ratio, an optical condenser system with fixed lenses was created. As the light source, LEDs with precise spectral emission were used. InGaAsSb sensors, which have a fast response time and good sensitivity to the NIR band, were also employed to detect diffusely reflected light. The suggested equipment was tested on water-adulterated milk samples. The findings revealed high coefficients of determination, greater than 0.99.

Aditya Dave et al. (2016) proposed a non-contact approach for detecting milk adulteration while maintaining the consistency and quality of the milk sample and making it reusable for future testing. An embedded system with an AVR microcontroller combined with an optical sensor, LCD, and keypad was created. The main characteristic involved in detecting adulteration was the refractive index. The refractive index fluctuated as the amount of water adulteration changed, and these

relationships were used to construct the system for detecting adulteration of a random milk sample. The results obtained detected the adulteration with an accuracy of roughly 95±1%, indicating a 200% increase in accuracy over older methods such as spectroscopy.

Lucas da Silva Dias et al. (2018) produced a prototype for raw milk analysis that identifies the added water. To avoid significant scattering of infrared light by fat globules, a sample preparation process was developed. The sampling method is based on diffuse reflectance and a low-cost integrating sphere, avoiding costly commercial alternatives. In the near-infrared response area, the created sphere has a reflectance index of 88%. LEDs are used as infrared light sources, and an In-Ga-As-Sb photodiode is used for detection. The calibration was done with a set of samples with distinct adulterations, and then a new set was examined to validate the estimator's model. The coefficient of determination (R^2) was found to be 0.9562. The root-mean-squared error of prediction in the validation stage was 0.01794 .

A paper proposed by N. Swomya et al. (2021) covers the design and development of a low-cost, portable, multispectral, AI-based, non-destructive spectroscopic sensor system that can identify milk adulterants in real time. They used three different bands Ultraviolet (UV), visible, and infrared (IR) wavelengths ranging from (410-940nm). The steps involved in the research were AI-enabled multispectral spectroscopic sensor design, sample preparation, spectral data collection and processing, and neural network methods. To the spectrum data, several machine learning algorithms such as Naive Bayes, Linear discriminant analysis, support vector machine, decision tree, and neural network model are applied, yielding accuracy of 90%, 88.1%, 90%, 91.7%, and 92.7% respectively. The Genetic algorithm framework is used to perform optimal

parameter selection/parameter modification of the neural network. The neural network performance is enhanced from 92.7% to 100% by using optimal parameter settings.

A team led by S.J. Dutta (2022) researched measuring the urea content in milk. They utilized silver nanoparticles, both uncapped and citrate-capped, along with ultraviolet-visible (UV-Vis) spectroscopy. The team characterized the nanoparticles using Ultraviolet-Visible, scanning electron microscope (SEM), X-ray diffraction (XRD), and Fourier transform infrared (FTIR) spectroscopy. When the nanoparticles interacted with urea, they observed a color change from yellow to blue. The developed technique demonstrated an average accuracy and sample percent relative standard deviation (% RSD) ranging from 62.67 to 121.85% and 0.11-1.21%, respectively.

In a study conducted by Khan Mohammad Khan et.al, (2014) near-infrared Raman spectroscopy was proposed as a potential method for accurately determining urea levels in milk. By implementing the Raman technique in combination with the partial least squares algorithm, the team achieved an impressive accuracy rate of over 97% for urea concentrations above 100 mg/dl. For levels between 50 and 100 mg/dl, accuracy remained high at 90% to 95%. However, accuracy decreased to 60% for urea concentrations lower than 50 mg/d.

Yan cheng et al. (2010) Built a portable compact Raman spectrometric system to detect melamine adulterants in milk powder. melamine fortified in milk powder was identified with high reproducibility using two distinct vibration modes at 673 and 982 cm1. The intensity of the first mode was used to calculate the amount of melamine in milk powder. A detection limit (DL) of 0.13% and a good partial least squares (PLS) analysis model were obtained.

Kamboj, Uma, et al. (2020) established the detection of the presence of sugar as an adulterant in milk using Near Infrared Spectroscopy. They used chemometric software (CAMO Unscrambler version X 10.3) to analyze the data. The chemometrics model was developed through multivariate analysis of obtained data using the Principal Component Analysis (PCA) and Partial Least Squares (PLS) regression methods. The partial least square regression model performed well in predicting sugar-adulterated milk samples, with a coefficient of correlation greater than 0.9 and a root mean square error of validation (RMSEV) of 0.04.

Olgun Cirak et al. (2017) aimed to develop a rapid spectroscopic technique for milk source classification and discrimination using Fourier transform infrared spectroscopy (FTIR). The study utilized Hierarchical cluster and principal component analyses to achieve milk species classification and successfully detected milk sample adulteration using the FTIR technique. The chemometric method employed amide-I (1700-1600/cm) and amide-II (1565-1520/cm) spectral bands.

Vinod Kumar Verma, et al. (2019) used the ultrasonic technique to examine the adulteration of a pure milk sample with artificial (synthetic) milk. The ultrasonic wave of frequency 0.5 MHz from the transmitter was passed through the sample and received on the other end by the receiver. The signal received at the receiver is analyzed using a digital storage oscillator (DSO). As a result, the voltage of the received signal increases with the percentage increase in adulteration. As a result, the voltage of the received signal increases with the percentage increase in adulteration.

Medha Khenwar et al. (2022) designed an IoT model to assess the quality of milk by integrating multiple sensors including bacterial activity monitored by a gas sensor, pH value monitored by a pH sensor, Viscosity by a Viscosity sensor, and temperature by the temperature sensor. The IoT model ensures milk quality with the help of these

sensors, and the overall performance of this IoT model is evaluated using LabVIEW. The results of this model guarantee milk quality by 90%.

Mabrook, et al. (2003) found A novel method to detect added water to full-fat milk has been developed using single-frequency electrical conductance measurements. The characteristics at 100 kHz and 8 ◦C for all skimmed milk samples revealed a linear decrease in conductance with increasing water content over the entire range of water concentrations. In contrast, full-fat milk's conductance decreased only at added water concentrations higher than 10%. At lower added water concentrations, the full-fat milk exhibited an anomalous conductivity maximum at 2–3% added water.

Sumaporn Kasemsumran, et al. (2007) employed Near-infrared spectroscopy (NIRS) to detect the adulteration of milk, non-destructively. Two adulteration types of cow milk with water and whey were prepared, respectively. NIR spectra of milk adulterations and natural milk samples in the region of 1100 - 2500 nm were collected. The classification of milk adulterations and natural milk was conducted by using discriminant partial least squares (DPLS) and soft independent modeling of class analogy (SIMCA) methods. PLS calibration models for determining water and whey contents in milk adulteration were also developed, individually.

Mauricio Moreira et al. (2016) developed an innovative digital photometer, which can detect water presence in milk. The device is compact, equipped with a microcontroller, and incorporates three NIR-emitting LEDs. Unlike traditional photometers, it doesn't require lenses, filters, or moving parts. By calculating the transmittance of IR radiation, the photometer can accurately determine the amount of added water in milk samples. The researchers conducted various experiments and found that the prototype had a mean absolute error of under 1% in measuring added water percentage. Moreover, the absolute deviations from the average were less than 0.7% in two sets of

10 measurements. The device proved to be as responsive as a commercial cryoscope but provided faster results(Sumaporn Kasemsumran, et al).

Ram et al. (2022) have developed a cost-effective, uncomplicated, and reliable paper-based microfluidic device was developed for the detection of starch concentration in milk. The device was designed to accommodate a 10 µl milk sample that was introduced into the inlet zone. Following a 5-minute waiting period, the resulting color transition length was captured using a smartphone. The starch concentration was then measured by a specialized app, "starch-app," developed in-house. The correlation between the values of the starch concentration measured by the device and spectrophotometer was found to be high ($R2 = 0.9981$) within the range of 0–10% w/v. The developed device and app have the potential to serve as a useful tool for detecting milk adulteration, thereby ensuring the quality and safety of milk products.

Sharifi, Fatemeh, et al. (2023) aimed to evaluate the efficacy of a previously developed photoacoustic spectroscopy system that employs light sources in the visible to short-wave near-infrared range (Vis-SWNIR, 395–940 nm) for detecting various adulterants in cow's milk, including formalin, urea, hydrogen peroxide, starch, sodium hypochlorite, and detergent powder. The outcomes of principal component analysis (PCA) revealed a visually distinct separation between different types of adulterations. The artificial neural networks (ANN) exhibited the highest accuracy rate of classification, almost 97.6%, in identifying the type and level of adulteration. In conclusion, the Vis-SWNIR photoacoustic spectroscopy system appears to be a reliable and efficacious tool for detecting different forms of milk adulterations.

Balan et al. (2020) demonstrated Fourier transform infrared (FTIR) spectroscopy, combined with multivariate chemometrics, as an efficient method for the qualitative and quantitative analysis of formalin in milk. The spectra of pure and adulterated milk

(0.5-5% v/v) were obtained using ATR-FTIR in the range of 4000-400 cm−1. Principal component analysis (PCA) was used to separate pure samples from adulterated samples, resulting in well-defined clusters. Soft Independent Modelling of Class Analogy (SIMCA) was used to classify test samples with 100% classification efficiency. Partial least squares (PLS) regression and principle component regression (PCR) models were developed using the normal, first derivative, and second derivative spectra to quantify the level of formalin in milk.

A paper published by Fazal Mabood et. al (2017) utilized near-infrared (NIR) spectroscopy in the absorption mode, encompassing the wavelength range from 700 to 2500 nm. The study implemented a 2 cm-1 resolution and utilized a sealed CaF2 cell with a path length of 0.2 mm. Multivariate methods such as Principal Component Analysis (PCA), Partial Least Discriminant Analysis (PLS-DA), and Partial Least Regression Analysis (PLS) were employed for the statistical analysis of the acquired NIR spectral data. The PLS regression model had an R-square of 93%, with a good prediction as evidenced by an RMSECV of 1.38. Additionally, the model had a RMSEP value of 1.50 and a correlation of 0.95.

The research paper by Veríssimo et al. (2020) "A new formaldehyde optical sensor: Detecting milk adulteration" presents a novel sensor that can detect formaldehyde in milk. The sensor uses an optical fiber coated with polyoxometalate salt, which changes its UV-Vis spectrum upon contact with formaldehyde. The sensor had a detection limit of 0.2 mg/L for formaldehyde, which is consistent with conventional spectrophotometric methods. The sensor was tested on milk samples for formaldehyde quantification The results obtained from this optical sensor were consistent with those from traditional methods, with no statistically significant differences (α=0.05).

Vipin K Gupta et al. (2015) study presents a low-cost and rapid colorimetric technology for determining formaldehyde in milk samples using a smartphone. The method involves spot-test reaction and digital image analysis with R-G-B approach, and has a limit of detection of 0.31 ppm. The analytical curves showed linearity ranging from 0.25-4 ppm with $R2 > 0.99$.

Carvalho et al. (2015) devised a swift technique to detect and measure the adulteration of milk powder through the addition of whey. This method involved assessing glycomacropeptide protein using mid-infrared spectroscopy (MIR). After drying fluid milk samples and spiking them with varying concentrations of GMP and whey, calibration models were created using multivariate techniques based on spectral data. Excellent percentages of correct classification were achieved in principal component analysis and discriminant analysis as the proportion of whey samples increased. In the best model of partial least squares regression analysis, the correlation coefficient (r) and root mean square error of prediction (RMSEP) were 0.9885 and 1.17, respectively. The rapid analysis, cost-effectiveness, and high throughput of samples tested per unit time suggest that MIR spectroscopy has the potential to serve as a rapid and reliable method for detecting milk powder frauds involving cheese whey.

In a research paper by Madhusudan G. Lanjewar et al (2024). Using the same setup water adulteration in milk was detected The spectroscopic data was pre-processed using various techniques such as SG filter, MSC, and SNV method. Wavelength/feature selection and PCA were used to select the most informative features and reduce their dimensions. Different ML models were employed to predict water concentration in milk. The KNN model performed the best in regression analysis with R2, RMSE, SEP, MAE, RPD, LOOCV-R2, and LOOCV-RMSE values of 0.999, 0.399 mL (% v/v), 0.096 mL (% v/v), 0.227 mL (% v/v), 33.005, 0.999,

and 0.353 mL (% v/v), respectively. On the other hand, RF achieved 100% accuracy and MCC in classification analysis.

Samaneh Ehsani et al. (2022) aimed to investigate the potential of using a portable near infrared (NIR) spectrometer, operating in the spectral range of 900-1700 nm, in combination with ensemble methods as a rapid, nondestructive, and simple technique for detecting water in bovine milk samples in the concentration range of 1% to 30% (v/v). The developed model showed reliability and robustness, with satisfactory values for calibration, cross-validation, and prediction sets. The performance of the RSDE method was compared to other common classification techniques, including PLS-DA and SVM, and was found to outperform these methods in terms of accuracy and reliability. Additionally, boosted regression tree (BRT) was used to quantify the level of water adulterant in milk, achieving a high level of accuracy. These results indicate the potential of using portable NIR spectrometers and ensemble methods for detecting water adulteration in milk. The ensemble regression model's performance was assessed using the regression coefficient (R2) and root mean square error (RMSE), with the BRT method achieving values of 0.95 and 0.58, respectively, in the prediction set.

In a study, Bruno G. Botelho et al. (2015) proposed a screening method for detecting five common adulterants in raw cow milk. The method utilized attenuated total reflectance (ATR) mid infrared spectroscopy in combination with multivariate supervised classification (partial least squares discrimination analysis - PLSDA) to simultaneously detect the presence of water, starch, sodium citrate, formaldehyde, and sucrose in milk samples containing up to five of these analytes in the range of 0.5-10% w/v. A multivariate qualitative validation was performed to estimate specific

figures of merit, including false positive and false negative rates, selectivity, specificity, efficiency rates, accordance, and concordance.

Y. Etzion et al. (2004) conducted a study to determine the protein concentration in raw cow milk using attenuated total reflectance spectroscopy in the mid-infrared range. The method relied on the characteristic absorbance of milk proteins, which include two absorbance bands in the 1500-1700 cm−1 range and absorbance in the 1060-1100 cm−1 range. An optimized automatic procedure for accurate water subtraction was applied to minimize the influence of the strong water band. Three methods were used to analyze the spectra: simple band integration, partial least squares (PLS), and neural networks. The neural network approach produced the most accurate results, with prediction errors of 0.20% protein when based on PCA scores only and 0.08% protein when lactose and fat concentrations were included in the model. The study suggests that Fourier transform infrared/attenuated total reflectance spectroscopy could be a useful technique for the rapid and potentially online determination of protein concentration in raw milk.

In a study conducted by Flavia Borges de Freitas Rezende et al., (2012) a high-performance liquid chromatography method with UV detection (HPLC-UV) was developed and validated for the detection of formaldehyde in bovine milk. The method involved a formaldehyde derivatization reaction with 2,4-dinitrophenylhydrazine at pH 4.0, allowing for the detection of formaldehyde in milk at 360 nm. The analytical curves ranged from 10.0 to 400.0 μg L−1 in aqueous solutions and milk samples, with an R2 value greater than 0.99, indicating the linearity of the method. The limit of quantification of 20.0 μg L−1 demonstrated the high sensitivity of the method for formaldehyde residues in milk. The method was validated using milk samples fortified with formaldehyde at three different

concentrations and demonstrated a mean overall recovery of 102.2 ± 1.3% (n = 9). The accuracy of the method was evaluated using the Student t-test, and comparable results were obtained at a 95% confidence level, demonstrating the usefulness and effectiveness of the proposed method.

A new analytical method by Maha Ibrahim Alkhalf & Elwathig, 2017 using FTIR spectroscopy was developed to determine formaldehyde in cheese. The method was accurate with a coefficient of determination (R2) of 0.986 and an average standard error of calibration of 2.24 mg/100g. The validation using the "leave-one-out" cross-validation method resulted in an R2 of 0.9662, with standard errors of prediction and standard deviation being 4.07 mg/100g and 4.61, respectively. These results suggest that FTIR spectroscopy is a precise and rapid technique for detecting formaldehyde in cheese samples.

Sandeep Choudary et al. (2022) developed a fluorescence-based method with a point-of-use colorimetric sensing system to test milk quality in real-time using fluorophores. The fluorescence intensities of the fluorophores were optimized to predict the pH of milk samples using a color sensor device (CSD) and were cross-referenced using a fiber optic spectrophotometer (FOS). The CSD and FOS measured pH and adulteration in a linear range of 4-9 pH units and 0-70 mm urea in milk, with a quick response time of 30 seconds and 5 minutes. The interday variability for pH sensing by the CSD and FOS was evaluated and expressed as a percentage relative standard deviation (%RSD), which was found to be 1.89 for CSD and 4.72 for FOS.

E. HOP et al. (1993) conducted a study using FT-IR spectrometry to determine the water content of milk. Calibration was performed in the mid-infrared using a specific water band at 2110 cm and a reference region at 2590 cm. Multiple linear regression

resulted in a prediction error of 0.14% water for milk samples with a water content between 84.9% and 88.0% w/w.

Masataka Kawasaki et al. (2008) developed a near-infrared (NIR) spectroscopic sensing system to obtain NIR spectra of raw milk automatically in a milking robot system. Calibration models were developed to determine major milk constituents (fat, protein, and lactose), somatic cell count (SCC), and milk urea nitrogen (MUN) in unhomogenized milk, and the precision and accuracy of the models were validated. The validation set for fat had a coefficient of determination (r2) of 0.95 and a standard error of prediction (SEP) of 0.25%. Lactose had r2 and SEP values of 0.83 and 0.26%, respectively, while protein had r2 and SEP values of 0.72 and 0.15%. For SCC, the r2 and SEP values were 0.68 and 0.28 log SCC/mL, respectively, and for MUN, they were 0.53 and 1.50mg/dL, respectively. These results indicate that the NIR spectroscopic system can be used in real-time to assess milk quality in an automatic milking system.

Mohammed Musa et al. (2021) developed a new procedure to quickly classify and quantify fresh milk adulteration. Fresh cow milk samples were collected from eight farms in China and were adulterated with tap water at ten percentage levels. NIR spectroscopy was used to scan the samples, and chemometric tools like SIMCA and PLS were applied for statistical analysis. The developed PLS regression model had a standard error of prediction (SEP) of 5.33 g/L for estimating the levels of adulteration with water. This method is non-destructive, low-cost, and requires minimal sample preparation, making it fast and simple for raw milk control in a dairy industry or quality inspection of commercialized milk.

Thiago R.L.C. Paixão et al. (2009) developed a disposable electronic tongue with all necessary electrodes integrated into a single device. The device was constructed with

gold CD-R and copper sheets substrates, and the sensing elements were gold, copper, and gold surface modified with a layer of Prussian Blue. The performance and capability of the device were evaluated with taste substances, different types of milk, and adulterated samples. The results showed good separation between different samples in the principal component analysis (PCA) score plots. The relative standard deviation for signals obtained from the electrodes was below 3.5% .

Noor Aidawati Salleha et al. (2020) conducted a study to differentiate between milk from different goat breeds using Fourier transform infrared spectroscopy (FTIR) and multivariate analysis. the results showed clear discrimination between the breeds using Partial Least Square Discriminant Analysis. The chemical composition analysis revealed that milk had superior protein, fat, and lactose content compared to the other breeds, with values of 3.7%, 4.20%, and 5.30%, respectively. This study suggests that Jamnapari goat milk is different from the other two breeds and can be identified using FTIR and multivariate analysis.

In a study conducted by Chirantan Das et al. (2018) to analyze the possibility of detecting soap as an adulterant in cow milk using Electrical Impedance Spectroscopy (EIS). The technique provides a simple, rapid, precise, and cost-effective platform for monitoring milk quality. The study analyzed the variation of electrical parameters, including impedance, capacitance, conductance, and current, for different concentrations of soap adulteration in milk. The results showed that capacitance, conductance, and current increased, while impedance decreased with increasing soap content in milk. The study also extracted the coefficient of sensitivity for soap-adulterated milk samples and explained it in terms of the measured conductance values .

A recent study by Moupali Chakraborty et al. (2018) aimed to identify the minimum detectable limit for five common milk adulterants using an impedance sensor. The LOD for adulterants in milk with varying fat percentages was studied using commercial packet milk, UHT milk, and raw milk. Statistical analysis was applied to verify the results and ensure data consistency.

A paper by A. Ravindran et al. (2018) With their research concluded that The use of harmful chemicals in food adulteration is common for extra profits. Traditional laboratory methods can detect adulterants but are not user-friendly, time-consuming, and destructive. Spectroscopy offers a non-destructive, fast, and accurate method for detecting adulteration in food. It can be used for detecting adulteration in various food products like spices, natural oil, juice, honey, milk products, and wines .

C H A P T E R 3

Materials and Methods

3.1 Sample Collection and Dataset creation

In this study, Raw milk samples were sourced from namely a buffalo and a cow, both of which were classified under the Bos taurus species. The samples were collected from the southern regions of Goa and were investigated at Goa University. Figure 3.1 illustrates the comprehensive flow of the project. Stringent measures were taken during the transportation of the milk to ensure that it was stored below 40°F (4.4°C) to preserve the optimal quality of the milk. The rationale behind this is that warmer temperatures expedite bacterial growth and spoilage, which could compromise the quality of the milk. A batch of these milk was utilized to capture spectral data for pure milk and milk adulterated with formalin. Adulteration levels ranged from 0.5% to 50%, with variations of 0.5% for the range of 1% to 5% (precisely,0.5%, 1%, 1.5%, 2%, 2.5%, 3%, 3.5%, 4%, 4.5%, and 5%) and 10% variations for the range of 10% to 50% (including 10%, 20%, 30%, 40% and 50%). In total, 60 samples were created with various adulteration levels of each milk. The volume of each milk sample used for spectral recording remained constant at 10 ml. The samples were uniformly mixed using a sonicator (Q Sonica, Model-Q500) fig 3.2 at 50% intensity for 15 minutes to distribute fat globules consistently. This step is crucial to avoid dispersion during the spectral analysis. The entire datasets was meticulously conducted, with careful consideration given to the temperature at which the milk was stored throughout the spectral analysis to maintain the integrity of the results. The database in this study was constructed based on the % v/v ratio of milk and formalin. Milk was measured using a pipette of 10 ml, and formalin was measured using a micropipette fig 3.3 with a maximum count of 100 microliters. This

method was chosen to ensure accuracy and consistency in the measurement of the samples. Fig 3.5 depicts a test tube containing pure milk and another containing milk adulterated with 50% formalin. The use of precise instruments and standardized measurement techniques is crucial in developing reliable data for scientific research.

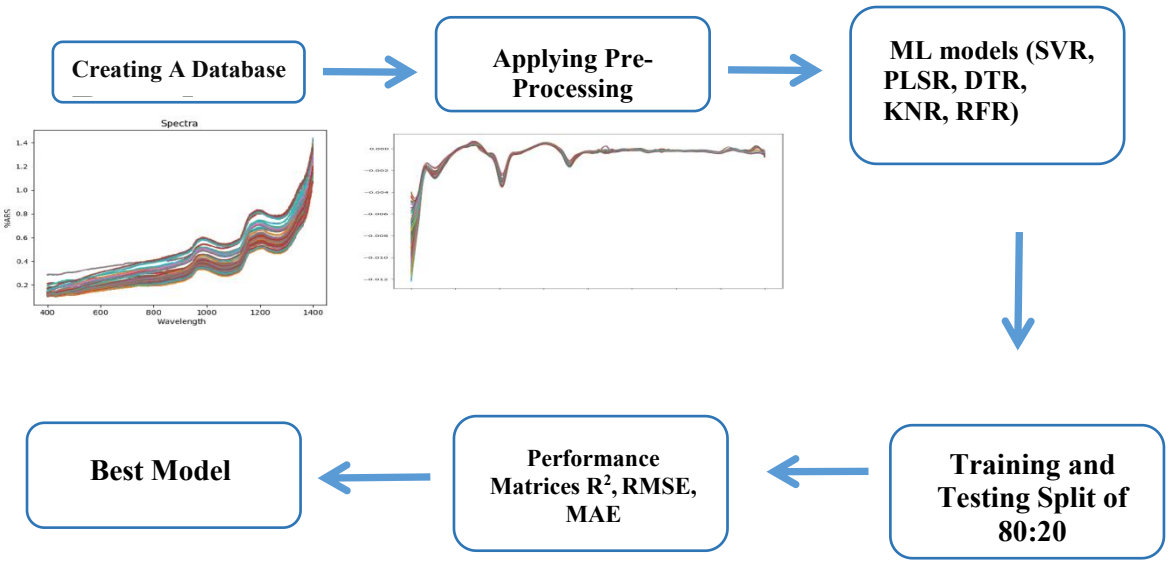

Fig 3.1:- Block Diagram of the Project

Fig 3.2:- Sonication device

Fig 3.3 :- Micropipette with Microtip

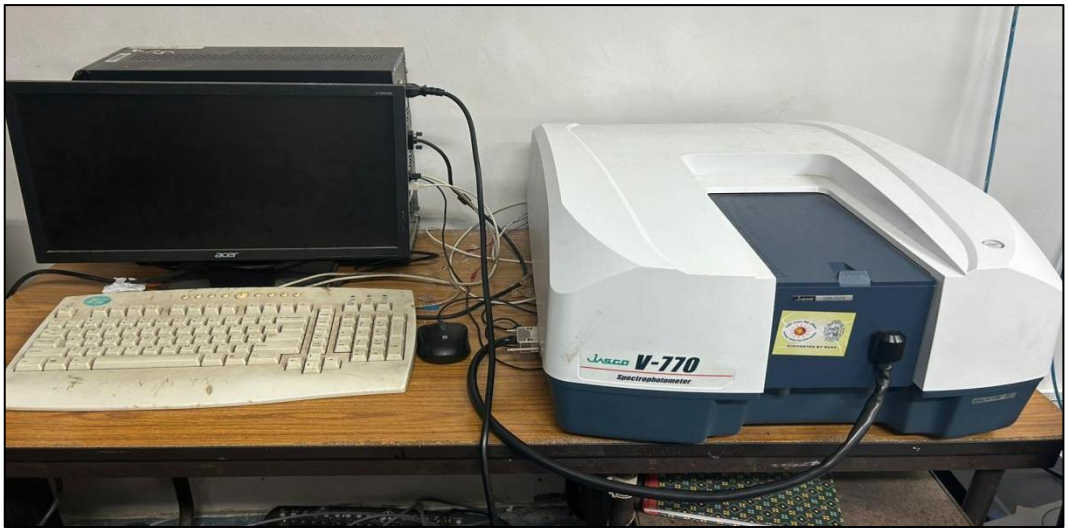

Fig 3.4:- The figure displays the setup of the spectrophotometer, which was used for capturing the spectral data in the experiment.

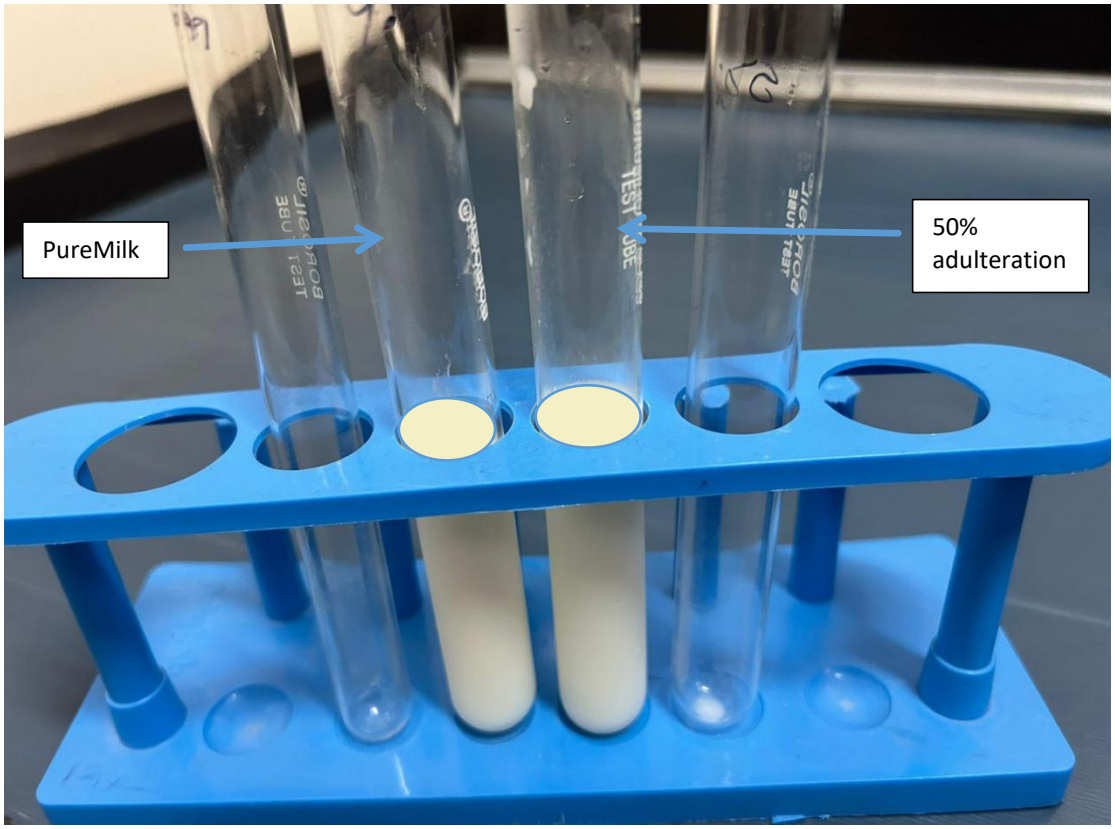

Fig 3.5:- The figure depicts a test tube containing pure milk and another containing milk adulterated with 50% formalin

Table 2.1:- Dataset

Sr. No.	Formalin (% v/v)	Milk (% v/v)	Formalin (ml)	Milk (ml)
1	0	100	0	10
2	0.5	99.5	0.05	9.95
3	1	99	0.1	9.9
4	1.5	98.5	0.15	9.85
5	2	98	0.2	9.8
6	2.5	97.5	0.25	9.75
7	3	97	0.3	9.7
8	3.5	96.5	0.35	9.65
9	4	96	0.4	9.6
10	4.5	95.5	0.45	9.55
11	5	95	0.5	9.5
12	10	90	1	9
13	20	80	2	8
14	30	70	3	7
15	40	60	4	6
16	50	50	5	5

A 10 ml sample of Jersey cow milk and Buffalo was collected and divided into four equal parts, each containing 2.5 ml of milk. Spectroscopic analysis was then performed on each of these four parts, with four spectra taken for each part, resulting in a total of 16 spectra. The entire dataset consisted of 256 spectras of each milk. The dataset comprised a combination of spectra from both Jersey cow milk and buffalo milk, resulting in a total of 512 spectras, constituting the third dataset.This approach

was used to assess the consistency and variability of the spectral data across multiple portions of the same milk sample.

3.2 Spectral Acquisition System and Spectra pre-processing

The experimental setup for the detection of formalin adulteration in milk using ML and NIR spectroscopy involved the use of a spectrophotometer JASCO V770 Fig 3.4 to capture the spectra, ranging from 200 to 1700 nm. The parameters for the acquisition process were set as Photometric mode: Abs, Data interval: 0.5 nm, UV/Vis bandwidth: 2.0 nm, NIR bandwidth: 8.0 nm, UV/Vis response: 0.06 sec, NIR response: 0.06 sec, Scan speed: 1000 nm/miMn, Change source at 340 nm, Change grating at 850 nm. These parameters were kept constant throughout the experiment for JASCO Fig3.5.

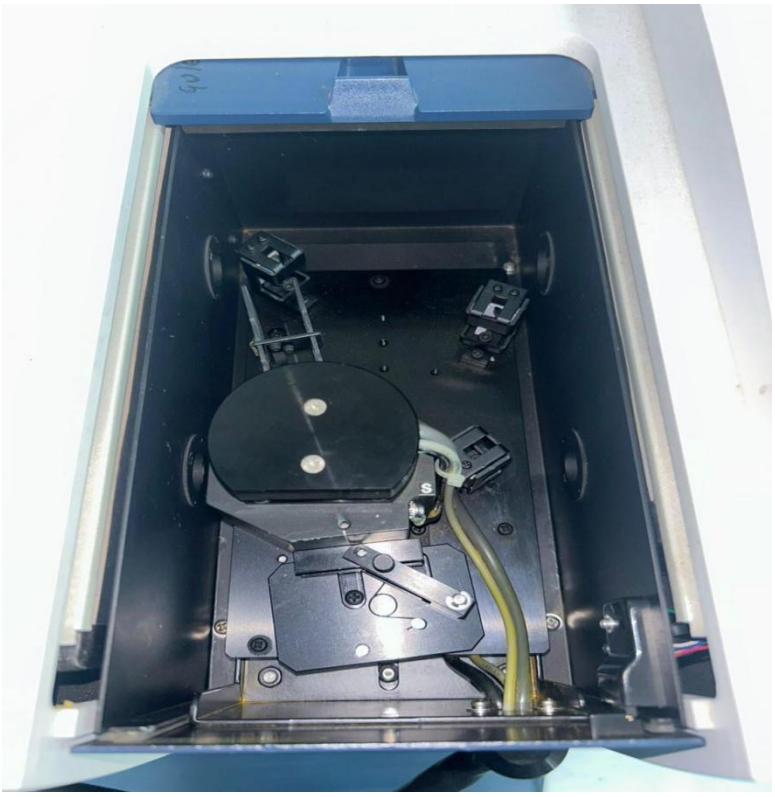

Fig 3.5:- The figure depicts the integrating sphere used in reflectance mode for capturing spectral data in the experiment.

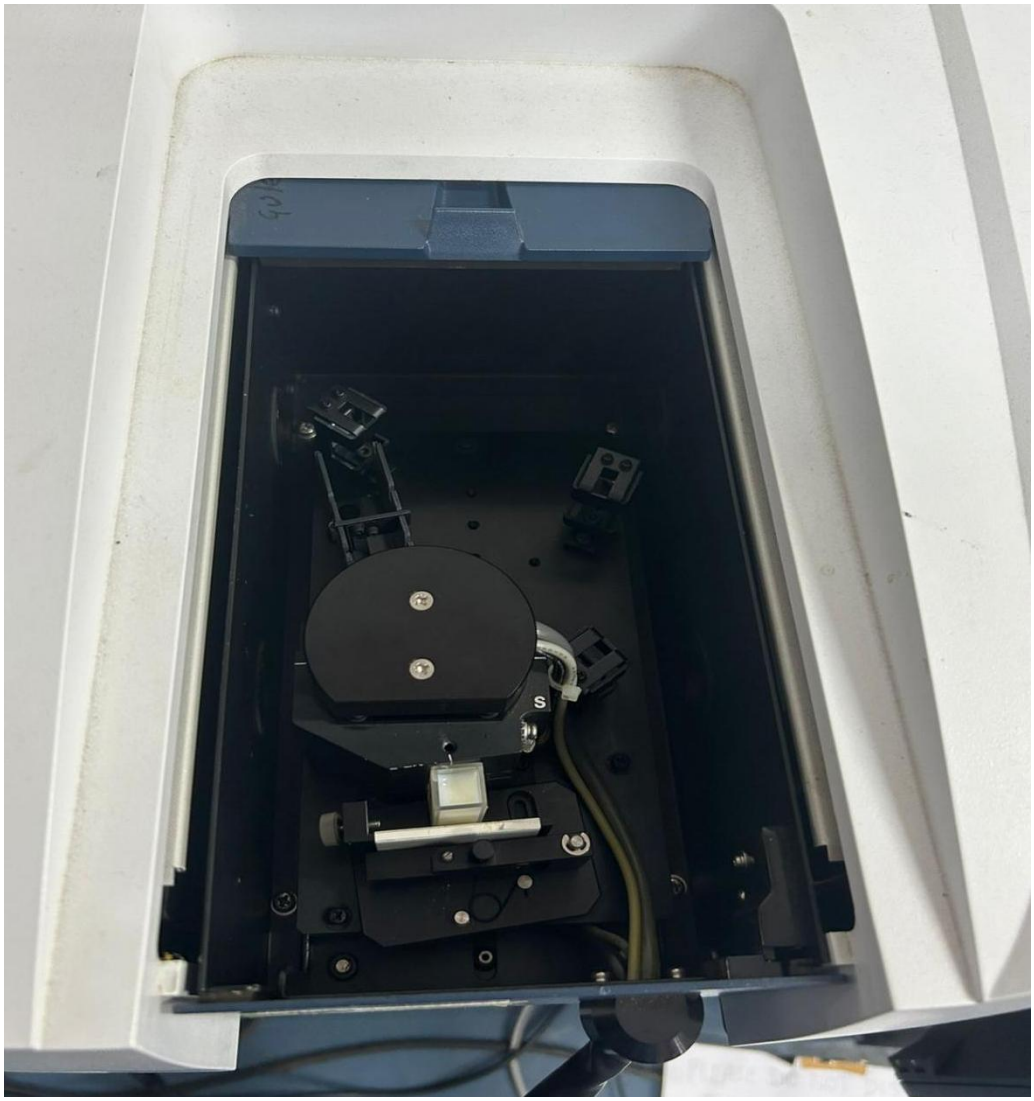

Fig 3.6:- The figure shows milk in a cuvette and how the setup of the spectrophotometer is kept with the lid closed during spectral data collection.

A spectrophotometer in reflectance mode Fig 3.6, specifically with an integrating sphere like the Jasco V770, measures the reflectance properties of a sample by integrating scattered light.

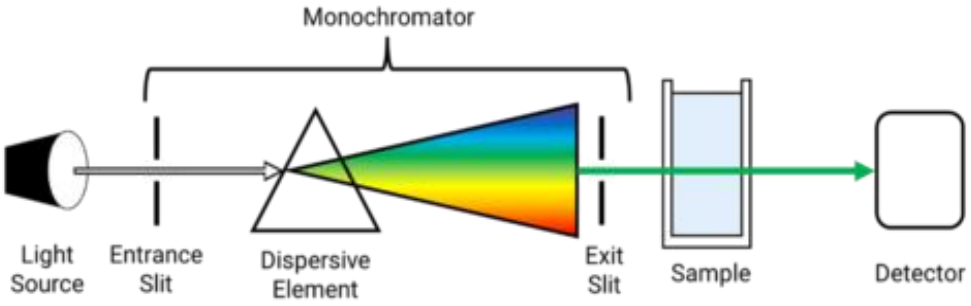

Fig 3.7 Block diagram of Spectrophotometer

At the core is the light source, typically a xenon or tungsten-halogen lamp, which emits a broad spectrum of light. This light is directed through a monochromator Fig 3.7, usually composed of diffraction gratings and mirrors, which selects a specific wavelength or range of wavelengths. The monochromatic output is then directed to the integrating sphere, a spherical chamber with a diffuse white interior that evenly reflects the light. The Reference Detector Captures light that doesn't interact with the sample (reference beam) to establish a baseline. Inside the sphere, the sample is positioned to reflect light into the sphere's interior, where it's uniformly scattered. This scattered light is collected by a detector, often a photodiode array or photomultiplier tube, and its intensity is measured to determine the reflectance properties of the sample. This data is then analyzed to understand the reflectance properties of the sample at different wavelengths, helping determine its composition or characteristics.

The recorded spectra (x) were then corrected using the SG filter with with a window length of 91, polynomial order of 3, and first derivative was apllied to (x). The dataset underwent a 10-fold cross-validation process, where it was divided into 10 equal parts.

The ML model was then trained and validated 10 times, with each iteration using a different part as the validation set and the remaining parts for training. Further, the pre-processed spectra were standardized using a Min-Max scaler method to bring them to a common scale. The SG algorithm used in the pre-processing step fits each polynomial to windows in the region of each point in the spectrum. These polynomials were then used to smooth the data, and the resulting pre-processed spectra were used for further analysis. The use of the SG filter method for noise reduction has been a popular technique in many fields of data processing, as discussed in previous literature (Hasar et al., 2023). This experimental setup and pre-processing technique were crucial in obtaining accurate and reliable results for the detection of formalin adulteration in milk using ML and NIR spectroscopy.

3.3 Operating Principle

Beer-Lambert Law

The Beer-Lambert law is a fundamental principle in the field of spectroscopy. It describes the linear relationship between the absorbance and the concentration of an absorber, which is typically a sample solution. According to Beer's law, two external assumptions are made in the experiment.

Firstly, it is assumed that the absorbance is directly proportional to the concentration of the sample. This means that as the concentration of the sample increases, so does the absorbance of light passing through it. Secondly, it is assumed that the absorbance is directly proportional to the length of the light path or the width of the container. This means that the longer the light path or the wider the container, the higher the absorbance will be.

In the Beer-Lambert law, light is passed from the sides of the container Fig 3.8, and the intensity of the light changes after it passes through the sample solution. By measuring the change in intensity, the absorbance of the sample can be calculated. The Beer-Lambert law is a powerful tool for scientists and researchers in many fields, including chemistry, biology, and environmental science, as it allows them to determine the concentration of a sample with great accuracy.

The Beer Lambert's Law is given by;

$$A=-\log (I/Io)$$

- I is the light intensity
- Io is the initial light intensity

$$A= \varepsilon L C$$

- A is the amount of light absorbed for a particular wavelength by the sample
- ε is the molar extinction coefficient
- L is the distance covered by the light through the solution
- c is the concentration of the absorbing species

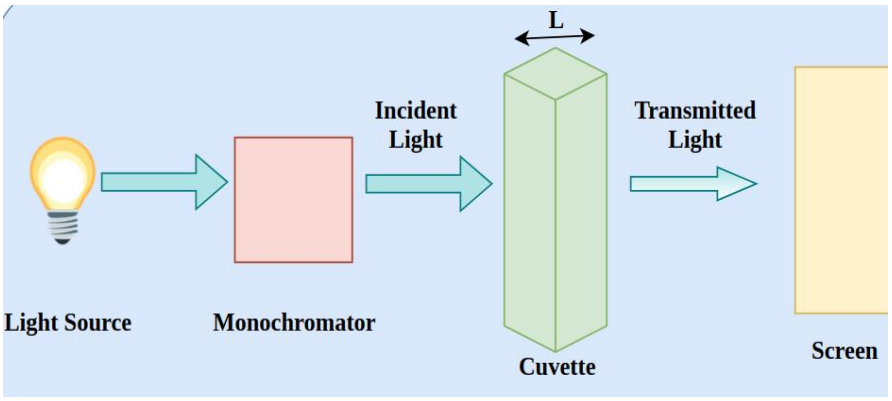

Fig 3.8 Beer Lambert Law

3.4 NIR Spectroscopy

NIR (Near-Infrared) spectroscopy has emerged as a powerful technique for detecting milk adulteration, offering a rapid, non-destructive, and reliable method to assess the quality and purity of milk. This technology operates in the near-infrared range Fig 3.9, typically from 700 nm to 2500 nm, allowing it to identify a wide range of chemical compounds and physical properties in milk samples. By analyzing the unique spectral patterns generated when NIR light interacts with the molecular bonds in milk components, such as proteins, fats, and lactose, it becomes possible to detect adulterants and contaminants. Common adulterants in milk include water, urea, starch, melamine, and vegetable oils, each of which has distinct spectral signatures that NIR spectroscopy can detect and quantify. The technique's speed and accuracy make it ideal for routine screening in dairy production and quality control environments, allowing producers and regulatory bodies to ensure milk safety and integrity without extensive sample preparation or hazardous chemicals.

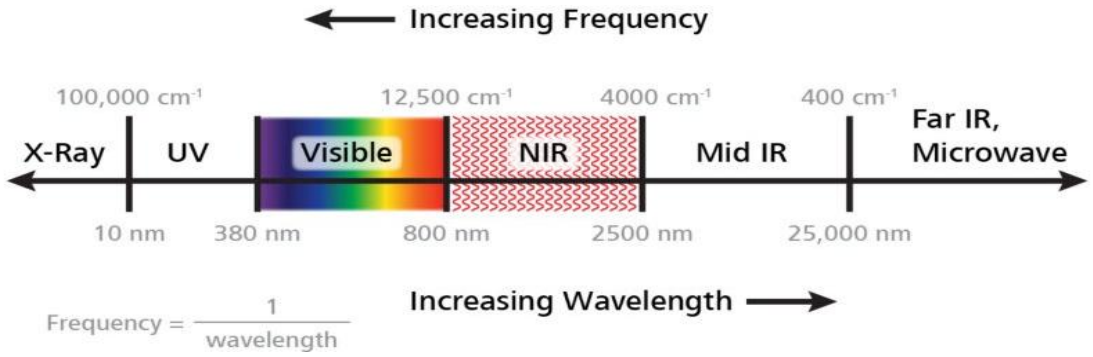

Fig 3.9 NIR spectral Band

C H A P T E R 4

Machine learning models

Support Vector Regression (SVR), Decision Tree Regression (DTR), Random Forest Regression (RFR), Partial Least Squares Regression (PLSR), and K-Nearest Neighbors Regression (KNNR). Each of these models has unique characteristics and underlying principles. Below is a brief explanation of each, including their basic block diagrams and how they work.

4.1 Support Vector Regression (SVR)

Support Vector Regression (SVR) is an advanced machine learning technique derived from the foundational principles of Support Vector Machines (SVM), designed specifically for regression tasks. Unlike traditional regression models that aim to minimize the error between predicted and actual values, SVR focuses on finding a function that approximates the data within a certain threshold, known as the epsilon-insensitive zone. This approach ensures that only significant deviations from the predicted function are penalized, thus offering robust performance even in the presence of noise. SVR employs kernel functions to map input features into a higher-dimensional space where a linear regression can be performed more effectively, thereby capturing complex, non-linear relationships between the input variables and the target. The flexibility of choosing different kernel functions, such as linear, polynomial, and radial basis function (RBF), allows SVR to adapt to a variety of data structures and distributions. Hyperplane is a separation line between two data classes in a higher dimension than the actual dimension as shown in Fig 4.1. In SVR it is defined as the line that helps in predicting the target value. By maximizing the margin between the support vectors (the critical data points that define the regression line)

and the epsilon boundary, SVR minimizes overfitting and enhances generalization to unseen data. This is particularly beneficial in high-dimensional spaces where traditional regression models often struggle. One of the key advantages of SVR is its ability to handle multicollinearity among predictors, a common issue in many real-world datasets. The model's robustness to outliers further ensures reliable predictions by focusing on the global trend rather than local fluctuations. In practical applications, SVR has been widely adopted across various domains, including financial time series forecasting, where it predicts stock prices with impressive accuracy, and bioinformatics, where it models complex biological interactions. In engineering, SVR is used to predict outcomes in manufacturing processes, ensuring optimal quality control. The implementation of SVR involves selecting appropriate hyperparameters, such as the regularization parameter (C), the kernel type, and the epsilon value, which can significantly impact the model's performance. Grid search and cross-validation techniques are commonly employed to fine-tune these parameters, ensuring the model's robustness and reliability. Despite its computational intensity, SVR's ability to balance bias and variance makes it a powerful tool for regression analysis. Furthermore, advancements in computational power and optimization algorithms have mitigated the computational challenges, making SVR more accessible for large-scale applications. The theoretical foundation of SVR is grounded in convex optimization, ensuring that a unique global solution is achieved, unlike neural networks that might converge to local minima. This mathematical rigor provides a solid basis for SVR's reliability and effectiveness. Moreover, the interpretability of SVR models, particularly when using linear kernels, offers insights into the underlying relationships within the data, making it a preferred choice for researchers and practitioners who require both accuracy and explainability. In summary, Support Vector Regression

stands out as a versatile, robust, and theoretically sound method for regression tasks, capable of handling a wide range of data complexities and ensuring high predictive accuracy across various fields. Its unique approach to minimizing error within a defined threshold, combined with the flexibility of kernel functions, positions SVR as a powerful tool for modern data analysis and predictive modeling (Sharp, 2020).

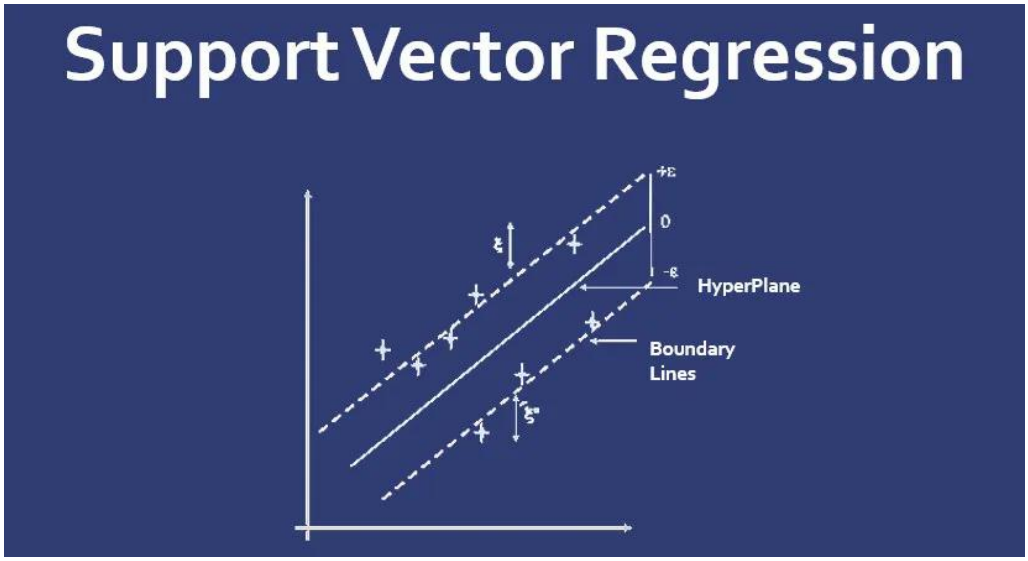

Fig 4.1 Support Vector Regression

4.2 Decision Tree Regression (DTR)

Decision Tree Regression (DTR) is a powerful and intuitive machine learning algorithm used for predicting continuous values by recursively partitioning the data space into smaller, homogenous regions. Unlike linear regression models that assume a linear relationship between the independent and dependent variables, DTR can model complex, non-linear relationships by splitting the data into subsets based on the values of input features. Each node in the decision tree represents a decision point as shown in FIg 4.2, where the dataset is divided into two branches according to a threshold value of a selected feature that minimizes the mean squared error (MSE) or another suitable loss function. The leaves of the tree represent the predicted values,

which are the average outcomes of the observations in those regions. This hierarchical structure of DTR allows for easy visualization and interpretation, making it a valuable tool for understanding and explaining the underlying patterns in the data. Additionally, DTRs handle both numerical and categorical data and are robust to outliers since the splits are based on the median of the data rather than being influenced by extreme values. They also do not require feature scaling, which simplifies preprocessing. Despite these advantages, decision trees can be prone to overfitting, especially when they are deep and complex, capturing noise in the training data. Techniques such as pruning, where parts of the tree that provide little power in predicting the target variable are removed, and setting constraints like maximum depth or minimum samples per leaf, can help mitigate overfitting. Furthermore, ensemble methods like Random Forest and Gradient Boosting have been developed to enhance the performance of DTR by aggregating the predictions of multiple trees, thereby reducing variance and improving generalization to new data. Random Forest, for instance, builds multiple decision trees using random subsets of features and data samples and averages their predictions, leading to a more robust model. Gradient Boosting, on the other hand, builds trees sequentially, where each tree tries to correct the errors of its predecessor, optimizing the performance iteratively. These ensemble techniques harness the strengths of individual decision trees while addressing their weaknesses, resulting in state-of-the-art predictive performance. In practical applications, DTR and its ensemble variants are widely used across various fields. In finance, they predict stock prices and assess credit risk; in healthcare, they are employed to predict patient outcomes and identify risk factors for diseases; in marketing, they help segment customers and forecast sales. The ease of implementation and interpretability of decision trees make them an attractive choice

for both researchers and practitioners. Tools and libraries like Scikit-learn in Python provide user-friendly interfaces for implementing DTR and tuning hyperparameters, facilitating their application in real-world scenarios. Despite their simplicity, the theoretical underpinnings of decision trees are grounded in solid mathematical concepts of entropy and information gain, ensuring their reliability as predictive models. In summary, Decision Tree Regression is a versatile and interpretable method for regression tasks that excels in capturing complex relationships in data without assuming any specific form of the relationship. Its susceptibility to overfitting can be effectively managed through pruning and ensemble methods, which enhance its predictive power and robustness. The widespread use of DTR in various industries attests to its effectiveness and practicality, cementing its place as a fundamental tool in the machine learning arsenal. The continuous advancements in ensemble learning and optimization algorithms further augment the capabilities of DTR, ensuring its relevance and utility in tackling increasingly complex predictive modeling challenges (Prasad, 2021).

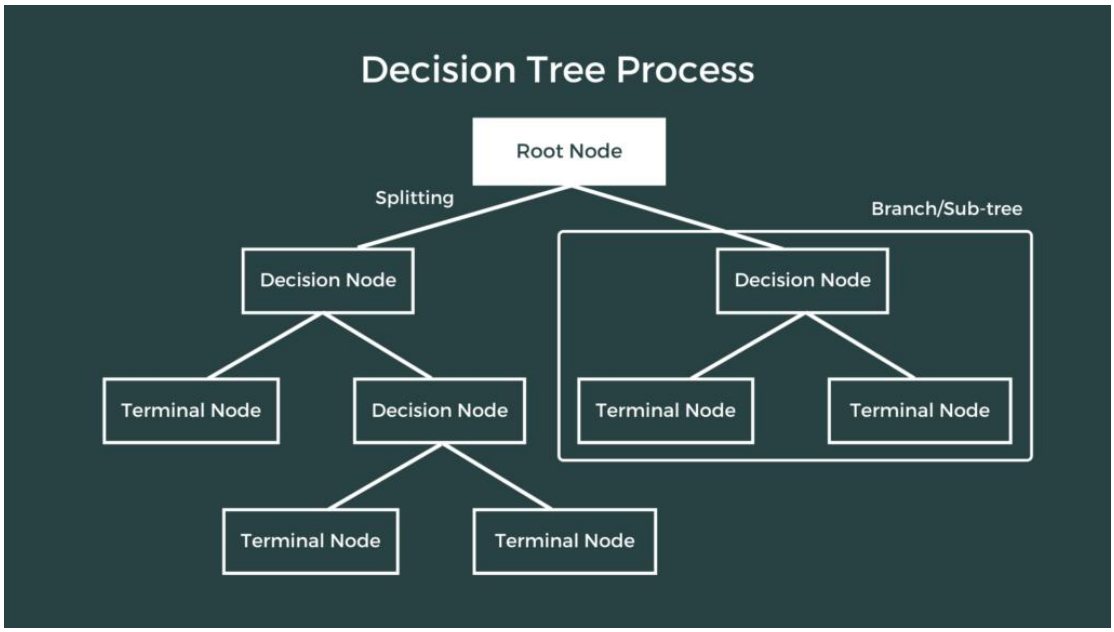

Fig 4.2 Decision Tree Regression

4.3 Random Forest Regression (RFR)

Random Forest Regression is a versatile machine-learning technique for predicting numerical values. It combines the predictions of multiple decision trees to reduce overfitting and improve accuracy. Python's machine-learning libraries make it easy to implement and optimize this approach.

Ensemble Learning

Ensemble learning is a machine learning technique that combines the predictions from multiple models to create a more accurate and stable prediction. It is an approach that leverages the collective intelligence of multiple models to improve the overall performance of the learning system.

Types of Ensemble Methods

There are various types of ensemble learning methods, including:

1. **Bagging (Bootstrap Aggregating):** This method involves training multiple models on random subsets of the training data. The predictions from the individual models are then combined, typically by averaging.
2. **Boosting:** This method involves training a sequence of models, where each subsequent model focuses on the errors made by the previous model. The predictions are combined using a weighted voting scheme.
3. **Stacking:** This method involves using the predictions from one set of models as input features for another model. The final prediction is made by the second-level model.

Random Forest

A random forest is an ensemble learning method that combines the predictions from multiple decision trees to produce a more accurate and stable prediction. It is a type

of supervised learning algorithm that can be used for both classification and regression tasks.

Every decision tree has high variance, but when we combine all of them in parallel then the resultant variance is low as each decision tree gets perfectly trained on that particular sample data, and hence the output doesn't depend on one decision tree but on multiple decision trees. In the case of a classification problem, the final output is taken by using the majority voting classifier. In the case of a regression problem, the final output is the mean of all the outputs. This part is called **Aggregation**.

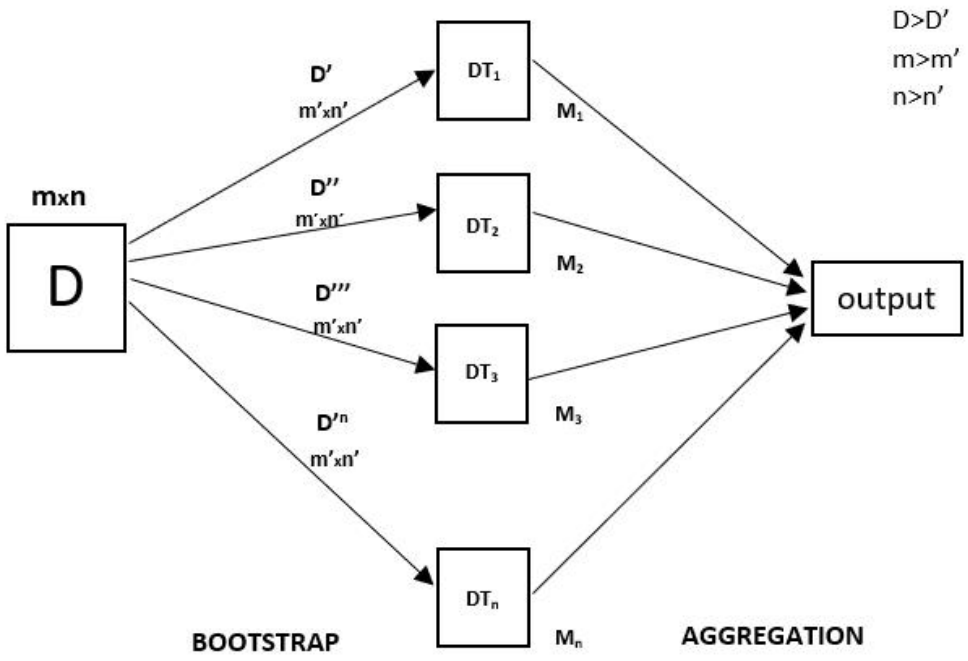

Fig 4.3 Random Forest Regression Model Working

What is Random Forest Regression?

Random Forest Regression in machine learning is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap (Dutta, 2019).

4.4 Partial Least Squares Regression (PLSR)

Partial Least Squares (PLS) Regression is a powerful statistical method used for modeling relationships between sets of observed variables, particularly when the predictor matrix has more variables than observations and those variables are highly collinear. Developed in the 1960s by Herman Wold, PLS regression is particularly useful in scenarios where traditional regression techniques fail due to multicollinearity or when the dataset contains missing values. PLS works by projecting the predictor variables and the response variables into a new space, finding linear combinations of the original variables that capture the most significant variance while also ensuring these new variables (called latent variables) have the highest possible correlation with the response variables. This method balances the goals of explaining the variance in the predictors and the covariance between predictors and responses. The first step in PLS regression involves standardizing the data, followed by the iterative extraction of the latent variables. Each extracted component is orthogonal to the others, ensuring independence and minimizing redundancy in the model. Unlike Principal Component Regression (PCR), which focuses solely on variance within the predictors, PLS regression ensures that the components extracted are those that are most relevant for predicting the response variables, hence improving predictive accuracy. This dual focus on variance and correlation makes PLS particularly suitable for chemometrics, genomics, and other fields dealing with high-dimensional data. Model evaluation in PLS involves cross-validation techniques to determine the optimal number of

components, preventing overfitting and ensuring generalizability. Furthermore, the interpretability of PLS models is enhanced through the inspection of variable importance in projection (VIP) scores, which indicate the contribution of each predictor to the model. This approach enables researchers to not only make accurate predictions but also to understand the underlying relationships within the data. Advanced versions of PLS, such as PLS-DA (Discriminant Analysis) and sparse PLS, extend its utility to classification problems and variable selection, respectively. The robustness of PLS regression against noise and missing data, combined with its flexibility and interpretability, underscores its enduring popularity across various scientific disciplines. Despite its advantages, PLS regression does have limitations, including the potential for overfitting if not properly validated and the assumption of a linear relationship between the latent variables and the responses.

4.5. K-Nearest Neighbors Regression (KNNR)

K-Nearest Neighbor (KNN) regression is a non-parametric, instance-based learning algorithm used for predicting continuous outcomes. Unlike traditional regression methods that assume a specific form for the relationship between input features and the target variable, KNN regression makes predictions based on the similarity of input features in the training dataset. The algorithm works by identifying the 'k' closest data points (neighbors) in the feature space to a given query point and then computing the average of their corresponding output values to make the prediction for the query point.

As an example, consider the following table of data points containing two features:

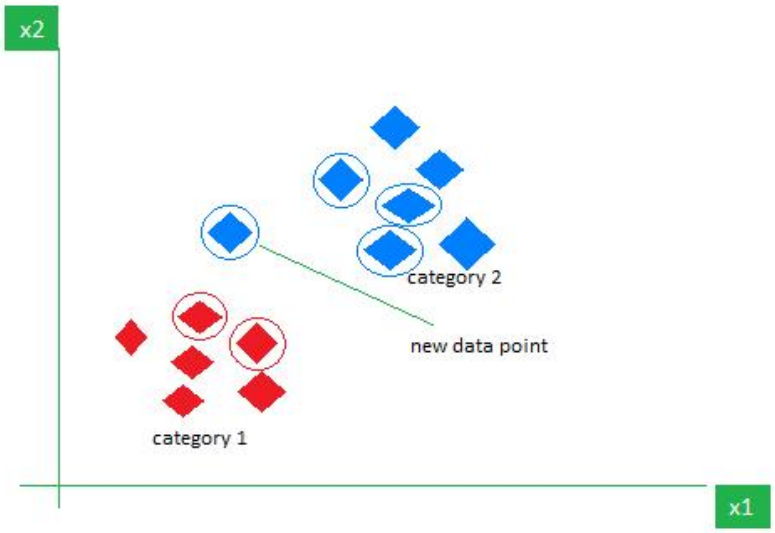

Fig 4.4 KNN Algorithm working visualization

Now, given another set of data points (also called testing data), allocate these points to a group by analyzing the training set.

(K-NN) algorithm is a versatile and widely used machine learning algorithm that is primarily used for its simplicity and ease of implementation. It does not require any assumptions about the underlying data distribution. It can also handle both numerical and categorical data, making it a flexible choice for various types of datasets in classification and regression tasks. It is a non-parametric method that makes predictions based on the similarity of data points in a given dataset. K-NN is less sensitive to outliers compared to other algorithms.

The K-NN algorithm works by finding the K nearest neighbors to a given data point based on a distance metric, such as Euclidean distance. The class or value of the data point is then determined by the majority vote or average of the K neighbors. This approach allows the algorithm to adapt to different patterns and make predictions based on the local structure of the data.

Distance Metrics Used in KNN Algorithm

As we know that the KNN algorithm helps us identify the nearest points or the groups for a query point. But to determine the closest groups or the nearest points for a query point we need some metric. For this purpose, we use below distance metrics:

Euclidean Distance

This is nothing but the Cartesian distance between the two points which are in the plane/hyperplane. Euclidean distance can also be visualized as the length of the straight line that joins the two points which are into consideration. This metric helps us calculate the net displacement done between the two states of an object.

$$\text{distance}(x, X_i) = \sqrt{\sum_{j=1}^{d} (x_j - X_{i_j})^2}]$$

Manhattan Distance

Manhattan Distance metric is generally used when we are interested in the total distance traveled by the object instead of the displacement. This metric is calculated by summing the absolute difference between the coordinates of the points in n-dimensions.

$$d\left(x, y\right) = \sum_{i=1}^{n} \left|x_i - y_i\right|$$

Minkowski Distance

We can say that the Euclidean, as well as the Manhattan distance, are special cases of the Minkowski distance.

$$d\left(x, y\right) = \left(\sum_{i=1}^{n} \left(x_i - y_i\right)^p\right)^{\frac{1}{p}}$$

From the formula above we can say that when p = 2 then it is the same as the formula for the Euclidean distance and when p = 1 then we obtain the formula for the Manhattan distance.

The above-discussed metrics are most common while dealing with a Machine Learning problem but there are other distance metrics as well like Hamming Distance which come in handy while dealing with problems that require overlapping comparisons between two vectors whose contents can be Boolean as well as string values (GeeksforGeeks, 2018).

How to choose the value of k for KNN Algorithm?

The value of k is very crucial in the KNN algorithm to define the number of neighbors in the algorithm. The value of k in the k-nearest neighbors (k-NN) algorithm should be chosen based on the input data. If the input data has more outliers or noise, a higher value of k would be better. It is recommended to choose an odd value for k to avoid ties in classification. Cross-validation methods can help in selecting the best k value for the given dataset.

Workings of KNN algorithm

The K-Nearest Neighbors (KNN) algorithm operates on the principle of similarity, where it predicts the label or value of a new data point by considering the labels or values of its K nearest neighbors in the training dataset.

Step 1: Selecting the optimal value of K

- K represents the number of nearest neighbors that needs to be considered while making prediction.

Step 2: Calculating distance

- To measure the similarity between target and training data points, Euclidean distance is used. Distance is calculated between each of the data points in the dataset and target point.

Step 3: Finding Nearest Neighbors

- The k data points with the smallest distances to the target point are the nearest neighbors.

Step 4: Voting for Classification or Taking Average for Regression

- In the classification problem, the class labels of are determined by performing majority voting. The class with the most occurrences among the neighbors becomes the predicted class for the target data point.

C H A P T E R 5

Results

In this study, we utilized Near Infrared (NIR) spectroscopy to identify the presence of formalin in Jersey Cow milk, Buffalo Milk and Combining the datasets of both the milks samples at differing concentrations. A spectrophotometer was employed to capture the essential spectral data. The subsequent analysis involved the application of five distinct predictive models to evaluate the presence and concentration of formalin within the milk samples. The ML model was then trained and cross validated 10 times, with each iteration using a different part as the validation set and the remaining parts for training. This technique was employed to ensure a robust performance evaluation of the model. The outputs from these models were evaluated based on key metrics such as the R-squared (R^2) score, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). Through these results, we can better understand the utility of this method in ensuring the safety and quality of milk, potentially leading to improved food safety practices and regulations.

R^2 score measures how well a regression model explains the variability in the dependent variable. It ranges from 0 to 1, with 1 indicating perfect prediction and 0 indicating no predictive power.

$$\text{Coefficient of determination } (R^2) = 1 - \frac{\text{sum of squares of residuals}}{\text{total sum of squares}}$$

RMSE measures the average distance between predicted and actual values. It quantifies the standard deviation of the errors or residuals, and is calculated from the squared differences between predictions and actual observations.

$$\text{RMSE (ml v/v)} = \sqrt{\frac{\sum_{i=1}^{n} (\bar{Y}_i - Y_i)^2}{n}}$$

MAE measures the average magnitude of the errors made in a set of predictions, irrespective of their direction. It is obtained by averaging the absolute differences between predicted and actual values. A lower MAE indicates a more accurate model, with zero indicating perfect prediction.

$$\mathbf{MAE\ (ml\ v/v)} = \frac{1}{n}\sum_{i=1}^{n} |\bar{Y}_i - Y_i|$$

In this study, we applied several reprocessing techniques to enhance the quality of our spectral data and improve the performance of our predictive models. Reprocessing method includes the SG filter designed to correct or reduce noise and other distortions in the data. PCA is a dimensionality reduction technique that transforms data into a set of orthogonal components capturing the most variance which was applied to the dataset.

Spectral readings were obtained from each sample, and these data were analyzed using five different predictive models.

Using SG- Filter, we observed a significant enhancement in the accuracy of the predictive models, as demonstrated by key metrics R^2, RMSE and MAE. These results indicate that the SG Filter reprocessing technique was effective in refining the dataset, contributing to more reliable detection of formalin in milk samples. The following sections will present the detailed results of these analyses, underscoring the impact of SG-Filter pre-processing on the overall performance of our models.

5.1 Buffalo Milk

Buffalo milk samples were intentionally adulterated with formalin at varying levels to assess the efficacy of Near Infrared (NIR) spectroscopy in detecting this chemical. The adulteration percentages ranged from 0% to 5% in 0.5% increments, followed by larger increments of 10%, 20%, 30%, 40%, and 50% to test a wider spectrum of contamination. Figure 5.1 displays the spectra plot of buffalo milk, while Figure 5.2 shows the plot after the application of the Savitzky-Golay filter.

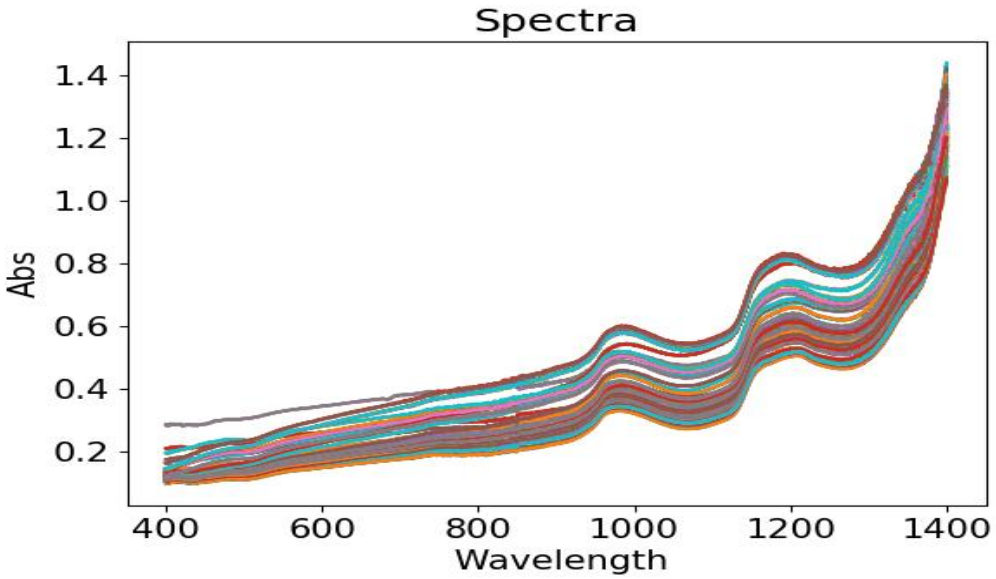

Fig 5.1:- Dataset for buffalo Milk

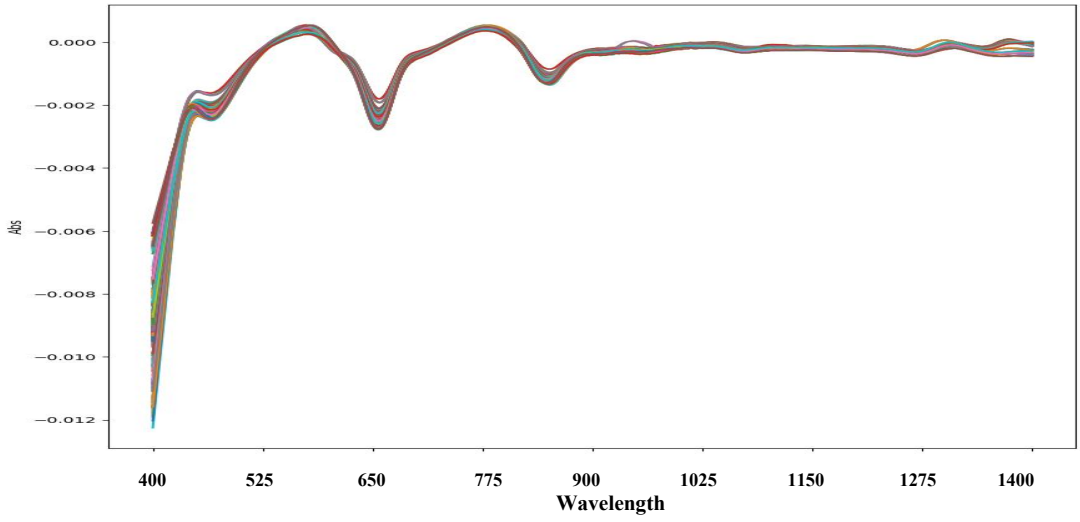

Fig 5.2 Savitzky Golay for Jersey Cow Milk

The Buffalo milk dataset was analyzed using five different models with 10 k-folds. Among these models KNN showed the best performance, as shown in the figures below. A detailed summary of all the results can be found in Table 5.1.

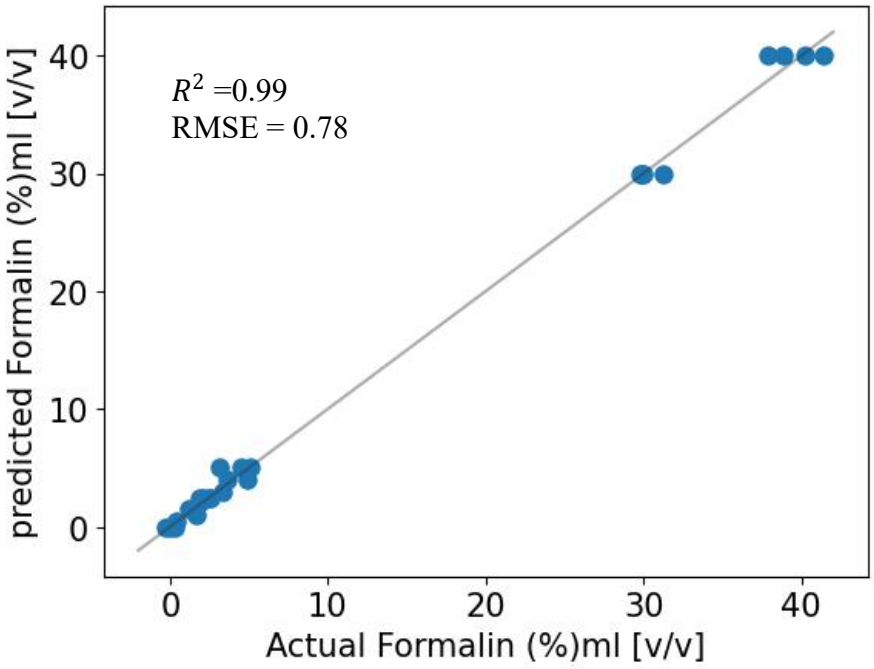

Fig 5.3 Graph for Buffalo Milk (SVR)

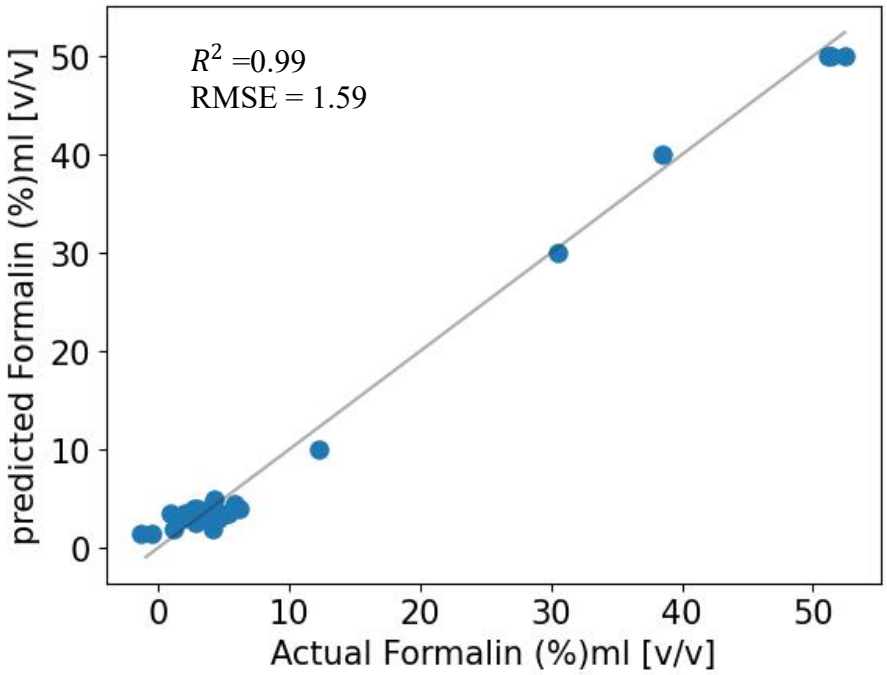

Fig 5.4 Graph for Buffalo Milk (PLSR)

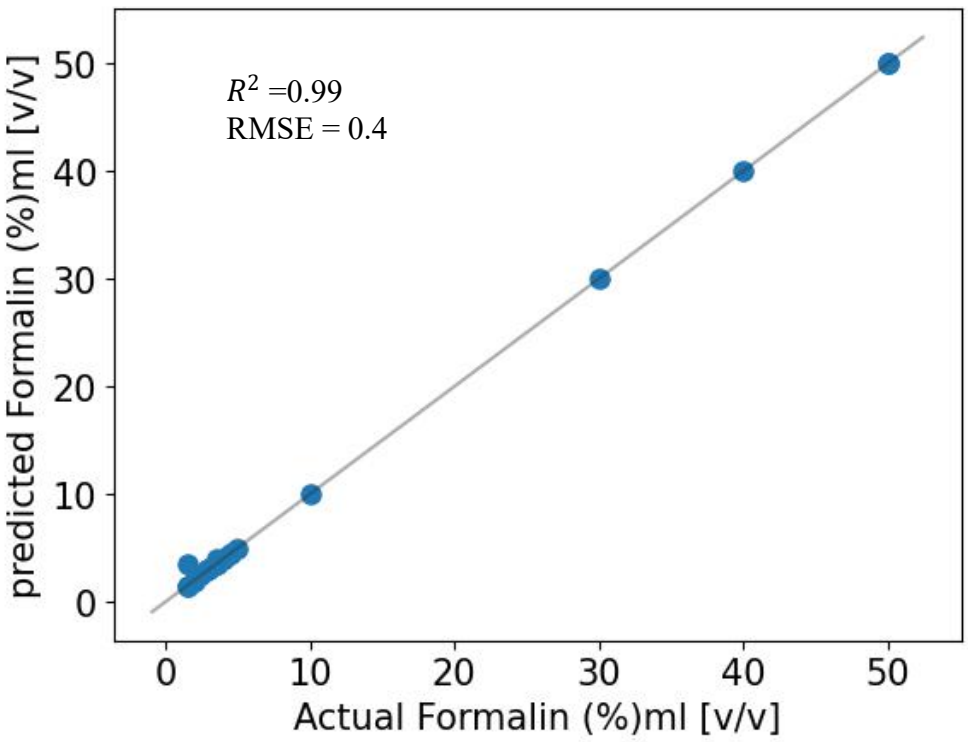

Fig 5.5 Graph for Buffalo Milk (DTR)

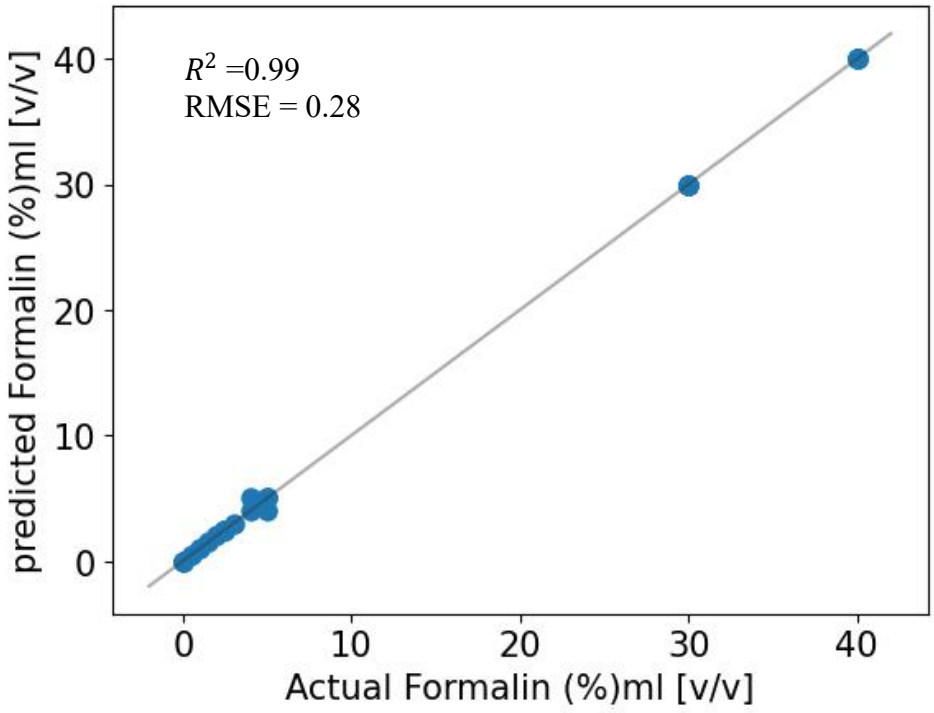

Fig 5.6 Graph for Buffalo Milk (KNR)

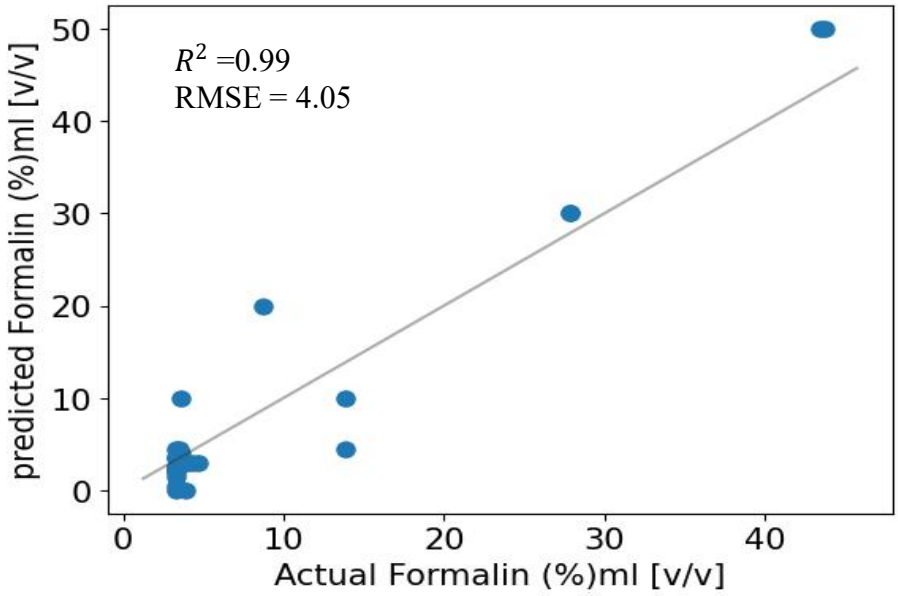

Fig 5.7 Graph for Buffalo Milk (RFR)

Table 5.1 performance of ML with SG performed spectra on Buffalo Milk

	Best fold				
Models	R^2	MSE	RMSE	MAE	Average RMSE
SVR	0.99	0.61	0.78	0.53	1.98
PLSR	0.99	2.54	1.59	1.43	2.15
DTR	0.99	0.16	0.4	0.09	3.72
KNR	0.99	0.08	0.28	0.08	1.08
RFR	0.91	16.45	4.05	2.86	6.60

5.2 Jersey Cow Milk

Jersey Cow Milk samples were intentionally adulterated with formalin at varying levels to assess the efficacy of Near Infrared (NIR) spectroscopy in detecting this chemical. The adulteration percentages ranged from 0% to 5% in 0.5% increments, followed by larger increments of 10%, 20%, 30%, 40%, and 50% to test a wider spectrum of contamination. Spectral readings were obtained from each sample fig 5.8 while Figure 5.9 shows the plot after the application of the Savitzky-Golay filter.

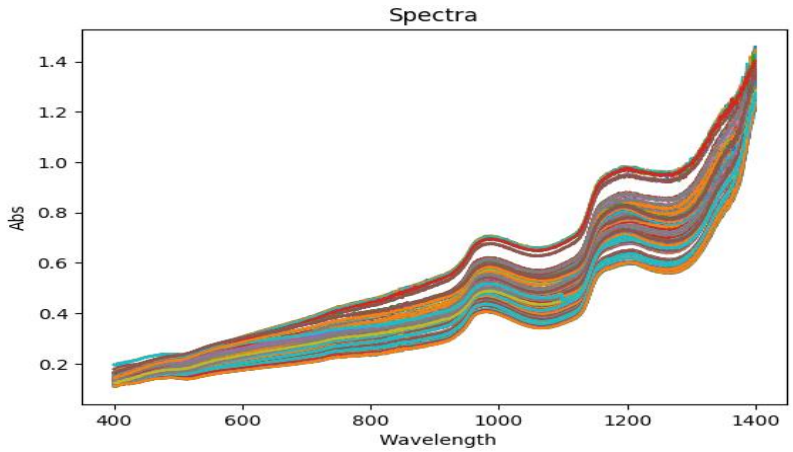

Fig 5.8 Plot for Jersey Cow Milk Dataset

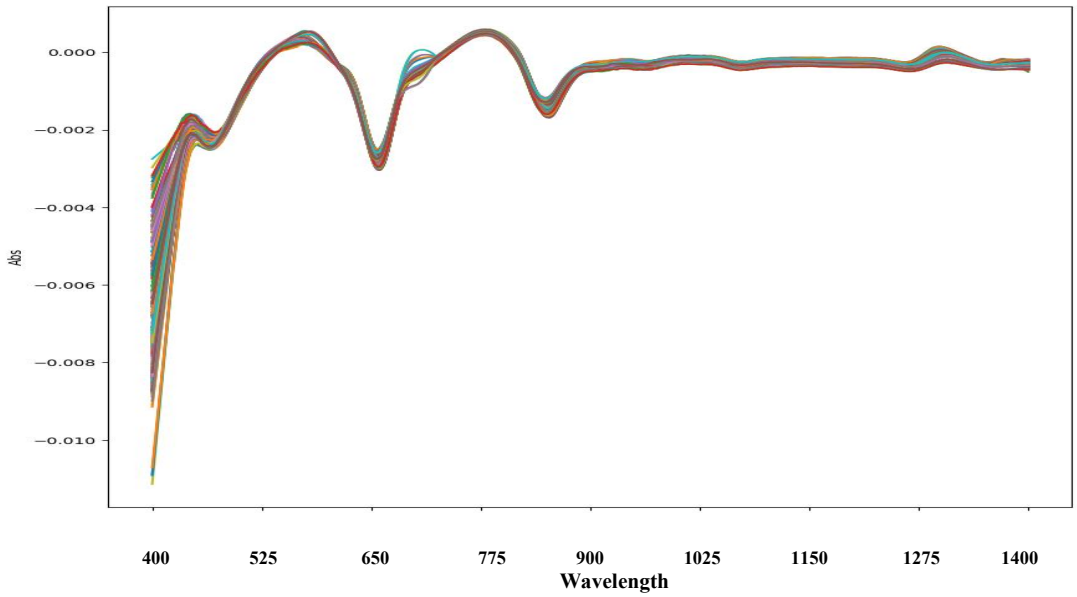

Fig 5.9 Savitzky Golay for Jersey Cow Milk

The Jersey Cow milk dataset was analyzed using five different models with 10 k-folds, and KNN showed the best performance among these models, as shown in the figures below. A detailed summary of all the results can be found in Table 5.2.

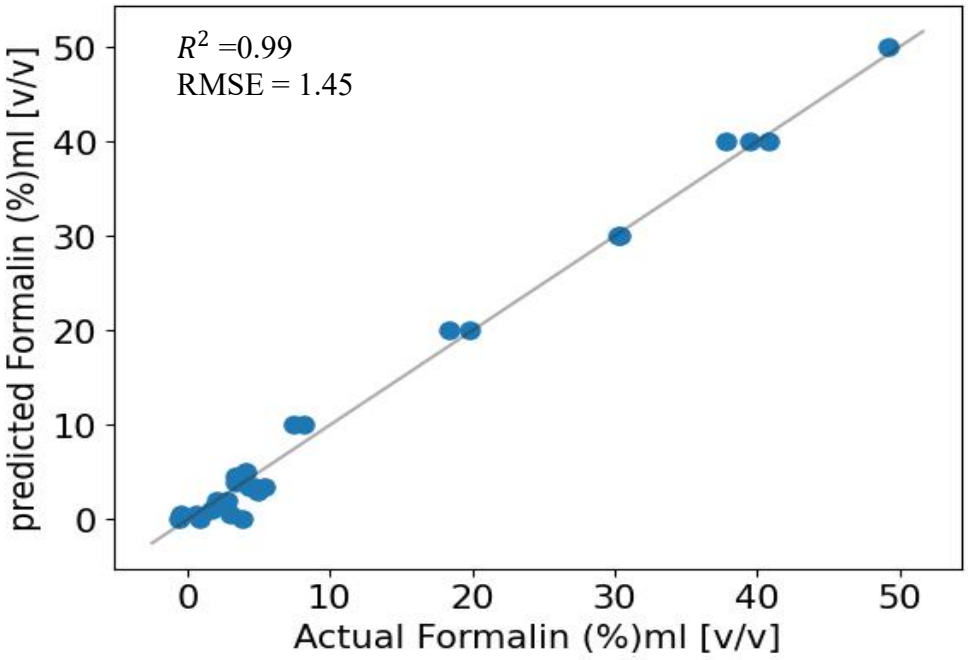

Fig 5.10 Graph for Jersey Cow Milk (SVR)

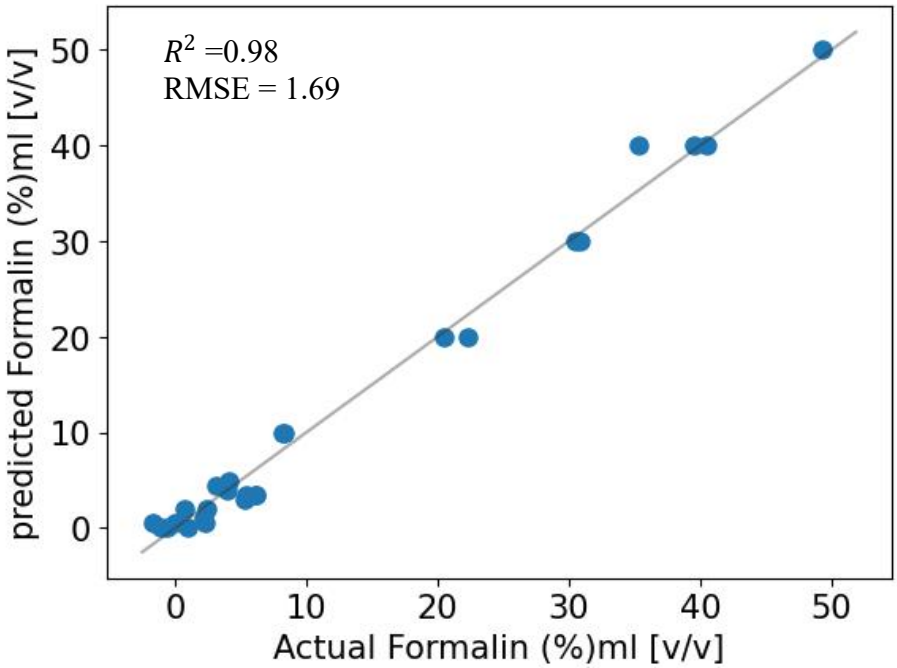

Fig 5.11 Graph for Jersey Cow Milk (PLSR)

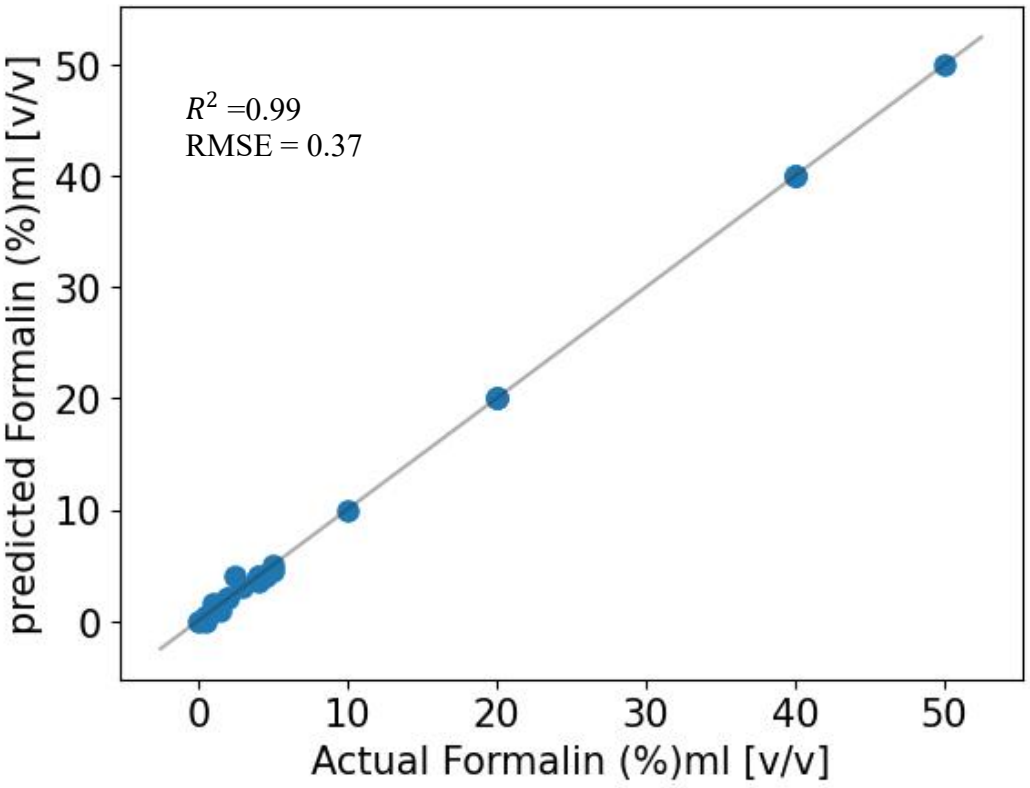

Fig 5.12 Graph for Jersey Cow Milk (DTR)

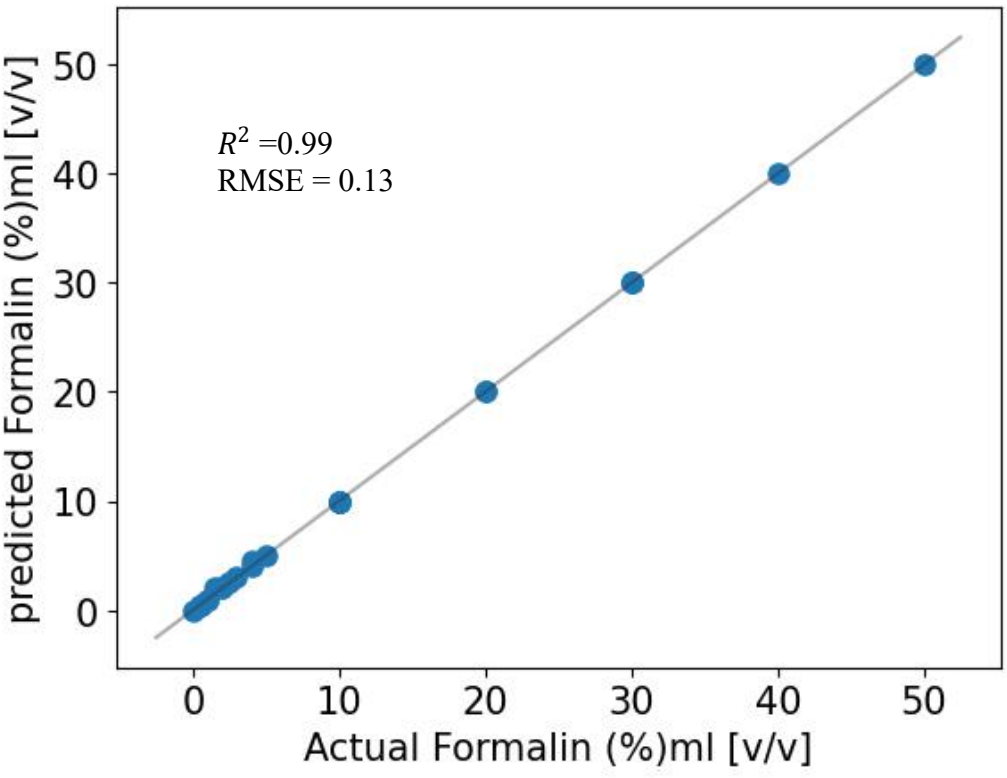

Fig 5.13 Graph for Jersey Cow Milk (KNR)

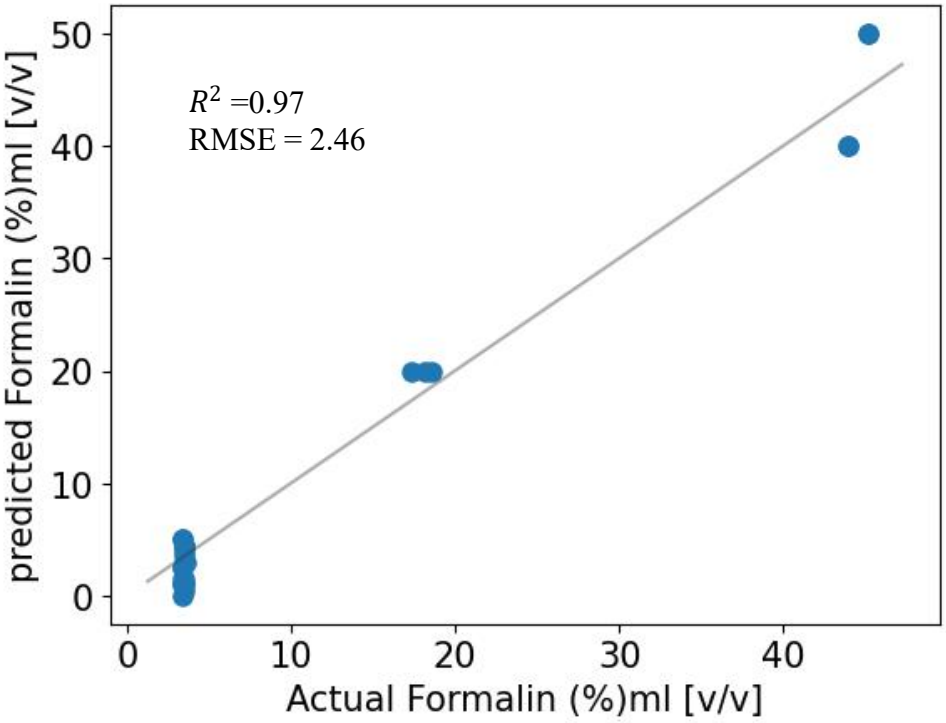

Fig 5.14 Graph for Jersey Cow Milk (KNR)

Table 5.2 performance of ML with SG performed spectra on Jersey Cow Milk

	Best fold				
Models	R^2	MSE	RMSE	MAE	Average RMSE
SVR	0.99	2.11	1.45	1.15	1.97
PLSR	0.98	2.88	1.69	1.37	2.18
DTR	0.99	0.14	0.37	0.17	2.25
KNR	0.99	0.01	0.13	0.03	0.48
RFR	0.97	6.06	2.46	2.10	4.38

5.3 Combination of two milks

Both Milk that is Cow milk and Jersey Cow Milk database were combined together which were intentionally adulterated with formalin at varying levels to assess the efficacy of Near Infrared (NIR) spectroscopy in detecting this chemical. The adulteration percentages ranged from 0% to 5% in 0.5% increments, followed by larger increments of 10%, 20%, 30%, 40%, and 50% to test a wider spectrum of contamination. Spectral readings were obtained from each sample fig 5.15 while Figure 5.16 shows the plot after the application of the Savitzky-Golay filter.

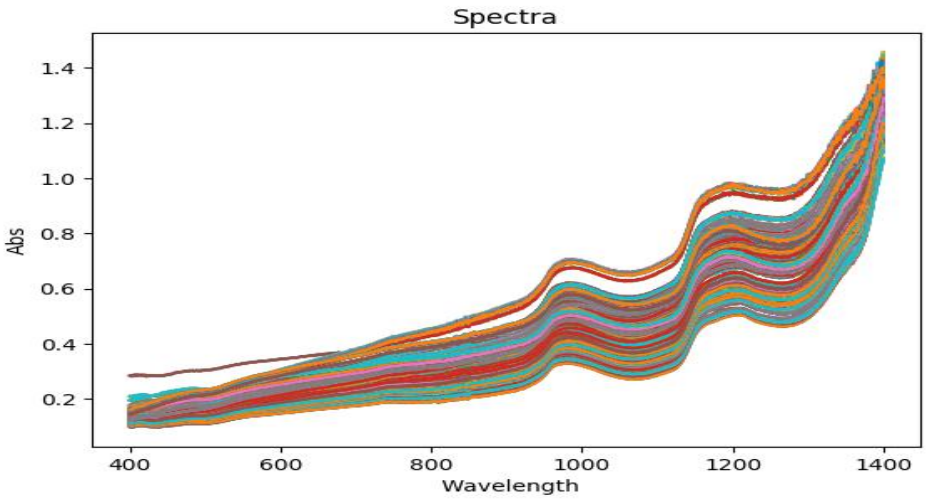

Fig 5.15 Plot for Datasets Combining Both Milks

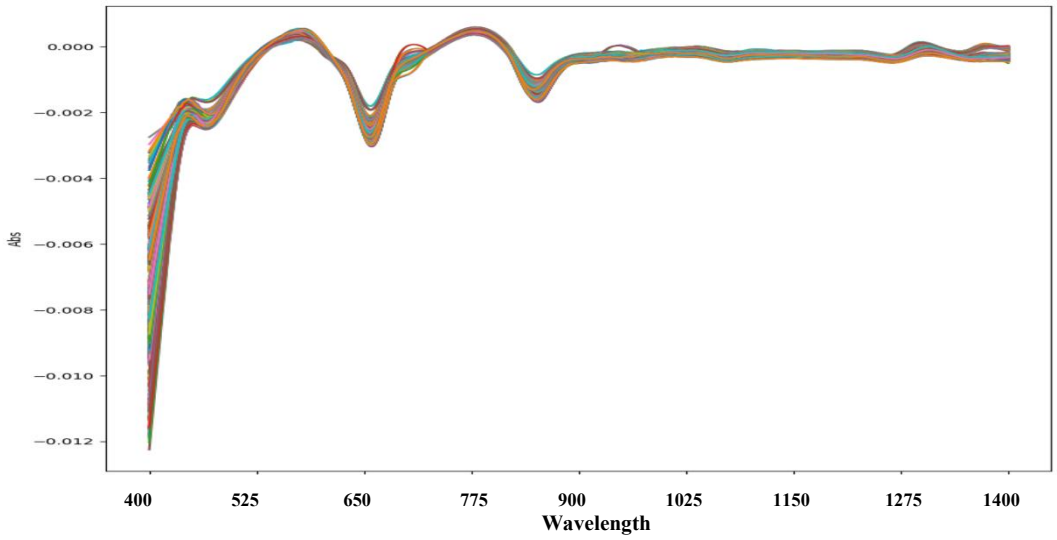

Fig 5.16 Savitzky Golay for Combination of Both the Milk

Five distinct models were utilized on the Combination of both the data sets with 10 k-folds, and KNN demonstrated superior performance among these five, as depicted in the figures below. Table 5.3 presents a comprehensive summary of all the results.

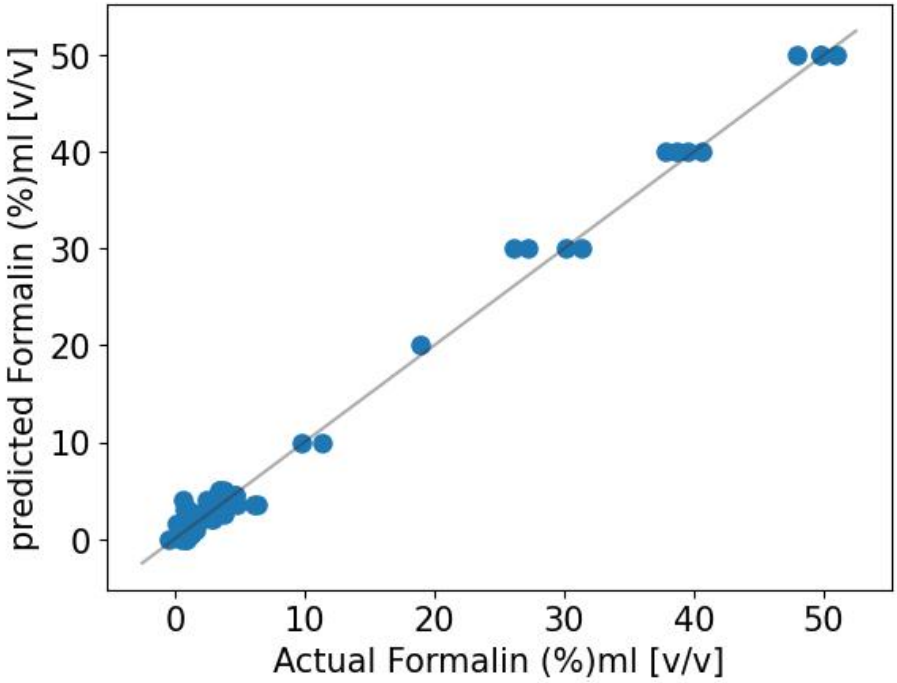

Fig 5.17 Graph for the combination (SVR)

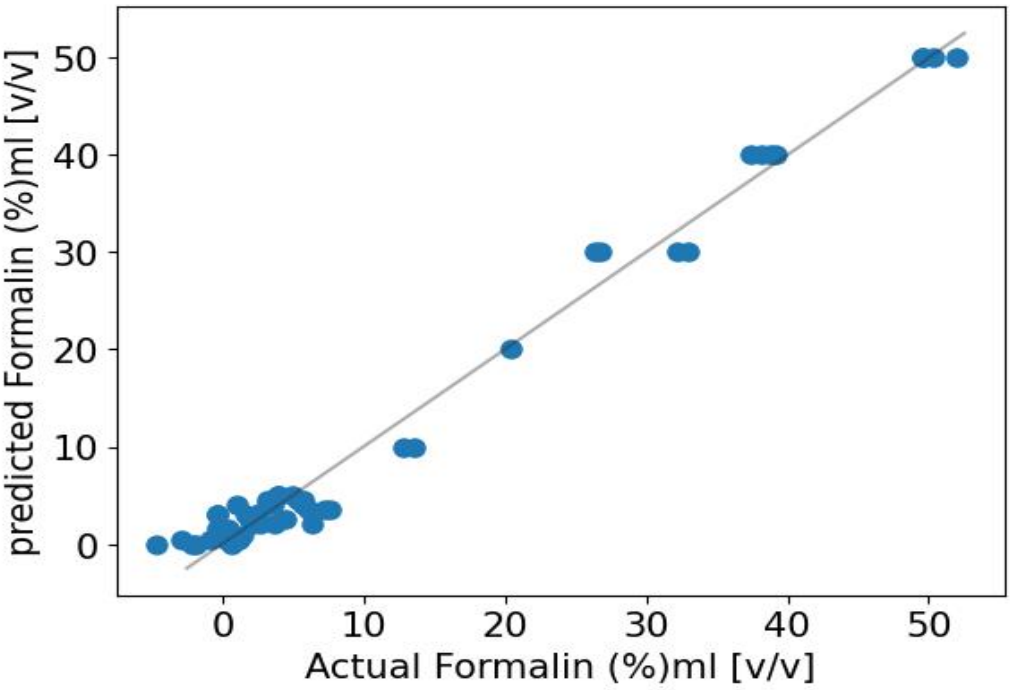

Fig 5.18 Graph for combination (PLSR)

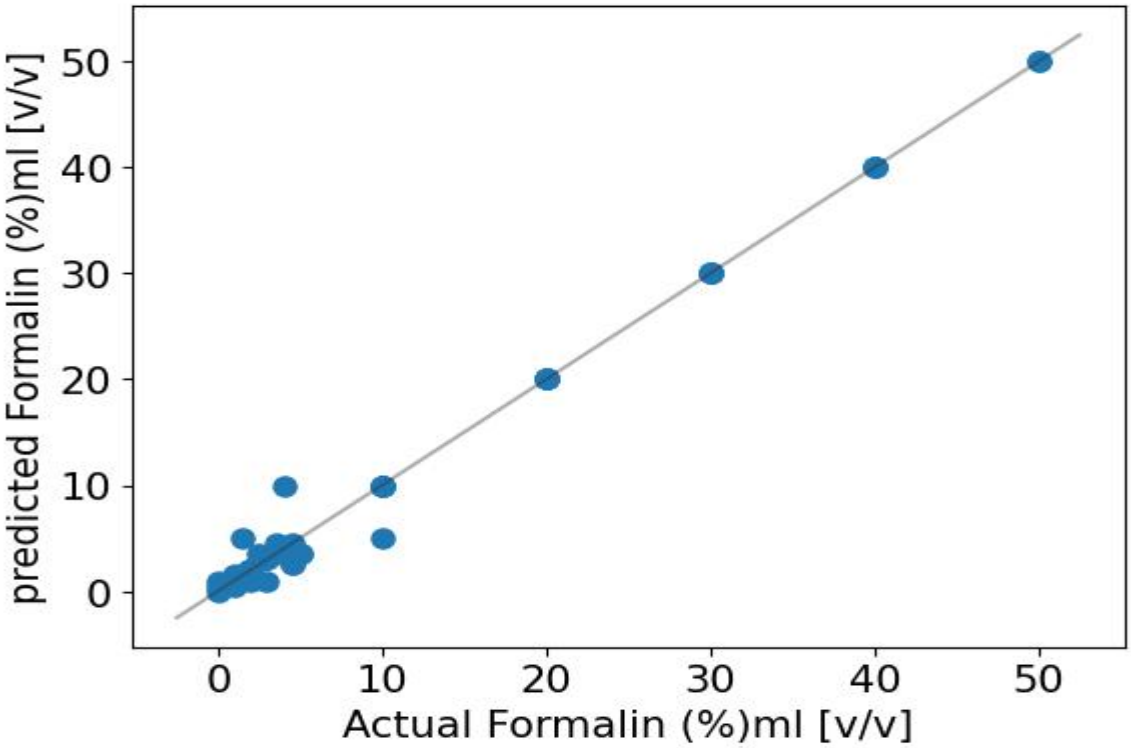

Fig 5.19 Graph for combination (DTR)

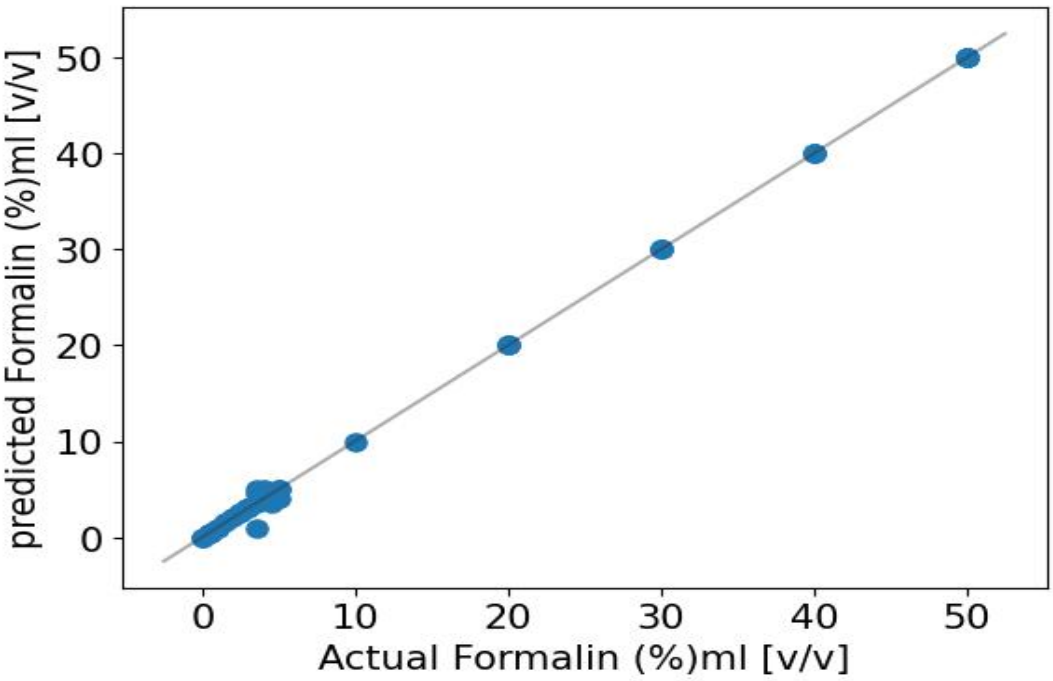

Fig 5.20 Graph for combination (KNR)

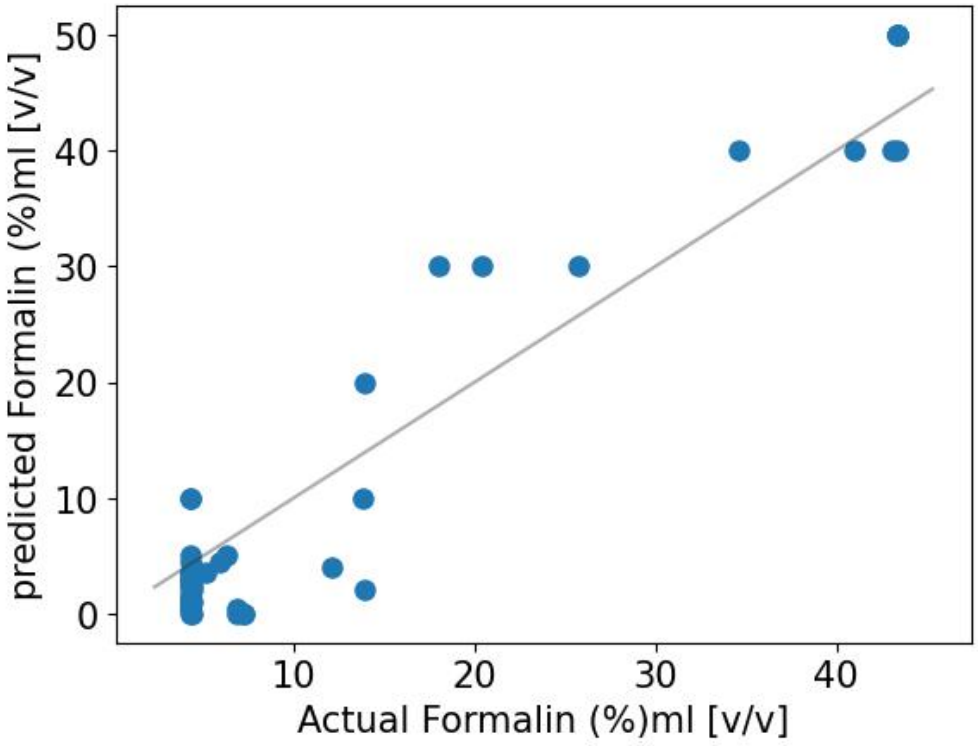

Fig 5.21 Graph for Combinations (RFR)

Table 5.3 performance of ML with SG performed spectra on the combination

	Best fold				
Models	R^2	MSE	RMSE	MAE	Average RMSE
SVR	0.99	1.93	1.39	1.06	2.12
PLSR	0.98	4.51	2.12	1.70	2.52
DTR	0.99	0.14	0.37	0.17	3.01
KNR	0.99	0.25	0.5	0. 18	1.06
RFR	0.97	6.06	2.46	2.10	5.78

\

5.4 Conclusion

The widespread issue of food fraud poses serious threats to consumer health and well-being, with common adulterants like water, starch, urea, detergents, and chemical preservatives leading to severe health impacts, including digestive issues, toxicity, and long-term diseases like cancer. To address the adulteration of milk with formalin, a non-destructive machine learning-based system using near-infrared (NIR) spectroscopy was developed. The research involved creating a comprehensive database by combining variable quantities of formalin in two types of milk and recording spectra using a Jasco spectrophotometer. The k-Nearest Neighbors (KNN) algorithm excels in detecting formalin in buffalo milk, Jersey cow milk, and their combination when analyzed using spectroscopy. After applying PCA to reduce the data to 5 principal components and utilizing the Savitzky-Golay filter with a window length of 91, polynomial order of 3, and first derivative, KNN consistently achieved the highest accuracy, precision, and recall among the evaluated machine learning models. KNN effectively identifies the presence of formalin, reduces false positives, and accurately classifies true negatives, proving its robustness in handling spectral data and detecting milk adulteration, thus making it the optimal choice for this application. KNN demonstrated superior performance for buffalo milk, Jersey cow milk, and the combination of both, with impressive R2, RMSE, and MAE values. Specifically, for buffalo milk, KNN achieved an R^2 of 0.999, RMSE of 0.28 ml (% v/v), and MAE of 0.08 ml (% v/v) with an average RMSE of 1.08. For Jersey cow milk, KNN resulted in an R^2 value of 0.999, RMSE of 0.13 ml (% v/v), and MAE of 0.03 ml (% v/v) with an average RMSE of 0.48. Additionally, for the combination of both milks, KNN attained an R^2 value of 0.999, RMSE of 0.5 ml (% v/v), and MAE of 0.18 ml (% v/v) with an average RMSE of 1.06.

References

Agharkar, M., & Mane, S. (2021). Utilization of gold nanoparticles to detect formalin adulteration in milk. *Materials Today: Proceedings*, *45*, 4421–4423. https://doi.org/10.1016/j.matpr.2020.12.233

Ai, K., Liu, Y., & Lu, L. (2009). Hydrogen-Bonding Recognition-Induced Color Change of Gold Nanoparticles for Visual Detection of Melamine in Raw Milk and Infant Formula. *Journal of the American Chemical Society*, *131*(27), 9496–9497. https://doi.org/10.1021/ja9037017

Albanell, E., Miñarro, B., & Carrasco, N. (2012). Detection of low-level gluten content in flour and batter by near infrared reflectance spectroscopy (NIRS). *Journal of Cereal Science*, *56*(2), 490–495. https://doi.org/10.1016/j.jcs.2012.06.011

Ambrose, A., & Cho, B.-K. (2014). A Review of Technologies for Detection and Measurement of Adulterants in Cereals and Cereal Products. *Journal of Biosystems Engineering*, *39*(4), 357–365. https://doi.org/10.5307/JBE.2014.39.4.357

Azad, T., & Ahmed, S. (2016). Common milk adulteration and their detection techniques. *International Journal of Food Contamination*, *3*(1). https://doi.org/10.1186/s40550-016-0045-3

Balan, B., Dhaulaniya, A. S., Jamwal, R., Amit, Sodhi, K. K., Kelly, S., Cannavan, A., & Singh, D. K. (2020). Application of Attenuated Total Reflectance-Fourier Transform Infrared (ATR-FTIR) spectroscopy coupled with chemometrics for detection and quantification of formalin in cow milk. *Vibrational Spectroscopy*, *107*, 103033. https://doi.org/10.1016/j.vibspec.2020.103033

Bezuayehu Gutema Asefa, Hagos, L., Kore, T., & Shimelis Admassu Emire. (2022). Feasibility of Image Analysis Coupled with Machine Learning for Detection and Quantification of Extraneous Water in Milk. *Food Analytical Methods*, *15*(11), 3092–3103. https://doi.org/10.1007/s12161-022-02352-w

Botelho, B. G., Reis, N., Oliveira, L. S., & Sena, M. M. (2015). Development and analytical validation of a screening method for simultaneous detection of five adulterants in raw milk using mid-infrared spectroscopy and PLS-DA. *Food Chemistry*, *181*, 31–37. https://doi.org/10.1016/j.foodchem.2015.02.077

Bunaciu, A. A., Aboul-Enein, H. Y., & Hoang, V. D. (2016). RETRACTED: Vibrational spectroscopy used in milk products analysis: A review. *Food Chemistry*, *196*, 877–884. https://doi.org/10.1016/j.foodchem.2015.10.016

Chakraborty, M., & Biswas, K. (2018). Limit of Detection for Five Common Adulterants in Milk: A Study With Different Fat Percent. *IEEE Sensors Journal*, *18*(6), 2395–2403. https://doi.org/10.1109/jsen.2018.2794764

Cheng, Y., Dong, Y., Wu, J., Yang, X., Bai, H., Zheng, H., Ren, D., Zou, Y., & Li, M. (2010). Screening melamine adulterant in milk powder with laser Raman spectrometry. *Journal of Food Composition and Analysis*, *23*(2), 199–202. https://doi.org/10.1016/j.jfca.2009.08.006

Choudhary, S., & Joshi, A. (2022). Development of an embedded system for real-time milk spoilage monitoring and adulteration detection. *International Dairy Journal*, *127*, 105207. https://doi.org/10.1016/j.idairyj.2021.105207

Cirak, O., Icyer, N. C., & Durak, M. Z. (2018). Rapid detection of adulteration of milks from different species using Fourier Transform Infrared Spectroscopy (FTIR). *Journal of Dairy Research*, *85*(2), 222–225. https://doi.org/10.1017/s0022029918000201

Cristian Olguín, Nicolás Laguarda-Miró, Pascual, L., García-Breijo, E., Ramón Martínez-Mañez, & Soto, J. (2014). An electronic nose for the detection of Sarin, Soman and Tabun mimics and interfering agents. *Sensors and Actuators. B, Chemical*, *202*, 31–37. https://doi.org/10.1016/j.snb.2014.05.060

D Maheswara Reddy, K Venkatesh , & C Venkata Sesha Reddy. (2017). Adulteration of milk and its detection: A review. *Chemijournal*.

Dai, X., Fang, X., Su, F., Yang, M., Li, H., Zhou, J., & Xu, R. (2010). Accurate analysis of urea in milk and milk powder by isotope dilution gas chromatography–mass spectrometry. *Journal of Chromatography B*, *878*(19), 1634–1638. https://doi.org/10.1016/j.jchromb.2010.04.005

Das, C., Chakraborty, S., Anupam Karmakar, & Chattopadhyay, S. (2018). *On-chip detection and quantification of soap as an adulterant in milk employing electrical impedance spectroscopy*. https://doi.org/10.1109/isdcs.2018.8379634

Dave, A., Banwari, D., Srivastava, S., & Sadistap, S. (2016). *Optical sensing system for detecting water adulteration in milk*. IEEE Xplore. https://doi.org/10.1109/GHTC.2016.7857345

de Carvalho, B. M. A., de Carvalho, L. M., dos Reis Coimbra, J. S., Minim, L. A., de Souza Barcellos, E., da Silva Júnior, W. F., Detmann, E., & de Carvalho, G. G. P. (2015). Rapid detection of whey in milk powder samples by spectrophotometric and multivariate calibration. *Food Chemistry*, *174*, 1–7. https://doi.org/10.1016/j.foodchem.2014.11.003

de Freitas Rezende, F. B., de Souza Santos Cheibub, A. M., Pereira Netto, A. D., & Marques, F. F. de C. (2017). Determination of formaldehyde in bovine milk

using a high sensitivity HPLC-UV method. *Microchemical Journal*, *134*, 383–389. https://doi.org/10.1016/j.microc.2017.07.003

De, S., & Jirankalgikar, N. (2014). Detection of tallow adulteration in cow ghee by derivative spectrophotometry. *Journal of Natural Science, Biology and Medicine*, *5*(2), 317. https://doi.org/10.4103/0976-9668.136174

Di Domenico, M., Di Giuseppe, M., Wicochea Rodríguez, J. D., & Cammà, C. (2017). Validation of a fast real-time PCR method to detect fraud and mislabeling in milk and dairy products. *Journal of Dairy Science*, *100*(1), 106–112. https://doi.org/10.3168/jds.2016-11695

Dutta, A. (2019, June 14). *Random Forest Regression in Python - GeeksforGeeks*. GeeksforGeeks. https://www.geeksforgeeks.org/random-forest-regression-in-python/

Ehsani, S., Dastgerdy, E. M., Yazdanpanah, H., & Parastar, H. (2022). Ensemble classification and regression techniques combined with portable near infrared spectroscopy for facile and rapid detection of water adulteration in bovine raw milk. *Journal of Chemometrics*. https://doi.org/10.1002/cem.3395

Etzion, Y., Linker, R., Cogan, U., & Shmulevich, I. (2004). Determination of Protein Concentration in Raw Milk by Mid-Infrared Fourier Transform Infrared/Attenuated Total Reflectance Spectroscopy. *Journal of Dairy Science*, *87*(9), 2779–2788. https://doi.org/10.3168/jds.s0022-0302(04)73405-0

Ewida, R. M., & El-Magiud, D. S. M. A. (2018). Species adulteration in raw milk samples using polymerase chain reaction-restriction fragment length polymorphism. *Veterinary World*, *11*(6), 830–833. https://doi.org/10.14202/vetworld.2018.830-833

Farzaneh Shalileh, Hossein Sabahi, Mehdi Dadmehr, & Hosseini, M. (2023). Sensing approaches toward detection of urea adulteration in milk. *Microchemical Journal*, *193*, 108990–108990. https://doi.org/10.1016/j.microc.2023.108990

Fernández Pierna, J. A., Vermeulen, P., Amand, O., Tossens, A., Dardenne, P., & Baeten, V. (2012). NIR hyperspectral imaging spectroscopy and chemometrics for the detection of undesirable substances in food and feed. *Chemometrics and Intelligent Laboratory Systems*, *117*, 233–239. https://doi.org/10.1016/j.chemolab.2012.02.004

Filazi, A., Sireli, U. T., Ekici, H., Can, H. Y., & Karagoz, A. (2012). Determination of melamine in milk and dairy products by high performance liquid chromatography. *Journal of Dairy Science*, *95*(2), 602–608. https://doi.org/10.3168/jds.2011-4926

Finete, V. de L. M., Gouvêa, M. M., Marques, F. F. de C., & Netto, A. D. P. (2013). Is it possible to screen for milk or whey protein adulteration with melamine, urea and ammonium sulphate, combining Kjeldahl and classical spectrophotometric methods? *Food Chemistry*, *141*(4), 3649–3655. https://doi.org/10.1016/j.foodchem.2013.06.046

Francis, A., Dhiman, T., & Mounya, K. S. (2020). *Adulteration of milk: A review. J. Sci. Technol, 5, 37-41*.

Garcia, J. S., Sanvido, G. B., Sérgio Henriques Saraiva, Jorge Jardim Zacca, Cosso, R. G., & Eberlin, M. N. (2012). Bovine milk powder adulteration with vegetable oils or fats revealed by MALDI-QTOF MS. *Food Chemistry*, *131*(2), 722–726. https://doi.org/10.1016/j.foodchem.2011.09.062

GeeksforGeeks. (2018, November 13). K-Nearest Neighbours - GeeksforGeeks. GeeksforGeeks. https://www.geeksforgeeks.org/k-nearest-neighbours/

Gupta, V. K., Aulakh, R. S., & Tomar, S. S. (2019). Novel Method for The Determination of Preservative (Formaldehyde) in Bovine Milk Through Smart Phone-Based Colorimetric Technology. *THE INDIAN JOURNAL of VETERINARY SCIENCES and BIOTECHNOLOGY*, *15*(02), 30–33. https://doi.org/10.21887/ijvsbt.15.2.8

Hazra, T., Sharma, V., Sharma, R., & Arora, S. (2017). A species specific simplex polymerase chain reaction-based approach for detection of goat tallow in heat clarified milk fat (ghee). *International Journal of Food Properties*, *20*(sup1), S69–S75. https://doi.org/10.1080/10942912.2017.1289542

Hemanth Singuluri, & Sukumaran, M. K. (2014). Milk Adulteration in Hyderabad, India – A comparative study on the levels of different adulterants present in milk. *Indian Journal of Dairy Science*, *68*(2), 190–192. https://doi.org/10.5146/ijds.v68i2.44300

Hop, E., H.-J. Luinge, & H. Van Hemert. (1993). Quantitative Analysis of Water in Milk by FT-IR Spectrometry. *Applied Spectroscopy*, *47*(8), 1180–1182. https://doi.org/10.1366/0003702934067865

Invest India. (n.d.). Dairy Industry In India - Growth, FDI, Companies, Exports. Www.investindia.gov.in. Retrieved November 28, 2023, from https://www.investindia.gov.in/sector/animal-husbandry-and-dairying/dairy#:~:text=India%20is%20the%20highest%20milk

Inaba, A., Yoo, G., Takei, Y., Matsumoto, K., & I. Shimoyama. (2013). *A graphene FET gas sensor gated by ionic liquid*. https://doi.org/10.1109/memsys.2013.6474408

Jablonski, J. E., Moore, J. C., & Harnly, J. M. (2014). Nontargeted Detection of Adulteration of Skim Milk Powder with Foreign Proteins Using UHPLC–UV.

Journal of Agricultural and Food Chemistry, *62*(22), 5198–5206. https://doi.org/10.1021/jf404924x

Jawaid, S., Talpur, F. N., Sherazi, S. T. H., Nizamani, S. M., & Khaskheli, A. A. (2013). Rapid detection of melamine adulteration in dairy milk by SB-ATR–Fourier transform infrared spectroscopy. *Food Chemistry*, *141*(3), 3066–3071. https://doi.org/10.1016/j.foodchem.2013.05.106

Kamboj, U., Kaushal, N., & Jabeen, S. (2020). Near Infrared Spectroscopy as an efficient tool for the Qualitative and Quantitative Determination of Sugar Adulteration in Milk. *Journal of Physics: Conference Series*, *1531*, 012024. https://doi.org/10.1088/1742-6596/1531/1/012024

Kaminski, J., Atwal, A. S., & Mahadevan, S. (1993). High Performance Liquid Chromatographic Determination of Formaldehyde in Milk. *Journal of Liquid Chromatography*, *16*(2), 521–526. https://doi.org/10.1080/10826079308020929

Kandpal, S. D., Srivastava, A. K., & Negi, K. S. (2012). *ESTIMATION OF QUALITY OF RAW MILK (OPEN & BRANDED) BY MILK ADULTERATION TESTING KIT. 24*(3), 188–192.

KASEMSUMRAN, S., THANAPASE, W., & KIATSOONTHON, A. (2007). Feasibility of Near-Infrared Spectroscopy to Detect and to Quantify Adulterants in Cow Milk. *Analytical Sciences*, *23*(7), 907–910. https://doi.org/10.2116/analsci.23.907

Kawasaki, M., Kawamura, S., Tsukahara, M., Morita, S., Komiya, M., & Natsuga, M. (2008). Near-infrared spectroscopic sensing system for on-line milk quality assessment in a milking robot. *Computers and Electronics in Agriculture*, *63*(1), 22–27. https://doi.org/10.1016/j.compag.2008.01.006

Khan, K. M., Krishna, H., Majumder, S. K., & Gupta, P. K. (2014). Detection of Urea Adulteration in Milk Using Near-Infrared Raman Spectroscopy. *Food Analytical Methods*, *8*(1), 93–102. https://doi.org/10.1007/s12161-014-9873-z

Kim, A., Barcelo, S. J., Williams, R. S., & Li, Z. (2012). Melamine Sensing in Milk Products by Using Surface Enhanced Raman Scattering. *Analytical Chemistry*, *84*(21), 9303–9309. https://doi.org/10.1021/ac302025q

Lanjewar, M. G., Parab, J. S., & Kamat, R. K. (2024). Machine Learning based Technique to Predict the Water Adulterant in Milk Using Portable Near Infrared Spectroscopy. *Journal of Food Composition and Analysis*, 106270–106270. https://doi.org/10.1016/j.jfca.2024.106270

Lindmark-Månsson, H., & Åkesson, B. (2000). Antioxidative factors in milk. *British Journal of Nutrition*, *84*(S1), 103–110. https://doi.org/10.1017/s0007114500002324

Lucas, Fabio Augusto Gentilin, Alexandre, J., Lúcia, A., & Bernadete, M. (2016). Development of a Hardware Platform for Detection of Milk Adulteration Based on Near-Infrared Diffuse Reflection. *IEEE Transactions on Instrumentation and Measurement*, *65*(7), 1698–1706. https://doi.org/10.1109/tim.2016.2540946

Lucas, Silva, Lúcia, A., & Alexandre, J. (2018). A NIR Photometer Prototype With Integrating Sphere for the Detection of Added Water in Raw Milk. *IEEE Transactions on Instrumentation and Measurement*, *67*(12), 2812–2819. https://doi.org/10.1109/tim.2018.2829398

M. Czauderna, & Kowalczyk, J. (2009). Easy and accurate determination of urea in milk, blood plasma, urine and selected diets of mammals by high-performance

liquid chromatography with photodiode array detection preceded by pre-column derivatization. *Chemia Analityczna*, *54*(5), 919–937.

Mabood, F., Hussain, J., MOO, A. N., Gilani, S. A., Farooq, S., Naureen, Z., Jabeen, F., Ahmed, M., Hussain, Z., & Harrasi, A. A. (2017). Detection and Quantification of Formalin Adulteration in Cow Milk Using Near Infrared Spectroscopy Combined with Multivariate Analysis. *Advances in Dairy Research*, *05*(01). https://doi.org/10.4172/2329-888x.1000167

Mabrook, M. F., & Petty, M. C. (2003). A novel technique for the detection of added water to full fat milk using single frequency admittance measurements. *Sensors and Actuators B: Chemical*, *96*(1-2), 215–218. https://doi.org/10.1016/s0925-4005(03)00527-6

Maha Ibrahim Alkhalf, & Elwathig, M. (2017). *Detection of formaldehyde in cheese using FTIR spectroscopy. 24*.

Matabaro, E., Ishimwe, N., Uwimbabazi, E., & Lee, B. H. (2017). Current Immunoassay Methods for the Rapid Detection of Aflatoxin in Milk and Dairy Products. *Comprehensive Reviews in Food Science and Food Safety*, *16*(5), 808–820. https://doi.org/10.1111/1541-4337.12287

Mailagaha Kumbure, M., & Luukka, P. (2021). A generalized fuzzy k-nearest neighbor regression model based on Minkowski distance. *Granular Computing*. https://doi.org/10.1007/s41066-021-00288-w

Mathaweesansurn, A., & Detsri, E. (2022). A new colorimetric method for determination of formaldehyde in sea food based on anti-aggregation of gold nanoparticles. *Journal of Food Composition and Analysis*, *114*, 104802. https://doi.org/10.1016/j.jfca.2022.104802

Medha Khenwar, Swati Vishnoi, & Ankur Sisodia. (2022). *An Assessment of Milk Adulteration IoT Based Model to Identify the Quality of Milk using Lab View*. https://doi.org/10.1109/smart55829.2022.10047364

Milk Production in India. (n.d.). Pib.gov.in. https://pib.gov.in/FeaturesDeatils.aspx?NoteId=151137&ModuleId%20=%202

Mishra, G. K., Mishra, R. K., & Bhand, S. (2010). Flow injection analysis biosensor for urea analysis in adulterated milk using enzyme thermistor. *Biosensors and Bioelectronics*, *26*(4), 1560–1564. https://doi.org/10.1016/j.bios.2010.07.113

Mohammed, A. M., & Shuming, Y. (2021). Detection and quantification of cow milk adulteration using portable near-infrared spectroscopy combined with chemometrics. *African Journal of Agricultural Research*, *17*(2), 198–207. https://doi.org/10.5897/ajar2020.15321

Moore, D. A., Kirk, J. H., Klingborg, D. J., Garry, F. B., Wailes, W., Dalton, J. C., Busboom, J. R., Sams, R. W., Poe, M., Payne, M. A., Marchello, J. A., Looper, M. L., Falk, D., & Wright, T. (2004). DairyBeef: Maximizing Quality and Profits—A Consistent Food Safety Message. *Journal of Dairy Science*, *87*(1), 183–190. https://doi.org/10.3168/jds.s0022-0302(04)73157-4

Moreira, M., de Franca, J. A., de Oliveira Toginho Filho, D., Beloti, V., Yamada, A. K., de M. Franca, M. B., & de Souza Ribeiro, L. (2016). A Low-Cost NIR Digital Photometer Based on InGaAs Sensors for the Detection of Milk Adulterations With Water. *IEEE Sensors Journal*, *16*(10), 3653–3663. https://doi.org/10.1109/jsen.2016.2530873

Motta, T. M. C., Hoff, R. B., Barreto, F., Andrade, R. B. S., Lorenzini, D. M., Meneghini, L. Z., & Pizzolato, T. M. (2014). Detection and confirmation of

milk adulteration with cheese whey using proteomic-like sample preparation and liquid chromatography–electrospray–tandem mass spectrometry analysis. *Talanta*, *120*, 498–505. https://doi.org/10.1016/j.talanta.2013.11.093

N. Bamiedakis, Hutter, T., Penty, R. V., White, I. H., & Elliott, S. R. (2013). PCB-Integrated Optical Waveguide Sensors: An Ammonia Gas Sensor. *Journal of Lightwave Technology*, *31*(10), 1628–1635. https://doi.org/10.1109/jlt.2013.2255582

Nagraik, R., Sharma, A., Kumar, D., Chawla, P., & Kumar, A. P. (2021). Milk adulterant detection: Conventional and biosensor based approaches: A review. *Sensing and Bio-Sensing Research*, *33*, 100433. https://doi.org/10.1016/j.sbsr.2021.100433

Nascimento, C. F., Santos, P. M., Pereira-Filho, E. R., & Rocha, F. R. P. (2017). Recent advances on determination of milk adulterants. *Food Chemistry*, *221*, 1232–1244. https://doi.org/10.1016/j.foodchem.2016.11.034

None Jyoti, None Kavita, & Verma, R. K. (2022). Selective detection of urea as milk adulterant using LMR based Fiber Optic Probe. *Journal of Food Composition and Analysis*, *114*, 104825–104825. https://doi.org/10.1016/j.jfca.2022.104825

Noor, R., & Bhuiyan, A. A. (2020). General Perspectives on Water and Fluid Borne Microorganisms in Bangladesh. *Applied Microbiology: Theory & Technology*. https://doi.org/10.37256/amtt.122020480

Ntakatsane, M. P., Liu, X. M., & Zhou, P. (2013). Short communication: Rapid detection of milk fat adulteration with vegetable oil by fluorescence spectroscopy. *Journal of Dairy Science*, *96*(4), 2130–2136. https://doi.org/10.3168/jds.2012-6417

Okazaki, S., Hiramatsu, M., Gonmori, K., Suzuki, O., & Tu, A. T. (2009). Rapid nondestructive screening for melamine in dried milk by Raman spectroscopy. *Forensic Toxicology*, *27*(2), 94–97. https://doi.org/10.1007/s11419-009-0072-3

Pereira, P. C. (2014). Milk nutritional composition and its role in human health. *Nutrition*, *30*(6), 619–627. https://doi.org/10.1016/j.nut.2013.10.011

Poonia, A., Jha, A., Sharma, R., Singh, H. B., Rai, A. K., & Sharma, N. (2016). Detection of adulteration in milk: A review. *International Journal of Dairy Technology*, *70*(1), 23–42. https://doi.org/10.1111/1471-0307.12274

Prasad, A. (2021, August 8). *Regression Trees | Decision Tree for Regression | Machine Learning*. Analytics Vidhya. https://medium.com/analytics-vidhya/regression-trees-decision-tree-for-regression-machine-learning-e4d7525d8047

Qin, J., Chao, K., & Kim, M. S. (2013). Simultaneous detection of multiple adulterants in dry milk using macro-scale Raman chemical imaging. *Food Chemistry*, *138*(2-3), 998–1007. https://doi.org/10.1016/j.foodchem.2012.10.115

Ram, R., Gautam, N., Paik, P., Kumar, S., & Sarkar, A. (2022). A novel and low-cost smartphone integrated paper-based sensor for measuring starch adulteration in milk. *Microfluidics and Nanofluidics*, *26*(12). https://doi.org/10.1007/s10404-022-02607-2

Rani, A., Sharma, V., Arora, S., Lal, D., & Kumar, A. (2013). A rapid reversed-phase thin layer chromatographic protocol for detection of adulteration in ghee (clarified milk fat) with vegetable oils. *Journal of Food Science and Technology*, *52*(4), 2434–2439. https://doi.org/10.1007/s13197-013-1208-3

Raturi et al. (2022). Study Of Adulteration in Milk and Milk Products And Their Adverse Health Effects. *Octa Journal of Biosciences*.

Ravindran, A., Princess, F., & D. Nirmal. (2018). *A Study on the use of Spectroscopic Techniques to Identify Food Adulteration*. https://doi.org/10.1109/iccsdet.2018.8821197

Ruchira Nandeshwar, Mandal, P., & Siddharth Tallur. (2023). Portable and Low-Cost Colorimetric Sensor for Detection of Urea in Milk Samples. *IEEE Sensors Journal*, *23*(14), 16287–16292. https://doi.org/10.1109/jsen.2023.3282810

Rui-cheng, W., Wang, R., Zeng, Q., Ming, C., & Liu Tie-zheng. (2009). High-Performance Liquid Chromatographic Method for the Determination of Cyromazine and Melamine Residues in Milk and Pork. *Journal of Chromatographic Science*, *47*(7), 581–584. https://doi.org/10.1093/chromsci/47.7.581

Sadhan Kumar Dutta, Chakraborty, G., Chauhan, V., Singh, L., Vijay Singh Sharanagat, & Vijay Kumar Gahlawat. (2022). *Development of a predictive model for determination of urea in milk using silver nanoparticles and UV–Vis spectroscopy*. *168*, 113893–113893. https://doi.org/10.1016/j.lwt.2022.113893

Salgó, A., & Gergely, S. (2012). Analysis of wheat grain development using NIR spectroscopy. *Journal of Cereal Science*, *56*(1), 31–38. https://doi.org/10.1016/j.jcs.2012.04.011

Salleh, N. A., Selamat, J., Meng, G. Y., Abas, F., Jambari, N. N., & Khatib, A. (2019). Fourier transform infrared spectroscopy and multivariate analysis of milk from different goat breeds. *International Journal of Food Properties*, *22*(1), 1673–1683. https://doi.org/10.1080/10942912.2019.1668803

Santos, P. M. dos, Costa, L. F. B., & Pereira-Filho, E. R. (2012). Study of Calcium and Sodium Behavior to Identify Milk Adulteration Using Flame Atomic Absorption Spectrometry. *Food and Nutrition Sciences*, *03*(09), 1228–1232. https://doi.org/10.4236/fns.2012.39161

Santos, P. M., Pereira-Filho, E. R., & Rodriguez-Saona, L. E. (2013). Application of Hand-Held and Portable Infrared Spectrometers in Bovine Milk Analysis. *Journal of Agricultural and Food Chemistry*, *61*(6), 1205–1211. https://doi.org/10.1021/jf303814g

Sharifi, F., Mojtaba Naderi-Boldaji, Mahdi Ghasemi-Varnamkhasti, Kamran Kheiralipour, Ghasemi, M., & Maleki, A. (2023). Feasibility study of detecting some milk adulterations using a LED-based Vis-SWNIR photoacoustic spectroscopy system. *Food Chemistry*, *424*, 136411–136411. https://doi.org/10.1016/j.foodchem.2023.136411

Sharp, T. (2020, May 6). *An Introduction to Support Vector Regression (SVR)*. Medium. https://towardsdatascience.com/an-introduction-to-support-vector-regression-svr-a3ebc1672c2

Shabir Barham, G. (2014). Detection and Extent of Extraneous Water and Adulteration in Milk Consumed at Hyderabad, Pakistan. *Journal of Food and Nutrition Sciences*, *2*(2), 47. https://doi.org/10.11648/j.jfns.20140202.15

Simone, Cruz, A. G., Eduardo H.M. Walter, Rocha, M., Maria, R., Ferreira, C., & Sant'Ana, A. S. (2011). *Monitoring the authenticity of Brazilian UHT milk: A chemometric approach*. *124*(2), 692–695. https://doi.org/10.1016/j.foodchem.2010.06.074

Singuluri, H. (2014). Milk Adulteration in Hyderabad, India – A Comparative Study on the Levels of Different Adulterants Present in Milk. *Journal of*

Chromatography & Separation Techniques, *05*(01). https://doi.org/10.4172/2157-7064.1000212

Sneha, S., Surjith, S., & Alex Raj, S. M. (2023, May 1). *A Review on Food Adulteration Detection Techniques: Methodologies, Applications, and Challenges*. IEEE Xplore. https://doi.org/10.1109/ICCC57789.2023.10165065

Soomro, & Abdul Aziz. (2014). *"Study on adulteration and composition of milk sold at Badin."* . Google Scholar. Intl J Res Appl Nat Social Sci 2.9 (2014): 57-70.

Sowmya, N., & Ponnusamy, V. (2021). Development of Spectroscopic Sensor System for an IoT Application of Adulteration Identification on Milk Using Machine Learning. *IEEE Access*, *9*, 53979–53995. https://doi.org/10.1109/access.2021.3070558

Thiago R.L.C. Paixão, & Bertotti, M. (2009). Fabrication of disposable voltammetric electronic tongues by using Prussian Blue films electrodeposited onto CD-R gold surfaces and recognition of milk adulteration. *Sensors and Actuators B: Chemical*, *137*(1), 266–273. https://doi.org/10.1016/j.snb.2008.10.045

Tittlemier, S. A. (2010). Methods for the analysis of melamine and related compounds in foods: a review. *Food Additives & Contaminants: Part A*, *27*(2), 129–145. https://doi.org/10.1080/19440040903289720

Tsai, T.-H., Thiagarajan, S., & Chen, S.-M. (2010). Detection of Melamine in Milk Powder and Human Urine. *Journal of Agricultural and Food Chemistry*, *58*(8), 4537–4544. https://doi.org/10.1021/jf904554s

Uysal, R. S., Boyaci, I. H., Genis, H. E., & Tamer, U. (2013). Determination of butter adulteration with margarine using Raman spectroscopy. *Food Chemistry*, *141*(4), 4397–4403. https://doi.org/10.1016/j.foodchem.2013.06.061

van der Avoort, C. M. T., van Loon, L. J. C., Verdijk, L. B., Poyck, P. P. C., Thijssen, D. T. J., & Hopman, M. T. E. (2021). Acute Effects of Dietary Nitrate on Exercise Tolerance, Muscle Oxygenation, and Cardiovascular Function in Patients With Peripheral Arterial Disease. *International Journal of Sport Nutrition and Exercise Metabolism*, *31*(5), 385–396. https://doi.org/10.1123/ijsnem.2021-0054

Vaz, E. B., Santos, M. F. C., Jesus, E. G. de, Vieira, K. M., Osório, V. M., & Menini, L. (2022). Development of Methodology for Detection of Formaldehyde-DNPH in Milk Manager by Central Composite Rotational Design and GC/MS. *Research, Society and Development*, *11*(9), e16411931575. https://doi.org/10.33448/rsd-v11i9.31575

Venkatasami, G., & Sowa, J. R. (2010). A rapid, acetonitrile-free, HPLC method for determination of melamine in infant formula. *Analytica Chimica Acta*, *665*(2), 227–230. https://doi.org/10.1016/j.aca.2010.03.037

Veríssimo, M. I. S., Gamelas, J. A. F., Fernandes, A. J. S., Evtuguin, D. V., & Gomes, M. T. S. R. (2020). A new formaldehyde optical sensor: Detecting milk adulteration. *Food Chemistry*, *318*, 126461. https://doi.org/10.1016/j.foodchem.2020.126461

Vinod Kumar Verma, Pervez Mustajab, & Sadat, A. (2019). *Determination of Adulteration in Milk using Ultrasonic Technique*. https://doi.org/10.1109/upcon47278.2019.8980234

Wu, Y., & Zhang, Y. (2013). Analytical chemistry, toxicology, epidemiology and health impact assessment of melamine in infant formula: Recent progress and developments. *Food and Chemical Toxicology*, *56*, 325–335. https://doi.org/10.1016/j.fct.2013.02.044

Yadav, A. Kr., Gattupalli, M., Dashora, K., & Kumar, V. (2022). Key Milk Adulterants in India and their Detection Techniques: a Review. *Food Analytical Methods*. https://doi.org/10.1007/s12161-022-02427-8

Yang, R., Liu, R., & Kexin, X. (2013). Detection of adulterated milk using two-dimensional correlation spectroscopy combined with multi-way partial least squares. *Food Bioscience*, *2*, 61–67. https://doi.org/10.1016/j.fbio.2013.04.005

Yang, S., Ding, J.-H., Zheng, J., Hu, B., Li, J., Chen, H., Zhou, Z., & Qiao, X. (2009). Detection of Melamine in Milk Products by Surface Desorption Atmospheric Pressure Chemical Ionization Mass Spectrometry. *Analytical Chemistry*, *81*(7), 2426–2436. https://doi.org/10.1021/ac900063u

Zhang, L., & Tian, F.-C. (2014). A new kernel discriminant analysis framework for electronic nose recognition. *Analytica Chimica Acta*, *816*, 8–17. https://doi.org/10.1016/j.aca.2014.01.049

Zhang, X., Zou, M., Qi, X., Liu, F., Zhu, X., & Zhao, B. (2010). *Detection of melamine in liquid milk using surface-enhanced Raman scattering spectroscopy*. *41*(12), 1655–1660. https://doi.org/10.1002/jrs.2629

Zhu, L., Gamez, G., Chen, H., Konstantin Chingin, & Zenobi, R. (2009). *Rapid detection of melamine in untreated milk and wheat gluten by ultrasound-assisted extractive electrospray ionization mass spectrometry (EESI-MS)*. *5*, 559–561. https://doi.org/10.1039/b818541g

CODE

```
from google.colab import drive
drive.mount('/content/drive')
import pandas as pd
import numpy as np
import numpy
import math
import matplotlib.pyplot as plt
from matplotlib import pyplot
from scipy import signal
from sklearn.svm import SVR
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.cross_decomposition import PLSRegression
from sklearn.model_selection import KFold, cross_val_score
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error
from sklearn.metrics import mean_absolute_error
dataset=pd.read_csv('/content/drive/MyDrive/final dataset files/machine
larning/Dataset.csv')
dataset
x = dataset
x
x.shape
x = dataset.drop(['per'], axis=1).to_numpy()
```

```
wavelength = x[0,0:1001]

wavelength

x= x[1:85,0:1001]

x

y= dataset.per.to_numpy()

y = y[1:85]

y

y.shape

# plotting the signal

pyplot.plot(wavelength, x.T)

pyplot.xlabel(' Wavelength')

pyplot.ylabel('Abs')

pyplot.title("Spectra")

pyplot.show()

#train and test

x_train, x_test, y_train, y_test  = train_test_split(x, y, test_size=0.2, random_state=0)

#standard scalar

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

scaler.fit(x_train)

x_train= scaler.transform(x_train)

x_test= scaler.transform(x_test)

k_folds = KFold(n_splits = 10)

# Calculate first derivative applying a Savitzky-Golay filter

X = signal.savgol_filter(x, window_length=91, polyorder=3, deriv=1)
```

```
f = pyplot.figure()
f.set_figwidth(7)
f.set_figheight(8)


print("Plot after re-sizing: ")
pyplot.plot(wavelength, x.T)
pyplot.xlabel(' Wavelength')
pyplot.ylabel('Abs')
pyplot.title("Spectra")
pyplot.show()
SUPPORT VECTOR REGRESSION
mod =  SVR(C=10, epsilon=0.2)
cv_scores = cross_val_score(mod, x_train, y_train, cv=k_folds)
mod.fit(x_train, y_train)
y_pred = mod.predict(x_test)
r = r2_score(y_test, y_pred)
print("Root Square:")
print(r)
MSE= mean_squared_error(y_test, y_pred)
print("Mean Square Error:")
print(MSE)
y_test = np.array(y_test).astype(float)
y_pred = np.array(y_pred).astype(float)
MSE = np.square(np.subtract(y_test, y_pred)).mean()
RMSE = math.sqrt(MSE)
```

```
print("Root Mean Square Error:")
print(RMSE)
y_pred
import numpy as np
import matplotlib.pyplot as plt
fig,ax = plt.subplots(1)
# plot the data
ax.scatter(y_test,y_pred,color="red", marker="o",)
m, b = np.polyfit(y_test, y_pred, 1)
#add linear regression line to scatterplot
plt.plot(y_test, m*y_test+b)
plt.xlabel('Actual')
plt.ylabel('Predicted concentraton in mg/L')
plt.title("Prediction of chlorophyll A")
plt.show()
from sklearn.datasets import make_regression
make_regression(n_features=4, n_informative=2,random_state=0, shuffle=False)
clf = SVR(C=10, epsilon=0.2)
kf = KFold(n_splits=10, shuffle=True, random_state=42)
rmse = 0
for i, (train_index, test_index) in enumerate(kf.split(x_pca)):
    print(f"Fold {i}:")
    #print("TRAIN:", train_index, "TEST:", test_index)
    X_train, X_test = x_pca[train_index], x_pca[test_index]
    y_train, y_test = y[train_index], y[test_index]
```

```
    X_train = X_train.astype(np.float64)
    X_test = X_test.astype(np.float64)
    y_test = y_test.astype(np.float64)
    y_train = y_train.astype(np.float64)
    clf.fit(X_train, y_train)
    ypred = clf.predict(X_test)
    ypred=np.array(ypred).flatten()
    aa=np.array(y_test).flatten()
    mat_plot(aa,ypred)
    rmse = rmse + np.sqrt(mean_squared_error(aa,ypred))
#df = pd.DataFrame(clf.cv_results_)
#df



rmse = rmse /10
print("average RMSE",rmse)


#partial least square regression
from sklearn.cross_decomposition import PLSRegression


pls2 = PLSRegression(n_components=5)
pls2.fit(x_train, y_train)
PLSRegression()
Y_pred = pls2.predict(x_test)
from sklearn.metrics import r2_score
```

```
r = r2_score(y_test, Y_pred)
print("Root Square:")
print(r)
from sklearn.metrics import mean_squared_error
MSE= mean_squared_error(y_test, Y_pred)
print("Mean Square Error:")
print(MSE)
import math
y_test = np.array(y_test).astype(float)
y_pred = np.array(y_pred).astype(float)

MSE = np.square(np.subtract(y_test, Y_pred)).mean()
RMSE = math.sqrt(MSE)

print("Root Mean Square Error:")
print(RMSE)
fig,ax = plt.subplots(1)

# plot the data
ax.scatter(y_test, Y_pred,color="blue", marker="o",)
m, b = np.polyfit(y_test, Y_pred, 1)
#add linear regression line to scatterplot
plt.plot(y_test, m*y_test+b)
plt.xlabel('Actual')
plt.ylabel('Predicted concentraton in mg/L')
```

```
plt.title("Prediction of chlorophyll A")
plt.show()
from sklearn.datasets import make_regression
make_regression(n_features=4, n_informative=2,random_state=0, shuffle=False)
clf = PLSRegression(n_components=5)
kf = KFold(n_splits=10, shuffle=True, random_state=42)
rmse = 0
for i, (train_index, test_index) in enumerate(kf.split(x_pca)):
    print(f"Fold {i}:")
    #print("TRAIN:", train_index, "TEST:", test_index)
    X_train, X_test = x_pca[train_index], x_pca[test_index]
    y_train, y_test = y[train_index], y[test_index]
    X_train = X_train.astype(np.float64)
    X_test = X_test.astype(np.float64)
    y_test = y_test.astype(np.float64)
    y_train = y_train.astype(np.float64)
    clf.fit(X_train, y_train)
    ypred = clf.predict(X_test)
    ypred=np.array(ypred).flatten()
    aa=np.array(y_test).flatten()
    mat_plot(aa,ypred)
    rmse = rmse + np.sqrt(mean_squared_error(aa,ypred))
#df = pd.DataFrame(clf.cv_results_)
#df
rmse = rmse /10
```

```
print("average RMSE",rmse)
DECISION TREE REGRESSION
import matplotlib.pyplot as plt
import numpy as np
from sklearn import tree
from sklearn.tree import DecisionTreeRegressor
clf = tree.DecisionTreeRegressor()
from sklearn.datasets import load_diabetes
from sklearn.model_selection import cross_val_score
from sklearn.tree import DecisionTreeRegressor
regressor = DecisionTreeRegressor(random_state=0)
cross_val_score(regressor,x_test, y_test, cv=3)
reg = DecisionTreeRegressor(max_depth=50)
clf = tree.DecisionTreeRegressor()
clf = clf.fit(x_train, y_train)
D_pred = clf.predict(x_test)
from sklearn.metrics import r2_score
r = r2_score(y_test, D_pred)
print("Root Square:")
print(r)
from sklearn.metrics import mean_squared_error
MSE= mean_squared_error(y_test, D_pred)
print("Mean Square Error:")
print(MSE)
import math
```

```
y_test = np.array(y_test).astype(float)
y_pred = np.array(y_pred).astype(float)
MSE = np.square(np.subtract(y_test, D_pred)).mean()
RMSE = math.sqrt(MSE)
print("Root Mean Square Error:")
print(RMSE)
fig,ax = plt.subplots(1)
# plot the data
ax.scatter(y_test, D_pred,color="blue", marker="o",)
, b = np.polyfit(y_test, D_pred, 1)
#add linear regression line to scatterplot
plt.plot(y_test, m*y_test+b)
plt.xlabel('Actual')
plt.ylabel('Predicted concentraton in mg/L')
plt.title("Prediction of chlorophyll A")
plt.show()
from sklearn.datasets import make_regression
make_regression(n_features=4, n_informative=2,random_state=0, shuffle=False)
clf = tree.DecisionTreeRegressor()
kf = KFold(n_splits=10, shuffle=True, random_state=42)
rmse = 0
for i, (train_index, test_index) in enumerate(kf.split(x_pca)):
    print(f'Fold {i}:")
    #print("TRAIN:", train_index, "TEST:", test_index)
    X_train, X_test = x_pca[train_index], x_pca[test_index]
```

```
    y_train, y_test = y[train_index], y[test_index]
    X_train = X_train.astype(np.float64)
    X_test = X_test.astype(np.float64)
    y_test = y_test.astype(np.float64)
    y_train = y_train.astype(np.float64)
    clf.fit(X_train, y_train)
    ypred = clf.predict(X_test)
    ypred=np.array(ypred).flatten()
    aa=np.array(y_test).flatten()
    mat_plot(aa,ypred)
    rmse = rmse + np.sqrt(mean_squared_error(aa,ypred))
#df = pd.DataFrame(clf.cv_results_)
#df
rmse = rmse /10
print("average RMSE",rmse)
```

RANDOM FOREST REGRESSION

```
from sklearn.ensemble import RandomForestRegressor
from sklearn.datasets import make_regression
make_regression(n_features=4, n_informative=2,random_state=0, shuffle=False)
regr = RandomForestRegressor(max_depth=2, random_state=0)
regr.fit(x_train, y_train)
R_pred=regr.predict(x_test)
from sklearn.metrics import r2_score
r = r2_score(y_test, R_pred)
print("Root Square:")
```

```
print(r)
from sklearn.metrics import mean_squared_error
MSE= mean_squared_error(y_test, R_pred)
print("Mean Square Error:")
print(MSE)
import math
y_test = np.array(y_test).astype(float)
y_pred = np.array(y_pred).astype(float)
MSE = np.square(np.subtract(y_test, R_pred)).mean()
RMSE = math.sqrt(MSE)
print("Root Mean Square Error:")
print(RMSE)
fig,ax = plt.subplots(1)
# plot the data
ax.scatter(y_test, R_pred,color="blue", marker="o",)
m, b = np.polyfit(y_test, R_pred, 1)
#add linear regression line to scatterplot
plt.plot(y_test, m*y_test+b)
plt.xlabel('Actual')
plt.ylabel('Predicted concentraton in mg/L')
plt.title("Prediction of chlorophyll A")
plt.show()
from sklearn.datasets import make_regression
make_regression(n_features=4, n_informative=2,random_state=0, shuffle=False)
clf = RandomForestRegressor(max_depth=2, random_state=0)
```

```
kf = KFold(n_splits=10, shuffle=True, random_state=42)
rmse = 0
for i, (train_index, test_index) in enumerate(kf.split(x_pca)):
    print(f'Fold {i}:")
    #print("TRAIN:", train_index, "TEST:", test_index)
    X_train, X_test = x_pca[train_index], x_pca[test_index]
    y_train, y_test = y[train_index], y[test_index]
    X_train = X_train.astype(np.float64)
    X_test = X_test.astype(np.float64)
    y_test = y_test.astype(np.float64)
    y_train = y_train.astype(np.float64)
    clf.fit(X_train, y_train)
    ypred = clf.predict(X_test)
    ypred=np.array(ypred).flatten()
    aa=np.array(y_test).flatten()
    mat_plot(aa,ypred)
    rmse = rmse + np.sqrt(mean_squared_error(aa,ypred))
#df = pd.DataFrame(clf.cv_results_)
#df
rmse = rmse /10
print("average RMSE",rmse)
```

K-NEAREST NEIGHBOR REGRESSION

```
# Import necessary libraries
from sklearn.decomposition import PCA
'''
```

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x_smooth, y, test_size=0.1,
random_state=0)'''
# Apply PCA
pca = PCA(n_components = 5)
X_pca = pca.fit_transform(X)
# Explained variance
explained_variance = pca.explained_variance_
total_explained_variance = explained_variance.sum()
# Print results
print(f"Explained Variance:\n{explained_variance}")
print(f"Total Explained Variance: {total_explained_variance:.4f}")
# Explained variance ratio
explained_variance_ratio = pca.explained_variance_ratio_
total_explained_variance_ratio = explained_variance_ratio.sum()
# Print results
print(f"\nExplained Variance Ratio:\n{explained_variance_ratio}")
print(f"Total Explained Variance Ratio: {total_explained_variance_ratio:.4f}")
# Import necessary libraries
import numpy as np
import matplotlib.pyplot as plt
# Plot explained variance ratio
cumulative_variance_ratio = np.cumsum(explained_variance_ratio)
f = plt.figure()
f.set_figwidth(10)
```

```
f.set_figheight(10)
plt.plot(cumulative_variance_ratio, marker='o')
plt.xlabel('Number of Principal Components')
plt.ylabel('Cumulative Explained Variance Ratio')
plt.title('Cumulative Explained Variance Ratio by Principal Components')
plt.show()
!pip install matplotlib
def mat_plot(a,b): # qq1 is the actual readings andthe b is the predictd
  from sklearn.metrics import mean_absolute_error
  from sklearn.metrics import mean_squared_error
  from sklearn.metrics import r2_score
  import numpy as np
  import matplotlib.pyplot as plt
  print("The R2 ", (r2_score( a, b)))
  print("RMSE:", np.sqrt(mean_squared_error(  a, b)))
  #print("MAPE%:", mean_absolute_percentage_error(  a, b))
  print("MAE",mean_absolute_error(a,b))
  print("MSE",mean_squared_error(a,b))
  print("RMSE",np.sqrt(mean_squared_error(a,b)))
  fig, ax = plt.subplots()
  ax.plot(b, a, linewidth=0,  marker="o", color='C0', markersize=8)
  #plot(x,  y,  color='green',  linestyle='dashed',  marker='o',  markerfacecolor='blue',
markersize=12).
  low_x, high_x = ax.get_xlim()
  low_y, high_y = ax.get_ylim()
```

```
  low = max(low_x, low_y)
  high = min(high_x, high_y)
  ax.plot([low, high], [low, high], ls="-", c=".2", alpha=.4)
  #ax.set_title('R2 score')
  plt.rcParams.update({'font.size': 20})
  ax.set_xlabel("Actual")
  ax.set_ylabel("Predicted ")
  plt.show()
import matplotlib
import numpy as np
from sklearn.model_selection import KFold
from sklearn.neighbors import KNeighborsRegressor
from sklearn.metrics import mean_squared_error
# Apply PCA
pca = PCA(n_components =  2)
x_pca = pca.fit_transform(X)
rmse = 0
kf = KFold(n_splits=10, shuffle=True, random_state=42)
for i, (train_index, test_index) in enumerate(kf.split(x_pca)):
    print(f"Fold {i}:")
    #print("TRAIN:", train_index, "TEST:", test_index)
    X_train, X_test = x_pca[train_index], x_pca[test_index]
    y_train, y_test = y[train_index], y[test_index]
    X_train = X_train.astype(np.float64)
    X_test = X_test.astype(np.float64)
```

```
    y_test = y_test.astype(np.float64)
    y_train = y_train.astype(np.float64)
    neigh = KNeighborsRegressor( n_neighbors = 1, p = 1, weights = 'distance',
algorithm = 'auto' )
    neigh.fit(X_train, y_train)
    y_pred=neigh.predict(X_test)
    y_pred=np.array(y_pred).flatten()
    qq1=np.array(y_test).flatten()
    print( "***********")
    #plt.plot(qq1,y_pred)
    mat_plot(qq1,y_pred)
    rmse = rmse + np.sqrt(mean_squared_error(qq1,y_pred))
    print( "Result for each fold " + str(i))
    print( "***********")
rmse = rmse /10
print("average RMSE",rmse)
from sklearn.model_selection import cross_val_score
from sklearn.neighbors import KNeighborsRegressor
from sklearn.metrics import mean_squared_error
# Import necessary libraries
from sklearn.decomposition import PCA
# Apply PCA
pca = PCA(n_components =  5)
x_pca = pca.fit_transform(X)
'''
```

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x_pca, y, test_size=0.1,
random_state=42, shuffle= True)'''
neigh = KNeighborsRegressor()
x_pca = x_pca.astype(np.float64)
y = y.astype(np.float64)
'''
y_train = y_train.astype(np.float64)
y_test = y_test.astype(np.float64)
'''
from sklearn.model_selection import KFold
from sklearn.model_selection import GridSearchCV
parameters = {'n_neighbors':list(range(1, 5)) , 'p':list(range(1, 5)) ,
'weights':['distance'], 'algorithm': ['auto']}
from sklearn.model_selection import GridSearchCV
kf = KFold(n_splits=10, shuffle=True, random_state=42)
clf = GridSearchCV(neigh, parameters, cv = kf , return_train_score=False, scoring=
'neg_mean_absolute_error') # scoring='neg_mean_squared_error
clf.fit(x_pca, y)
'''
#clf.cv_results_
ypred = clf.predict(x_test)
ypred=np.array(ypred).flatten()
aa=np.array(y_test).flatten()
mat_plot(aa,ypred)
```

```
'''
kf = KFold(n_splits=10, shuffle=True, random_state=42)
rmse = 0
for i, (train_index, test_index) in enumerate(kf.split(x_pca)):
    print(f'Fold {i}:')
    #print("TRAIN:", train_index, "TEST:", test_index)
    X_train, X_test = x_pca[train_index], x_pca[test_index]
    y_train, y_test = y[train_index], y[test_index]
    X_train = X_train.astype(np.float64)
    X_test = X_test.astype(np.float64)
    y_test = y_test.astype(np.float64)
    y_train = y_train.astype(np.float64)
    clf.fit(X_train, y_train)
    ypred = clf.predict(X_test)
    ypred=np.array(ypred).flatten()
    aa=np.array(y_test).flatten()
    #plt.plot(qq1,y_pred)
    mat_plot(aa,ypred)
    rmse = rmse + np.sqrt(mean_squared_error(aa,ypred))
df = pd.DataFrame(clf.cv_results_)
df
rmse = rmse /10
print("average RMSE",rmse)
```

APPENDIX

Features of Jasco- V-770 Spectrophotometer:

A wide range UV-Visible/Near Infrared Spectrophotometer with a unique optical design featuring a single monochromator and dual detectors for the wavelength range from 190 to 2700nm (3200nm option).

The V-770's single monochromator design provides for maximum light throughput with excellent absorbance linearity. A PMT detector is used for the UV to visible region and a Peltier-cooled PbS detector for the NIR region.

The V-770 UV-Visible/NIR spectrophotometer is operated using Spectra Manager™ Suite. This innovative cross-platform spectroscopy software is compatible with Windows 7 Pro (32- and 64-bit) and Windows 8.1 operating systems. For simple operation, the handheld iRM has a great look and feel with a colour touch sensitive screen. Data can also be downloaded to Spectra Analysis on a PC further PC data processing.

The V-700 Series has a growing list of software applications for both Spectra Manager™ and iRM. If you have an application which you don’t see listed, please let us know as we may already have it or we can prepare an application designed specifically for your requirements.

Optical System	Czerny-Turner grating mount Single monochomator Fully symmetrical double beam
Light source	Halogen lamp, Deuterium lamp
Wavelength range	190 to 2700 nm (3200 nm option)
Wavelength accuracy	+/-0.3 nm (at 656.1 nm) +/-1.5 nm (at 1312.2 nm)
Wavelength repeatability	+/-0.05 nm (UV-Vis), +/-0.2 nm (NIR)
Spectral bandwidth (SBW)	UV-Visible: 0.1, 0.2, 0.5, 1, 2, 5, 10 nm L2, L5, L10 nm (low stray light mode) M1, M2 nm (micro cell mode) NIR: 0.4, 0.8, 1, 2, 4, 8, 20, 40 L8, L20, L40 nm (low stray light mode) M4, M8 nm (micro cell mode)
Stray light	1 % (198 nm KCL) 0.0005 % (220 nm NaI) 0.0005 % (340 nm NaNO2) 0.0005 % (370 nm NaNO2)72 SBW: L2 nm 0.04 % (1420 nm: H2O) 0.1 % (1690 nm: CH2Br2) SBW: L8 nm
Photometric range	UV-Visible: -4~4 Abs NIR: -3~3 Abs
Photometric accuracy	+/-0.0015 Abs (0 to 0.5 Abs) +/-0.0025 Abs (0.5 to 1 Abs) +/-0.3 %T Tested with NIST SRM 930D
Photometric repeatability	+/-0.0005 Abs (0 to 0.5 Abs) +/-0.0005 Abs (0.5 to 1 Abs) Tested with NIST SRM 930D
Scanning speed	10~4000 nm/min (8000 nm/min in preview mode)
Slew speed	UV-Vis: 12,000 nm/min NIR: 48,000 nm/min
RMS noise	0.00003 Abs (0 Abs, wavelength: 500 nm, measurement time: 60 sec, SBW: 2 nm)
Baseline stability	0.0003 Abs/hour (Wavelength: 250 nm, response: slow and SBW: 2 nm)
Baseline flatness	+/-0.0002 Abs (200 - 2500 nm)
Detector	PMT, Peltier cooled PbS
Standard functions	IQ accessories, Start button, Analog output
Standard programs	Abs/%T meter, Quantitative analysis, Spectrum measurement,

	Time course measurement, Fixed wavelength measurement, Validation, Daily check, Dual wavelength time course measurement
Dimensions and weight	460(W) x 602(D) x 268(H) mm, 29 kg
Power requirements	150 VA
Installation requirements	Room temperature: 15-30 Celsius, humidity: below 85%