University of Kentucky
UKnowledge

Theses and Dissertations--Biosystems and Agricultural Engineering

Biosystems and Agricultural Engineering

2020

FOURIER TRANSFORM INFRARED SPECTROSCOPY (AS A RAPID METHOD) COUPLED WITH MACHINE LEARNING APPROACHES FOR DETECTION AND QUANTIFICATION OF GLUTEN CONTAMINATIONS IN GRAIN-BASED FOODS

Abuchi Godswill Okeke
University of Kentucky, okekeag@gmail.com
Author ORCID Identifier:
https://orcid.org/0000-0001-8684-1115
Digital Object Identifier: https://doi.org/10.13023/etd.2020.353

Right click to open a feedback form in a new tab to let us know how this document benefits you.

Recommended Citation

Okeke, Abuchi Godswill, "FOURIER TRANSFORM INFRARED SPECTROSCOPY (AS A RAPID METHOD) COUPLED WITH MACHINE LEARNING APPROACHES FOR DETECTION AND QUANTIFICATION OF GLUTEN CONTAMINATIONS IN GRAIN-BASED FOODS" (2020). *Theses and Dissertations--Biosystems and Agricultural Engineering*. 73.
https://uknowledge.uky.edu/bae_etds/73

This Master's Thesis is brought to you for free and open access by the Biosystems and Agricultural Engineering at UKnowledge. It has been accepted for inclusion in Theses and Dissertations--Biosystems and Agricultural Engineering by an authorized administrator of UKnowledge. For more information, please contact UKnowledge@lsv.uky.edu.

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Abuchi Godswill Okeke, Student

Dr. Akinbode Adedeji, Major Professor

Dr. Donald Colliver, Director of Graduate Studies

FOURIER TRANSFORM INFRARED SPECTROSCOPY (AS A RAPID METHOD) COUPLED WITH MACHINE LEARNING APPROACHES FOR DETECTION AND QUANTIFICATION OF GLUTEN CONTAMINATIONS IN GRAIN-BASED FOODS

THESIS

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science in
Biosystems and Agricultural Engineering in the College of Engineering and the College of Agriculture, Food and Environment
at the University of Kentucky

By

Abuchi Godswill Okeke

Lexington, Kentucky

Director: Dr. Akinbode A. Adedeji, Associate Professor of Food Process Engineering
Lexington, Kentucky

2020

Copyright © Abuchi G. Okeke 2020
https://orcid.org/0000-0001-8684-1115

ABSTRACT OF THESIS

FOURIER TRANSFORM INFRARED SPECTROSCOPY
(AS A RAPID METHOD) COUPLED WITH MACHINE LEARNING APPROACHES
FOR DETECTION AND QUANTIFICATION OF GLUTEN CONTAMINATIONS IN
GRAIN-BASED FOODS

Cross-contamination between food grains during harvesting, transportation, and/or food processing is still a major issue in the food industry. Due to cross-contact with gluten-rich grains (wheat, barley, and rye grains), gluten can get into food that's naturally free from gluten and thus may not be safe for consumption for people susceptible to gluten-related disorders such as celiac disease, wheat allergy, gluten intolerance or sensitivity. The conventional method of gluten detection is cumbersome, time-consuming, and requires well-trained personnel. Therefore, there is a need for a rapid and equally effective technique to authenticate gluten contamination in foods. This research work explored the use of a Fourier transform infrared (FTIR) spectroscopy coupled with machine learning approaches to detect and quantify gluten contamination in grain-based foods. The research was divided into three different phases including the use of FTIR with supervised machine learning (ML) approaches to authenticate cross-contact between non-gluten and gluten flours, the use of FTIR with ML approaches to detect and quantify wheat flour contamination in gluten-free bread (cornbread), and finally, the use of Enzyme-linked immunosorbent assay (ELISA) as a complementary test to estimate and establish a gluten-free threshold of $\leq$ 20 ppm for the amount of gluten in wheat contaminated flour and cornbread.

Different machine learning algorithms such as linear discriminant analysis (LDA), partial least square regression (PLSR), k-nearest neighbor (KNN), support vector machine, decision tree, and ensemble learning method were used for the development of ML models. The results obtained for the first phase of the research show that FTIR with LDA and PLSR has the potential to detect and quantify cross-contact between a non-gluten (corn flour, CF) and gluten-rich (wheat flour, WF, barley flour, BF, and rye flour, RF) flours, at contamination levels of 0.5% - 10% (w/w), with 0.5% increments. For the second phase, a majority voting-based ensemble learning (stack of random forest, k-nearest neighbor (KNN) and support vector classifier) model was able to detect WF contamination in a cornbread at the true-positive rate and false-negative rate of 1.0, respectively. The ELISA tests for both phases (the raw flour samples and the baked bread) showed a threshold limit of $\leq$0.5% contamination level for CF contaminated with WF to be labeled gluten-free and $\leq$ 3.5% for the cornbread contaminated with the WF to be gluten-free. This research is still in its development stage and has the potential to contribute towards artificial intelligence applications in ensuring food safety, and to food quality inspection.

KEYWORDS: Gluten, Wheat Allergy, Celiac Disease, Cross-Contamination, Machine learning, FTIR Spectroscopy

Abuchi G. Okeke

(Name of Student)

08/04/2020

Date

FOURIER TRANSFORM INFRARED SPECTROSCOPY (AS A RAPID METHOD) COUPLED WITH MACHINE LEARNING APPROACHES FOR DETECTION AND QUANTIFICATION OF GLUTEN CONTAMINATIONS IN GRAIN-BASED FOODS

By
Abuchi G. Okeke

Dr. Akinbode A. Adedeji
Director of Thesis

Dr. Donald Colliver
Director of Graduate Studies

08/04/2020
Date

DEDICATION

I dedicate and give all glory to God Almighty for his strength, wisdom, knowledge and understanding throughout the period of my studies and the success of this journey. Also, I dedicate this thesis to my parents (Mr. and Mrs. Abuchi Gabriel Okeke) and to all my family members. Finally, I dedicate this thesis to all those who have contributed and supported me throughout the struggle of writing this thesis.

ACKNOWLEDGMENTS

My profound gratitude goes to my advisor, Dr. Akinbode Adedeji, and all my committee members, Dr. Michael Sama, and Dr. Daniel Lau for their immense support and guidance throughout this study and writing of my thesis.

I would like to appreciate the department chair, Dr. Mike Montross for his support and help with the review of my thesis. To all the members of the Biosystems Engineering Department, to all my friends and to my fellow graduate students who made this journey to be a smooth one, I remain grateful to you all.

My thanks go to Mr. Namal Wanninayake for all his assistantship with the Fourier transform infrared spectrometer and to Annet Kyomuhangi for all her help with the enzyme-linked immunosorbent test. You guys were of great help to me.

In addition, I would like to acknowledge the support and love from my parents, my siblings, and all my family members. They all kept me going with constant motivation and prayers. Finally, to all those who contributed to the success of this study, I salute you all, and I appreciate each one of you.

TABLE OF CONTENTS

ACKNOWLEDGMENTS iii

TABLE OF CONTENTS iv

LIST OF TABLES viii

LIST OF FIGURES x

CHAPTER 1. Introduction 1

1.1 General Background 1

1.1.1 Gluten Contamination 1

1.1.2 FTIR Spectroscopy method coupled with the machine learning process 2

1.2 Specific Objectives 4

CHAPTER 2. Literature Review 5

2.1 Gluten Overview 5

2.1.1 Effects of Processing on Gluten Proteins 6

2.1.2 Gluten-Related Disorders 7

2.2 Method of Fourier transformed infrared spectroscopy (FTIR) 11

2.3 Machine Learning Approaches 16

2.3.1 An Overview on Machine Learning 16

2.3.2 Types of Machine Learning (ML) Approaches 17

CHAPTER 3. Detection and Quantification of Cross-Contact of a Non-Gluten and Gluten-Rich Flours by Fourier Transform Infrared (FTIR) Spectroscopy Coupled with Machine Learning Approaches 32

3.1 Introduction 33

3.2 Material and Methods 36

3.2.1 Sample preparation and FTIR-Spectroscopy 36

3.2.2 Selection of the region of interest 37

3.2.3 Spectra pre-processing 38

3.2.4 Model development and evaluation 38

3.3 Results and Discussion 41

3.3.1 Spectra characteristics of the flour samples 41

3.3.2 Spectra preprocessing and spectra models 45

3.3.3 Classification modeling results 47

3.3.4 PLSR prediction model 50

CHAPTER 4. Fourier Transform Infrared (FTIR) Spectroscopy with Machine Learning Approaches for Detection and Quantification of Wheat Flour Contamination in a Non-Gluten Bread 60

4.1 Introduction 62

4.2 Materials and methods 64

4.2.1 Basic ingredients for bread 64

4.2.2 Laboratory baking 65

4.2.3 Spectra Data and pre-processing 65

4.2.4 Models development 66

4.3 Results and Discussion 69

4.3.1 Spectra characteristics of the ground bread samples 69

4.3.2 Classification modeling results 72

4.3.3 Prediction model 75

CHAPTER 5. Enzyme Linked Immunosorbent Assay (ELISA) test for Quantification of amount of Gluten present in the Contamination Levels in Chapter three and four 84

5.1 Introduction 84

5.2 Materials and Methods 85

5.2.1 Materials 85

5.2.2 Methods 86

Equipment 86

5.3 Results and Discussion 89

CHAPTER 6. General Conclusion And Recommendation 96

APPENDIX 99

Appendix A: MATLAB Code 99

A.1 Spectra Data Analysis Code 99

A.2 Function for loading FTIR (.SPA) data into set of arrays in MATLAB 101

A.3 Function for splitting data using Kennard Stone algorithm 103

A.4 Function for averaging the spectra data 105

Appendix B: Python Code 105

B1. Python Library 105

B.2 Classification models 106

B.3 Predictive/Regression models 106

BIBLIOGRAPHY 108

VITA 123

LIST OF TABLES

TABLE 3.1: TRAINING MODEL CONFUSION MATRIX FOR THE LDA + 4-FOLD CROSS-VALIDATION + BAGGING 48

TABLE 3.2: LDA TRAINING MODEL CONFUSION MATRIX PARAMETERS FOR CLASSIFICATION OF CONTAMINATION BETWEEN GLUTEN-RICH (BF (CLASS 1), WF (CLASS 2), RF (CLASS 3)) AND GLUTEN-FREE (CF (CLASS 4)) FLOURS. 48

TABLE 3.3: TEST MODEL CONFUSION MATRIX FOR LDA + 4-FOLD CV+ BAGGING 49

TABLE 3.4: RESULTS FOR THE EVALUATION OF THE EACH OF THE LDA TEST MODEL CLASSES (GLUTEN-RICH: BF (CLASS 1), WF (CLASS2), RF (CLASS 3)) AND GLUTEN-FREE (CF (CLASS 4)) FLOURS). 49

TABLE 3.5: PLSR MODEL RESULTS FOR CORN FLOUR CONTAMINATED WITH WHEAT FLOUR SAMPLES USING DIFFERENT PRE-PROCCESSING METHODS. 51

TABLE 3.6: RESULTS OF PLSR MODELS FOR CORN FLOUR CONTAMINATED WITH BARLEY FLOUR SAMPLES USING DIFFERENT PRE-PROCESSING METHODS. 52

TABLE 3.7: PLSR MODEL RESULTS AFTER DIFFERENT PRE-PROCESSING METHODS FOR CORN FLOUR CONTAMINATED WITH RYE FLOUR. 53

TABLE 4.1: CONFUSION MATRIX PARAMETERS FOR THE MAJORITY VOTING-BASED LEARNING CLASSIFICATION TRAINING MODEL (CLASS 1: NO CONTAMINATION, CLASS2: CONTAMINATED WITH WHEAT) 74

TABLE 4.2: CONFUSION MATRIX TABLE FOR THE MAJORITY VOTING-BASED ENSEMBLE LEARNING CLASSIFICATION TRAINING MODEL 74

TABLE 4.3: CONFUSION MATRIX PARAMETERS FOR THE CLASSIFICATION TEST MODEL (CLASS 1: NO CONTAMINATION, CLASS2: CONTAMINATED WITH WHEAT) 74

TABLE 4.4: CONFUSION MATRIX TABLE FOR THE MAJORITY VOTING-BASED ENSEMBLE LEARNING CLASSIFICATION TEST MODEL 75

TABLE 4.5: PREDICTION ANALYSIS ON A DIFFERENT LEARNING ALGORITHM..... 79

TABLE 5.1: CONTENT (REAGENTS PROVIDED) OF EACH ELISA KIT 86

TABLE 5.2: QUANTIFICATION OF THE AMOUNT OF GLUTEN IN PPM FOR THE RAW FLOUR SAMPLES CONTAMINATED WITH WF 92

TABLE 5.3: QUANTIFICATION OF THE AMOUNT OF GLUTEN IN PPM FOR THE PROCESSED FLOUR (BREAD) SAMPLES CONTAMINATED WITH WHEAT FLOUR (WF)..... 93

LIST OF FIGURES

FIGURE 2.1: SUPERVISED MACHINE LEARNING APPROACH TO REAL LIFE EXPERIENCE (KOTSIANTIS ET AL., 2007) 18

FIGURE 3.1: THE MEAN SPECTRA OF THE DIFFERENT PURE FLOUR SAMPLES. 43

FIGURE 3.2: (A) RAW SPECTRA OF CORN FLOUR (CF) CONTAMINATED WITH BARLEY FLOUR (BF), (B) RAW SPECTRA OF CORN FLOUR (CF) CONTAMINATED WITH WHEAT FLOUR (WF), AND (C) RAW SPECTRA OF CORN FLOUR (CF) CONTAMINATED WITH RYE FLOUR (RF). THE DIFFERENT CONTAMINATION LEVELS OF 0.5% - 10% AT 0.5% INCREMENT IS REPRESENTED BY THE DIFFERENT COLORED SPECTRUM. 44

FIGURE 3.3: (A) SPECTRA PRE-PROCESSED BY SMOOTHING (1ST DERIVATIVES) (B) SPECTRA PRE-PROCESSED BY STANDARD NORMAL VARIATE (SNV) AND (C) SPECTRA PRE-PROCESSED BY MULTIPLICATIVE SCATTER CORRECTION (MSC). 46

FIGURE 4.1: RAW SAMPLE OF FTIR-SPECTRA OF THE CORN-FLOUR CONTAMINATED WITH 0.5% WHEAT FLOUR 70

FIGURE 4.2: BAKED SAMPLE OF FTIR-SPECTRA AFTER THE CORN-BREAD CONTAMINATED WITH 0.5% WHEAT FLOUR 71

FIGURE 4.3: PLOT OF NUMBER OF PRINCIPAL COMPONENTS (PCS) AND VARIANCE EXPLAINED IN THE SAMPLES. 73

FIGURE 4.4: VALIDATION CURVE FOR DECISION TREE REGRESSOR. 76

FIGURE 4.5: VALIDATION CURVE FOR K-NEAREST NEIGHBORS REGRESSOR 76

FIGURE 4.6: VALIDATION CURVE FOR PARTIAL LEAST SQUARE REGRESSION (PLSR) 77

FIGURE 4.7: VALIDATION CURVE FOR RANDOM FOREST REGRESSOR. 77

FIGURE 4.8: VALIDATION CURVE FOR SUPPORT VECTOR MACHINE 78

FIGURE 5.1: ELISA STANDARD CURVE 91

CHAPTER 1. INTRODUCTION

1.1 General Background

1.1.1 Gluten Contamination

Foods containing gluten are not suitable for people with gluten-related health implications such as celiac disease, wheat allergy, and gluten intolerance. People suffering from these disorders need to strictly stay away from gluten-containing foods to avoid any kind of health complications such as a bloated stomach, extreme fatigue, bone pain, muscle pain, headaches, etc., and in some critical cases, there can be an occurrence of anaphylaxis, a life-threatening allergic response (Elli et al., 2015). Thus, they have to depend on a "gluten-free" diet as the only means of dealing with these health issues.

Food rich in gluten can be described as any food containing the three major gluten-rich grains, which are wheat, barley, and rye grain. For food to be considered as "gluten-free", it must have a limit of $\leq$ 20 ppm of gluten from any of these grains or their crossbreeds (Lacorn et al., 2017). However, food or diet completely free from gluten would be hard to sustain. Higher trace amounts of gluten may be found in gluten-free products on the market due to cross-contact during the growing of grains alongside the gluten-rich grains, harvesting, transporting or food processing that requires the use of the same food-processing equipment or kitchen space for both non-gluten and gluten-rich grains (Thompson, 2003; Thompson et al., 2010). Contamination of foods with gluten has been a major challenge of "gluten-free" products which are required to attain the regulatory (the

U.S Food and Drug Administration) threshold limit. Valdés et al. (2003) tested over 3,000 products and reported that in Europe, one-third of gluten-free foods may contain over 20 ppm of gluten. Also, in the United States and Sweden, gluten contamination is a major issue. According to reports by Størsrud et al. (2003) and Thompson (2004), there was high contamination of gluten in the majority of the oats-based foods purchased from the market. In another study, Lee et al. (2014) analyzed 78 samples of foods in the U.S. market with gluten-free label on them and reported that 16 samples (20.5%) of the total samples have gluten levels of > 20 ppm, varying between 20.3-60.3 ppm. Specifically, five out of eight cereal food samples for breakfast showed gluten contents above 20 ppm. The results obtained justify a need for more reliable rapid means of checking for gluten contamination in foods and a need to ensure that food labeled "gluten-free" is safe for consumption for the people susceptible to gluten. Therefore, the use of a non-destructive method, Fourier transform infrared (FTIR) spectroscopy was explored in this study.

1.1.2 FTIR Spectroscopy method coupled with the machine learning process

Spectroscopy is a study that uses optical technology to evaluate or measure the interaction between electromagnetic radiation and material, or samples, involved at different wavelengths (spectrum). This has become a major contactless means of carrying out precise quality control and examination of food constituents such as sugars, protein, lipids, and other different chemical compounds (El-Mesery et al., 2019). The principle of Fourier transform infrared spectroscopy is that of molecular bond absorption of light energy frequency in the electromagnetic continuum which depends upon the state

(vibrational, electronic, or rotational). The intensity of the absorption is estimated at different uniform wavenumbers. The FTIR spectroscopy technique uses the Fourier transformation principle to generate or convert the readings at the detector to a frequency spectrum depicting a molecular "fingerprint" of a sample or material being measured. FTIR spectroscopy collects high-resolution spectra data simultaneously and covers a wide range of spectra features. The spectra data generated contains highly correlated features including noise and redundant features at each wavenumber. The data are arranged in an array format making it applicable to being assessed using different chemometrics (the science of using data-driven means to extract information from chemical systems).

The machine learning (ML) approach is a data-driven process that involves using computer algorithms to learn from previous or past information without explicitly being programmed. The procedure starts with the collection or observation of data through examples, direct experience, experimental setup to identify a pattern in the data collected or observed, and make an informed decision in the future depending on the domain of the information. The basic motive is to allow a computer to learn automatically without human interference or aid and calibrate actions accordingly.

The overall goal of this study was to integrate the principle of FTIR spectroscopy to obtain spectra data with unique chemical and structural information about the samples used, learn or identify patterns in the spectra data using supervised ML approaches, and thus, prototype ML models that can be utilized to authenticate gluten contamination based on what has been learned from the spectra features. The study is divided into three phases: authentication of cross-contamination of gluten-rich and non-gluten raw flour samples,

authentication of cross-contamination of wheat-flour in processed samples (baked bread), and execution of a complementary analysis using enzyme-linked immunosorbent assay with the following objectives listed below.

1.2 Specific Objectives

The specific objectives of this research include:

1. The detection and quantification of gluten-rich flour (wheat flour, barley flour and rye flour) contamination in a non-gluten flour (corn flour) using FTIR coupled with machine learning approaches at the contamination levels of 0-10% with 0.5% increments.
2. Application of FTIR with machine learning approaches for quantification of wheat flour contamination in non-gluten bread (cornbread) at the contamination levels of 0-10% with 0.5% increments.
3. Enzyme-linked immunosorbent assay (ELISA) to establish the regulatory gluten-free labeling threshold (≤ 20 ppm) for the wheat flour (WF) contamination levels in objectives 1 and 2.

At the end of the research, the expectation is to obtain ML models that can be integrated into a software system (e.g. mobile or computer application) with the ability to detect and quantify cross-contamination between a non-gluten and gluten flour sample. Also, if possible, estimate the amount of gluten present in the contaminated sample within the domain used.

CHAPTER 2. LITERATURE REVIEW

This part of the study gives a comprehensive review of gluten and the health implications (symptoms) associated with gluten-related disorders. Also, Fourier transformed infrared (FTIR) spectroscopy with its application to food analysis, quality control, and inspection, and machine learning approaches is briefly discussed below.

2.1 Gluten Overview

Gluten is a type of protein family that mainly exists in wheat, rye, barley, and their crossbred varieties. Gluten also exists in food products that contain extracted or pure gluten as a source of protein or binding agent (Biesiekierski, 2017). In some cases, due to cross-contamination, gluten-free products may also contain gluten in the process of harvesting, transporting, storage of grain and/or during the process of manufacturing a gluten-free food products (Sharma et al., 2015). Gluten composites are prolamins and glutelins. Prolamins are poorly soluble in water but highly soluble in alcohol. Prolamins can be extracted using 40-70% ethanol. In barley, rye, wheat and oats grains, prolamins are referred to as hordeins, secalins, gliadins and avenins, respectively. Glutelin fraction is soluble in dilute acids or alkali solutions, and the wheat glutelins are called glutenin (Shewry et al., 2002).

Also, gluten protein can be grouped based on the amount of sulphur they contain, their structural size, or properties (Kanerva, 2011). For example, prolamins are monomeric and are characterized by weak hydrogen bonds, intramolecular disulfide bonds, and easily soluble in water-alcohol mixtures (Waga, 2004). However, glutelins are polymerics and

also contain intermolecular bonds that adjust them to form a network when cooked or heated.

2.1.1 Effects of Processing on Gluten Proteins

Different conversion processes affect and modify gluten protein in a variety of ways. For example, baking and cooking processes denature gluten protein by forming new disulfide bonds and aggregates which makes it more difficult to extract the gluten proteins. This may result in lower gluten protein solubility and lead to a lower rate of detection that will require modification of the extraction protocol (Hayta & Alpaslan, 2001). The extrusion process has a major effect on the solubility of the protein structure. Extrusion process involves a redox reaction that modifies the protein's secondary structure caused by heat and shear of the extrusion. These modifications in the structure of proteins and starches are crucial for the final properties of the product (Camire, 1998). The process of fermentation and hydrolysis has been reported by Kanerva (2011) to break down protein into smaller fragments resulting in a decrease and difficulty in the identification and estimation of the proteins. Other processes that can affect the solubility and detectability of protein include deamination, transamination, mixing, sheeting, drying, etc. (Hayta & Alpaslan, 2001). Therefore, these processes should be considered when testing for gluten proteins on these products as they affect the result obtained.

2.1.2 Gluten-Related Disorders

Most people can tolerate gluten-rich foods but for some people, it causes several kinds of immune responses and other physiological reactions. Some of the gluten-related disorders such as celiac disease, wheat allergy, gluten intolerance, dermatitis herpetiformis, and gluten ataxia are discussed below.

2.1.2.1 Celiac disease

Celiac disease (CD) is an auto-immune reaction that causes damages to the intestinal villi (small intestine) that can lead to inflammation and less nutrients absorption when foods containing gluten are consumed by susceptible individuals (Meresse et al., 2012). Lebwohl et al. (2015a) reported that the symptoms of celiac disease vary widely including both intestinal and extra-intestinal. The symptoms of celiac disease are often similar to the symptoms of other gluten-reltated disorders or diseases such as lactose intolerance which complicates diagnosis. In adults, these symptoms range from diarrhea, weight loss, bloating, abdominal pain, infertility, neurological or psychiatric problems, to vitamin deficiencies. Additionally, infants and children usually have symptoms of diarrhea, and abnormal stretching of the abdomen, dental defects, anemia, developmental delay (Lebwohl et al., 2015a). Furthermore, other symptoms may vary and include a bloated stomach, breathing difficulties, mouth ulcers, extreme fatigue, bone pain, hives, nausea, inability to focus, and in some critical cases there can be an occurrence of anaphylaxis, a life-threatening allergic response (Nordqvist, 2018).

According to a study by Rubio-Tapia et al. (2012), the common cases of CD in the United States was 0.71% (1 in 141) and similar with that of many European countries, while the widely accepted prevalence of CD as of recent in the United States is at 1% (1 in 133 of average healthy people) (Fueyo-Díaz et al., 2019). Also, the incidence of diagnosed CD has been reported by Lebwohl et al. (2015a) to be increasing with data from a North American country indicating a steady rise in occurrences from 1950s reaching 17 per 100,000 people each year from 2008 to 2011. Several factors that contribute and affect the prevalence and incidence of CD include genetics, exposure to gluten, infant feeding patterns, awareness of the disease (among medical practitioners and patients), frequency of testing and other environmental risk factors (Lebwohl et al., 2015b).

2.1.2.2 Wheat Allergy

Inomata (2009) defines a wheat allergy as adverse immunological reactions that are caused by proteins found in wheat. These reactions are not only due to gluten but may be triggered by other proteins found in wheat including albumins (dissolvable in water and harden by heat) and globulins (dilute in a solution of salt) (Tatham & Shewry, 2008). The symptoms associated with the reactions may rapidly progress from tolerable to acute symptoms. In children, wheat consumption can cause bronchial obstruction, urticaria, nausea, angioedema, and abdominal pain, or in acute manifestation systemic anaphylaxis. Impeded supersensitivity symptoms may appear within 24 hours after the consumption of food-containing-wheat and include gastrocolic symptoms and aggravation of atopic dermatitis (Majamaa et al., 1999; Varjonen et al., 2000). In adults, allergies of food related

to wheat ingestion seem to be rare and can be described as an anaphylactic reaction typically caused by workout activities (Crespo & Rodriguez, 2003). These allergic reactions can be triggered in a few minutes or hours of the food consumption and if not managed properly, can result in a critical condition or state.

2.1.2.3 Gluten sensitivity or intolerance

Gluten sensitivity (GS) or non-celiac gluten sensitivity (NCGS) is a non-allergic or non-autoimmune response to gluten (Schuppan et al., 2015). Any response or reaction that is not triggered by the body's immune system when gluten-containing foods are consumed is termed NCGS. People susceptible to NCGS also experience gastrointestinal symptoms such as fatigue, constipation, abdominal pain, diarrhea, skin rashes, muscle pain, headaches, eczema, bloating, anemia, and depression (Fasano et al., 2015). These symptoms usually appear after gluten has been consumed and then disappear when gluten is no longer being consumed. Opposing to CD and wheat allergy, there is no clear histopathologic basis for physicians to confirm the pronouncement of NCGS (Elli et al., 2015). Also, in the United States, the prevalence of NCGS has been estimated to be up 6% of the American population (Mooney et al., 2013).

2.1.2.4 Dermatitis Herpetiformis

Dermatitis herpetiformis (DH) is an autoimmune-related chronic skin condition that occurs in genetically susceptible people when exposed to gluten-rich foods. DH affects about 10-15% of patients with gluten-sensitive enteropathy (celiac disease). The presence

of digestive symptoms is not frequent in DH, people of all ages can be affected by DH but usually seen in those between the ages of 30 and 40 for the first time. Additionally, the development of DH tends to show up more in northern Europeans than in Africans or Asians (Celiac Disease Foundation, 2020). The clinical conditions of DH are characterized by grouped polymorphic lesions comprising of erythema, papules, and urticarial plaques, including the extensor surfaces of the elbows, knees, hindquarters, sacral locale, shoulders, neck, face, and scalp (Antiga & Caproni, 2015). Furthermore, Antiga and Caproni (2015) reported that sometimes patients may show erythema or serious pruritus alone, in this manner making the diagnosis of DH more difficult.

2.1.2.5 Gluten Ataxia

Hadjivassiliou et al. (2003) reported gluten ataxia as an immunologically intervened illness, gluten sensitivity spectrum part, and records for up to 40% of instances of idiopathic sporadic ataxia. Gluten ataxia is related to celiac disease but it mainly affects the brain and central nervous system with no gastrointestinal symptoms (Hadjivassiliou et al., 2002). Specifically, the cerebellum is attacked by the antibodies produced because of the response of the immune system when food-containing gluten is consumed (Hadjivassiliou et al., 2015), which may lead to certain effects such as fatigue, inability to balance, nausea, vomiting, loss of coordination, speaking difficulties, swallowing difficulties, and abnormal gait or difficulty walking (Gluten Free Society, 2020).

2.2 Method of Fourier transformed infrared spectroscopy (FTIR)

Infrared spectroscopy involves the interaction of electromagnetic radiation in the infrared region of a spectrum (infrared light) with a molecule. The infrared region is usually between 4000-400 cm^{-1} where the cm^{-1} unit is the wavenumber scale and is given by 1/frequency (wavelength in cm). Excitation of the vibration of the covalent bonds within a molecule is trigerred by the infrared radiation (IR) and can incorporate stretching and bending modes. Fourier transformed infrared spectroscopy (FTIR) applies the principle of infrared spectroscopy and using a mathematical method called Fourier transform (FT) to change over time space domain to traditional frequency domain spectrum to decode all the reading or recording (interferogram) from the spectroscopy detector (Baravkar et al., 2011; Doyle, 1992).

Ismail et al. (1997) described an infrared spectrometer to be essentially made up of a steady source of infrared light energy, a technique for changing or transforming the infrared radiation into its component wavelengths (a fixed or moving mirror or a beam splitter), and a detector. The equation below mathematically defines the process of obtaining an IR range from a sample.

$$T(\overline{v}) = {I(\overline{v})}\Big/{I_0(\overline{v})} \tag{2.1}$$

where T designates the transmittance, I is defined as the intensity of the IR in contact with the detector when the sample is placed between the source and the detector, I_0 is the

intensity reaching the detector without any sample in between the beam, and $\overline{v}$ designates the wavenumber of the IR. Theoretically, the spectrum is obtained by measuring the transmittance at equally spaced wavenumber intervals, $\Delta\,\overline{v}$, where, $\Delta\,\overline{v}$ is defined as the resolution. Usually, the y-axis of the spectrum is converted from units of percent transmittance ($\%T = 100 \times T$) to absorbance (A) units using the relational equation below.

$$A = -log\,T \tag{2.2}$$

Furthermore, the advantages of the FTIR method include being a fast, label-free, and non-destructive method that provides several spatially settled infrared spectra containing chemical and structural information of a molecular compound presented in an array format. The rich information that is contained in the spectra data obtained allows multiple functional groups in the molecular compound to be tracked using the intensity of the peaks formed. Different mathematical analysis or modeling can be carried out due to the data formation in an array (Kazarian & Chan, 2013).

Many researchers have shown evidence of using FTIR and some other spectroscopy methods with ML algorithms as fast and non-destructive means of food safety, quality inspection, and control. Recently, Sujka et al. (2017) examined the use of FTIR spectroscopy for quality assessment of flours acquired from Polish producers. In the study, 11 flour types from various grains (wheat, rye, spelt, triticale, and spelt bran) were investigated. There physical and chemical composition were obtained and FTIR spectra data were correlated with reference results using classical square regression (CSR) and partial least square regression (PLSR). The author noted high linear correlations between

the real and estimated or predicted values of the parameters examined. The simpler CSR procedure was noted to produced preferred outcomes over the PLS strategy. A quick strategy for an examination of potential wheat flour added to oat flour was created by Wang et al. (2014) using FT-NIR spectrometry and chemometrics. FTIR-spectra data of samples of unadulterated oat and wheat flours were obtained with adulteration levels of 5- 50% at 5% increment measured within the working range of 4000–12000 cm^{-1} and partial least squares regression (PLSR) models were created on both raw and pre-processed (standard normal variate) data with Monte Carlo cross-validation. For all of the PLS models, the differences between root mean square error of prediction (RMSEP = 1.921) and root mean squared error of Monte Carlo cross-validation (RMSEMCCV = 1.975). Three or four component PLS models were highlighted to accurately predict the levels of wheat flour in oat flour. Amir et al. (2013) applied FTIR spectroscopy for the identification of wheat assortments. Four economically accessible wheat assortments were studied for their physical, chemical, and rheological properties using standard method and advanced FTIR technique. It was observed that FTIR provided an excellent means to visualize the chemical composition of the different wheat varieties with an added advantage of being very quick, reliable, and cheaper over the use of the standard method. In another study, Duarte et al. (2002) used FTIR spectroscopy to quantify the amount of sugars (as a function of ripening) in mango juice. Mango juices obtained from the Tommy Atkins mango cultivar grown in Brazil were used and a six concentration levels of different types of sugar (glucose, fructose, and sucrose) arranged by a triangular test design (1 solution of each sugar, 9 binary mixtures, and 10 ternary mixtures) as the calibration set were utilized. The author

concluded that FTIR coupled with partial least squares (PLS) regression and calibrated by triangular model of standard sugar solutions has the potential to authenticate the amount of sucrose (1.4 prediction error), fructose (1.4 prediction error), and glucose (4.9% prediction error) in mango juices got from the fruits at different ripening degrees. FTIR spectroscopy was also utilized to analyze the defilement of extra virgin oil with palm oil by Rohman and Man (2010). Samples of pure extra virgin olive oil (EVOO) and those defiled with palm oil (PO) in accurately weighted proportions of 1-50% were classified using discriminant analysis. The quantification method explored the use of PLS and principal component regression (PCR) at FTIR wavenumber region ranging from 1500–1000 cm^{-1}. The performance metrics (R^2 = 0.999 and RMSE of cross-validation of 0.285 (PLS) and 0.373 (PCR)) obtained from the study indicates the effectiveness of using FTIR spectroscopy for the evaluation of PO in EVOO. Other spectroscopy methods including near-infrared reflectance and Fourier transform (FT) Raman spectroscopy have shown great potential as a reliable, cheaper and rapid non-destructive method of food quality analysis. BAŞLAR and Ertugay (2011) used the capability of near-infrared reflectance spectroscopy (NIRS) for the assurance of protein, dry and wet gluten contents and Zeleny sedimentation of wheat flour. Wheat bread samples (120 samples) were used and their NIRS data were recorded at 2nm intervals from 1100–2500 nm using NIRS systems 6500 scanning spectrophotometer (Foss NIRSystems Inc., USA) in reflectance mode. The results acquired showed that the execution of the NIRS for wet and dry gluten appears to be strongly subjected on the correlation to protein content. Czaja et al. (2016a) evaluated FT-Raman spectroscopy for quantification of gluten in wheat flour. FT-Raman spectra were collected for three groups

of samples including pure wheat, wheat adulterated with 2-5% starch or 2-4% gluten, and wheat adulterated with starch, dietary fiber, and corn oil in the region of 100–3700 cm^{-1} at a resolution of 8 cm^{-1}. Partial least squares regression (PLSR) models were implemented with principal component analysis, and pre-treated using mean value normalization, multiplicative scatter correction algorithms, and standard normal variate (SNV). FT-Raman spectroscopy was observed to have a very high potential for gluten evaluation in wheat flour. Furthermore, FTIR spectroscopy has also been increasingly applied in some other areas of food research, such as nuts (Ciemniewska-Żytkiewicz et al., 2015; Dogan et al., 2007), quality assessment of fats and dampness determination in butter (Van de Voort et al., 1992a), oils (Quiñones-Islas et al., 2013), cakes and flakes (Reder et al., 2014), honey adulteration (Gallardo-Velázquez et al., 2009) and meat (Rohman et al., 2011; Xu et al., 2012).

In this study, FTIR coupled with machine learning approaches was used to inspect and assess gluten contamination in grain-based foods. From the previous studies, FTIR has been widely used and can be seen to be effective for food inspection and quality assessment. The potential of using FTIR coupled with different chemometrics can also be justified.

2.3 Machine Learning Approaches

2.3.1 An Overview on Machine Learning

Machine learning (ML) involves the technique of learning from a set of data to execute a task. The data learned from include data from a previous experience, this could be historical data collected over a period of time or an organized set of data which can be collected through an experimental design or laboratory setup. The data is made up of a set of examples, each example is characterized by a set of attributes, otherwise called features or variables. This can be represented in the form of nominal (e.g. gender, age, race, etc.), binary (accepts two possible values, true or false, represented by 0 or 1), ordinal (an ordered form of categorical values e.g. educational level: elementary school, high school, college, graduate school), or numeric (measurable data e.g height, width, heart rate, etc) (Liakos et al., 2018). Ayodele (2010a) defined machine learning as a process of developing a system of computer that automatically learns and improves with experience. Furthermore, the focal point of many scientific disciplines is to model a function that relates between a lot of observables features (inputs) and another arrangement of features that are identified with these (outputs). The mathematical model created would then be able to be utilized and to potentially anticipate the estimation of the desired variables by estimating the observables. In reality, some genuine encounters are too compounded to be directly modeled as a closed system of input to output relationship. Therefore, machine learning gives strategies that can naturally construct and alter a computational model that fits into these complex

connections through augmenting information from the past experience and executing assessment to limit mistakes or minimize error (Baştanlar & Özuysal, 2014).

2.3.2 Types of Machine Learning (ML) Approaches

According to Ayodele (2010b) the different types of approaches in ML algorithms include supervised learning, unsupervised learning, and reinforcement learning. Also, the ML algorithms can be dependent on the type of their input and output data, and the intended type of task to be executed or problem to be solved. Supervised learning instances use labeled examples – known input X and the corresponding output Y is used to train a learning algorithm that would predict the relationship between X and Y (P(Y|X)). This is typically utilized for developing classification and regression models (Lee et al., 2018a). In contrast, unsupervised learning uses unmarked or unlabeled examples (the input of X value only) to learn and make predictions (P(X)) and it's mainly applicable to clustering tasks, compression, feature extraction, etc. When some training examples are missing training labels, an approach of semi-supervised learning which uses unlabeled examples in conjunction with labeled examples to help gain proficiency on the probability distribution over the input space P(X) is utilized in order to produce a considerable improvement in the learning accuracy.

Reinforcement learning combines the learning process of the input X with an acting phase (critic (C)) to simultaneously learn and achieve a self-optimizing feature. The training information that is made available to the learning algorithm by the environment (external trainer) is a scalar fortification sign that comprises a proportion of the operational

accuracy of the system. The learning algorithm isn't coordinated to which actions to take, yet rather should find the activities that yield the best result, by attempting each activity in a steady progression (Baştanlar & Özuysal, 2014; Kotsiantis et al., 2007).

This research focuses on the use of supervised machine learning approaches. The flow chart in Figure 1 below displays a supervised machine learning (ML) application process towards model prototyping and was utilized in developing our models throughout this work.

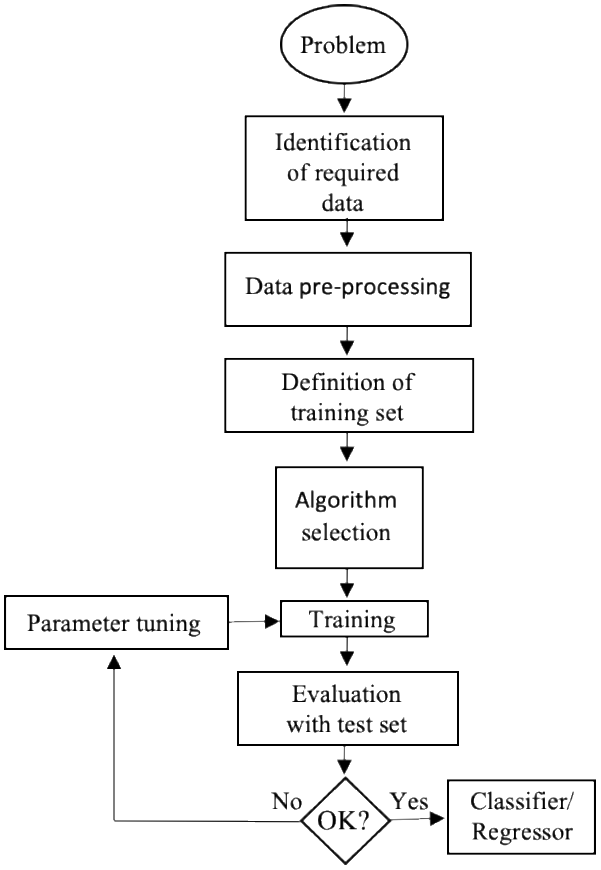


Figure 2.1: Supervised machine learning approach to real-life experience (Kotsiantis et al., 2007)

The process begins with obtaining the required dataset from the area or region of interest, which requires identifying the most informative fields, features, or attributes. This could either be done by an expert with vast knowledge in that area or by the least complex strategy of using "brute-force," which involves estimating and considered every feature with the expectation that the important or relevant features can be confined. However, the process of "brute-force" dataset collection does not directly work well with induction. In most cases, it tends to contain noise, redundant features, or missing feature values, and in this way would require huge pre-processing (Zhang et al., 2003).

Data pre-processing is the next step after the required data have been identified. In real-world experience, data representation is often very complex and has too many features with only a few related to the targeted objective(s). There are usually redundant variables, where only a few features are correlated and needed for modeling; and interdependence, where at least two features collected pass on significant information that is unclear if one feature is incorporated without the other (Guyon & Elisseeff, 2003). Data pre-processing can help remove redundant data or eliminate noise. It can also be used to select the most informative features that would significantly affect speculation execution of a supervised ML algorithm. It incorporates data preparation exacerbated by integration, cleaning, standardization, and data transformation; and data reduction tasks such as instance selection, feature selection, discretization, etc. The outcome of a dependable and successful execution of data pre-processing task is a useful and adequate final dataset selected for subsequent data analysis such as classification and predictive modeling (García et al., 2015). In spectra transformation or preprocessing, several pre-processing methods can be

used including smoothing, multiplicative scatter correction (MSC), standard normal variate (SNV), derivatives (Savitzky–Golay), normalization, etc. The functions of these techniques differ and are based on the circumstances, for example smoothing by Savitzky–Golay method and first-derivative transforms can be used to eliminate noise and baseline offset discrepancies from a set of spectra data respectively, while the second-derivative transforms are useful in separating protruding peaks and tapered spectra features (Cen & He, 2007; Wu et al., 1995). The impacts of non-uniform dissipating obstructions and particle size throughout a spectrum can be eliminated using MSC (by using calculated mean spectrum of the dataset) and SNV (by normalizing every spectrum utilizing just the information from that specific range) (Barbin et al., 2012; Maleki et al., 2007).

Kotsiantis et al. (2007) reported that the algorithm selection is a critical step done by preliminarily testing different algorithms and once the evaluation criteria are satisfied, the best performing algorithm can be selected for routine use. Over a decade, various supervised ML algorithms have been shown in studies to be effective in the classification of protein structure such as support vector machine (Cai et al., 2001; Shamim et al., 2007), decision trees (Çamoglu et al., 2005), neural networks (Chung et al., 2003; Ding & Dubchak, 2001), ensemble learning methods (Saha et al., 2014; Tan et al., 2003), random forest (Dehzangi et al., 2010), partial least square regression and others.

The training procedure usually involves splitting the dataset by using about 70-80% for training (training set) and the other 20-30% for evaluating performance (test set). Another strategy, known as cross-validation which may involve dividing the training set into fundamentally unrelated and equivalent measured subsets and for every subset, the

classifier is trained on the combination of all the other subsets. The error rate of the regressor/classifier is then estimated by averaging the rate of the error of each subset (Kotsiantis et al., 2007).

The performances of these algorithms are often evaluated based on the purpose of usage (classification or prediction). The classifier's assessment is regularly founded on obtaining the confusion matrix (CM) parameters by computing the true positives (TP); the number of effectively perceived class tests, true negatives (TN); the quantity of accurately perceived examples that are not part of the class, and false positives (FP); samples that were either erroneously allotted to the class or false negatives (FN); that were not perceived as class samples (Sokolova & Lapalme, 2009). Other measures can be calculated based on the scores of TP, FP, FP and FN such as the precision, sensitivity or recall (true positive rate); extent of positive cases that were accurately recognized, and specificity (true negative rate); extent of negatives cases that were accurately recognized by the algorithm (Forbes, 1995). The assessment of the performance of the regressor model is usually done by computing the statistical parameters such as the coefficient of determination (R^2) and root mean square error (RMSE). After the evaluation process using the test data, if the outcome of the analysis satisfies our desired result(s), the classifier or regressor is then selected or deployed for future classification or prediction.

References

Amir, R. M., Anjum, F. M., Khan, M. I., Khan, M. R., Pasha, I., & Nadeem, M. (2013). Application of Fourier transform infrared (FTIR) spectroscopy for the identification of wheat varieties. *Journal of food science and technology, 50*(5), 1018-1023.

Antiga, E., & Caproni, M. (2015). The diagnosis and treatment of dermatitis herpetiformis. *Clinical, cosmetic and investigational dermatology, 8*, 257.

Ayodele, T. O. (2010a). Machine learning overview. *New Advances in Machine Learning*, 9-19.

Ayodele, T. O. (2010b). Types of machine learning algorithms. *New Advances in Machine Learning*, 19-48.

Baravkar, A., Kale, R., & Sawant, S. (2011). FTIR Spectroscopy: principle, technique and mathematics. *International Journal of Pharma and Bio Sciences, 2*(1), 513-519.

Barbin, D. F., ElMasry, G., Sun, D.-W., & Allen, P. (2012). Predicting quality and sensory attributes of pork using near-infrared hyperspectral imaging. *Analytica Chimica Acta, 719*, 30-42.

BAŞLAR, M., & Ertugay, M. F. (2011). Determination of protein and gluten quality-related parameters of wheat flour using near-infrared reflectance spectroscopy (NIRS). *Turkish Journal of Agriculture and Forestry*, *35*(2), 139-144.

Baştanlar, Y., & Özuysal, M. (2014). Introduction to machine learning. In *miRNomics: MicroRNA Biology and Computational Analysis* (pp. 105-128): Springer.

Biesiekierski, J. R. (2017). What is gluten? *Journal of gastroenterology and hepatology, 32*, 78-81. doi:10.1111/jgh.13703.

Cai, Y.-D., Liu, X.-J., Xu, X.-b., & Zhou, G.-P. (2001). Support vector machines for predicting protein structural class. *BMC bioinformatics, 2*(1), 3.

Camire, M. E. (1998). Chemical changes during extrusion cooking. In *Process-induced chemical changes in food* (pp. 109-121): Springer.

Çamoglu, O., Can, T., Singh, A. K., & Wang, Y.-F. (2005). Decision tree based information integration for automated protein classification. *Journal of Bioinformatics and Computational Biology, 3*(03), 717-742.

Celiac Disease Foundation. (2020). Dermatitis Herpetiformis. *[Online]*.

Cen, H., & He, Y. (2007). Theory and application of near infrared reflectance spectroscopy in determination of food quality. *Trends in food science & technology, 18*(2), 72-83.

Chung, I.-F., Huang, C.-D., Shen, Y.-H., & Lin, C.-T. (2003). Recognition of structure classification of protein folding by NN and SVM hierarchical learning architecture. In *Artificial Neural Networks and Neural Information Processing—ICANN/ICONIP 2003* (pp. 1159-1167): Springer.

Ciemniewska-Żytkiewicz, H., Bryś, J., Sujka, K., & Koczoń, P. (2015). Assessment of the hazelnuts roasting process by pressure differential scanning calorimetry and MID-FT-IR spectroscopy. *Food Analytical Methods, 8*(10), 2465-2473.

Crespo, J. F., & Rodriguez, J. (2003). Food allergy in adulthood. *Allergy, 58*(2), 98-113. doi:10.1034/j.1398-9995.2003.02170.x

Czaja, T., Mazurek, S., & Szostak, R. (2016). Quantification of gluten in wheat flour by FT-Raman spectroscopy. *Food Chemistry, 211*, 560-563. doi:https://doi.org/10.1016/j.foodchem.2016.05.108

Dehzangi, A., Phon-Amnuaisuk, S., & Dehzangi, O. (2010). Using random forest for protein fold prediction problem: an empirical study. *J. Inf. Sci. Eng., 26*(6), 1941-1956.

Ding, C. H., & Dubchak, I. (2001). Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics, 17*(4), 349-358.

Dogan, A., Siyakus, G., & Severcan, F. (2007). FTIR spectroscopic characterization of irradiated hazelnut (Corylus avellana L.). *Food Chemistry, 100*(3), 1106-1114.

Doyle, W. M. (1992). Principles and applications of Fourier transform infrared (FTIR) process analysis. *Process Control Qual, 2*(1), 11-41.

Duarte, I. F., Barros, A., Delgadillo, I., Almeida, C., & Gil, A. M. (2002). Application of FTIR spectroscopy for the quantification of sugars in mango juice as a function of ripening. *Journal of agricultural and food chemistry, 50*(11), 3104-3111.

Elli, L., Branchi, F., Tomba, C., Villalta, D., Norsa, L., Ferretti, F., . . . Bardella, M. T. (2015). Diagnosis of gluten related disorders: Celiac disease, wheat allergy and non-celiac gluten sensitivity. *World journal of gastroenterology: WJG, 21*(23), 7110.

Fasano, A., Sapone, A., Zevallos, V., & Schuppan, D. J. G. (2015). Nonceliac gluten and wheat sensitivity. *148*(6), 1195-1204.

Forbes, A. D. (1995). Classification-algorithm evaluation: Five performance measures based onconfusion matrices. *Journal of Clinical Monitoring, 11*(3), 189-206.

Fueyo-Díaz, R., Magallón-Botaya, R., Masluk, B., Palacios-Navarro, G., Asensio-Martínez, A., Gascón-Santos, S., . . . Sebastián-Domingo, J. J. (2019). Prevalence of celiac disease in primary care: the need for its own code. *BMC Health Services Research, 19*(1), 578. doi:10.1186/s12913-019-4407-4

Gallardo-Velázquez, T., Osorio-Revilla, G., Zuñiga-de Loa, M., & Rivera-Espinoza, Y. (2009). Application of FTIR-HATR spectroscopy and multivariate analysis to the quantification of adulterants in Mexican honeys. *Food Research International, 42*(3), 313-318.

García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining*: Springer.

Gluten Free Society. (2020). Ataxia – Another Symptom of Gluten Induced Damage. *Gluten free society blog, nerve damage, nutritional deficiencies.* .

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research, 3*(Mar), 1157-1182.

Hadjivassiliou, M., Boscolo, S., Davies–Jones, G., Grünewald, R., Not, T., Sanders, D., . . . Woodroofe, N. (2002). The humoral response in the pathogenesis of gluten ataxia. *Neurology, 58*(8), 1221-1226.

Hadjivassiliou, M., Davies-Jones, G., Sanders, D., & Grünewald, R. (2003). Dietary treatment of gluten ataxia. *Journal of Neurology, Neurosurgery & Psychiatry, 74*(9), 1221-1224.

Hadjivassiliou, M., Sanders, D., & Aeschlimann, D. (2015). Gluten-related disorders: gluten ataxia. *Digestive Diseases, 33*(2), 264-268.

Hayta, M., & Alpaslan, M. (2001). Effects of processing on biochemical and rheological properties of wheat gluten proteins. *Food/Nahrung, 45*(5), 304-308.

Inomata, N. (2009). Wheat allergy. *Current opinion in allergy and clinical immunology, 9*(3), 238-243.

Ismail, A. A., van de Voort, F. R., & Sedman, J. (1997). Fourier transform infrared spectroscopy: principles and applications. In *Techniques and instrumentation in analytical chemistry* (Vol. 18, pp. 93-139): Elsevier.

Kanerva, P. (2011). *Immunochemical analysis of prolamins in gluten-free foods*: University of Helsinki.

Kazarian, S. G., & Chan, K. A. (2013). ATR-FTIR spectroscopic imaging: recent advances and applications to biological systems. *Analyst, 138*(7), 1940-1951.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering, 160*, 3-24.

Lebwohl, B., Ludvigsson, J. F., & Green, P. H. (2015a). Celiac disease and non-celiac gluten sensitivity. *Bmj, 351*, h4347.

Lebwohl, B., Ludvigsson, J. F., & Green, P. H. J. B. (2015b). Celiac disease and non-celiac gluten sensitivity. *351*, h4347.

Lee, J. H., Shin, J., & Realff, M. J. (2018). Machine learning: Overview of the recent progresses and implications for the process systems engineering field. *Computers & Chemical Engineering, 114*, 111-121.

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors, 18*(8), 2674.

Majamaa, H., Moisio, P., Majamaa, H., Turjanmaa, K., & Holm, K. (1999). Wheat allergy: diagnostic accuracy of skin prick and patch tests and specific IgE. *Allergy, 54*(8), 851-856.

Maleki, M., Mouazen, A., Ramon, H., & De Baerdemaeker, J. (2007). Multiplicative scatter correction during on-line measurement with near infrared spectroscopy. *Biosystems engineering, 96*(3), 427-433.

Meresse, B., Malamut, G., & Cerf-Bensussan, N. (2012). Celiac disease: an immunological jigsaw. *Immunity, 36*(6), 907-919.

Mooney, P., Aziz, I., & Sanders, D. (2013). Non- celiac gluten sensitivity: clinical relevance and recommendations for future research. *Neurogastroenterology & Motility, 25*(11), 864-871.

Nordqvist, C. B., N. . (2018). What is a wheat allergy? . *Retrieved from Medical News Today Website: https://www.medicalnewstoday.com/articles/174405.php*.

Quiñones-Islas, N., Meza-Márquez, O. G., Osorio-Revilla, G., & Gallardo-Velazquez, T. (2013). Detection of adulterants in avocado oil by Mid-FTIR spectroscopy and multivariate analysis. *Food Research International, 51*(1), 148-154.

Reder, M., Koczoń, P., Wirkowska, M., Sujka, K., & Ciemniewska-Żytkiewicz, H. (2014). The application of FT-MIR spectroscopy for the evaluation of energy value, fat content, and fatty acid composition in selected organic oat products. *Food Analytical Methods, 7*(3), 547-554.

Rohman, A., Erwanto, Y., & Man, Y. B. C. (2011). Analysis of pork adulteration in beef meatball using Fourier transform infrared (FTIR) spectroscopy. *Meat Science, 88*(1), 91-95.

Rohman, A., & Man, Y. C. (2010). Fourier transform infrared (FTIR) spectroscopy for analysis of extra virgin olive oil adulterated with palm oil. *Food Research International, 43*(3), 886-892.

Rubio-Tapia, A., Ludvigsson, J. F., Brantner, T. L., Murray, J. A., & Everhart, J. E. (2012). The prevalence of celiac disease in the United States. *American Journal of Gastroenterology, 107*(10), 1538-1544.

Saha, I., Zubek, J., Klingström, T., Forsberg, S., Wikander, J., Kierczak, M., . . . Plewczynski, D. (2014). Ensemble learning prediction of protein–protein interactions using proteins functional annotations. *Molecular BioSystems, 10*(4), 820-830.

Schuppan, D., Pickert, G., Ashfaq-Khan, M., & Zevallos, V. (2015). Non-celiac wheat sensitivity: differential diagnosis, triggers and implications. *Best Practice & Research Clinical Gastroenterology, 29*(3), 469-476.

Shamim, M. T. A., Anwaruddin, M., & Nagarajaram, H. A. (2007). Support vector machine-based classification of protein folds using the structural properties of

amino acid residues and amino acid residue pairs. *Bioinformatics, 23*(24), 3320-3327.

Sharma, G. M., Pereira, M., & Williams, K. M. (2015). Gluten detection in foods available in the United States – A market survey. *Food Chemistry, 169*, 120-126. doi:https://doi.org/10.1016/j.foodchem.2014.07.134

Shewry, P. R., Halford, N. G., Belton, P. S., & Tatham, A. S. (2002). The structure and properties of gluten: an elastic protein from wheat grain. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 357*(1418), 133-142. doi:10.1098/rstb.2001.1024

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management, 45*(4), 427-437.

Sujka, K., Koczoń, P., Ceglińska, A., Reder, M., & Ciemniewska-Żytkiewicz, H. (2017). The application of FT-IR spectroscopy for quality control of flours obtained from polish producers. *Journal of Analytical Methods in Chemistry, 2017*.

Tan, A. C., Gilbert, D., & Deville, Y. (2003). Multi-class protein fold classification using a new ensemble machine learning approach. *Genome Informatics, 14*, 206-217.

Tatham, A., & Shewry, P. (2008). Allergens to wheat and related cereals. *Clinical & Experimental Allergy, 38*(11), 1712-1726.

Van de Voort, F., Sedman, J., Emo, G., & Ismail, A. (1992). A rapid FTIR quality control method for fat and moisture determination in butter. *Food Research International, 25*(3), 193-198.

Varjonen, E., Vainio, E., & Kalimo, K. (2000). Antigliadin IgE–indicator of wheat allergy in atopic dermatitis. *Allergy*, *55*(4), 386-391.

Waga, J. (2004). Structure and allergenicity of wheat gluten proteins-a review. *Polish journal of food and nutrition sciences*, *13*(4), 327-338.

Wang, N., Zhang, X., Yu, Z., Li, G., & Zhou, B. (2014). Quantitative analysis of adulterations in oat flour by FT-NIR spectroscopy, incomplete unbalanced randomized block design, and partial least squares. *Journal of Analytical Methods in Chemistry*, *2014*.

Wu, W., Walczak, B., Massart, D., Prebble, K., & Last, I. (1995). Spectral transformation and wavelength selection in near-infrared spectra classification. *Analytica Chimica Acta*, *315*(3), 243-255.

Xu, L., Cai, C.-B., Cui, H.-F., Ye, Z.-H., & Yu, X.-P. (2012). Rapid discrimination of pork in Halal and non-Halal Chinese ham sausages by Fourier transform infrared (FTIR) spectroscopy and chemometrics. *Meat Science*, *92*(4), 506-510.

Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. *Applied artificial intelligence*, *17*(5-6), 375-381.

CONNECTING STATEMENT

In this part of the study, we explored how Fourier transformed infrared (FTIR) spectrometer with different machine learning (ML) algorithms could be used to develop models to authenticate gluten-related cross-contamination in raw-flour foods (uncooked food). One of the advantages of building models with raw foods is that the food chemical structures are still intact and there was no form of deformation from conversion processes yet. Because of this, it was easier to understand what is happening within the FTIR-spectra data obtained when visualized and it helps to make better intuition during ML model prototyping. Therefore, we studied how we can detect and quantify cross-contamination between a non-gluten flour (corn-flour) and the three major gluten-rich flours including wheat flour, barley flour, and rye flour.

CHAPTER 3. Detection and Quantification of Cross-Contact of a Non-Gluten and Gluten-Rich Flours by Fourier Transform Infrared (FTIR) Spectroscopy Coupled with Machine Learning Approaches

Abstract

Gluten-related disorders can result in serious health issues if not managed properly by maintaining a 100% gluten-free diet. In this study, FTIR coupled with supervised machine learning approaches (linear discriminant analysis and partial least squares regression) were evaluated for the detection and quantification of cross-contact between a non-gluten (corn flour (CF)) and gluten-rich (wheat flour (WF), barley flour and rye flour) flours, at contamination levels of 0.5% - 10% (w/w), with 0.5% increments. The F1-scores (0.963, 0.949, 0.963 and 1.0), R^2p (0.96, 0.94, and 0.98), and RMSEP (0.82, 0.99, and 0.53) obtained for the best results show that the methods used have the potential to authenticate the cross-contact of non-gluten and gluten-rich flours within the defined contamination levels.

KEYWORDS - Celiac Disease, FTIR, Gluten, Machine learning, Wheat flour

3.1 Introduction

Gluten is a family of proteins mainly present in wheat, barley, rye, and their cross-breeds. Gluten provides nutritional benefits and impacts important functionality in processed foods like bread with the viscoelastic property it imparts. However, it causes several health-related disorders that can lead to some severe health issues if not managed properly. These gluten related-disorders have no cure and the only effective treatment is to avoid any food that contains any of the gluten-rich grains and their cross-contacts (Mena & Sousa, 2015). Due to cross-contact with these grains during food processing or packaging, foods that are non-gluten may be contaminated with gluten. Therefore, there is a need for more fast and effective techniques or methods to ensure that gluten-free foods are safe for consumption for people with these disorders.

Some of the major health disorders related to gluten consumption in foods are celiac disease, wheat allergy, and gluten sensitivity, or gluten intolerance. Albanell et al. (2012) reported celiac disease as an immune system intervened enteropathy that is brought about by the response of consuming gluten-containing grains in food such as wheat, rye, barley, and oat in genetically susceptible people. These reactions from the response of the immune system affect the villi of the small digestive tract and if left untreated can lead to other critical health issues. Keeping up a diet without gluten is the best way to prevent symptoms of celiac disease (Albanell et al., 2012; Feighery, 1999). In non-celiac gluten sensitivity, there is no response from the body's autoimmune system – rather it is triggered by the body's intolerance to gluten and it has similar symptoms to celiac disease when gluten-

containing foods are consumed (Tanveer & Ahmed, 2019). The abnormal immune system response to at least one of the proteins found in wheat is what triggers wheat allergy and this might not necessarily be gluten (Tatham & Shewry, 2008). Several serious health symptoms associated with reaction to gluten include but are not limited to a bloated stomach, fatigue, diarrhea, stomach pain, breathing difficulties, hives, inability to focus, as well as pain in the bones and joints (Nordqvist, 2018).

To ensure foods (raw and processed) are gluten-free, several studies have explored the use of different chemical and/or non-destructive methods for detecting, visualizing and quantifying gluten with the overall goal of ensuring that gluten-free foods do not contain gluten above the regulated limit. The standard wet chemical analytical method approved by the Association of Official Analytical Chemists (AOAC International) for identifying and measuring gluten in food is by enzyme-linked immunosorbent assay (ELISA) and for food to be marked as gluten-free it must contain 20 ppm gluten or less (Lacorn et al., 2017). The steps involved in this method are cumbersome and time-consuming especially when an enormous number of samples are to be examined. However, non-destructive methods have the added advantage of being rapid, less laborious, efficient, and reliable.

Fourier transform infrared spectroscopy (FTIR) is a reliable, fast, and non-destructive method with next to zero sample preparation needed. It uses the principle of infrared light energy interaction with the molecular vibration of substances to obtain chemical and structural information from samples (Glassford et al., 2013). Such information from FTIR can be used to make informed decisions in food processing, inspection, and analysis, and this has been broadly utilized for food quality assessment and

food adulteration control (Rodriguez-Saona & Allendorf, 2011). Previous studies by Sujka et al. (2017) reported FTIR spectroscopy to have the potential for quality assessment of flours obtained from different producers in Poland ("Strzelce" company (Borowo, Poland), with a Quadrumat Senior mill (Brabender), Jelonki Ltd (Ostr´ow Mazowiecka, Poland), Młyny Wodne Ltd. (Korczew, Poland)). Supervised machine learning (ML) statistical models such as classical square and partial least square regression (PLSR) with the leave-one-out cross-validation techniques were explored. A range of coefficient of determination between (R^2) 0.94 to 0.97 was obtained for the best performing results indicating the accuracy and effectiveness of the methodology. In another study, a quick means of analyzing measurable wheat flour added to oat flour was developed by Wang et al. (2014) using FT-NIR spectrometry and chemometrics, FTIR-spectra data of samples of unadulterated oat and wheat flours were obtained with adulteration levels of 5% - 50% at 5% increment measured within the working range of 4000 cm^{-1} – 12000 cm^{-1} and PLSR models were developed on both raw and pre-processed (standard normal variate) data with Monte Carlo cross-validation (MCCV). PLSR models were highlighted to precisely estimate the levels of wheat flour in oat flour. FTIR spectroscopy has also been increasingly applied in other areas of food research, such as nuts using principal components discriminant approach (Ciemniewska-Żytkiewicz et al., 2015; Dogan et al., 2007), oils using Soft Independent Modeling Class Analogy and PLSR approach (Quiñones-Islas et al., 2013), cakes and flakes (Reder et al., 2014), and meat (Rohman et al., 2011; Xu et al., 2012).

For this study, the overall goal is to use FTIR coupled with machine learning approaches, for the evaluation of cross-contact of non-gluten and gluten-rich flours. Specifically, to obtain FTIR-data at different contamination levels, to evaluate multiple data pre-processing methods for effective analysis, to develop classification and regression models based on the pre-processing methods.

Accomplishing these objectives will lead to the development of detection and quantification models that can be deployed to systems (online-application, mobile application, and other software applications) for rapid and effective non-destructive food inspection and quality assessment in the grain and food processing industries. It will enhance the inspection and authentication of gluten contamination in grain-based foods.

3.2 Material and Methods

3.2.1 Sample preparation and FTIR-Spectroscopy

Gluten-rich flours, including wheat flour (WF), rye flour (RF), barley flour (BF) and a non-gluten flour (cornflour (CF)) were purchased from Bob's Red Mill Natural Foods (Milwaukie Oregon, USA). The gluten-rich flours (WF, RF, and BF) were used to contaminate the non-gluten flour (CF) in the range of 0% – 10% (w/w) with a 0.5% increment. Approximately 20 g of the mixture were prepared for each treatment. The gluten-rich flours were thoroughly mixed and homogenized (using mini-vortexer) with the CF at the different contamination levels. For each of the preparations, 220 samples (10 samples per contaminant level × 20 levels with the addition of 10 × 2 for pure samples of

each gluten-rich flours and CF) were obtained. Based on this, 640 samples (200 × 3 for contaminated samples and 10 × 4 pure samples) were prepared.

Attenuated total reflectance (ATR) spectra of the samples were recorded on a Fourier transform infrared (FTIR) spectrometer (Nicolet iS 50 Massachusetts, USA) in the frequency range of 4000 - 450 cm^{-1} with a resolution of 4 cm^{-1} and a total accumulation of 32 scans. The spectral data were then read into MATLAB R2018b (Mathworks Inc., Natick, MA, USA) for further analysis. The data were then divided into calibration and prediction sets at a ratio of 8:2 using Kennard-Stone (KS) algorithm (Galvao et al., 2005), that is 80% of the data were used for training and 20% for testing or validation of the models.

3.2.2 Selection of the region of interest

The spectral region between 1860-1480 cm^{-1} (C-N, C-C, C=O stretching vibrations) was selected as our region of interest for the classification models. Within this region are the two significant groups of the protein infrared spectrum, amide I and amide II bands (Jabs, 2005). Amide I is between the frequency of about 1690 cm^{-1} and 1600 cm^{-1} and it is the most intense absorption band among the proteins present. The amide II is more complex than amide I and it is found in the region of wavelength or frequency between 1580 cm^{-1} to 1480 cm^{-1} (Makarenko et al., 2002).

3.2.3 Spectra pre-processing

Attenuated total reflectance-FTIR spectra data in their raw form have highly correlated variables, which comprise of both informative and uninformative regions. Noise and correlated wavenumbers could decrease the capability of several multivariate techniques associated with exploratory and classification purposes (Lee et al., 2018b). Therefore, the aim of the spectra preprocessing is to eliminate or decrease undesired signals from the spectra before modeling. In addition to non-pre-processed data, the spectra data of the samples were pre-treated by the following methods: Savitzky Golay (SG) derivative (1st derivative, 2 order polynomials, 7 points window), mean-centering (MC), double centering (DC), smoothing (1^{st} derivate and 2^{nd} derivative), standard normal variate (SNV), multiplicative scatter correction (MSC), scaling, auto-scaling and robust auto-scaling methods. These different pre-processing methods were tested, and the methods which produced the best results were determined dependent on the model prediction coefficient of determination (R^2p), and the lowest prediction root means square error (RMSEP).

3.2.4 Model development and evaluation

All models were developed using MATLAB R2019 (Mathworks Inc., Natick, MA, USA).

3.2.4.1 Classification Model

Classification models for the contaminated samples' data were developed using the different spectra pre-processing methods and classification techniques including k-nearest neighbors, decision trees, and linear discriminant analysis (LDA) method on the selected

region of interest (1860 cm^{-1} – 1480 cm^{-1}). Based on performance predicated on R^2p and RMSEP, LDA with SG preprocessing was selected as the best classifier and was used for further development. The LDA function includes a linear combination of features and classification of samples based on the function's value obtained (Duda et al., 2012), and this method comes with the added advantage of being simple to implement (Theodoridis & Koutroumbas, 2003). Bootstrap aggregation (bagging) was applied to the LDA in order to improve the performance of the learning algorithm. The bagging process generally involves training various *M*-base models by a cluster of different subsets of data of size n selected from a dataset *T* of size *N* where $n < N$. The sample size of *n* is made by drawing arbitrary samples with replacement from the original training set *T* (Oza, 2005). This has the advantage of reducing variance, decreasing overfitting and handling higher dimensionality data (Kotsiantis & Pintelas, 2004). In this study, the datasets were divided into a training set (80%), and a testing set (20%), then the training set was divided into four subsets and each subset was used in training an LDA model using four-fold cross-validation. Classifications were made on the test samples using a voting method on the four bagged LDA classifiers obtained (Rady & Adedeji, 2018; Varmuza & Filzmoser, 2016). A confusion matrix was then used to evaluate the model performance. The confusion matrix summarizes how successful the classification model is at predicting samples belonging to the different classes. The performance metrics calculated by the confusion matrix are the true negative (TN), true positive (TP), false negative (FN), false positive (FP), precision (P), recall (R), true negative rate (TNR), false-negative rate (FNR), true positive rate (TPR), false-positive rate (FPR), misclassification error (err), and F1_score (a measure of

test's accuracy with value at 1 signifying best performance and 0 indicates model's worst performance). The higher the performance metrics (TPR, TNR, P, and F1-score) value (near 1), the more the accuracy of a model. The precision is the proportion of true positive prediction to the general number of the positive predictions.

$$precision\ (P) = \frac{TP}{TP+FP} \tag{3.1}$$

The extent of positive cases that were accurately measured (sensitivity), calculated exactly as recall (R) TPR:

$$\frac{TP}{(TP+FN)} \tag{3.2}$$

The extent of negatives cases that were inaccurately delegated positive, FPR:

$$\frac{FP}{(FP+TN)} \tag{3.3}$$

The extent of negatives cases that were classified accurately (Specificity), TNR:

$$\frac{TN}{(TN+FP)} \tag{3.4}$$

The extent of positive cases that were inaccurately classified as negative, FNR:

$$\frac{FN}{(FN+TP)} \tag{3.5}$$

Misclassification error: the extent of the samples which were inaccurately classified, Err:

$$1 - accuracy = 1 - \frac{(\mathrm{FP+FN})}{(TP+TN+FP+FN)} \tag{3.6}$$

A measure of the model's accuracy, F1-Score:

$$\mathrm{F1_score} = 2\frac{PR}{P+R} \tag{3.7}$$

3.2.4.2 Contamination quantification models

Partial least squares regression (PLSR) was applied to the full spectra region to obtain the quantitative prediction models for the levels of each contaminant (wheat, barley, and rye). The PLSR was conducted based on the SIMPLS algorithm developed by (Jong, 1993). The input data were first pre-processed and were divided into a training set (80%) and prediction or test set (20%). Cross-validation (four-fold) were implemented on the training set. Root mean square error of the cross-validation (RMSECV) was used to select the best optimal training model based on the different pre-processing method before subjecting the models to the prediction or test set. Then prediction's coefficient of determination (R^2p) and root mean square error of prediction (RMSEP) was used to evaluate the overall performance of the models. In this study, the development of the models involved the use of the spectra in the data matrix (X) as explanatory variables to estimate or predict the different or contamination levels in CF given in the dependent variables column vector (Y). The number of latent variables was chosen in this study as 20 based on the lowest RMSECV (Rady & Adedeji, 2020).

3.3 Results and Discussion

3.3.1 Spectra characteristics of the flour samples

Visual inspection of the mean ATR-FTIR spectra obtained for the pure samples (Figure 3.1) shows a similar spectrum pattern for each of the individual samples which signifies similarity in chemical compositions. However, on closer inspection, differences

between the non-gluten CF and gluten-rich flours (BF, WF, and RF) can be seen at the absorbance peak of 1707 cm^{-1} within the region between 1860 cm^{-1} – 1480 cm^{-1}, including most of the amide I (1690 cm^{-1} – 1600 cm^{-1}), and amide II (1580 cm^{-1} – 1480 cm^{-1}) characteristic bands that are susceptible to the protein's secondary structure content. The peak distinctively differentiates both types of flours and can be used as a basis for discrimination of the flours (Czaja et al., 2016b). Figure 3.2(a), (b) and (c) indicate that the proportions of the CF contamination from 0.5% to 10% each of the spectra has similarities in peak, trend, and trough with different intensities. These differences and peak variations can be used for pattern recognition in classification model developments.

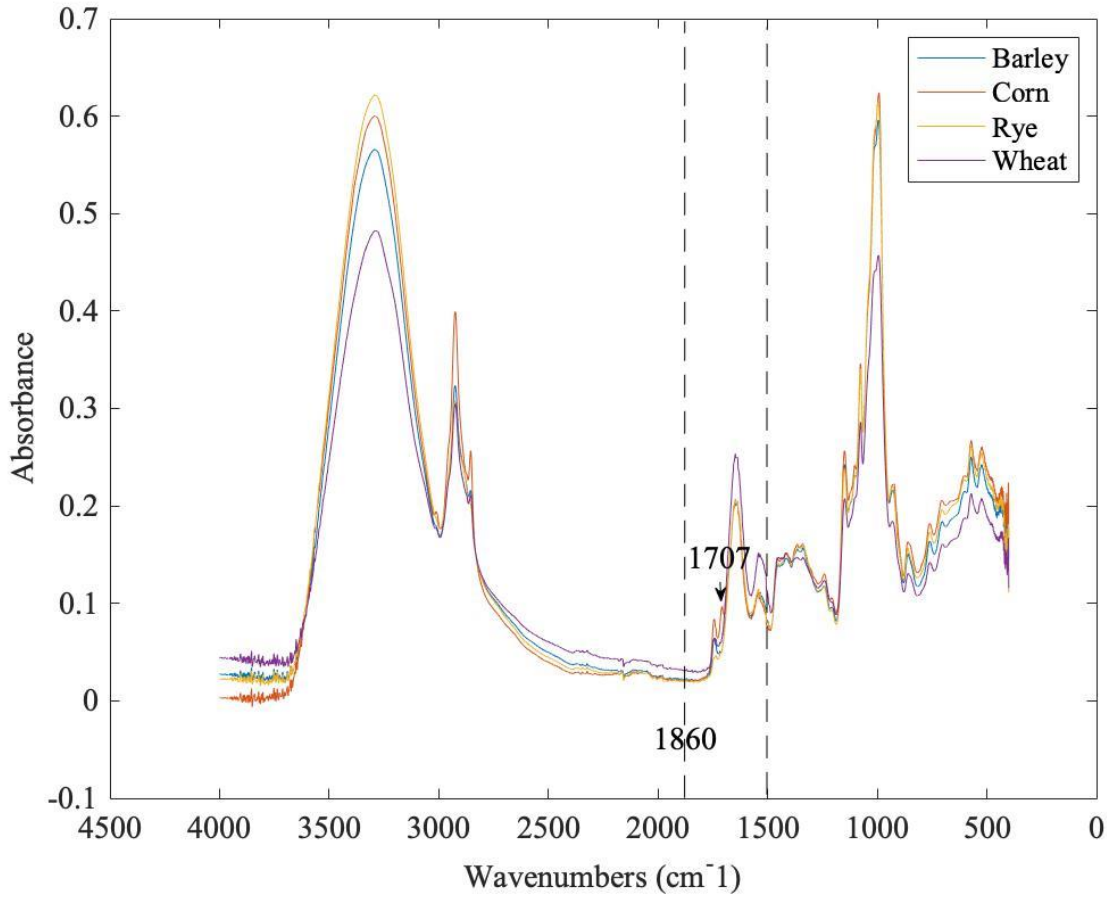
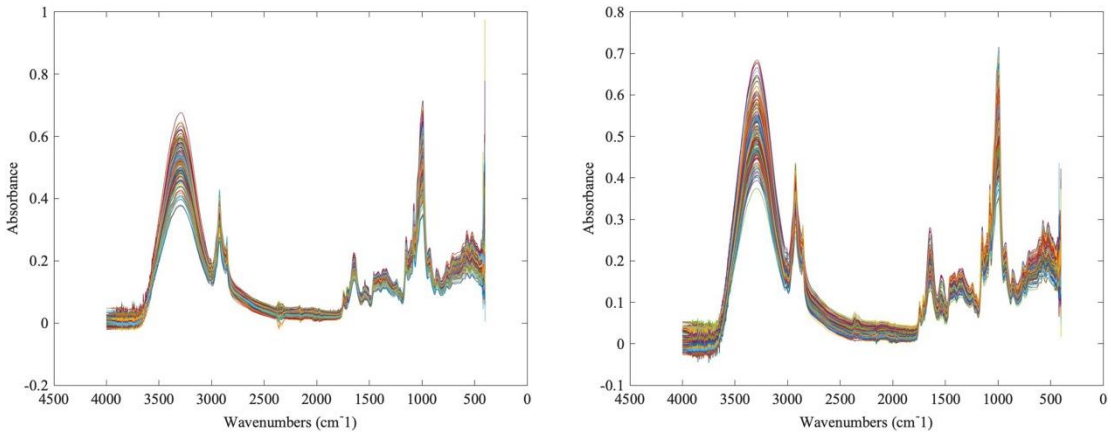


Figure 3.1: The mean spectra of the different pure flour samples.

(a) (b)



(c)

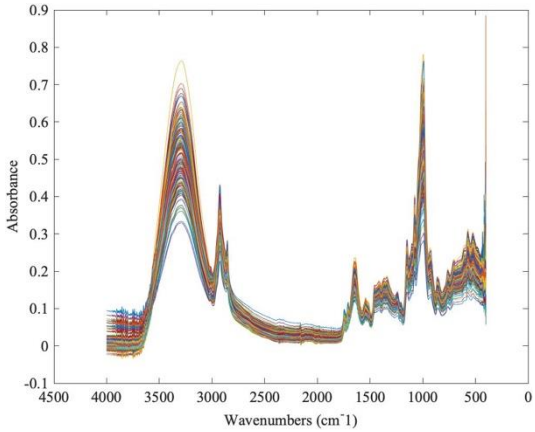


Figure 3.2: (a) Raw spectra of corn flour (CF) contaminated with barley flour (BF), (b) Raw spectra of corn flour (CF) contaminated with wheat flour (WF), and (c) Raw spectra of corn flour (CF) contaminated with rye flour (RF). The different contamination levels of 0.5% - 10% at 0.5% increment is represented by the different colored spectrum.

3.3.2 Spectra preprocessing and spectra models

Several preprocessing methods were examined for comparison purposes. Figure 3.3 shows how some of the corresponding spectra data were transformed by the pre-processing algorithms. The details of pre-processing methods used for the spectra treatment and other statistical parameters of the LDA and PLSR models are presented in Tables 3.4, 3.5, 3.6, and 3.7 below. Savitzky-Golay (SG) was selected for the classification models as it has the best performing test confusion matrix evaluation parameters (F1-score ranging from 0.949 to 1.0 in Table 3.4). Futhermore, as indicated in the results obtained for the regression models evaluation, the performances of most of the pre-processing methods were good and have a less significant difference from each other, but for this study, mean centering, smoothing (second derivative) and robust auto-scaling was selected as the best based on their R^2p (0.96, 0.94, 0.98 respectively) and RMSEP (0.82, 0.99, 0.53 respectively) for all of the developed models (Zhao et al., 2019).

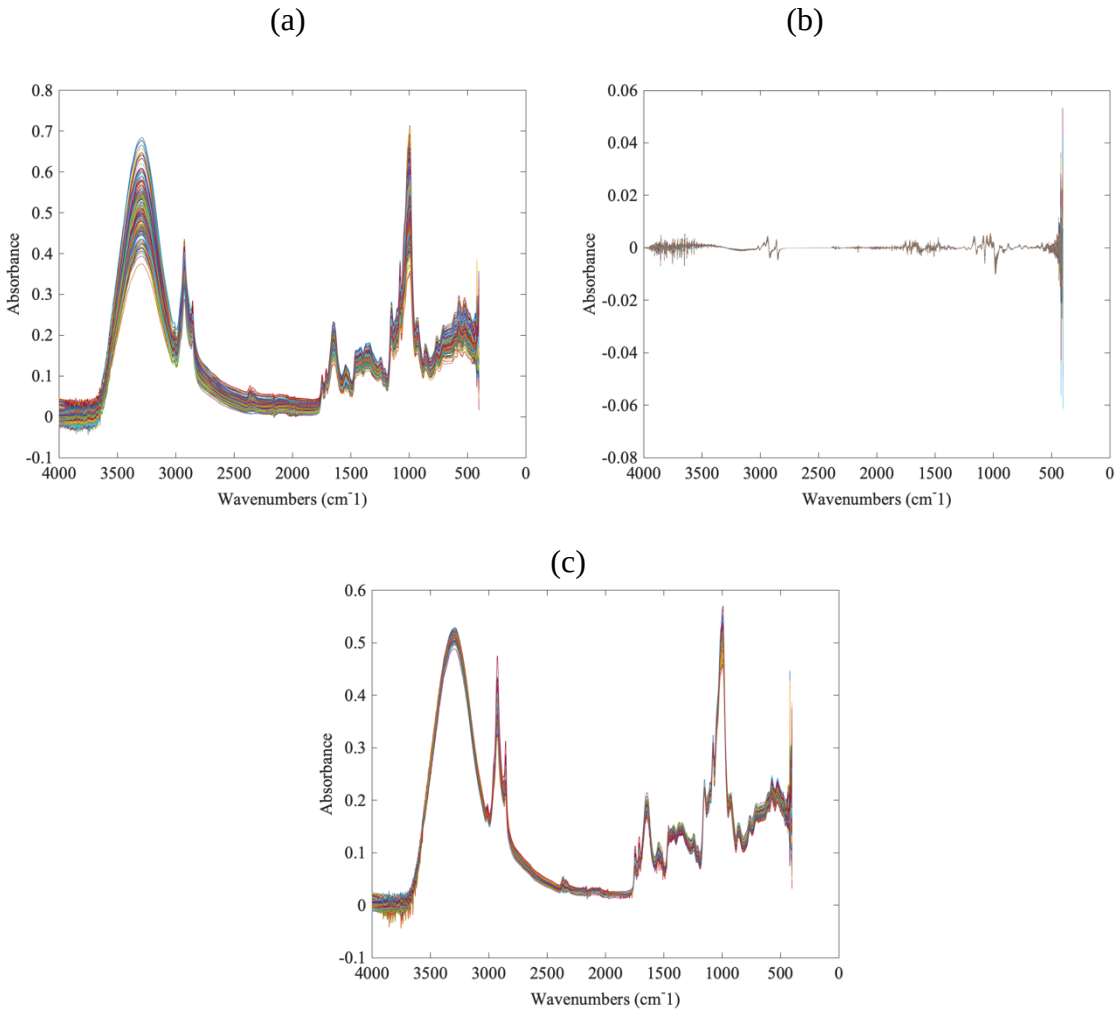


Figure 3.3: (a) Spectra pre-processed by smoothing (1st derivatives) (b) Spectra pre-processed by standard normal variate (SNV) and (c) Spectra pre-processed by multiplicative scatter correction (MSC).

3.3.3 Classification modeling results

The LDA models' performance was determined based on the result of the confusion matrix obtained for both the training and test data. The aim was to classify the spectra to two different groups made up of four different model classes: gluten-rich (BF (class 1), WF (class 2), RF (class 3) and non-gluten (CF (class 4)) flours. Tables 3.1 to Table 3.4 summarized how successful the classification models are at predicting the spectrum belonging to the various classes. The results of the confusion matrices are presented in Table 3.1 and Table 3.3 for each of the classes in training and test models. For example, in Table 3.3, out of 40 samples that are are not contaminated in class 4 (pure CF samples) for the test samples, the model correctly classified all the samples that truly belong to the class indicating a 100% classification capacity. Also, class 4 in Table 3.4 shows that the model has a TPR of 1 with an F1-score value of 1 which indicates a measure of 100% accuracy. However, in real-world situation 100% accuracy might be hard to achieved due to certain limitations such as real-world data complexity, missing features, unbalance data, etc. The higher the F1 score the better the model and value near 1 indicates a reliable and good model while value closer to zero may indicate a poor model. The models have a TPR ranging from 0.951 to 1 and an F1-score ranging from 0.949 to 1. This shows that ATR-FTIR with LDA has the potential to detect the cross-contact of CF with BF, WF or RF within the contamination levels.

Table 3.1: Training model confusion matrix for the LDA + 4-fold Cross-validation + Bagging

	Actual Class			
	Class 1	Class 2	Class 3	Class 4
Classified as Class 1	40	0	0	0
Classified as Class 2	0	39	1	0
Classified as Class 3	0	1	39	0
Classified as Class 4	0	0	0	40
Classified as Unassigned	0	0	0	0

Class 1: CF contaminated with BF, Class 2: CF contaminated with WF, Class 3: CF contaminated with RF and Class 4: Pure CF (BF: Barley Flour, WF: Wheat Flour, RF: Rye Flour, CF: Corn Flour), LDA: Linear discriminant analysis.

Table 3.2: LDA training model confusion matrix parameters for classification of contamination between gluten-rich (BF (class 1), WF (class 2), RF (class 3)) and gluten-free (CF (class 4)) flours.

Class	TPR	FPR	TNR	FNR	Err	P	F1_score
Class 1	1.000	0.000	1.000	0.000	0.000	1.000	1.000
Class 2	0.975	0.008	0.992	0.025	0.013	0.975	0.975
Class 3	0.975	0.008	0.992	0.025	0.013	0.975	0.975
Class 4	1.000	0.000	1.000	0.000	0.000	1.000	1.000

BF: Barley flour, CF: Corn flour, WF: wheat flour, RF: Rye flour, TPR: True positive rate, FPR: False positive rate, TNR: True negative rate, FNR: False negative rate, Err: Error, P: Precision, F1: scores for a measure of accuracy, LDA: Linear discriminant analysis.

Table 3.3: Test model confusion matrix for LDA + 4-fold CV+ Bagging

	Actual Class			
	Class 1	Class 2	Class 3	Class 4
Classified as Class 1	39	2	0	0
Classified as Class 2	0	37	1	0
Classified as Class 3	1	1	39	0
Classified as Class 4	0	0	0	40
Classified as Unassigned	0	0	0	0

Class 1: CF contaminated with BF, Class2: CF contaminated with WF, Class 3: CF contaminated with RF and Class 4: Pure CF (BF: Barley Flour, WF: Wheat Flour, RF: Rye Flour, CF: Corn Flour), LDA: Linear discriminant analysis.

Table 3.4: Results for the evaluation of the each of the LDA test model classes (gluten-rich: BF (class 1), WF (class2), RF (class 3)) and gluten-free (CF (class 4)) flours).

Class	TPR	FPR	TNR	FNR	Err	P	F1_score
Class 1	0.951	0.008	0.992	0.049	0.019	0.975	0.963
Class 2	0.974	0.025	0.975	0.026	0.025	0.925	0.949
Class 3	0.951	0.008	0.992	0.049	0.019	0.975	0.963
Class 4	1.000	0.000	1.000	0.000	0.000	1.000	1.000

BF: Barley flour, CF: Corn flour, WF: wheat flour, RF: Rye flour, TPR: True positive rate, FPR: False positive rate, TNR: True negative rate, FNR: False negative rate, Err: Error, P: Precision, F1: scores for a measure of accuracy, LDA: Linear discriminant analysis.

3.3.4 PLSR prediction model

The coefficient of determination (R^2) and RMSE was used to evaluate the performance of each of the PLSR models based on different pre-processing methods. The results obtained are presented in Table 3.5, Table 3.6, and Table 3.7 for each of the contaminants BF, WF, and RF respectively. For CF contaminated with WF, PLSR with MC was chosen as the best model with R^2cv, RMSECV, and R^2p, RMSEP to be 0.98, 0.37 and 0.96, 0.82 respectively. For CF contaminated with BF, PLSR with smoothing (second derivative) was chosen as the best model with R^2cv, RMSECV, and R^2p, RMSEP, to be 0.97, 0.53, and 0.94, 0.99 respectively while for CF contaminated with RF, PLSR with robust auto-scaling was chosen as the best model with R^2cv, RMSECV, and R^2p, RMSEP to be 0.99, 0.37 and 0.98, 0.53 respectively. Su and Sun (2017) reported that generally, it is best to obtain RMSEs near 0 and R^2 approaching 1, where R^2 greater than 0.90 indicates exceptional performance and lower than 0.82 might indicate low performance of the model. Also, the similarity between the different model performances could indicate the consistency and effectiveness of PLSR. Therefore, it can be concluded that the prediction models are good and adequate to correctly predict the percentage (%) of contamination of the gluten-rich flours (BF, WF, and RF) in the non-gluten flour (CF).

Furthermore, all the results obtained demonstrate that the application of the method proposed in this study to be feasible on the real-time application and can be deployed to be used alongside compact and portable FTIR systems such as the Agilent 4100 ExoScan FTIR with diamond ATR head in other to make an informed decision. However, the study

was developed in non-real-time or non-on-line computing environment. For such models to be deployed to an on-line or real-time environment, it requires the development of a software system that will be able to integrate the models while meeting the needs to produce high-quality processes in time-sensitive situations. Therefore, future studies will explore the use of handheld FTIR devices and carry-out more research on the best methods of deploying the ML models into a software system for authentication of cross-contact of gluten-rich and non-gluten flours.

Table 3.5: PLSR model results for corn flour contaminated with wheat flour samples using different pre-processing methods.

Pre-processing method	**No. of LV**	**Cross-validation**		**Prediction**	
		R^2_{cv}	**RMSECV**	**R^2_p**	**RMSEP**
Non	20	0.98	0.37	0.96	0.82
Mean Centering	20	0.98	0.37	0.96	0.82
Scaling	20	0.98	0.41	0.94	1.01
Auto Scaling	20	0.98	0.41	0.94	1.01
Robust Auto-Scaling	20	0.98	0.41	0.94	1.01
Double Centering	20	0.98	0.37	0.96	0.82
SNV	20	0.96	0.60	0.86	1.58
Smoothing using Savitzky-Golay	20	0.98	0.42	0.95	0.94
Smoothing 1st Derivative	20	0.98	0.44	0.95	0.98
Smoothing 2nd Derivative	20	0.98	0.46	0.92	1.14
MSC	20	0.94	0.78	0.51	4.49

LV: Latent variables, R2cv: Cross-validation's coefficient of determination, RMSECV: Root mean square error of cross-validation, R2p: Prediction's coefficient of determination, RMSEP: Root mean square error of prediction, MSC: Multiplicative scatter correction, SNV: Standard normal variate.

Table 3.6: Results of PLSR models for corn flour contaminated with barley flour samples using different pre-processing methods.

Pre-processing method	**No. of LV**	**Cross-validation**		**Prediction**	
		R^2_{cv}	**RMSECV**	**R^2_p**	**RMSEP**
Non	20	0.96	0.58	0.92	1.27
Mean Centering	20	0.96	0.58	0.92	1.27
Scaling	20	0.96	0.64	0.87	1.53
Auto Scaling	20	0.96	0.64	0.87	1.53
Robust Auto-Scaling	20	0.96	0.65	0.87	1.56
Double Centering	20	0.96	0.58	0.92	1.27
SNV	20	0.90	1.00	0.85	1.87
Smoothing using Savitzky-Golay	20	0.96	0.62	0.91	1.29
Smoothing 1st Derivative	20	0.98	0.46	0.94	1.01
Smoothing 2nd Derivative	20	0.97	0.53	0.94	0.99
MSC	20	0.87	1.09	0.69	3.10

LV: Latent variables, R2cv: Cross-validation's coefficient of determination, RMSECV: Root mean square error of cross-validation, R2p: Prediction's coefficient of determination, RMSEP: Root mean square error of prediction, RPDP: Ratio between performance to deviation of prediction, MSC: Multiplicative scatter correction, SNV: Standard normal variate.

Table 3.7: PLSR model results after different pre-processing methods for corn flour contaminated with rye flour.

Pre-processing method	No. of LV	Cross-validation		Prediction	
		R^2_{cv}	RMSECV	R^2_p	RMSEP
Non	20	0.99	0.31	0.98	0.62
Mean Centering	20	0.99	0.31	0.98	0.62
Scaling	20	0.99	0.37	0.98	0.54
Auto Scaling	20	0.99	0.37	0.98	0.54
Robust Auto-Scaling	20	0.99	0.37	0.98	0.53
Double Centering	20	0.99	0.31	0.98	0.62
SNV	20	0.98	0.40	0.97	0.73
Smoothing using Savitzky-Golay	20	0.99	0.32	0.98	0.63
Smoothing 1st Derivative	20	0.99	0.24	0.97	0.70
Smoothing 2nd Derivative	20	1.00	0.18	0.97	0.67
MSC	20	0.98	0.47	0.89	1.32

LV: Latent variables, R2cv: Cross-validation's coefficient of determination, RMSECV: Root mean square error of cross-validation, R2p: Prediction's coefficient of determination RMSEP: Root mean square error of prediction, MSC: Multiplicative scatter correction, SNV: Standard normal variate.

Conclusion

The present study indicated the feasibility of using Fourier transformed infrared (FTIR) spectroscopy coupled with machine learning methods to detect and quantify the cross-contact of gluten-rich and gluten-free flours. Linear discriminant analysis (LDA) showed strong potential for detecting the defined contamination levels (0% - 10% at 0.5% increment) of WF, BF, and RF in CF. The best model for the predictive analyses emerged by PLSR with MC, smoothing (second derivatives), and robust auto-scaling methods respectively for CF contaminated with WF, BF and RF. The proposed methods are simple, rapid and have high efficiency. The results obtained show that they could have great potentials in the food industry to compliment or add to the analytical methods used for detection and quantification of gluten cross-contamination in grain-based foods thus reducing the test time drastically.

References

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. doi:https://doi.org/10.1016/j.jcs.2012.06.011

Ciemniewska-Żytkiewicz, H., Bryś, J., Sujka, K., & Koczoń, P. (2015). Assessment of the hazelnuts roasting process by pressure differential scanning calorimetry and MID-FT-IR spectroscopy. *Food Analytical Methods, 8*(10), 2465-2473.

Czaja, T., Mazurek, S., & Szostak, R. (2016). Quantitative analysis of solid samples using modified specular reflectance accessory. *Talanta, 161,* 655-659.

Dogan, A., Siyakus, G., & Severcan, F. (2007). FTIR spectroscopic characterization of irradiated hazelnut (Corylus avellana L.). *Food Chemistry, 100*(3), 1106-1114.

Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern classification*: John Wiley & Sons.

Feighery, C. (1999). Coeliac disease. *Bmj, 319*(7204), 236-239.

Galvao, R. K. H., Araujo, M. C. U., Jose, G. E., Pontes, M. J. C., Silva, E. C., & Saldanha, T. C. B. (2005). A method for calibration and validation subset partitioning. *Talanta, 67*(4), 736-740.

Glassford, S. E., Byrne, B., & Kazarian, S. G. (2013). Recent applications of ATR FTIR spectroscopy and imaging to proteins. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, 1834*(12), 2849-2858. doi:https://doi.org/10.1016/j.bbapap.2013.07.015

Goutte, C., & Gaussier, E. (2005). *A probabilistic interpretation of precision, recall and F-score, with implication for evaluation.*

Jabs, A. (2005). Determination of secondary structure in proteins by fourier transform infrared spectroscopy (FTIR). *Jena Library of Biologica Macromolecules.*

Jong, S. D. (1993). PLS fits closer than PCR. *Journal of chemometrics, 7*(6), 551-557.

Kotsiantis, S., & Pintelas, P. (2004). Combining bagging and boosting. *International Journal of Computational Intelligence, 1*(4), 324-333.

Lacorn, M., Siebeneicher, S., & Weiss, T. (2017). Measurement of Gluten in Food Products: Proficiency- -Testing Rounds as a Measure of Precision and Applicability. *Celiac Disease and Non-Celiac Gluten Sensitivity*, 27.

Lee, L. C., Liong, C.-Y., & Jemain, A. A. (2018). *Iterative random vs. Kennard-Stone sampling for IR spectrum-based classification task using PLS2-DA.*

Makarenko, S. P., Trufanov, V. A., & Putilina, T. E. (2002). Infrared Spectroscopic Study of the Secondary Structure of Wheat, Rye, and Barley Prolamins. *Russian Journal of Plant Physiology, 49*(3), 326-331. doi:10.1023/a:1015584700841

Mena, M., & Sousa, C. (2015). Analytical tools for gluten detection: Policies and regulation. *OmniaScience Monographs*.

Nordqvist, C. B., N. . (2018). What is a wheat allergy? . *Retrieved from Medical News Today Website:* https://www.medicalnewstoday.com/articles/174405.php.

Oza, N. C. (2005). *Online bagging and boosting.* Paper presented at the 2005 IEEE international conference on systems, man and cybernetics.

Quiñones-Islas, N., Meza-Márquez, O. G., Osorio-Revilla, G., & Gallardo-Velazquez, T. (2013). Detection of adulterants in avocado oil by Mid-FTIR spectroscopy and multivariate analysis. *Food Research International, 51*(1), 148-154.

R-Biopharm, R. (2009). Instructions, Gliadin R7001. In.

Rady, A., & Adedeji, A. (2018). Assessing different processed meats for adulterants using visible-near-infrared spectroscopy. *Meat science, 136*, 59-67.

Rady, A., & Adedeji, A. A. (2020). Application of Hyperspectral Imaging and Machine Learning Methods to Detect and Quantify Adulterants in Minced Meats. *Food Analytical Methods*, 1-12.

Reder, M., Koczoń, P., Wirkowska, M., Sujka, K., & Ciemniewska-Żytkiewicz, H. (2014). The application of FT-MIR spectroscopy for the evaluation of energy value, fat content, and fatty acid composition in selected organic oat products. *Food Analytical Methods, 7*(3), 547-554.

Rodriguez-Saona, L., & Allendorf, M. (2011). Use of FTIR for rapid authentication and detection of adulteration of food. *Annual review of food science and technology, 2*, 467-483.

Rohman, A., Erwanto, Y., & Man, Y. B. C. (2011). Analysis of pork adulteration in beef meatball using Fourier transform infrared (FTIR) spectroscopy. *Meat Science, 88*(1), 91-95.

Rostami, K., Bold, J., Parr, A., & Johnson, M. W. (2017). Gluten-free diet indications, safety, quality, labels, and challenges. In: Multidisciplinary Digital Publishing Institute.

Su, W.-H., & Sun, D.-W. (2017). Evaluation of spectral imaging for inspection of adulterants in terms of common wheat flour, cassava flour and corn flour in organic Avatar wheat (Triticum spp.) flour. *Journal of Food Engineering, 200*, 59-69.

Sujka, K., Koczoń, P., Ceglińska, A., Reder, M., & Ciemniewska-Żytkiewicz, H. (2017). The application of FT-IR spectroscopy for quality control of flours obtained from polish producers. *Journal of Analytical Methods in Chemistry, 2017*.

Tanveer, M., & Ahmed, A. (2019). Non-Celiac Gluten Sensitivity: A Systematic Review. *Journal of the College of Physicians and Surgeons Pakistan, 29*(1), 51-57.

Tatham, A., & Shewry, P. (2008). Allergens to wheat and related cereals. *Clinical & Experimental Allergy, 38*(11), 1712-1726.

Varmuza, K., & Filzmoser, P. (2016). *Introduction to multivariate statistical analysis in chemometrics*: CRC press.

Wang, N., Zhang, X., Yu, Z., Li, G., & Zhou, B. (2014). Quantitative analysis of adulterations in oat flour by FT-NIR spectroscopy, incomplete unbalanced randomized block design, and partial least squares. *Journal of Analytical Methods in Chemistry, 2014*.

Xu, L., Cai, C.-B., Cui, H.-F., Ye, Z.-H., & Yu, X.-P. (2012). Rapid discrimination of pork in Halal and non-Halal Chinese ham sausages by Fourier transform infrared (FTIR) spectroscopy and chemometrics. *Meat Science, 92*(4), 506-510.

Zhao, X., Wang, W., Ni, X., Chu, X., Li, Y.-F., & Lu, C. (2019). Utilising near-infrared hyperspectral imaging to detect low-level peanut powder contamination of whole wheat flour. *Biosystems engineering, 184*, 55-68. doi:10.1016/j.biosystemseng.2019.06.010

CONNECTING STATEMENT

After the completion of the first phase of the research work, the samples of the non-gluten flour (corn flour) contaminated with wheat flour were used bake bread. The samples were baked into bread with different contamination levels. The general purpose of this phase of this of the research is to visualize the effect of the baking process and also, to use improved machine learning techniques coupled with FTIR that can be used to authenticate cross-contamination from wheat flour in a non-gluten bread. The results are presented in the next chapter below.

CHAPTER 4. FOURIER TRANSFORM INFRARED (FTIR) SPECTROSCOPY WITH MACHINE LEARNING APPROACHES FOR DETECTION AND QUANTIFICATION OF WHEAT FLOUR CONTAMINATION IN A NON-GLUTEN BREAD

Abstract

This study evaluates the use of the Fourier transform infrared (FTIR) method coupled with machine learning (ML) approaches to detect and quantify wheat flour contamination in a non-gluten bread. Samples of corn-flour (CF) were contaminated with wheat flour (WF) in the range of 0% - 10% with a 0.5% increment. The flour samples were baked into loaves of bread using basic bread ingredients and then ground into finer particles in other to achieve a homogenous mixture. Spectra data of the ground samples were obtained using FTIR and then standardized before the modeling process. For the classification model, majority voting-based ensemble learning (stack of k-nearest neighbor (KNN), random forest, and support vector classifier) was developed to detect WF contamination in the samples. To quantify the percentage (%) level of wheat contamination in these samples, KNN regressor was selected as the best predictive model. From the confusion matrix parameters for the test classification models, F1_score, true-positive rate (TPR), false-negative rate (FNR) were obtained to be 1.0, 1.0, and 0.0, respectively. And for the quantification models, coefficient of determination (R^2_T) and root mean square error (RMSET) for the training set were obtained to be 0.9820 and 0.4062 respectively, and for the test or prediction (R^2_P and RMSEP) set to be 0.9871 and 0.3374 respectively. The F1_score, TPR, FNR, R^2_T, and $RMSE_T$, R^2_P and RMSEP obtained show that application of

FTIR with the supervised machine learning approaches has an effective capacity to efficiently detect and quantify the defined WF contamination in the corn-bread.

Keywords – Celiac Disease, Corn-bread, Ensemble learning, Gluten, Machine learning, Wheat flour, FTIR

4.1 Introduction

Gluten proteins in wheat (gliadin and glutenin) may induce different types of immunological or physiological issues, for example, celiac disease, wheat allergy, gluten intolerance or sensitivity, and others with a wide range of side effects or symptoms in susceptible people. For the gluten-related disorders, a strict diet containing no gluten (gluten-free diet) is essential to properly manage them. A gluten-free diet is recommended by the United States Food and Drug Administration (FDA) to be any food containing ≤ 20 ppm of gluten (Allred & Ritter, 2010). However, many factors can lead to gluten-free food to be contaminated and exceed the recommended level of 20 ppm. For example, during food processing, food naturally free-from gluten or non-gluten food may contain gluten due to cross-contact with gluten-rich grains including wheat, barley, rye, and their crossbred varieties. In the process of bread baking different flours from these grains, most especially wheat, are used because of their gluten contents that give the bread that stretchy, almost bouncy texture and a little bit of chew. And to make gluten-free bread involves using flours from non-gluten grains such as millet, corn, rice, chia, potato, almond, buckwheat, quinoa, and others. Most times the same equipment or kitchen is used during this process and thus, gluten-free bread may end up being contaminated with gluten-rich flour such as wheat if proper cleaning is not done or care is not taken due to human factors. Therefore, there is a need to ensure that bread labeled gluten-free is safe for consumption for people having gluten-related disorders.

Enzyme-linked immunosorbent assay (ELISA) is the endorsed method of testing for gluten-free bread. However, it is a cumbersome method and requires a highly skilled

chemist to be executed. Considering the demand to ensure the food safety of baked foods, it is progressively important to develop a fast and similarly dependable technique that can be used to accomplish food inspection and quality control. Several applications of different rapid methods, for example, Fourier-transform near-infrared (FT-NIR) in comparison to NIR spectroscopy instrumentation was proven to have effective and reliable performance in predicting grain and different wheat flours quality attributes (Armstrong et al., 2006). In another study, FTIR spectrometer was reported as a significant and viable alternative method for milk quality analysis to that of a commercial IR milk analyzer (filter-based, multi-spec MK1) (Van De Voort et al., 1992b). Also, attenuated total reflectance FTIR coupled with ML supervised learning methods including partial least-squares regression (PLSR) and principal component analysis (PCA) has indicated potential in determining the sugar content in honey for quality assessment. First-derivative spectra pre-processed with multiplicative scatter correction and straight-line subtraction yielded the best calibration results with R^2 ranging from 0.757 to 0.923 against the result for the test set validation (R^2 = 0.6046 to 0.8903) (Anjos et al., 2015). Furthermore, quantification of free fatty acid contents in palm olein as a means to take out the utilization and removal of hazardous solvents required by the chemical method has been established using FTIR with PLS models (R^2 = 0.997) (Man & Setiowaty, 1999). In addition, FTIR offers many possibilities to be used as a potential means of identifying adulterated foods, Lohumi et al. (2014), reported that FTIR and FT-NIR spectroscopy with PLSR approach can rapidly detect and quantify onion powder adulterated with cornstarch. Adulteration of cod-liver oil (Rohman & Che Man, 2009), pork in beef meatball (Rohman et al., 2011), lard content in cake

formulation (Syahariza et al., 2005), lotus root powder with potato starch (Liu et al., 2013), sugar cane ((Irudayaraj et al., 2003) and inverted beet sugar (Sivakesava & Irudayaraj, 2001) in honey have been authenticated using FTIR in combination with different learning algorithms.

In this study, FTIR spectroscopy combined with supervised machine learning approaches was used to detect and quantify wheat flour contamination in non-gluten bread. More sophisticated machine learning algorithms were explored and compared for effective analysis.

4.2 Materials and methods

4.2.1 Basic ingredients for bread

The corn-flour (CF) and wheat flour used were purchased from Bob's Red Mill Natural Foods (US food company). During the mixing process, the following basic bread ingredient formulation adopted from (Mondal & Datta, 2008) was used. The corn-bread formulation includes corn-flour (100%) and other ingredients based on the weight of the flour with the following percentages: water (70%), dried yeast (2%), salt (2%), sugar (2%), vegetable fat (3%) and 0 – 10% of wheat flour at 0.5% increment for the contamination levels.

4.2.2 Laboratory baking

The corn-flour was mixed with the aforementioned formulated ingredients at the different wheat flour contamination levels. The bread dough was mixed using a kitchen mixer (KitchenAid, Model KV25G0X, Benton Harbor, MI) with a variable speed ranging from 1 (60 rpm) to 10 (280 rpm). All the ingredients were mixed for 1 min at speed 1 (60 rpm) and a total 6 min at speed 2 (95 rpm). The process also involved scrapping the dough every 2 min while mixing. The dough was poured into aluminum baking pans and proofed for 35 min at 40°C and subsequently baked for 1 hr at 190.6°C in an oven (Hobart, HR202, OH, U.S.A). The baked loaves of bread were kept for 1 hr at room temperature (24°C) to cool and then blended to finer particles (for homogenous mixture) using a commercial laboratory blender (Waring Commercial 7010BU Lab Blender) for 40 seconds before measurements. In the end, 21 different bread samples (20 g each) were obtained.

4.2.3 Spectra Data and pre-processing

Spectra measurements were carried out on the ground bread samples using attenuated total reflectance Fourier transform infrared (ATR-FTIR) spectrometer (Nicolet iS 50 Massachusetts, USA) in the frequency range of 4000 - 450 cm^{-1} with a resolution of 4 cm^{-1} and a total accumulation of 32 scans. The data obtained were then pre-processed using standard scaling (SC) method or standardization. The SC is obtained by subtracting the mean of the feature vectors (μ) from every data point (X) and then divides each column by the corresponding element in the vector's standard deviation (σ). Generally, standard

scaling makes the data increasingly interpretable, because the normal estimation of Y when x (the mean or centered X) is zero represents the expected value of Y when X is at its mean with a standard deviation of 1. This transforms the data to have a resulting distribution of a mean of 0 and a standard deviation of 1.

$$x_{ij} = \frac{(X_{ij} - \mu_{ij})}{\sigma_{ij}} \tag{4.1}$$

In the process of developing the models, the data were split into a training set (70%) and test set (30%)

4.2.4 Models development

All models were developed using scikit learn 0.22.2 (machine learning in python). A robust library that provides a range of python-based supervised and unsupervised machine algorithms with the capability to deploy machine learning models from prototypes to a production system. It is also a free-open-source software with very huge support from the technology community and commonly used in the industry when compared to MATLAB.

4.2.4.1 Feature reduction

Spectra data obtained from the ATR-FTIR is a high dimensional feature data with a lot of redundant features. Due to the problem of overfitting the model, the data features were reduced using the method of principal component analysis (PCA). PCA method reconstructs features of a dataset into a new set of uncorrelated features called principal

components (PCs). The number of PCs is then selected based on the desired maximum amount of variance explained (Howley et al., 2005).

4.2.4.2 Classification model

To detect whether a bread sample is contaminated with wheat flour during the baking process, a classification model was developed by training different individual classifiers and using an ensemble learning technique or method. The ensemble learning method involves combining different learning algorithms to obtain a high-accuracy meta-model, and experimental evidence indicates this method to be often much more accurate than using a single learning algorithm (Dietterich, 2002). In this study, a voting-based ensemble learning was used. The method stacks different supervised machine learning classification algorithms including a random-forest (RF) classifier, support vector machine (SVM) classifier, and k-nearest neighbor (KNN) classifier. Each base model was trained using 70% of the dataset and then made a classification (vote) on the test (30% of the dataset) instances. The final output was the one that received more than half of the votes (majority voting). This was similar to the method used by Bouziane et al. (2011) to predict protein secondary structure which yielded more significant performance over the use of the best individual classifier. The model was evaluated based on the confusion matrix parameters obtained with emphasis on the false-negative rate (FNR), true positive rate (TPR), and the F1_score. Defined as:

$$TPR = \frac{TP}{(TP+FN)} \tag{4.2}$$

$$FNR = \frac{FN}{(FN+TP)} \tag{4.3}$$

$$\text{F1_score} = 2\frac{PR}{P+R} \tag{4.4}$$

Where: TP = true positive, FN = false positive, P = precision, and R = recall

4.2.4.3 Prediction model

The prediction model was based on developing several individual regression models, ensemble learning, and then selecting the best performing model just as in section 2.4.2. Supervised machine learning regression models including k-nearest neighbors (KNN) regressor, random forest (RF) regressor, decision tree (Dct) regressor, SVM regressor, and partial least square regressor (PLSR) were used for this purpose. The coefficient of determination and root mean square error of the training set (R^2_T, RMSET), and for the test or prediction set (R^2_P, RMSEP) was used to evaluate the performance of the models. Thus, the best model is characterized by higher R^2_T and R^2_P, and the lower root means square error RMSET and RMSEP. To improve each of the individual models, cross-validation was used to tune and determine the value of the model's hyper-parameter selected and their learning curve was obtained. This is to ensure that the model is not underfitting or overfitting the data in any way.

4.3 Results and Discussion

4.3.1 Spectra characteristics of the ground bread samples

Figure 4.1 and Figure 4.2 provide visualizes of the FTIR-spectra formation of the sample contaminated with 0.5% wheat flour when it is in raw form (flour) and after being processed (bread). Comparing these two Figures (4.1 and 4.2) at the region between 1860 cm^{-1} – 1480 cm^{-1}, which includes most of the amide I (1690 cm^{-1} – 1600 cm^{-1}), and amide II (1580 cm^{-1} – 1480 cm^{-1}) characteristic bands that are susceptible or sensitive to the secondary structure content of proteins. This region maintains a smooth formation with unique peaks due to CO and NH or other potential (CC and CN) stretching vibrations in Figure 4.1 (Jabs, 2005). However, in Figure 4.2 we could see some form of noisy deformation within the region and this may be due to protein denaturation during the heating process and other conversion processes such as mixing with other ingredients (e.g. salt) (Neill et al., 2012). Other differences could be seen in the intensities of the peak and trough. Therefore, this shows that the unnatural processes can cause changes in the formation of FTIR-spectra of a sample containing protein.

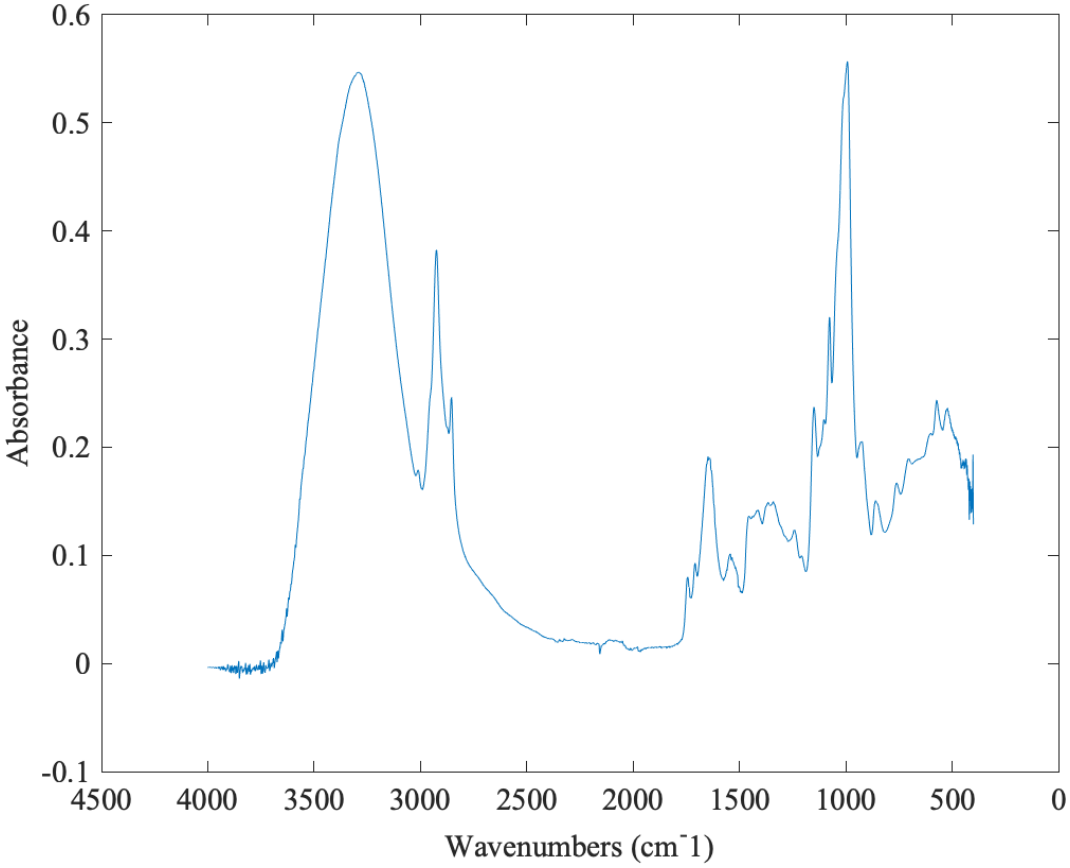


Figure 4.1: Raw sample of FTIR-spectra of the corn-flour contaminated with 0.5% wheat flour

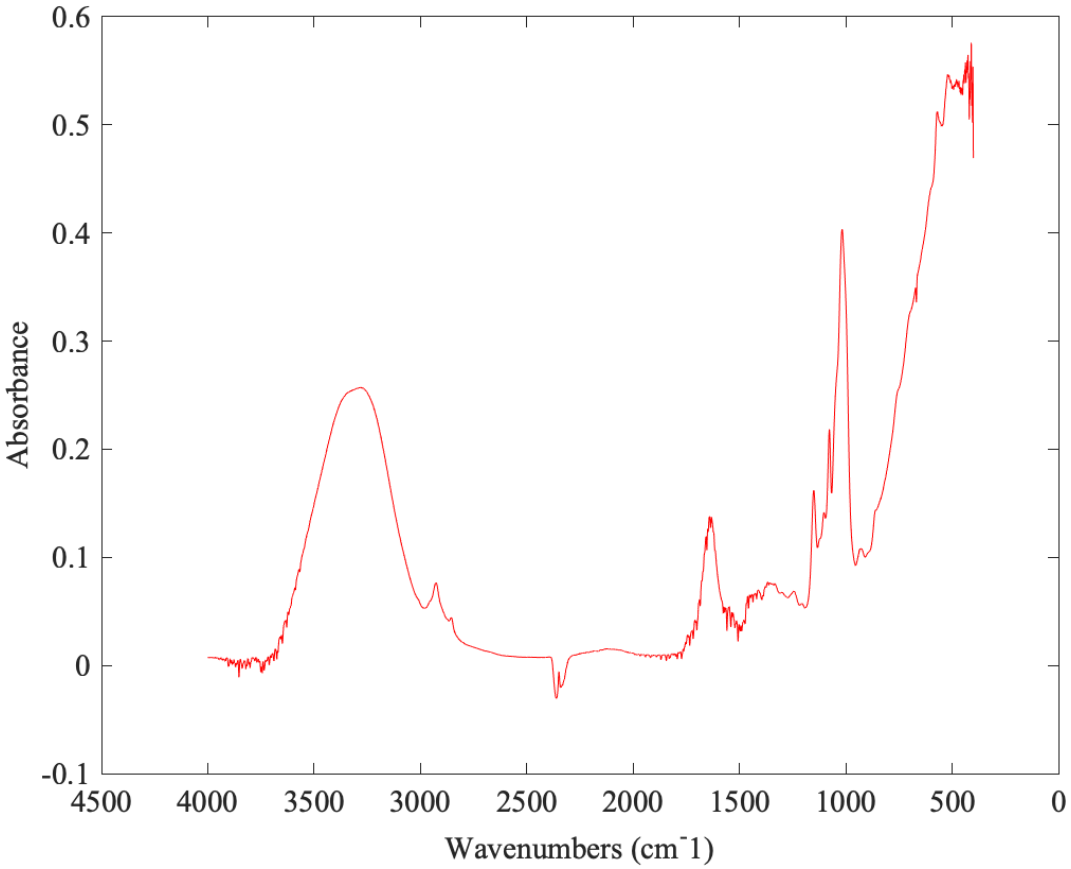


Figure 4.2: Baked sample of FTIR-spectra after the corn-bread contaminated with 0.5% wheat flour

4.3.2 Classification modeling results

Running the PCA reduced the number of features to 20 principal components (PCs), which explained about 100% of the variance in the data samples (Figure 4.3). The 20 PCs were utilized to develop the classification models based on two classes: class1 (No. Contamination) and class 2 (Contamination with Wheat). Among all the classifier methods used including RF classifier, SVM classifier, KNN classifier, and majority voting-based ensemble learning by stacking the individual learning algorithms, the ensemble method gave the best result based on the confusion matrix parameters obtained. Table 4.1 and Table 4.2 presents the confusion matrix and its parameters obtained during the training of the train set (70% of the dataset). The false-negative rate (0), true-positive rate (1.0), and F1-score (1.0) values obtained indicate a 100% rate performance of the model at all times. Table 4.3 and Table 4.4 present the confusion matrix parameters obtained after the model was subjected to a new test data (30% of the dataset). The model was able to accurately classify all samples belonging to each of the class with a TPR, FNR, and F1-score of 1.0, 0, and 1.0 respectively. This shows the ability of the ensembled classifiers to learn every feature in the binary classes of the samples used, also when subjected to the test sets the performance was reliable and consistent. This might not be the case in a real-life application (100% accuracy) due to the complexity of real-world data which always involves limitations such as missing data, unbalanced data, redundant variables, etc. But it justifies that the ensemble learning method is very efficient and has the great potential to

learn most features in classes towards the detection or classification of the wheat contamination levels in the cornbread used.

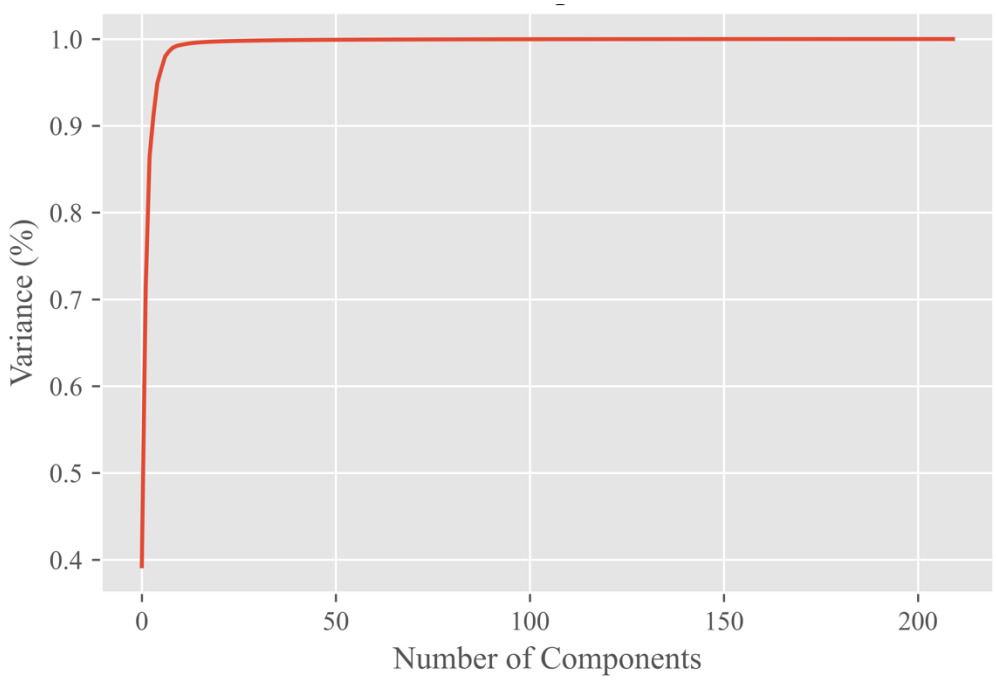


Figure 4.3: Plot of number of principal components (PCs) and variance explained in the samples.

Table 4.1: Confusion matrix parameters for the majority voting-based learning classification training model (Class 1: No Contamination, Class2: Contaminated with Wheat)

Class	TPR	FPR	TNR	FNR	Err	P	F1_score
Class 1	1.0	0.0	1.0	0.0	0.0	1.0	1.0
Class 2	1.0	0.0	1.0	0.0	0.0	1.0	1.0

TPR: True positive rate, FPR: False positive rate, TNR: True negative rate, FNR: False-negative rate, Err: Error, P: Precision, F1: scores for a measure of accuracy.

Table 4.2: Confusion matrix table for the majority voting-based ensemble learning classification training model

Actual Class		
	Class 1	Class 2
Classified as Class 1	138	0
Classified as Class 2	0	142
Classified as Unassigned	0	0

Class 1: No Contamination, Class2: Contaminated with Wheat

Table 4.3: Confusion matrix parameters for the classification test model (Class 1: No Contamination, Class2: Contaminated with Wheat)

Class	TPR	FPR	TNR	FNR	Err	P	F1_score
Class 1	1.0	0.0	1.0	0.0	0.0	1.0	1.0
Class 2	1.0	0.0	1.0	0.0	0.0	1.0	1.0

TPR: True positive rate, FPR: False positive rate, TNR: True negative rate, FNR: False-negative rate, Err: Error, P: Precision, F1: scores for a measure of accuracy.

Table 4.4: Confusion matrix table for the majority voting-based ensemble learning classification test model

Actual Class		
	Class 1	Class 2
Classified as Class 1	62	0
Classified as Class 2	0	58
Classified as Unassigned	0	0

Class 1: No Contamination, Class2: Contaminated with Wheat

4.3.3 Prediction model

Table 4.5 presents the evaluation parameters for the predictive learning algorithm used including RF regressor, KNN regressor, Decision trees, SVM regressor, PLSR, and ensemble learning. The results for KNN and PLSR are very close in performance with an R^2_T = 0.9820 (KNN), 0.9903 (PLSR), R^2_P = 0.9871 (KNN), 0.9694 (PLSR) and RMSET = 0.4062 (KNN), 0.0790 (PLSR), and RMSEP = 0.3314 (KNN), 05192 (PLSR), respectively, which indicates that both learning algorithms have the potential to quantify the level of the wheat flour contaminant in the bread samples within the percentage levels used. Based on the values of R^2_P (0.9871) and RMSEP (0.3314) for KNN, it was selected as the best performing learning algorithm. Figure 4.4 to 4.8 shows the learning curves obtained as a

function of the number of the hyper-parameter tuned for each of the individual algorithms.

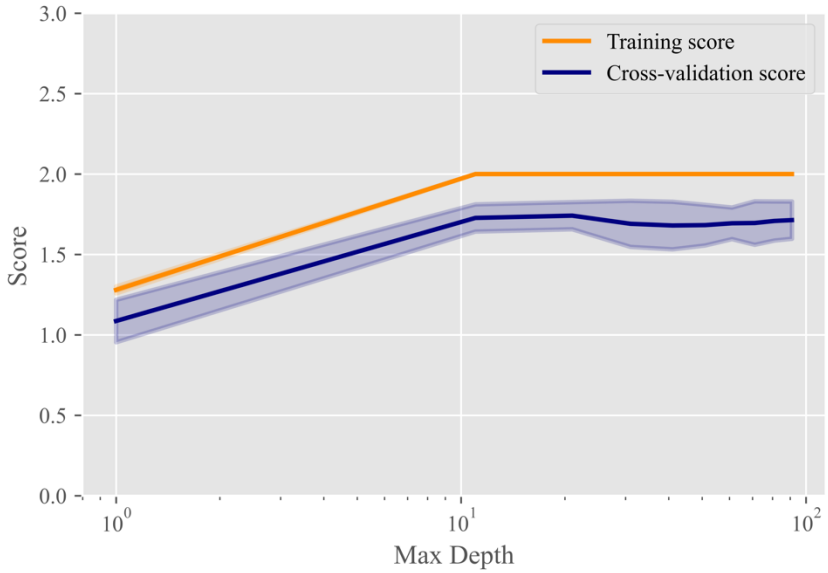


Figure 4.4: Validation curve for decision tree regressor.

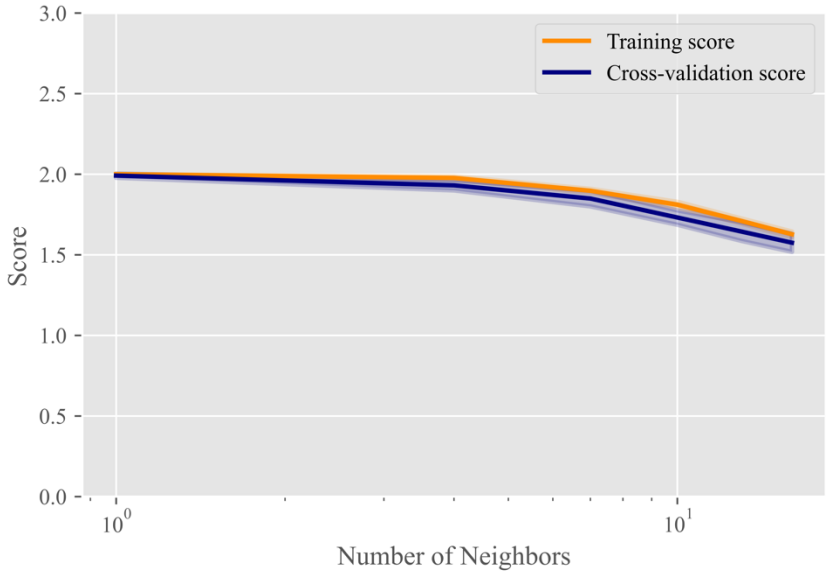


Figure 4.5: Validation curve for K-Nearest Neighbors regressor

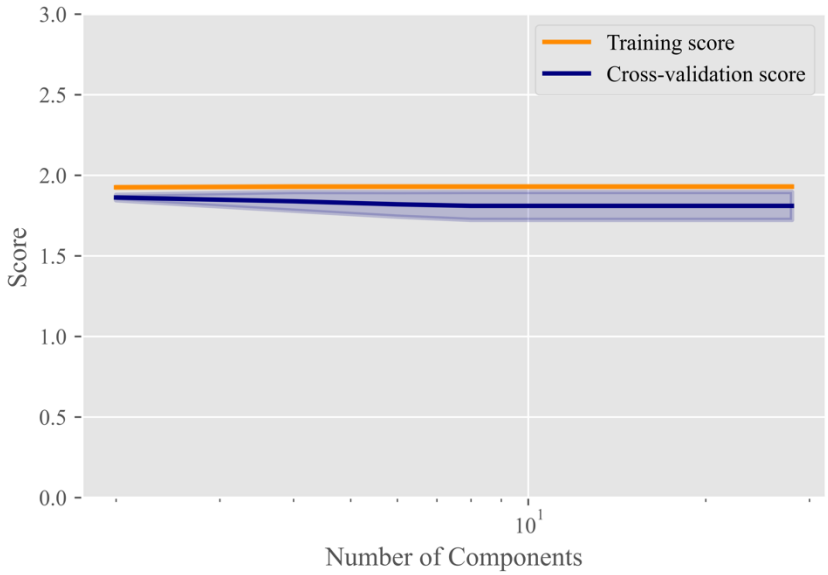


Figure 4.6: Validation curve for partial least square regression (PLSR)

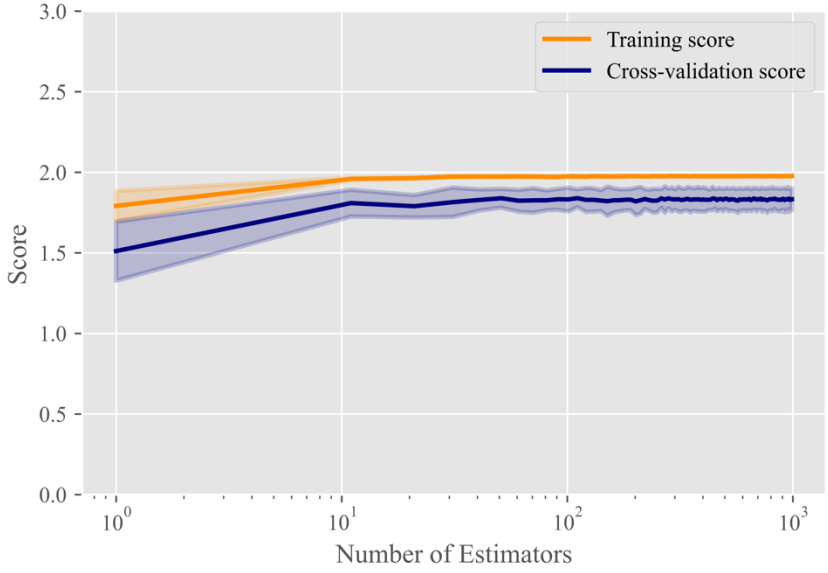


Figure 4.7: Validation curve for random forest regressor.

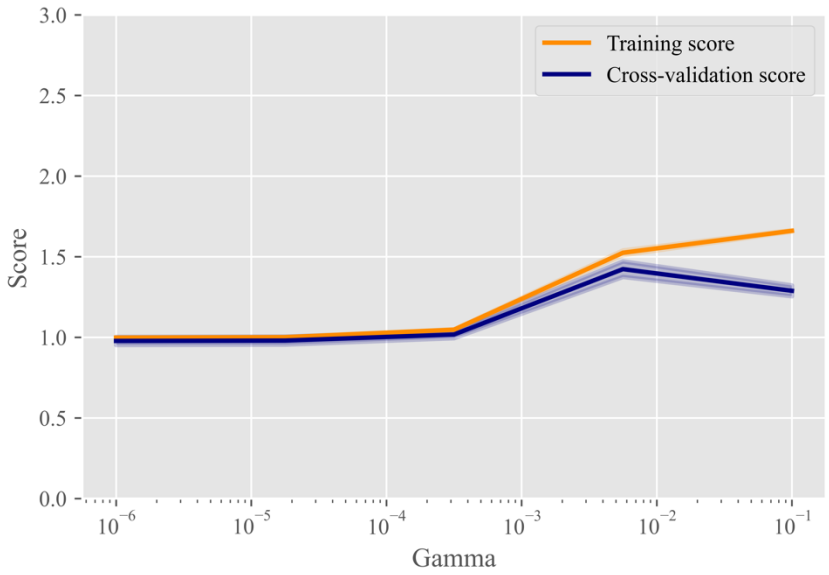


Figure 4.8: Validation curve for support vector machine

This was to choose a hyper-parameter that will strike balance between the bias and variance in other to prevent overfitting. As seen in Figure 4.5 for the best learning algorithm (KNN), the learning gap between the scores measured on the training and cross-validation set is very minimal and insignificant. It was observed that as we increase the number of neighbors for the algorithm the changes remain constant until approaching the value of 6 where the scores started dropping leading to lower accuracy of the model. Therefore, we can conclude from this that the number of neighbors ranging from 1 to 5 to be more effective for the KNN model to predict or quantify the percentage contamination of the wheat flour in the corn-bread.

Table 4.5: Prediction analysis on a different learning algorithm

Learning Algorithm	Hyper-parameter	Training		Prediction	
		R^2_T	RMSET	R^2_P	RMSEP
Random Forest (rf)	n_estimators = 991	1.0	0.0	0.5159	2.0643
K-Nearest Neighbors (knn)	n_neighbors = 4, metric = 'manhattan'	0.9820	0.4062	0.9871	0.3374
Decision Tree (dct)	max_depth = 6	0.9910	0.2874	0.5745	1.9354
Support Vector Machine (svr)	gamma = 0.03	0.8960	0.9767	0.7548	1.4692
Partial Least Square Regression (plsr)	n_components = 30	0.9993	0.0790	0.9694	0.5192
Ensemble Method (voting)	(rf, knn, dct, svr, weight = none)	0.9903	1.2899	0.8110	1.2899

Conclusion

FTIR spectroscopy has always played an important role in the food industry with regards to food safety inspection and quality assessment. In this study, we used FTIR spectroscopy coupled with supervised machine learning (ML) approaches to detect and quantify wheat flour (WF) contamination in the range of 0% - 10% at 0.5% increment in non-gluten bread (corn-bread It was observed that the use of ensemble learning method performed better than using individual supervised ML algorithms in detecting the cornbread samples contaminated with WF. The KNN regressor emerged the most promising technique in quantifying the percentage level of the WF contamination with the best prediction's coefficient of determination of 0.9871 and prediction's root mean squared

error of 0.337. Therefore, the results obtained from this study indicate the potential and the effectiveness of using an FTIR spectrometer with ML techniques in the authentication of WF contamination in a non-gluten bread.

References

Allred, L. K., & Ritter, B. W. (2010). Recognition of gliadin and glutenin fractions in four commercial gluten assays. *Journal of AOAC International, 93*(1), 190-196.

Anjos, O., Campos, M. G., Ruiz, P. C., & Antunes, P. (2015). Application of FTIR-ATR spectroscopy to the quantification of sugar in honey. *Food Chemistry, 169,* 218-223.

Armstrong, P., Maghirang, E., Xie, F., & Dowell, F. (2006). Comparison of dispersive and Fourier-transform NIR instruments for measuring grain and flour attributes. *Applied Engineering in Agriculture, 22*(3), 453-457.

Bouziane, H., Messabih, B., & Chouarfia, A. (2011). Profiles and majority voting-based ensemble method for protein secondary structure prediction. *Evolutionary Bioinformatics, 7,* EBO. S7931.

Dietterich, T. G. (2002). Ensemble learning. *The handbook of brain theory and neural networks, 2,* 110-125.

Howley, T., Madden, M. G., O'Connell, M.-L., & Ryder, A. G. (2005). *The effect of principal component analysis on machine learning accuracy with high dimensional spectral data.* Paper presented at the International Conference on Innovative Techniques and Applications of Artificial Intelligence.

Irudayaraj, J., Xu, R., & Tewari, J. (2003). Rapid determination of invert cane sugar adulteration in honey using FTIR spectroscopy and multivariate analysis. *Journal of food science, 68*(6), 2040-2045.

Jabs, A. (2005). Determination of secondary structure in proteins by fourier transform infrared spectroscopy (FTIR). *Jena Library of Biologica Macromolecules*.

Lacorn, M., Siebeneicher, S., & Weiss, T. (2017). Measurement of Gluten in Food Products: Proficiency- -Testing Rounds as a Measure of Precision and Applicability. *Celiac Disease and Non-Celiac Gluten Sensitivity*, 27.

Liu, J., Wen, Y., Dong, N., Lai, C., & Zhao, G. (2013). Authentication of lotus root powder adulterated with potato starch and/or sweet potato starch using Fourier transform mid-infrared spectroscopy. *Food Chemistry, 141*(3), 3103-3109.

Lohumi, S., Lee, S., Lee, W.-H., Kim, M. S., Mo, C., Bae, H., & Cho, B.-K. (2014). Detection of starch adulteration in onion powder by FT-NIR and FT-IR spectroscopy. *Journal of agricultural and food chemistry, 62*(38), 9246-9251.

Man, Y. C., & Setiowaty, G. (1999). Application of Fourier transform infrared spectroscopy to determine free fatty acid contents in palm olein. *Food Chemistry, 66*(1), 109-114.

Mondal, A., & Datta, A. (2008). Bread baking–A review. *Journal of Food Engineering, 86*(4), 465-474.

Neill, G., Ala'a, H., & Magee, T. (2012). Optimisation of time/temperature treatment, for heat treated soft wheat flour. *Journal of Food Engineering, 113*(3), 422-426.

R-Biopharm, R. (2009). Instructions, Gliadin R7001. In. https://food.r-biopharm.com/wp-content/uploads/sites/2/2016/05/R7001-Gliadin-15-10-09.pdf

Rohman, A., & Che Man, Y. B. (2009). Analysis of cod- liver oil adulteration using Fourier transform infrared (FTIR) spectroscopy. *Journal of the American Oil Chemists' Society*, *86*(12), 1149.

Rohman, A., Erwanto, Y., & Man, Y. B. C. (2011). Analysis of pork adulteration in beef meatball using Fourier transform infrared (FTIR) spectroscopy. *Meat Science*, *88*(1), 91-95.

Sivakesava, S., & Irudayaraj, J. (2001). Detection of inverted beet sugar adulteration of honey by FTIR spectroscopy. *Journal of the Science of Food and Agriculture*, *81*(8), 683-690.

Sørensen, L. (2009). Application of reflectance near infrared spectroscopy for bread analyses. *Food Chemistry*, *113*(4), 1318-1322.

Syahariza, Z., Man, Y. C., Selamat, J., & Bakar, J. (2005). Detection of lard adulteration in cake formulation by Fourier transform infrared (FTIR) spectroscopy. *Food Chemistry*, *92*(2), 365-371.

Van De Voort, F. R., Sedman, J., Emo, G., & Ismail, A. A. (1992). Assessment of Fourier transform infrared analysis of milk. *Journal of AOAC International*, *75*(5), 780-785.

CONNECTING STATEMENT

After the development of the ML models that can detect and quantify the wheat contamination levels in the raw flour samples and processed food (bread). This part of the study was carried out to estimate the amount of gluten present in the contamination levels. In the next chapter, Enzyme-linked immunosorbent assay (ELISA), an approved method by the United States Food and drug administration was used to authenticate gluten in the samples. Therefore, this method was used to complement the quantification models obtained in chapter 3 and 4 above to establish a threshold limit at which we can label the contamination level of our samples to be gluten-free. Furthermore, for the raw flour samples (chapter 3), only the samples with wheat flour contamination were considered because wheat is the most commonly used flour in the food industries in making grain-based foods.

CHAPTER 5. ENZYME LINKED IMMUNOSORBENT ASSAY (ELISA) TEST FOR QUANTIFICATION OF AMOUNT OF GLUTEN PRESENT IN THE CONTAMINATION LEVELS IN CHAPTER THREE AND FOUR

5.1 Introduction

The United States Food and Drug Administration (FDA) specifies a regulatory threshold of ≤ 20 ppm of gluten for any food to be labeled "gluten-free" or "no-gluten" (Allred & Ritter, 2010). Also, it is generally recommended by FDA that foods containing any of the gluten-rich grains including wheat, barley, and rye with a contamination level of gluten below 20 ppm to be considered safe for consumption for most people with gluten related-disorders (Lacorn et al., 2017). Therefore, it is important to inspect foods labeled "gluten-free" or "no-gluten" and the contamination level from the gluten-rich grains to validates that it meets the regulatory threshold limit. In order to ensure that it is safe for consumption for people with gluten-related health concerns.

In chapters 3 and 4, the percentage (%) level of contamination from the wheat flour (WF) in the different samples were detected and quantified using the FTIR with machine learning approaches. However, the amount of gluten present in these samples still needs to be established. Establishing the amount of the gluten in the samples will help determine the percentage limit of the WF contamination level at which the regulatory threshold is applicable.

The use of testing kits or methods that are fully approved and certified by the Association of Official Analytical Chemists (AOAC International) have been suggested by most international organizations and regulatory agencies including FDA. Today, gluten is

validated in foods using Enzyme-linked immunosorbent assay (ELISA) method as it met all the requirements for testing and estimating the amount of gluten in food by the regulatory agencies. Therefore, this part of the study used the ELISA test to established threshold (≤ 20 ppm) of the wheat flour (WF) contamination levels in the study's objective 1 (chapter 3) and objective 2 (chapter 4) for gluten-free labeling.

5.2 Materials and Methods

5.2.1 Materials

The samples from the part of the study in chapter three (raw samples of corn flours contaminated with wheat flour) and in chapter four (processed samples of cornbreads contaminated with wheat flour) were analyzed to quantify the amount of gluten (in ppm) present in each of the samples selected. RIDASCREEN® Gliadin (R7001) ELISA test kit (AOAC international approved) from R-Biopharm (Darmstadt, Germany) was used during the ELISA analysis. The detection limit of the kit is 0.5 ppm gliadin or 1ppm gluten based on the matrix and a quantification limit of 2.5 ppm gliadin or 5 ppm gluten. The specificity of the kit involves the reaction of the monoclonal antibody R5 with the gliadin-divisions from wheat and the corresponding prolamins from barley and rye. Table 5.1 below presents details of all the contents or materials provided in the kit and sufficient enough for 96 measurements (including standard analyses).

Table 5.1: Content (reagents provided) of each ELISA kit

Component	Cap color	Format		Volume
Microtiter plate		Ready to use		96 wells
Buffer	White	Concentrate	5x	60 ml
Standard 1	Transparent	Ready to use	0 ng / ml gliadin	1.3 ml
Standard 2	Transparent	Ready to use	5 ng / ml gliadin	1.3 ml
Standard 3	Transparent	Ready to use	10 ng / ml gliadin	1.3 ml
Standard 4	Transparent	Ready to use	20 ng / ml gliadin	1.3 ml
Standard 5	Transparent	Ready to use	40 ng / ml gliadin	1.3 ml
Standard 6	Transparent	Ready to use	0 ng / ml gliadin	1.3 ml
Wash buffer	Brown	Concentrate	10x	100 ml
Conjugate	Red	Concentrate		1.2 ml
Substrate	Green	Ready to use		7 ml
Chromogen	Blue	Ready to use		7 ml
Stop Solution	Yellow	Ready to use		14 ml

Source: (R-Biopharm, 2009)

5.2.2 Methods

The ELISA method used follows all the laboratory protocol (R-Biopharm, 2009) provided in the test kit. Some of the detailed procedure from the kit instructional manual has been outlined below.

Equipment

i. Microtiter plate spectrophotometer (450 nm)

ii. Centrifuge (Eppendorf, 5417R), centrifugal vials (Greiner centrifuge tube – 1.5 ml)

iii. Shaker or rotator (Rocker II, model: 260350)

iv. Laboratory mincer/grinder, ultra-turrax or homogenizer (Fisher vortex genie 2, Cat no. 12-812)

v. Water bath (50 °C / 122 °F)

vi. Graduated pipettes (Eppendorf)

vii. Variable 20 µl - 200 µl and 200 - 1000 µl micropipettes

Other Reagents Needed

i. Distilled or deionized water

ii. Gluten-free skim milk powder (food quality)

iii. Cocktail (patented) (R7006) and ethanol solution (80 %): i.e. add 120 ml ethanol p.a. to 30 ml distilled water and shake well.

5.2.2.1 Preparation of samples and supernatant extraction

To maintain a free contamination process, 40% ethanol or 2-propanol was used to clean or wiped all surfaces, vials, mincers, and other equipment. All work done was under a chemical hood because of ß-mercaptoethanol content in the Cocktail (patented). Homogenized sample (0.25 g of each) was weighed with the addition of skimmed milk powder (0.25 g), and Cocktail (patented) (2.5 ml). All samples were placed in a 1.5 ml vial and mixed well. After thorough mixing of the samples, they were incubated for 40 min at 50°C (122°F) and cooled down before mixing it with 80 % ethanol (7.5 ml). Then, using a rotator, vials containing the samples were shaken for 1 hour at room temperature (25°C / 77°F). At the end of this, the samples were centrifuged for 10 min, at 20,000 g, and room temperature (25°C / 77°F). The supernatants were then separated utilizing a pipette and extracted into a screw-top vial.

5.2.2.2 Test preparation and Implementation

Preparations

All reagents were brought to room temperature (25°C/77°F) before use. The needed buffer concentrate was diluted at 1:5 (1+4) with distilled water, the needed conjugate (bottle with red cap) concentrate was shaken carefully and then diluted at 1:11 (1+10) with distilled water for reconstitution. Also, the needed washing buffer concentrate was diluted at 1:10 (1+9) with distilled water.

Test procedure

All procedures provided in the instructional manual (R-Biopharm, 2009) from the kit were duly followed.

i. The wells were embedded into the microwell holder for all standards and the samples to be run in copies while recording their positions.
ii. Standard solution and sample of 100 µl each were added to a different copy well and then incubated for 30 min at room temperature (25 °C / 77 °F).
iii. The wells were drained of all liquid and tapped upside down vigorously (three times) against an absorbent paper for the total expulsion of the liquid from the wells. After this, 250 µl diluted washing buffer was poured into each of the wells and then the liquid was poured out again repeatedly twice.
iv. A diluted conjugate of 100 µl was poured into each well and incubated (30 min) at room temperature.
v. Then repeat step 3 (iii)

vi. Substrate and chromogen of 50 µl were added to each well, mixed gently by manually agitating the plate and incubate in the dark for 30 mins at room temperature.

vii. The stop solution (100 µl) was added to each well, mixed gently by manually agitating the plate. Then, the absorbances were measured at 450 nm, 30 mins after adding the stop solution.

After the readings were done, all calculations were carried out in M.S Excel (V. 16.37, 2020) using a cubic spline function.

5.3 Results and Discussion

The standard curve obtained from the six known standards is shown in Figure 5.1 below with R^2 of 0.9994, this was used to estimate the quantity of gluten present in the samples selected from the raw flour samples (in chapter three) and processed samples (in chapter four). The results obtained are presented in Tables 5.2 and 5.3. Some of the results obtained from the ELISA test were invalidated because they were out of range of the standard curve obtained. This might be due to some errors in sample preparation, or issues with equipment readings. Fortunately, this did not interfere with the main goal of running the test. Which is to estimate at what contaminant level (between 0.5-10% at 0.5% increment) is the recommended FDA's threshold limit of 20 ppm for food to be labeled "gluten-free" (Lacorn et al., 2017). The result (Table 5.2) for the raw flour samples from chapter three indicates this to be at 0.5% (15.10 ppm) and any threshold above 20 ppm is

labeled “gluten-contaminated”. Therefore, from this, we can conclude that CF contaminated with WF at a contamination level approaching 1% and above is likely to be more than 20 ppm and not gluten-free. In future works, the recommendation is that this is extended and related to other major sources of gluten (barley flour and rye flour). For the processed (bread) samples contaminated with WF, the result is presented in Table 5.3 below. The ELISA test estimation indicates that the threshold for the WF contamination level to be gluten-free is at 3.5% (19.84 ppm). Therefore, this concludes that cornbread contaminated with the WF at a contamination level below 3.5% to be less than 20 ppm and thus, can be labeled gluten-free. However, it can be observed that the threshold of the baked samples (3.5%) is higher when compared to the raw flour samples (0.5%). This could be due to the baking process (heating) that denatured the protein structures by forming new disulfide bonds and aggregation of the proteins. This makes it more difficult to extract the gluten proteins at a lower level. Furthermore, it might result in lower gluten protein solubility and lead to a lower rate of detection that will require modification of the extraction protocol (Hayta & Alpaslan, 2001). For this reason, the threshold of the raw flour samples (0.5%) is generally recommended as the contamination limit for gluten-free labeling.

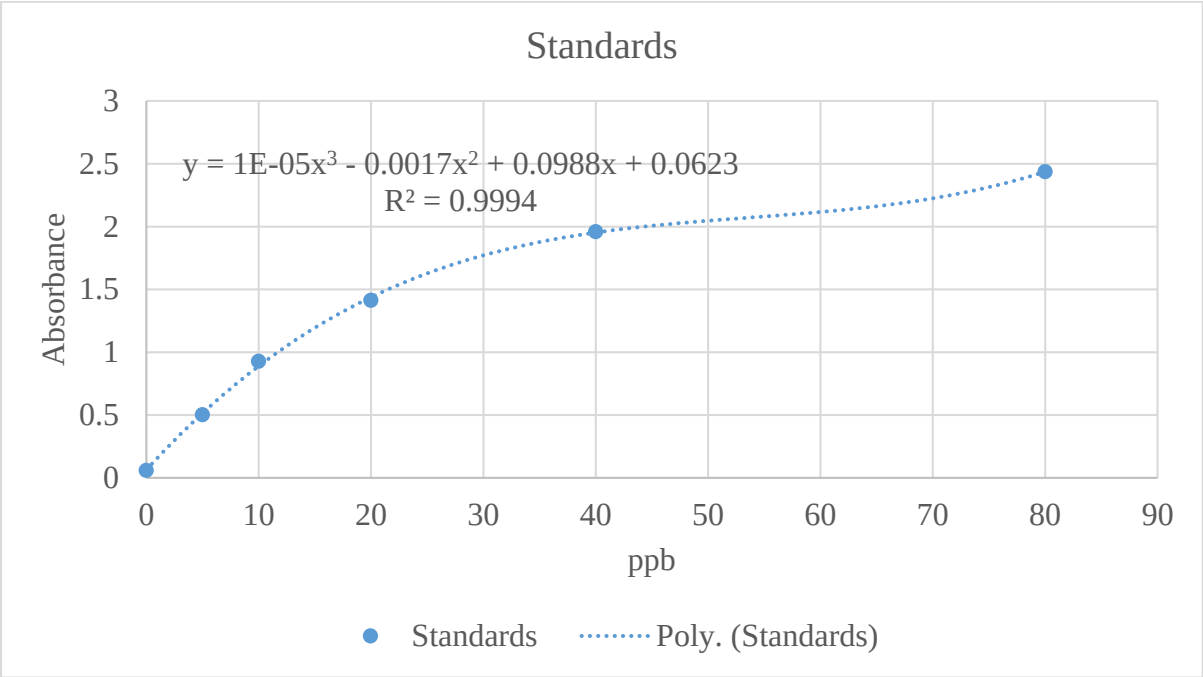


Figure 5.1: ELISA standard curve

Table 5.2: Quantification of the amount of gluten in ppm for the raw flour samples contaminated with WF

WF Contaminant level (%)	Gluten level (ppm)	Label
0.5	15.10	Gluten-Free
1	24.81	Gluten-Contaminated
1.5	45.87	Gluten-Contaminated
2	97.33	Gluten-Contaminated
2.5	83.79	Gluten-Contaminated
3	108.88	Gluten-Contaminated
3.5	***	***
4	***	***
4.5	***	***
5	***	***
5.5	***	***
6	***	***
6.5	***	***
7	***	***
7.5	***	***
8	***	***
8.5	***	***
9	190.18	Gluten-Contaminated
9.5	202.83	Gluten-Contaminated
10	217.36	Gluten-Contaminated

*** Indicates threshold higher values than expected; preparation error, WF: wheat flour.

Table 5.3: Quantification of the amount of gluten in ppm for the processed flour (bread) samples contaminated with wheat flour (WF)

Selected WF contaminated bread samples level (%)	Gluten level (ppm)	Label
0.5	3.04	Gluten-Free
1	4.89	Gluten-Free
1.5	10.54	Gluten-Free
2.5	14.81	Gluten-Free
3.5	19.84	Gluten-Free
4.5	***	***
5.5	38.20	Gluten-Contaminated
6.5	***	***
7.5	40.96	Gluten-Contaminated
8.5	***	***
9.5	47.43	Gluten-Contaminated
10	60.75	Gluten-Contaminated

*** Indicates threshold higher values than expected; preparation error.

Conclusion

This part of the study uses enzyme-linked immunosorbent assay (ELISA) test to estimate or determine the threshold ($\leq$ 20 ppm) of the amount of gluten in the contamination percentage level for the samples in chapter two (corn flour (CF) samples contaminated with wheat flour (WF)) and chapter three (Cornbread samples contaminated with WF) at which they can be labeled gluten-free. The results obtained for the raw CF contaminated with WF show that at less than 0.5% (15.10 ppm) the samples can be marked as gluten-free while the threshold for gluten-free labeling for the samples of the processed CF (corn-bread) contaminated with WF to be at 3.5% (19.84 ppm). This process will help to make more informed decisions about the amount of gluten present in the detection and quantification of the machine-learning models.

Reference

Allred, L. K., & Ritter, B. W. (2010). Recognition of gliadin and glutenin fractions in four commercial gluten assays. *Journal of AOAC International, 93*(1), 190-196.

Hayta, M., & Alpaslan, M. (2001). Effects of processing on biochemical and rheological properties of wheat gluten proteins. *Food/Nahrung, 45*(5), 304-308.

Lacorn, M., Siebeneicher, S., & Weiss, T. (2017). Measurement of Gluten in Food Products: Proficiency- -Testing Rounds as a Measure of Precision and Applicability. *Celiac Disease and Non-Celiac Gluten Sensitivity*, 27.

R-Biopharm, R. (2009). Instructions, Gliadin R7001. In.

CHAPTER 6. GENERAL CONCLUSION AND RECOMMENDATION

This thesis contains a comprehensive research work developed to use Fourier transformed infrared (FTIR) spectroscopy method coupled with machine learning (ML) approaches to detect and quantify gluten cross-contamination in grain-based foods. Gluten is a type of storage protein that is mainly present in wheat, rye and barley grains. Gluten helps food to maintain its shape by acting like a glue that binds and gives dough that stretchy structure during baking. Gluten poses danger to some people who are susceptible to gluten-related disorders such as celiac disease, wheat allergy, and gluten sensitivity when gluten-containing foods are consumed. This can cause many health implications and can be critical if not managed properly. Therefore, this thesis is sectioned into three different phases with specific objectives.

Phase I in chapter 3, FTIR spectrometer with ML approaches was used to detect and quantify cross-contamination between a non-gluten flour (corn flour (CF)) and gluten-rich flour (wheat (WF), barley (BF), and rye (RF)) at different contamination levels of 0-10% with 0.5% increment. Linear discriminant analysis (LDA) with F1-scores (0.963 (WF), 0.949 (BF), 0.963 (RF) and 1.0 (CF)) for the different classes, and partial least square regression (PLSR) with coefficient of determination (R^2_P) for the prediction or test set (0.96 (WF), 0.94 (BF), and 0.98 (RF)), and root mean square error (RMSEP) for the prediction or test set (0.82 (WF), 0.99 (BF), and 0.53 (RF)) emerged as the best performing approaches.

The phase II of the research in chapter 4 utilized FTIR with more advanced methods of ML approaches to detect and quantify cross-contamination processed food (baked) between a non-gluten bread (corn-bread) and wheat at the same contamination levels used in chapter 3. For this phase, majority voting-based ensemble learning (stack of random forest, k-nearest neighbor (KNN) and support vector classifier) ML approach was developed and evaluated for class detection or classification. The following performance metrics were obtained for the two classes (WF and CF) of the models: F1-score, true-positive rate (TPR), false-negative rate (FNR) were obtained to be 1.0, 1.0, and 0.0, respectively. And for the quantification or regression models, K-nearest neighbor was selected as the best performing learning algorithm with R^2_P and RMSEP set to be 0.9871 and 0.3374 respectively.

Chapter 5 discusses comprehensively, the Enzyme-linked immunosorbent assay (ELISA) analysis that was carried out to complement the quantification results obtained in chapter 3 and chapter 4. The ELISA test was used to establish the regulatory gluten threshold limit ($\leq$ 20 ppm) for the samples of WF contamination levels in samples from chapter 3 (raw-flour samples) and chapter 4 (processed samples) to be labeled gluten-free. For the raw-flour samples of WF in chapter 3, this limit was obtained to be at $\leq$ 0.5% while for the processed samples (corn-bread) in chapter 4, it was obtained at $\leq$ 3.5%. Generally, the results obtained from the approaches used in this research indicate a great potential of using a non-destructive method coupled with an ML approach to authenticating the cross-contact of gluten in grain-based foods. With further development and optimization, it could

be deployed as a useful intelligent analytical procedure for fast gluten determinations or estimation in flour and/or grain-based foods.

Copyright © Abuchi G. Okeke 2020

APPENDIX

Appendix A: MATLAB Code

A.1 Spectra Data Analysis Code

```
%Author: Abuchi Okeke
%Last Date modified: 03/28/2020

%Description:
%Calls function that Loads the spectra data in .SPA format to Matlab format .mat
%Splits data into sets using KernardStone algorithm
%Initiates the PLS Toolbox by Eigenvector Research for data analysis

%Prerequisite
%Install from https://eigenvector.com/software/pls-toolbox/

%clean up
clc; clear; close all;

tic
%1
%Load Raw Spectral Data
[Spectra, Wavenumbers, SpectraTitles, Filenames, ...
   SpectraComments] = LoadSpectra ();

%Read from CSV file
%filename1 = 'classification_resampled.csv';
%Spectra = csvread(filename1); %reads the specified worksheet.

%Spectra = Spectra';
ir = 10;              %number of data replication or duplicates
[m, n] = size(Spectra);
N = n/ir;



%Set data splitting ratio
cRatio=0.8;  % Eighty percent of the samples were selected as calibration set and twenty percent as prediction set
pRatio=1 - cRatio;
```

```
nC = round(n*cRatio);   %column size for calibration set
nP = round(n*pRatio);   %column size for prediction set
calibrationSet = zeros((nC),m);  %initiate calibration set
predictionSet = zeros(nP,m);     %initiate prediction set
yCal = zeros((nC),1);            %initiate calibration labels
yPred = zeros(nP,1);             %initiate prediction labels

%Initiate average/mean data sets
% meanSpectra = zeros(m,N);
mCalibrationSet = zeros(N,m);
mPredictionSet = zeros(N,m);

%Read labels for the examples (change filename)
filename = 'Yclassification_label.csv';
data = csvread(filename); %reads the specified worksheet.
yVar = data(:,1);
```

```
%2
%Splits data using KennardStone Algorithm
%Finds average Spectra Data

   j = ir;
   k = 1;
   kC = 1;
   kP =1;
   jC = round(ir*cRatio);
   jP = round(ir*pRatio);

   for i = 1:N

  [calibrationSet(kC:jC,:), predictionSet(kP:jP,:),yCal(kC:jC), yPred(kP:jP), ...
      mCalibrationSet(i,:), mPredictionSet(i,:), mYCal, mYPred] = callKennardStone(Spectra(:,k:j),cRatio,yVar(k:j));

   meanSpectra(:,i) = meanSpectrum (Spectra(:,k:j));  %%call function to get spectra mean

   k = k + ir;
   j = j + ir;
   kC = kC + round((ir*cRatio));
   kP = kP + round((ir*pRatio));
   jC = jC + round((ir*cRatio));
   jP = jP + round((ir*pRatio));

   end
```

```
X = meanSpectra';

%Visualize data
   plot(Wavenumbers(:,1:N),meanSpectra);          %Visualize samples mean
%  plot(Wavenumbers,Spectra);                    %Visualize all samples data
   set(gca,'xdir','reverse','fontsize', 18);
   xlabel('Wavenumbers (cm^-1)');
   ylabel('Absorbance');

   %Legend for pure samples
   legend('Barley', 'Corn', 'Rye', 'Wheat');

toc
```

```
%3
pls % launches analysis window with for PLS Toolbox by Eigenvector Research.
%All data-pre-processing and analyses can be done directly in the app provided by the PLS Toolbox when launched

%%%END%%%
```

Published with MATLAB® R2018b

A.2 Function for loading FTIR (.SPA) data into set of arrays in MATLAB

```
function [Spectra, Wavenumbers, SpectraTitles, Filenames, ...
   SpectraComments] = LoadSpectra ()

%
% LoadSpectra.m
%
% Imports the absorbance data in .SPA spectrum files into a set of arrays
% with data from the selected files stored in columns.
%
% Copyright Kurt Oldenburg - 06/28/16
%

[Filenames,pathname]=uigetfile({'*.spa','Thermo Spectrum (*.spa)'}, ...
      'MultiSelect','on','Select Spectra Files...');

cd (pathname);  % Change to directory where the spectrum files are.
```

```
if ischar(Filenames)== 1          % If only 1 file is selected, Filenames
   NumSpectra = 1;                % is a char instead of a cell of chars,
else                              % which messes up fopen.
   NumSpectra =length(Filenames);
end

for i = 1:NumSpectra

   DataStart=0;
   CommentStart=0;

   if NumSpectra == 1
      fid=fopen(Filenames,'r');
   else
      fid=fopen(Filenames{i},'r');
   end;

   fseek(fid,30,'bof');
   SpectraTitles(i)={char(nonzeros(fread(fid,255,'uint8'))')};

   fseek(fid,564,'bof');
   Spectrum_Pts=fread(fid,1,'int32');

   fseek(fid,576,'bof');
   Max_Wavenum=fread(fid,1,'single');
   Min_Wavenum=fread(fid,1,'single');

   % The Wavenumber values are assumed to be linearly spaced between
   % between the Min and Max values. The array needs to be flipped
   % around to get the order lined up with the absorbance data.

   Wavenumbers(:,i)=flipud(linspace(Min_Wavenum,...
      Max_Wavenum,Spectrum_Pts).')';


   % The starting byte location of the absorbance data is stored in the
   % header. It immediately follows a flag value of 3:

   Flag=0;

   fseek(fid,288,'bof');

   while Flag ~= 3
```

```
    Flag = fread(fid,1,'uint16');
  end;

  DataPosition=fread(fid,1,'uint16')';
  fseek(fid,DataPosition,'bof');

  Spectra(:,i)=fread(fid,Spectrum_Pts,'single');

  % Same story goes for the Comments section with a flag of 4.
  % The size of the section is the difference between the two.

  Flag=0;

  fseek(fid,288,'bof');

  while Flag ~= 4
    Flag = fread(fid,1,'uint16');
  end

  CommentPosition=fread(fid,1,'uint16')';
  SpectraComments(i)={char(nonzeros(fread(fid, ...
    (DataPosition-DataPosition), 'uint8'))')};

  fclose(fid);

end;
```

Published with MATLAB® R2018b

A.3 Function for splitting data using Kennard Stone algorithm

```
function [calibrationSet, predictionSet, yCal, yPred, mCalibrationSet, mPredictionSet, ...
  mYCal, mYPred] = callKennardStone (spectra,ratio,yVar)

%Author: Abuchi Okeke
%Sample spectra data
%Date modified: 07/06/2019
%Description: calls Kennard-Stone function for sample selection
%Matlab format
```

```
% %1
% %Load Raw Spectral Data
% [Spectra, Wavenumbers, SpectraTitles, Filenames, ...
%     SpectraComments] = LoadSpectra ();

[m, n] = size(spectra);

%2
%Perform Kennard-Stone to uniformly select samples
k = round(n*ratio);  %number of samples to select.
x = spectra'; %transpose data;
selSpectra = kennardstone(x, k);
calibrationSet = x(selSpectra,:);
% idx = find(selSpectra == 1);
predictionSet = x(~selSpectra,:);
y = yVar';

yCal = y(selSpectra);
yPred = y(~selSpectra);

[mCal, nCal] = size(calibrationSet);
mCalibrationSet = zeros(nCal,1);
mYCal = zeros(mCal,1);

[mPred, nPred] = size(predictionSet);
mPredictionSet = zeros(nPred,1);
mYPred = zeros(mPred,1);

for i = 1:nCal
  sumCalibrationSet= sum(calibrationSet(:,i));
  mCalibrationSet(i) = sumCalibrationSet/mCal;
%    sumYCal = sum(yCal(i));
%    mYCal(i) = sumYCal/mCal;
end


for i = 1:nPred
  sumPredictionSet= sum(predictionSet(:,i));
  mPredictionSet(i) = sumPredictionSet/mPred;
%    sumYPred = sum(yPred(i));
%    mYPred(i) = sumYPred/mPred;
end
```

```
end
```

Published with MATLAB® R2018b

A.4 Function for averaging the spectra data

```
function mSpectrum = meanSpectrum (spectra)

%Author: Abuchi Okeke
%Date: 06/30/2019
%Description: This function calculates average of spectra data

[m, n] = size(spectra);
mSpectrum = zeros(m,1);

for i = 1:m
  sumSpectra = sum(spectra(i,:));
  mSpectrum(i) = sumSpectra/n;
end

end
```

Published with MATLAB® R2018b

Appendix B: Python Code

B1. Python Library

Sci-kit Learn: https://scikit-learn.org/stable/

B.2 Classification models

Scan the QR code below with your smart phone camera or click on the link below it to access the python notebook for prototyping classification models



bit.ly/ml-classification

B.3 Predictive/Regression models

Scan the QR code below or click on the link below it to access the python notebook for prototyping the regression or predictive models.



bit.ly/ml-prediction

BIBLIOGRAPHY

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. doi:https://doi.org/10.1016/j.jcs.2012.06.011

Allred, L. K., & Ritter, B. W. (2010). Recognition of gliadin and glutenin fractions in four commercial gluten assays. *Journal of AOAC International, 93*(1), 190-196.

Amir, R. M., Anjum, F. M., Khan, M. I., Khan, M. R., Pasha, I., & Nadeem, M. (2013). Application of Fourier transform infrared (FTIR) spectroscopy for the identification of wheat varieties. *Journal of food science and technology, 50*(5), 1018-1023.

Anjos, O., Campos, M. G., Ruiz, P. C., & Antunes, P. (2015). Application of FTIR-ATR spectroscopy to the quantification of sugar in honey. *Food Chemistry, 169*, 218-223.

Antiga, E., & Caproni, M. (2015). The diagnosis and treatment of dermatitis herpetiformis. *Clinical, cosmetic and investigational dermatology, 8*, 257.

Armstrong, P., Maghirang, E., Xie, F., & Dowell, F. (2006). Comparison of dispersive and Fourier-transform NIR instruments for measuring grain and flour attributes. *Applied Engineering in Agriculture, 22*(3), 453-457.

Ayodele, T. O. (2010a). Machine learning overview. *New Advances in Machine Learning*, 9-19.

Ayodele, T. O. (2010b). Types of machine learning algorithms. *New Advances in Machine Learning*, 19-48.

Baravkar, A., Kale, R., & Sawant, S. (2011). FTIR Spectroscopy: principle, technique and mathematics. *International Journal of Pharma and Bio Sciences, 2*(1), 513-519.

Barbin, D. F., ElMasry, G., Sun, D.-W., & Allen, P. (2012). Predicting quality and sensory attributes of pork using near-infrared hyperspectral imaging. *Analytica Chimica Acta, 719*, 30-42.

BAŞLAR, M., & Ertugay, M. F. (2011). Determination of protein and gluten quality-related parameters of wheat flour using near-infrared reflectance spectroscopy (NIRS). *Turkish Journal of Agriculture and Forestry, 35*(2), 139-144.

Baştanlar, Y., & Özuysal, M. (2014). Introduction to machine learning. In *miRNomics: MicroRNA Biology and Computational Analysis* (pp. 105-128): Springer.

Biesiekierski, J. R. (2017). What is gluten? *Journal of gastroenterology and hepatology, 32*, 78-81. doi:10.1111/jgh.13703.

Bouziane, H., Messabih, B., & Chouarfia, A. (2011). Profiles and majority voting-based ensemble method for protein secondary structure prediction. *Evolutionary Bioinformatics, 7*, EBO. S7931.

Cai, Y.-D., Liu, X.-J., Xu, X.-b., & Zhou, G.-P. (2001). Support vector machines for predicting protein structural class. *BMC bioinformatics, 2*(1), 3.

Camire, M. E. (1998). Chemical changes during extrusion cooking. In *Process-induced chemical changes in food* (pp. 109-121): Springer.

Çamoglu, O., Can, T., Singh, A. K., & Wang, Y.-F. (2005). Decision tree based information integration for automated protein classification. *Journal of Bioinformatics and Computational Biology, 3*(03), 717-742.

Celiac Disease Foundation. (2020). Dermatitis Herpetiformis. *[Online]*.

Cen, H., & He, Y. (2007). Theory and application of near infrared reflectance spectroscopy in determination of food quality. *Trends in food science & technology, 18*(2), 72-83.

Chung, I.-F., Huang, C.-D., Shen, Y.-H., & Lin, C.-T. (2003). Recognition of structure classification of protein folding by NN and SVM hierarchical learning architecture. In *Artificial Neural Networks and Neural Information Processing—ICANN/ICONIP 2003* (pp. 1159-1167): Springer.

Ciemniewska-Żytkiewicz, H., Bryś, J., Sujka, K., & Koczoń, P. (2015). Assessment of the hazelnuts roasting process by pressure differential scanning calorimetry and MID-FT-IR spectroscopy. *Food Analytical Methods, 8*(10), 2465-2473.

Crespo, J. F., & Rodriguez, J. (2003). Food allergy in adulthood. *Allergy, 58*(2), 98-113. doi:10.1034/j.1398-9995.2003.02170.x

Czaja, T., Mazurek, S., & Szostak, R. (2016a). Quantification of gluten in wheat flour by FT-Raman spectroscopy. *Food Chemistry, 211*, 560-563. doi:https://doi.org/10.1016/j.foodchem.2016.05.108

Czaja, T., Mazurek, S., & Szostak, R. (2016b). Quantitative analysis of solid samples using modified specular reflectance accessory. *Talanta, 161*, 655-659.

Dehzangi, A., Phon-Amnuaisuk, S., & Dehzangi, O. (2010). Using random forest for protein fold prediction problem: an empirical study. *J. Inf. Sci. Eng., 26*(6), 1941-1956.

Dietterich, T. G. (2002). Ensemble learning. *The handbook of brain theory and neural networks, 2*, 110-125.

Ding, C. H., & Dubchak, I. (2001). Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics, 17*(4), 349-358.

Dogan, A., Siyakus, G., & Severcan, F. (2007). FTIR spectroscopic characterization of irradiated hazelnut (Corylus avellana L.). *Food Chemistry, 100*(3), 1106-1114.

Doyle, W. M. (1992). Principles and applications of Fourier transform infrared (FTIR) process analysis. *Process Control Qual, 2*(1), 11-41.

Duarte, I. F., Barros, A., Delgadillo, I., Almeida, C., & Gil, A. M. (2002). Application of FTIR spectroscopy for the quantification of sugars in mango juice as a function of ripening. *Journal of agricultural and food chemistry, 50*(11), 3104-3111.

Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern classification*: John Wiley & Sons.

El-Mesery, H. S., Mao, H., & Abomohra, A. E.-F. (2019). Applications of non-destructive technologies for agricultural and food products quality inspection. *Sensors, 19*(4), 846.

Elli, L., Branchi, F., Tomba, C., Villalta, D., Norsa, L., Ferretti, F., . . . Bardella, M. T. (2015). Diagnosis of gluten related disorders: Celiac disease, wheat allergy and

non-celiac gluten sensitivity. *World journal of gastroenterology: WJG*, *21*(23), 7110.

Fasano, A., Sapone, A., Zevallos, V., & Schuppan, D. J. G. (2015). Nonceliac gluten and wheat sensitivity. *148*(6), 1195-1204.

Feighery, C. (1999). Coeliac disease. *Bmj*, *319*(7204), 236-239.

Forbes, A. D. (1995). Classification-algorithm evaluation: Five performance measures based onconfusion matrices. *Journal of Clinical Monitoring, 11*(3), 189-206.

Fueyo-Díaz, R., Magallón-Botaya, R., Masluk, B., Palacios-Navarro, G., Asensio-Martínez, A., Gascón-Santos, S., . . . Sebastián-Domingo, J. J. (2019). Prevalence of celiac disease in primary care: the need for its own code. *BMC Health Services Research, 19*(1), 578. doi:10.1186/s12913-019-4407-4

Gallardo-Velázquez, T., Osorio-Revilla, G., Zuñiga-de Loa, M., & Rivera-Espinoza, Y. (2009). Application of FTIR-HATR spectroscopy and multivariate analysis to the quantification of adulterants in Mexican honeys. *Food Research International*, *42*(3), 313-318.

Galvao, R. K. H., Araujo, M. C. U., Jose, G. E., Pontes, M. J. C., Silva, E. C., & Saldanha, T. C. B. (2005). A method for calibration and validation subset partitioning. *Talanta, 67*(4), 736-740.

García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining*: Springer.

Glassford, S. E., Byrne, B., & Kazarian, S. G. (2013). Recent applications of ATR FTIR spectroscopy and imaging to proteins. *Biochimica et Biophysica Acta (BBA)* -

Proteins and Proteomics, 1834(12), 2849-2858. doi:https://doi.org/10.1016/j.bbapap.2013.07.015

Gluten Free Society. (2020). Ataxia – Another Symptom of Gluten Induced Damage. *Gluten free society blog, nerve damage, nutritional deficiencies.* .

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research, 3*(Mar), 1157-1182.

Hadjivassiliou, M., Boscolo, S., Davies–Jones, G., Grünewald, R., Not, T., Sanders, D., . . . Woodroofe, N. (2002). The humoral response in the pathogenesis of gluten ataxia. *Neurology, 58*(8), 1221-1226.

Hadjivassiliou, M., Davies-Jones, G., Sanders, D., & Grünewald, R. (2003). Dietary treatment of gluten ataxia. *Journal of Neurology, Neurosurgery & Psychiatry, 74*(9), 1221-1224.

Hadjivassiliou, M., Sanders, D., & Aeschlimann, D. (2015). Gluten-related disorders: gluten ataxia. *Digestive Diseases, 33*(2), 264-268.

Hayta, M., & Alpaslan, M. (2001). Effects of processing on biochemical and rheological properties of wheat gluten proteins. *Food/Nahrung, 45*(5), 304-308.

Howley, T., Madden, M. G., O'Connell, M.-L., & Ryder, A. G. (2005). *The effect of principal component analysis on machine learning accuracy with high dimensional spectral data.* Paper presented at the International Conference on Innovative Techniques and Applications of Artificial Intelligence.

Inomata, N. (2009). Wheat allergy. *Current opinion in allergy and clinical immunology, 9*(3), 238-243.

Irudayaraj, J., Xu, R., & Tewari, J. (2003). Rapid determination of invert cane sugar adulteration in honey using FTIR spectroscopy and multivariate analysis. *Journal of food science, 68*(6), 2040-2045.

Ismail, A. A., van de Voort, F. R., & Sedman, J. (1997). Fourier transform infrared spectroscopy: principles and applications. In *Techniques and instrumentation in analytical chemistry* (Vol. 18, pp. 93-139): Elsevier.

Jabs, A. (2005). Determination of secondary structure in proteins by fourier transform infrared spectroscopy (FTIR). *Jena Library of Biologica Macromolecules*.

Jong, S. D. (1993). PLS fits closer than PCR. *Journal of chemometrics, 7*(6), 551-557.

Kanerva, P. (2011). *Immunochemical analysis of prolamins in gluten-free foods*: University of Helsinki.

Kazarian, S. G., & Chan, K. A. (2013). ATR-FTIR spectroscopic imaging: recent advances and applications to biological systems. *Analyst, 138*(7), 1940-1951.

Kotsiantis, S., & Pintelas, P. (2004). Combining bagging and boosting. *International Journal of Computational Intelligence, 1*(4), 324-333.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering, 160*, 3-24.

Lacorn, M., Siebeneicher, S., & Weiss, T. (2017). Measurement of Gluten in Food Products: Proficiency- -Testing Rounds as a Measure of Precision and Applicability. *Celiac Disease and Non-Celiac Gluten Sensitivity*, 27.

Lebwohl, B., Ludvigsson, J. F., & Green, P. H. (2015a). Celiac disease and non-celiac gluten sensitivity. *Bmj, 351*, h4347.

Lebwohl, B., Ludvigsson, J. F., & Green, P. H. J. B. (2015b). Celiac disease and non-celiac gluten sensitivity. *351*, h4347.

Lee, H. J., Anderson, Z., & Ryu, D. (2014). Gluten contamination in foods labeled as "gluten free" in the United States. *Journal of food protection, 77*(10), 1830-1833.

Lee, J. H., Shin, J., & Realff, M. J. (2018a). Machine learning: Overview of the recent progresses and implications for the process systems engineering field. *Computers & Chemical Engineering, 114*, 111-121.

Lee, L. C., Liong, C.-Y., & Jemain, A. A. (2018b). *Iterative random vs. Kennard-Stone sampling for IR spectrum-based classification task using PLS2-DA.*

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors, 18*(8), 2674.

Liu, J., Wen, Y., Dong, N., Lai, C., & Zhao, G. (2013). Authentication of lotus root powder adulterated with potato starch and/or sweet potato starch using Fourier transform mid-infrared spectroscopy. *Food Chemistry, 141*(3), 3103-3109.

Lohumi, S., Lee, S., Lee, W.-H., Kim, M. S., Mo, C., Bae, H., & Cho, B.-K. (2014). Detection of starch adulteration in onion powder by FT-NIR and FT-IR spectroscopy. *Journal of agricultural and food chemistry, 62*(38), 9246-9251.

Majamaa, H., Moisio, P., Majamaa, H., Turjanmaa, K., & Holm, K. (1999). Wheat allergy: diagnostic accuracy of skin prick and patch tests and specific IgE. *Allergy, 54*(8), 851-856.

Makarenko, S. P., Trufanov, V. A., & Putilina, T. E. (2002). Infrared Spectroscopic Study of the Secondary Structure of Wheat, Rye, and Barley Prolamins. *Russian Journal of Plant Physiology, 49*(3), 326-331. doi:10.1023/a:1015584700841

Maleki, M., Mouazen, A., Ramon, H., & De Baerdemaeker, J. (2007). Multiplicative scatter correction during on-line measurement with near infrared spectroscopy. *Biosystems engineering, 96*(3), 427-433.

Man, Y. C., & Setiowaty, G. (1999). Application of Fourier transform infrared spectroscopy to determine free fatty acid contents in palm olein. *Food Chemistry, 66*(1), 109-114.

Mena, M., & Sousa, C. (2015). Analytical tools for gluten detection: Policies and regulation. *OmniaScience Monographs*.

Meresse, B., Malamut, G., & Cerf-Bensussan, N. (2012). Celiac disease: an immunological jigsaw. *Immunity, 36*(6), 907-919.

Mondal, A., & Datta, A. (2008). Bread baking–A review. *Journal of Food Engineering, 86*(4), 465-474.

Mooney, P., Aziz, I., & Sanders, D. (2013). Non- celiac gluten sensitivity: clinical relevance and recommendations for future research. *Neurogastroenterology & Motility, 25*(11), 864-871.

Neill, G., Ala'a, H., & Magee, T. (2012). Optimisation of time/temperature treatment, for heat treated soft wheat flour. *Journal of Food Engineering, 113*(3), 422-426.

Nordqvist, C. B., N. . (2018). What is a wheat allergy? . *Retrieved from Medical News Today Website: https://www.medicalnewstoday.com/articles/174405.php*.

Oza, N. C. (2005). *Online bagging and boosting.* Paper presented at the 2005 IEEE international conference on systems, man and cybernetics.

Quiñones-Islas, N., Meza-Márquez, O. G., Osorio-Revilla, G., & Gallardo-Velazquez, T. (2013). Detection of adulterants in avocado oil by Mid-FTIR spectroscopy and multivariate analysis. *Food Research International, 51*(1), 148-154.

R-Biopharm, R. (2009). Instructions, Gliadin R7001. In.

Rady, A., & Adedeji, A. (2018). Assessing different processed meats for adulterants using visible-near-infrared spectroscopy. *Meat science, 136*, 59-67.

Rady, A., & Adedeji, A. A. (2020). Application of Hyperspectral Imaging and Machine Learning Methods to Detect and Quantify Adulterants in Minced Meats. *Food Analytical Methods*, 1-12.

Reder, M., Koczoń, P., Wirkowska, M., Sujka, K., & Ciemniewska-Żytkiewicz, H. (2014). The application of FT-MIR spectroscopy for the evaluation of energy value, fat content, and fatty acid composition in selected organic oat products. *Food Analytical Methods, 7*(3), 547-554.

Rodriguez-Saona, L., & Allendorf, M. (2011). Use of FTIR for rapid authentication and detection of adulteration of food. *Annual review of food science and technology, 2*, 467-483.

Rohman, A., & Che Man, Y. B. (2009). Analysis of cod- liver oil adulteration using Fourier transform infrared (FTIR) spectroscopy. *Journal of the American Oil Chemists' Society, 86*(12), 1149.

Rohman, A., Erwanto, Y., & Man, Y. B. C. (2011). Analysis of pork adulteration in beef meatball using Fourier transform infrared (FTIR) spectroscopy. *Meat Science, 88*(1), 91-95.

Rohman, A., & Man, Y. C. (2010). Fourier transform infrared (FTIR) spectroscopy for analysis of extra virgin olive oil adulterated with palm oil. *Food Research International, 43*(3), 886-892.

Rubio-Tapia, A., Ludvigsson, J. F., Brantner, T. L., Murray, J. A., & Everhart, J. E. (2012). The prevalence of celiac disease in the United States. *American Journal of Gastroenterology, 107*(10), 1538-1544.

Saha, I., Zubek, J., Klingström, T., Forsberg, S., Wikander, J., Kierczak, M., . . . Plewczynski, D. (2014). Ensemble learning prediction of protein–protein interactions using proteins functional annotations. *Molecular BioSystems, 10*(4), 820-830.

Schuppan, D., Pickert, G., Ashfaq-Khan, M., & Zevallos, V. (2015). Non-celiac wheat sensitivity: differential diagnosis, triggers and implications. *Best Practice & Research Clinical Gastroenterology, 29*(3), 469-476.

Shamim, M. T. A., Anwaruddin, M., & Nagarajaram, H. A. (2007). Support vector machine-based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs. *Bioinformatics, 23*(24), 3320-3327.

Sharma, G. M., Pereira, M., & Williams, K. M. (2015). Gluten detection in foods available in the United States – A market survey. *Food Chemistry, 169,* 120-126. doi:https://doi.org/10.1016/j.foodchem.2014.07.134

Shewry, P. R., Halford, N. G., Belton, P. S., & Tatham, A. S. (2002). The structure and properties of gluten: an elastic protein from wheat grain. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 357*(1418), 133-142. doi:10.1098/rstb.2001.1024

Sivakesava, S., & Irudayaraj, J. (2001). Detection of inverted beet sugar adulteration of honey by FTIR spectroscopy. *Journal of the Science of Food and Agriculture, 81*(8), 683-690.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management, 45*(4), 427-437.

Størsrud, S., Yman, I. M., & Lenner, R. (2003). Gluten contamination in oat products and products naturally free from gluten. *European Food Research and Technology, 217*(6), 481-485.

Su, W.-H., & Sun, D.-W. (2017). Evaluation of spectral imaging for inspection of adulterants in terms of common wheat flour, cassava flour and corn flour in organic Avatar wheat (Triticum spp.) flour. *Journal of Food Engineering, 200,* 59-69.

Sujka, K., Koczoń, P., Ceglińska, A., Reder, M., & Ciemniewska-Żytkiewicz, H. (2017). The application of FT-IR spectroscopy for quality control of flours obtained from polish producers. *Journal of Analytical Methods in Chemistry, 2017*.

Syahariza, Z., Man, Y. C., Selamat, J., & Bakar, J. (2005). Detection of lard adulteration in cake formulation by Fourier transform infrared (FTIR) spectroscopy. *Food Chemistry, 92*(2), 365-371.

Tan, A. C., Gilbert, D., & Deville, Y. (2003). Multi-class protein fold classification using a new ensemble machine learning approach. *Genome Informatics, 14*, 206-217.

Tanveer, M., & Ahmed, A. (2019). Non-Celiac Gluten Sensitivity: A Systematic Review. *Journal of the College of Physicians and Surgeons Pakistan, 29*(1), 51-57.

Tatham, A., & Shewry, P. (2008). Allergens to wheat and related cereals. *Clinical & Experimental Allergy, 38*(11), 1712-1726.

Thompson, T. (2003). Oats and the gluten-free diet. *Journal of the American Dietetic Association, 103*(3), 376-379.

Thompson, T. (2004). Gluten contamination of commercial oat products in the United States. *New England Journal of Medicine, 351*(19), 2021-2022.

Thompson, T., Lee, A. R., & Grace, T. (2010). Gluten contamination of grains, seeds, and flours in the United States: a pilot study. *Journal of the American Dietetic Association, 110*(6), 937-940.

Valdés, I., García, E., Llorente, M., & Méndez, E. (2003). Innovative approach to low-level gluten determination in foods using a novel sandwich enzyme-linked immunosorbent assay protocol. *European journal of gastroenterology & hepatology, 15*(5), 465-747.

Van de Voort, F., Sedman, J., Emo, G., & Ismail, A. (1992a). A rapid FTIR quality control method for fat and moisture determination in butter. *Food Research International, 25*(3), 193-198.

Van De Voort, F. R., Sedman, J., Emo, G., & Ismail, A. A. (1992b). Assessment of Fourier transform infrared analysis of milk. *Journal of AOAC International, 75*(5), 780-785.

Varjonen, E., Vainio, E., & Kalimo, K. (2000). Antigliadin IgE–indicator of wheat allergy in atopic dermatitis. *Allergy*, *55*(4), 386-391.

Varmuza, K., & Filzmoser, P. (2016). *Introduction to multivariate statistical analysis in chemometrics*: CRC press.

Waga, J. (2004). Structure and allergenicity of wheat gluten proteins-a review. *Polish journal of food and nutrition sciences*, *13*(4), 327-338.

Wang, N., Zhang, X., Yu, Z., Li, G., & Zhou, B. (2014). Quantitative analysis of adulterations in oat flour by FT-NIR spectroscopy, incomplete unbalanced randomized block design, and partial least squares. *Journal of Analytical Methods in Chemistry, 2014*.

Wu, W., Walczak, B., Massart, D., Prebble, K., & Last, I. (1995). Spectral transformation and wavelength selection in near-infrared spectra classification. *Analytica Chimica Acta*, *315*(3), 243-255.

Xu, L., Cai, C.-B., Cui, H.-F., Ye, Z.-H., & Yu, X.-P. (2012). Rapid discrimination of pork in Halal and non-Halal Chinese ham sausages by Fourier transform infrared (FTIR) spectroscopy and chemometrics. *Meat Science*, *92*(4), 506-510.

Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. *Applied artificial intelligence, 17*(5-6), 375-381.

Zhao, X., Wang, W., Ni, X., Chu, X., Li, Y.-F., & Lu, C. (2019). Utilising near-infrared hyperspectral imaging to detect low-level peanut powder contamination of whole wheat flour. *Biosystems engineering, 184*, 55-68. doi:10.1016/j.biosystemseng.2019.06.010

VITA

Abuchi Godswill Okeke

PLACE OF BIRTH

Onitsha, Anambra State, Nigeria

EDUCATION

B.Sc. Agricultural and Environmental Engineering, University of Ibadan, Ibadan, Nigeria, February 2018

PROFESSIONAL EXPERIENCE

Graduate Research Assistant, Department of Biosystems and Agricultural Engineering, University of Kentucky, Lexington, Kentucky. January 2019 – Present. Advisor: Dr. Akinbode Adedeji.

Software Engineer, Teamapt Limited, Lagos, Nigeria. Feburary 2018 – December 2018

PROFESSIONAL SOCIETIES

American Society of Agricultural and Biological Engineers

National Society of Black Engineers

Minorities in Agriculture Natural Resources and Related Sciences

HONORARY SOCIETIES

Alpha Epsilon Engineering Honor Society

Omicron Delta Kappa Honor Society