DECODING TOFU QUALITY: AN INTEGRATIVE INVESTIGATION OF SOYBEAN SEED CHARACTERISTICS AND INNOVATIVE EVALUATION APPROACHES

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Amanda Malik

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Program:
Cereal Science

November 2023

Fargo, North Dakota

North Dakota State University

Graduate School

Title

DECODING TOFU QUALITY: AN INTEGRATIVE INVESTIGATION OF SOYBEAN SEED CHARACTERISTICS AND INNOVATIVE EVALUATION APPROACHES

By

Amanda Malik

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Minwei Xu

Chair

Bingcan Chen

Shahidul Islam

Changhui Yan

Approved:

November 21st, 23 — Richard Horsley

Date — Department Chair

ABSTRACT

This research unravels the intricate relationship between soybean seed characteristics, sources, and the resultant quality parameters of tofu. The study analyzed 178 soybean varieties from diverse sources, categorizing them into distinct clusters. Significant variations emerged, with soybeans from the United States exhibiting higher protein, while Chinese sources displayed higher moisture content.

Subsequently, the research delved into diverse tofu quality parameters using multivariate analysis. Distinct clusters were identified based on attributes including yield, texture, moisture content, and brix levels. These parameters exhibited complex interrelationships, providing insights into factors defining tofu sensory qualities. Furthermore, an innovative integration of Hyperspectral Imaging and machine learning accurately predicted tofu quality categories from soybean seeds with 96-99% precision.

The research underscores the multifaceted nature of factors influencing tofu quality, considering seed origin and composition. The pioneering use of advanced technologies sets the foundation for enhanced quality evaluation and improved production practices in the tofu industry.

ACKNOWLEDGMENTS

I would like to extend my heartfelt gratitude to several individuals who have played pivotal roles in my academic journey. First and foremost, I am thankful to Yahweh for granting me the opportunity to embark on this path of higher education.

I would like to express my deepest appreciation to my esteemed advisor, Dr. Minwei Xu, for his unwavering support and guidance throughout my graduate studies. His patience, mentorship, and encouragement have been invaluable, allowing me to become an independent researcher. Dr. Xu's dedication to answering my questions and providing counsel whenever needed have been instrumental in my academic growth. I consider myself fortunate to have had the privilege of working under his tutelage; he is undoubtedly one of the best professors one could hope for.

I am also indebted to my other esteemed graduate committee members: Dr. Bingcan Chen, Dr. Shahidul Islam, and Dr. Changhui Yan. Their patient guidance and commitment to my success have been deeply appreciated.

I extend my sincere thanks to all those who provided technical assistance during my academic journey. Special thanks to Allen Peckrul for his generous help with numerous laboratory inquiries and to Dr. Jae Ohm for his invaluable instructions and allowing me to use various instruments. I also want to express my gratitude to Dr. Frank Manthey, Dr. Zhao Jin, Dr. Anuradha Vegi, Dr. Jiajia Rao, and Dr. Kalidas Shetty for imparting their invaluable knowledge in the field of cereal science.

To my cherished family and friends who stood by me during the most challenging times, words cannot adequately convey my appreciation. I am immensely thankful to Liuyi Chang, Atoosa Bigvand, Parag Chaudhari and Julianna Miller for their countless invitations and unwavering support in my daily life, from grocery shopping to exploring Fargo and other cities. My heartfelt thanks go to my classmates, who provided companionship throughout the demanding semesters.

To my beloved brother, Moses Malik, whom I hold dear in my heart, thank you for your unending encouragement and unwavering belief in me.

Lastly, I want to express my profound gratitude to my parents, Manoj Kumar Malik and Reni Manoj Malik. Your boundless love and support in all my endeavors have been my greatest source of strength and motivation.

DEDICATION

I humbly dedicate this work to the most important figures in my life:

To my God, Yahweh, your divine guidance, and unwavering support have been the driving force behind my journey in graduate research. Without your grace, I would not have achieved this milestone. Thank you for blessing my life.

To my beloved brother, Moses Malik your ability to bring a smile to my face is a precious gift. Even in my most stressful moments, the thought of your bright future fills me with joy and helps me overcome all challenges.

To my father, Manoj K Malik, your sage counsel has been a constant source of wisdom, keeping me grounded throughout this journey. Your presence in my life is a reminder of the love that surrounds me.

To my dear mother, Dr. Reni Manoj Malik, your dreams have become my aspirations, and I am committed to turning them into reality. I live to fulfill the dreams you hold in your heart, and I cherish the unwavering support you have provided me.

TABLE OF CONTENTS

ABSTRACT iii

ACKNOWLEDGMENTS iv

DEDICATION vi

LIST OF TABLES x

LIST OF FIGURES xi

1. GENERAL INTRODUCTION 1

  1.1. Introduction 1

2. LITERATURE REVIEW 4

  2.1. Soybean 4

    2.1.1. Soybean foods 5

    2.1.2. Tofu process 6

    2.1.3. Evaluation of tofu quality 12

    2.1.4. Chemical composition of soybean seeds affecting tofu quality 13

  2.2. Hyperspectral imaging (HSI) 17

  2.3. Machine learning is a powerful tool on image processing 19

    2.3.1. Imaging processing 19

    2.3.2. Machine learning 20

3. UNRAVELING THE IMPACT OF DIFFERENT SOURCES ON SOYBEAN SEED ATTRIBUTES AND ASSOCIATED TOFU PROPERTIES 24

  3.1. Abstract 24

  3.2. Introduction 24

  3.3. Materials and methods 26

    3.3.1. Seeds and materials 26

    3.3.2. Water uptake capacity of soybean seeds 26

    3.3.3. Preparation of tofu 27

3.3.4. Chemical composition ... 28

3.3.5. Evaluation of tofu texture ... 28

3.3.6. Clustering of soybean seeds ... 28

3.3.7. Principal component analysis ... 29

3.3.8. Statistical analysis ... 29

3.4. Results and discussion ... 30

3.4.1. Sources of soybean seeds ... 30

3.4.2. Characteristics of soybean seeds collected from different sources in the United States and China ... 31

3.4.3. Characteristics of tofu collected from different sources in the United States and China ... 35

3.4.4. Application of principal component analysis to assess the impact of seed origin on tofu quality parameters ... 50

3.4.5. Correlation coefficients between soybean seed characteristics and tofu quality parameters derived from soybean seeds from United States and China ... 52

3.5. Conclusion ... 53

4. PREDICTING TOFU QUALITY FROM SOYBEAN SEEDS USING HYPERSPECTRAL IMAGING AND MACHINE LEARNING ... 55

4.1. Abstract ... 55

4.2. Introduction ... 55

4.3. Materials and methods ... 58

4.3.1. Seeds and materials ... 58

4.3.2. Water uptake capacity of soybean seeds ... 58

4.3.3. Preparation of tofu ... 58

4.3.4. Evaluation of tofu texture ... 59

4.3.5. Clustering of soybean seeds ... 59

4.3.6. Hyperspectral scanning of soybean seeds ... 59

4.3.7. Hyperspectral imaging (HSI) image processing 61

4.3.8. Feature selection of HSI 61

4.3.9. Model establishment using convolutional neural network (CNN)............................ 62

4.3.10. External validation of predictive machine learning model 63

4.3.11. Statistical analysis 64

4.4. Results and discussion 64

4.4.1. Clustering of soybean based on tofu quality 64

4.4.2. Hyperspectral imaging (HSI) of soybean seeds 69

4.4.3. Predicting tofu quality based on HSI with CNN model... 74

4.5. Conclusions 78

5. OVERALL SUMMARY AND CONCLUSION.. 80

5.1. Conclusion 80

5.2. Future research 81

REFERENCES 83

LIST OF TABLES

Table Page

3.1. Origins of soybean seeds for comparative evaluation of seeds from the United States and China 30

3.2. Evaluation of soybean seed protein and moisture across various soybean classes 34

3.3. Evaluation of tofu quality across various soybean classes 46

4.1. Evaluation of tofu and soybean seed quality across various soybean classes. 67

4.2. Performance of featured wavelength selected by different models. 72

4.3. Featured wavelengths and the corresponding bonds. 74

4.4. Parameters of CNN architecture used in this work 76

4.5. Prediction of tofu quality with CNN based on 10 featured wavelengths 76

LIST OF FIGURES

Figure Page

2.1. World soybean production quantity, area harvested, and yield: 1971–2020, Source: FAOSTAT : https://www.fao.org/faostat/en/#data/QCL/visualize 4

2.2. A flowchart for tofu produced using a multistep process with their nutrients and physiological benefits 12

2.3. Schematic representation of hyperspectral imaging (HSI) hypercube showing the relationship between spectral and spatial dimensions, Adapted from Gowen et al., (2007) 18

3.1. Clustering of soybean seeds considering their protein and moisture based on their sources using hierarchical clustering analysis (HCA). The color of Cluster i, ii, iii, iv, v and vi are indicated with red, green, blue, orange, dark green and purple respectively 35

3.2. Clustering of tofu quality based on their sources using hierarchical clustering analysis (HCA). The color of Cluster 1, 2, 3, 4, 5 and 6 are indicated with red, green, blue, orange, dark green and purple respectively.. 49

3.3. Clustering of soybean seeds based on tofu quality and characteristics of tofu from United States and China using principal component analysis (PCA)............................... 51

3.4. Pearson’s correlation analysis of soybean seed characteristics and tofu quality across diverse clusters... 53

4.1. The hyperspectral imaging (HSI) system 60

4.2. Clustering of soybean seeds based on tofu quality using (a) hierarchical clustering analysis (HCA) and (b) principal component analysis (PCA) and the loading score of each component. The color of Class I, II, III, and IV are indicated with red, yellow, green, and blue, respectively. 66

4.3. Hyperspectral imaging (HSI) profile of soybean seeds at the spectral range spanned from 900 to 1700 nm. (a) the HSI wavelength profile of all the soybeans; (b) the HSI wavelength profile of classified soybeans; (c) images of soybeans at ten featured wavelengths. The 10 featured wavelengths represented by the image planes were acquired by XGBoost with the feature importance listed. 70

4.4. Verification of soybean seeds with tofu quality (a) tofu yield, (b) water uptaking capacity, (c) firmness, (d) springiness, and (e) cohesiveness. Note: The circle symbol indicated the mean value of tested tofu quality. The line of each box from top to bottom indicates upper whisker, upper quartile, median, lower quartile, and lower whisker. The black dots indicate the parameter values in the training dataset. Different letters indicate statistically significant differences ($p < 0.05$)............. 77

1. GENERAL INTRODUCTION

1.1. Introduction

At present, soybeans are considered one of the most important food crops grown worldwide. While they were first domesticated in Asia and have been cultivated in that region for thousands of years, the Americas (North and South) are now home to seven of the top ten soybean growers (Colletti et al., 2020). In recent decades, soybeans have gained global popularity and have been incorporated into regional cuisines. Additionally, the use of soybeans in various human food products has increased.

Soybeans are processed into a range of products, including soymilk and tofu, which have become popular due to their nutritional benefits and versatility in cooking. Tofu, a soy milk-based food, has become a staple in many cuisines worldwide. It is a rich source of protein and essential nutrients, making it a popular choice for vegetarians and individuals seeking plant-based protein alternatives. The production of tofu involves several processing steps, including cleaning, soaking, grinding, coagulation, and pressing, which contribute to its texture and flavor (T. Cai & Chang, 1999; Hou & Chang, 2004)

The quality of tofu is evaluated through tofu processing, which can be labor-intensive and costly. However, there are current problems in tofu quality evaluation, including time-consuming processes and the lack of accuracy in the results (Hou & Chang, 2004; Hui & Xing, 2022; Poysa et al., 2006; Stanojevic et al., 2011; Zhu et al., 2016)

To address the current problems in tofu quality evaluation, researchers have been exploring new methods and techniques. For example, the use of transglutaminase pre-crosslinking treatment has been investigated to improve the physicochemical and digestive properties of tofu (Hui & Xing, 2022). Bench-scale tofu production methods have also been developed to analyze tofu-related traits and their relationship with agronomic traits in soybeans (Kurasch et al., 2018). Rekha and Vijayalakshmi, (2013) studied the influence of processing parameters on the quality of tofu. By using imaging techniques, such as thermal imaging or multispectral imaging, the texture and quality of tofu can be assessed during different stages of processing, such as coagulation, stirring, and molding. This real-time monitoring can help identify any deviations or inconsistencies in the process, allowing for timely adjustments and improvements.

Understanding the relationship between cultivation and tofu quality could help predict the tofu quality. The quality of tofu is influenced by factors such as protein, oil content, and protein composition (Kim et al., 2008). Different varieties of soybeans can affect these quality parameters. The cultivating environment of soybean seeds also plays a role in the quality of soybeans and tofu. While soybeans were initially domesticated in Asia, the Americas have emerged as major soybean producers. The sources of soybeans can impact their protein composition, lipid content, and overall quality (James & Yang, 2016; J. Zhang et al., 2017) and further affect the tofu quality.

Non-destructive technologies such as hyperspectral imaging and machine learning can also help efficiently predict the quality of tofu. Hyperspectral imaging combines spectroscopy with imaging to extract spectral and spatial information from an object, enabling non-destructive analysis of soybeans and tofu (Jurado et al., 2021). Machine learning algorithms, including convolutional neural networks, extreme gradient boost, and k-nearest neighbors, can be utilized to analyze hyperspectral data and improve the accuracy of quality evaluation (N. Parsa & Byrne, 2021; L. Qiao, 2022). By integrating hyperspectral imaging and machine learning techniques, researchers can gain a deeper understanding of the composition, texture, and overall quality of soybeans and tofu. This knowledge can contribute to the development of improved processing methods, quality control measures, and product innovations in the soybean and tofu industry.

In conclusion, soybeans and tofu hold substantial recognition worldwide as pivotal food resources, with soybean cultivation prevalent across diverse global regions. Tofu, a soy milk-based food, has garnered significant acclaim owing to its nutritional richness and adaptability in various culinary contexts. The quality of tofu is subject to the interplay of elements like protein content, water absorption capabilities, and texture, which are susceptible to variations arising from sources and distinct processing methodologies. Leveraging cutting-edge technologies such as hyperspectral imaging and machine learning presents a promising trajectory for bolstering the precision of soybean and tofu quality evaluation. Such advancements are poised to foster the refinement of production practices and drive innovation within the industry

2. LITERATURE REVIEW

2.1. Soybean

Soybean, a member of the Leguminosae family, is one of the most significant industrial plants of the world. Soybeans provide 1/3 of edible oils and 2/3 of protein sources. It was found in northern China 5,000 years ago. Up until the 1950s, the top producer of the globe for soybeans was China. But after this time, the United States rose to become the world leader in the production of soybeans. There has been a progressive growth in the global output of soybeans produced organically (**Figure 2.1**). This is mostly due to a rise in demand for organic soybean products, such as vegetable oil, soybean milk, edamame, and soybean tofu, as well as soybean meal for agriculture-based livestock feed (Hartman et al., 2016).

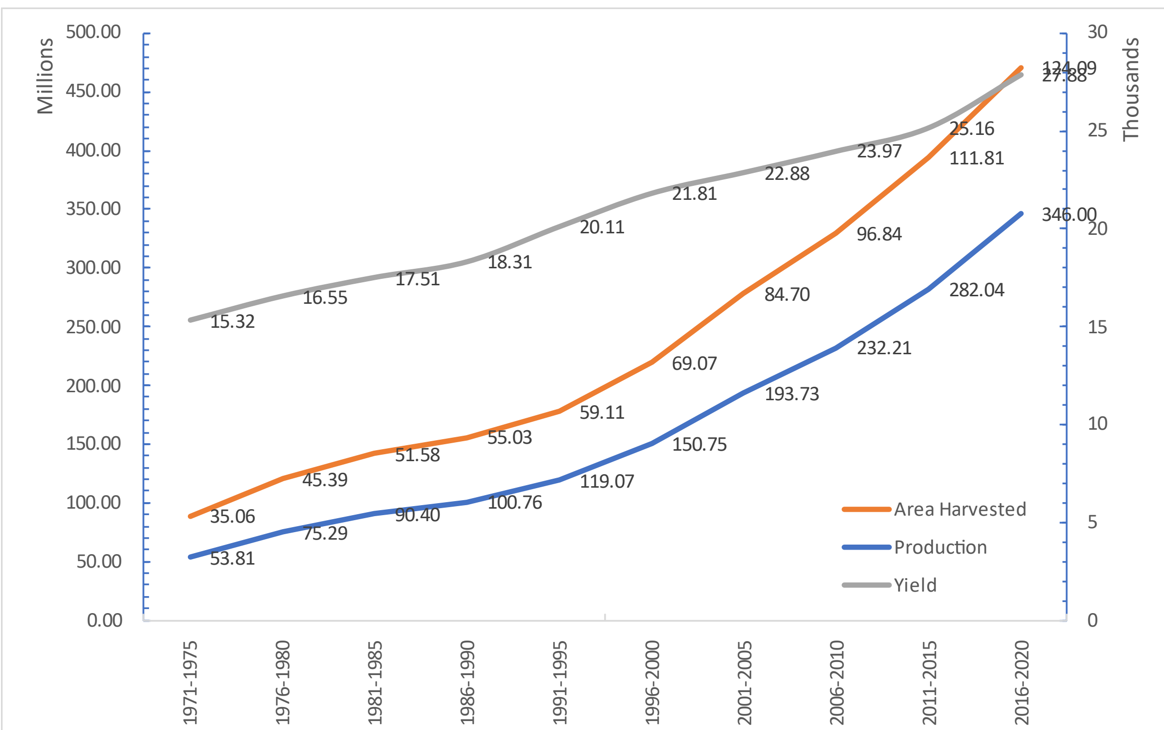


Figure 2.1 World soybean production quantity, area harvested, and yield: 1971–2020, Source: FAOSTAT : https://www.fao.org/faostat/en/#data/QCL/visualize

2.1.1. Soybean foods

At the moment, soybeans are considered to be one of the most important food crops grown all over the world. Even though they were first domesticated in Asia and have been produced in that region for millennia, seven of the top ten growers are today situated in the Americas (North and South) (Colletti et al., 2020). In the last few decades, not only have these foods achieved global popularity and been adapted to regional cuisines, but the use of soybeans in other human food products has also increased. According to the FDA, consuming 25 grams of soy protein per day reduces the risk of cardiovascular disease, which may have contributed to an increase in the consumption of soy-based foods.

Food-grade soybean cultivars contain greater protein (40 to 45% dry content) and less oil (18 to 20% dry matter). Increased protein levels increase product yield. Soy protein-based foods play a great role in modern food market. Daily foods, such as infant formulas, soy meat alternatives, canned foods, breakfast bars, breakfast cereals, baked goods, snack foods, and sports meals and drinks, can be derived from soy protein isolates. Food-grade soybean cultivars have high levels of sucrose or low levels of lipoxygenase as well. Higher sugar levels speed fermentation and boost manufacturing production.

Soybeans are processed into soymilk in liquid form. Soymilk is utilized in the production of soy drinks, yoghurts, cheeses, whipped toppings, and liquid or powdered nondairy frozen desserts. A study on soymilk found that newborns allergic to animal milk can choose soymilk to acquire the nutrients they need early in life (Muraro et al., 2002). Soymilk consumption in the US is growing due to recent research suggesting anti-cancer effects, the prospect that these products could serve as dairy substitutes for people with lactose intolerance, and as cholesterol-free milk/meat substitutes for people with cardiovascular disease.

Tofu is another important soy milk-based food. Tofu was invented by Liu An firstly recorded in the book Huainan Ztu in 139BC. In Indonesia, tofu is a staple food that makes up a significant portion of the diet of the people and can be found everywhere. Tofu has several essential components, including proteins and oils, to name just two of them (F. Wang et al., 2020). Tofu made from soybeans contains over 70% primary storage proteins (i.e., glycinin -11S and β-conglycinin -7S) (Joo & Cavender, 2020). Historically, Western interest in soy products like tofu and soymilk was primarily reserved for vegetarians. Tofu is a source of protein that is both inexpensive and simple to digest. It is an excellent provider of various micronutrients, and it has a high protein content. The global market for tofu is expected to reach USD 3527.1 million in 2028 with a 5.2% CAGR (Anjum et al. 2023).

2.1.2. Tofu process

Tofu production, texture, and flavor are controlled by processing methods and scenarios. Overall, cleaning, soaking, wet grinding, filtering, boiling, coagulation, and pressing are typical steps in the tofu process (**Figure. 2.2**).

2.1.2.1. Cleaning

Like all other grains, cleaning soybeans is the first technical phase in the production of tofu. This operation is intended to eliminate undesirable fractions such as loose husks, straw particles, weed seeds, and foreign items like sand, stones, metal particles, sticks, and dust. Cleaning is necessary to produce high-quality finished products and the preservation of processing equipment. Sieves with air suction are used to remove dust, plant tissue, pebbles, and other light impurities from soybeans, in addition to bigger contaminants (e.g., stones, stems, nails). Typically, destoners and magnetic iron separators are used to separate large contaminants.

Cleaned soybeans are weighed using automatic hopper scales to monitor the rate of feed and account for the overall number of raw materials (Riaz, 2006).

2.1.2.2. Soaking

Soaking soybeans changes their structure and grinding properties, which is an important step in the preparation of tofu. Soybean seeds absorb water from the environment during soaking. Soaking raw soybeans enhances protein extraction and produces tofu with an increased protein content (James & Yang, 2016). The rate of water absorption is accelerated at higher temperatures. In addition, grinding quality and soaking conditions are connected exclusively with the ultimate water content of the soybeans (Pan & Tangratanavalee, 2003). Soaking lowers the solid content of soybean seeds, making them more processable and so affecting the production and protein content of tofu. This solid loss may diminish the number of solids in soymilk, hence altering the texture of tofu. Choosing the proper soaking conditions, including water-to-soybean-seed ratio, duration, and temperature, to prepare tofu is crucial. (T. D. Cai et al., 1997) reported that soybeans steeped for 14 hours at 20ºC produced the most protein-rich soymilk and firmest tofu.

2.1.2.3. Grinding

Grinding is used to create a slurry from soaked soybean. Soybeans are turned into raw soymilk after being mashed, and an emulsion is generated because oil and water are mixed with dissolved protein as the emulsifier (Guan et al., 2021). Certain processing procedures, such as cooling and heating are combined with grinding to modify the particle size of soymilk and improve the flavor attributes (Q. Zhang et al., 2017).

- **Cool grind method:** Traditional soymilk production in China and Japan requires grinding soaked and rinsed soybeans with water in a stone mill while maintaining a cool grinding temperature. Using this method will result in soymilk with an excellent texture and a high yield. However, soymilk has a strong rancid-oil-like odor, which is particularly irritating when soymilk is consumed as a cold beverage as opposed to a hot beverage.

- **Hot grind method:** Typically, ground soybean slurry is produced by combining ground soybeans (with or without hulls, soaked or dry) with almost boiling water and, in some instances, steam injection. In many hot-grinding processes, sodium bicarbonate or caustic soda is added to the water to drastically modify its pH and convert it to an alkaline condition. High temperatures and pH levels, both of which inactivate the lipoxygenase enzyme, considerably mitigate the rancid oil-like flavor of soymilk. After soymilk has been extracted, its alkalinity is neutralized by adding hydrochloric acid or another acid. The grainy texture of soymilk produced by this method is owing to the adverse effect of heat on the capacity of the protein to dissolve, but the rancid oil-like flavor of soymilk is substantially diminished (Yadav et al., 2003)
- **Hot-Blanch method:** This is an improvement over the hot-grind method, which involves blanching the beans in boiling water or an alkaline solution for enough time to completely inhibit the activity of the lipoxygenase enzyme. This hot-blanch technique not only fully eliminates the rancid-oil-like flavor, but it also imparts a roasted nut flavor to the product and renders the protein practically insoluble in water. Commonly in a colloid mill, blanched soybeans are finely crushed in water, and the resulting fine slurry is homogenized under high pressure. The resulting soymilk is a suspension of small soy particles suspended

in water. Despite having a delicious taste, the soymilk has a very coarse texture. Soymilk is neutralized with an acid to obtain a pH range of 6.7 to 7.2 when using an alkaline solution.

2.1.2.4. Filtering

Filtering is the step that separates soymilk from insoluble soy residues (okara) in the slurry. The insoluble soy residues are removed from the soy slurry by a decanting centrifuge to improve flavor and mouth feel to reduce the oligosaccharides (Lusas et al., 1989). Soybean slurry has been treated to produce varied qualities of soymilk. Plate-and-frame filters, laboratory cheesecloth filters, and various centrifugation processes are utilized to separate soymilk from insoluble residue. There are frequently observed changes in both the filter pressure and the centrifugal force.

2.1.2.5. Boiling

Boiling occurs when a liquid is heated to its boiling point, which is at 100°C at atmospheric pressure. The boiling procedure denatures the soymilk protein and exposes the hydrophobic groups. A soy protein coagulated without heating (no polypeptide chain unfolding) has a globular form result in a softer texture of the gel than heated soybean protein (Saio et al., 1968). In addition to the texture, boiling can also kill food pathogens, denature lipoxygenase and other anti-nutritional components, like trypsin inhibitors and lectin. Soymilk is heated by convective heat transfer at atmospheric pressure in conventional soy food processing. Traditional cooking causes soymilk to heat unevenly in industrial manufacturing. By utilizing pressure cooking and increasing the cooking temperature above the boiling point of water, high-temperature pressure cooking (HTPC) increases the efficiency of heat transmission. After 10 minutes at 115°C, sensory assessment indicated that the texture of HTPC soymilk was smooth and creamy; thus, customer acceptance of HTPC soymilk was higher than that of conventional commodities (Guan et al., 2021).

2.1.2.6. Coagulation

The use of a coagulant is a vital step in the production of tofu, as it causes the soy protein to form a gel network structure that is reflected macroscopically in the coagulation of tofu. Heat processed soymilk creates hydrophobic groups, which aggregates soybean storage proteins (Kohyama & Nishinari, 1993; Peng et al., 2016). Choosing a coagulant to provide salt ions and pH during tofu making is critical. Salt, acid, enzymatic, and other coagulants are examples. These chemicals serve as coagulants and stimulate protein network development when creating tofu. Different coagulants can yield tofu with different eating characteristics. Calcium sulphate and magnesium chloride are often used as salt coagulants. People are most familiar with and favorable toward the flavor and aroma of this type of tofu products (Q. Zhang et al., 2017). However, to make tofu with salt coagulants need skillful labor. The remaining undissolved calcium sulphate in tofu or a rapid release of magnesium chloride can result in a gritty texture, which is an indicator of poor quality (M. Li et al., 2015; Ting et al., 2009). Acid coagulants have been widely researched as an acidifying agent to produce tofu with a homogeneous network (Bi et al., 2013; Chang et al., 2014).

Acid coagulants, include Glucono-δ-lactone (GDL), lactic acid, physalis, succinic acid, acetic acid, malic acid, citric acid, and tartaric acid, are well employed in tofu process, with GDL being the most utilized. Acid coagulants promote the isoelectric precipitation of protein because they release hydrogen ions that reduce the pH of soymilk to the isoelectric point of soy protein (Guan et al., 2021).

Coagulant enzymes, which are abundant in animal and plant tissues in addition to microbes, have great potential as soybean curd coagulants. Transglutaminase (TGase), acalase, pepsin, bromelain, and papain are now the most researched enzyme coagulants.

It hypothesized that most of the effect of TGase on the increase in the strength of tofu is due to 7S and 11S protein. The α’ and α subunits in 7S and the A3 peptide chain in 11S have the largest effect on the activity of TGase in this process, followed by the β subunit in 7S and the A peptide chain in 11S. They examined the amino acid composition of these subunits and peptide chains and observed that TGase activity in soy protein is closely associated with lysine (Guan et al., 2021).

2.1.2.7. Pressing

Tofu is classified into several market varieties based on product hardness. Firm tofu includes a pressing stage. An example of pressing tofu is to press the coagulated soymilk with 500 g initial weight for 15 minutes, followed by 1,000 g for 15 min. The whey of the tofu drains during the pressing, leaving the tofu cake (Rekha & Vijayalakshmi, 2013).

The protein content of tofu and its composition are influenced by the time and pressure of bean curd compression. According to T. Cai & Chang (1999) the protein content of tofu increased when compressed with little force over a shorter period. This is related to the carbohydrate elimination that occurs during pressing processes. The pressed tofu has a layer of skin that is harder than the interior. Therefore, the location from where test samples are extracted from a tofu cake has a direct impact on the textural profile (Yuan & Chang, 2007).

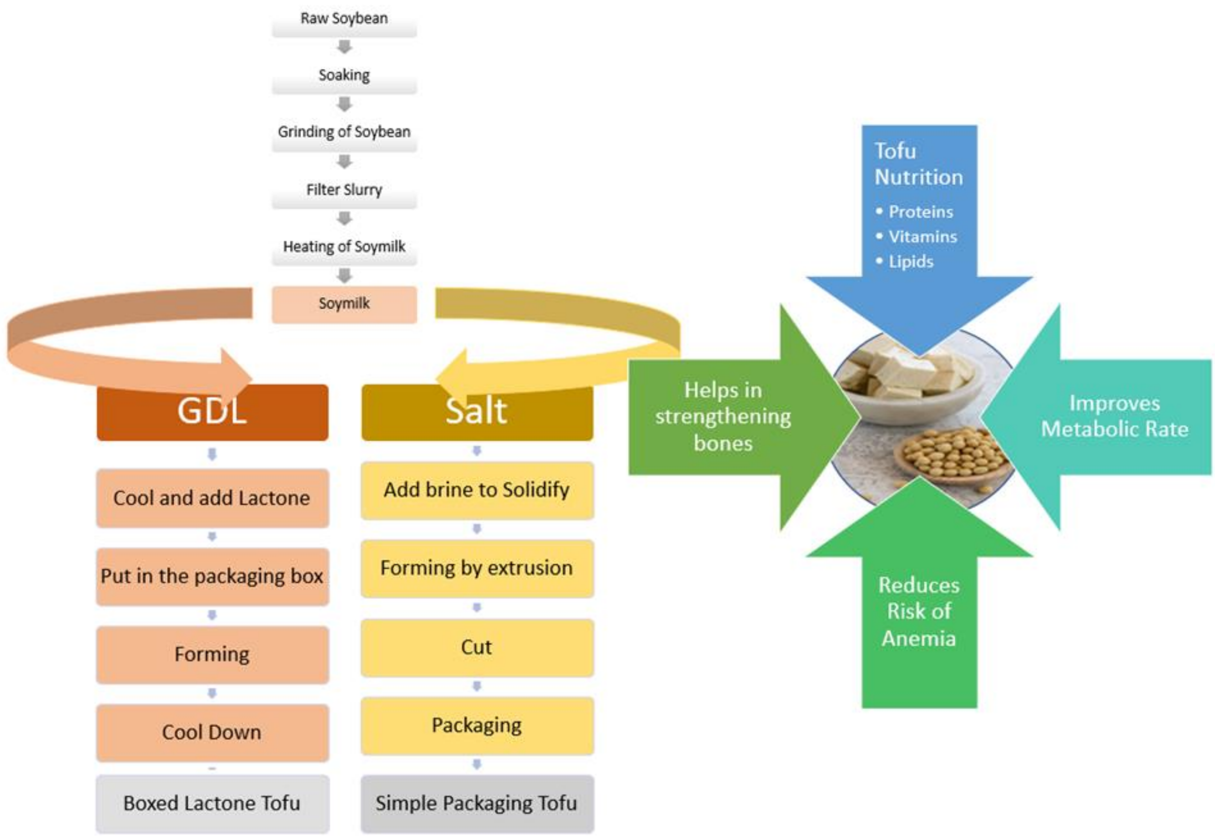


Figure 2.2 A flowchart for tofu produced using a multistep process with their nutrients and physiological benefits

2.1.3. Evaluation of tofu quality

Tofu quality can be evaluated based on its yield, protein, and texture (T. D. Cai et al., 1997). Tofu processors want to get a product with high protein content and a high yield of tofu. In addition, tofu should have a smooth and silky texture, with no gritty or grainy bits. Tofu acceptance is determined by yield and texture, and tofu manufacturers are primarily concerned with these two factors (Rekha & Vijayalakshmi, 2013).

A high yield is important in tofu production because it means that more tofu can be produced from the same number of soybeans, resulting in lower production costs and greater efficiency, as well as to have a better quality of tofu. The fresh tofu yield ranged from 4.45 to

5.26 kg/kg of soybeans. Due to its economic significance, the difference between the maximum and minimum tofu yield influences large tofu manufacturing facilities (Lim et al., 1990).

Texture is an important factor in determining the quality of tofu because it affects the overall eating experience. Fracturability, firmness, springiness, and elasticity are all physical parameters that can be used to evaluate the quality of tofu. Fracturability refers to the ability of tofu to break or fracture cleanly when cut or pressed. High-quality tofu should have a good fracturability, meaning it should break cleanly with minimal crumbing. Firmness refers to the resistance to deformation when pressed. Firm tofu has a higher resistance to deformation, while soft tofu will have less resistance. Springiness refers to the ability of tofu to return to its original shape after being pressed or deformed. High-quality tofu should have a good springiness, meaning it should return quickly to its original shape. Elasticity refers to the ability of tofu to stretch or bend without breaking. High-quality tofu should have a good elasticity, meaning it should be able to stretch and bend without breaking. All these parameters are related to the protein content and water content of the tofu, and how the tofu was processed. These parameters are used to give a general idea of the quality of the tofu and how it will behave when cooked (Schaefer & Love, 1992).

2.1.4. Chemical composition of soybean seeds affecting tofu quality

Chemical composition of soybean seeds plays key roles regarding tofu quality. The soybean seeds have approximately 40.3% protein, 21.0% fat, 4.9% ash and 33.9% carbohydrates (Perkins, 1995). In general, a higher protein content results in a larger production and an improved overall quality of the tofu while the lipid and carbohydrate contents inversely related to the textural characteristics of tofu (James & Yang, 2016; Zhang et al., 2017).

2.1.4.1. Protein

Soybeans have different types of proteins. The main ones are called globulins and albumins, with globulins making up the majority (around 70%) of soy protein. Within globulins, there are four types, and two of them, known as glycinin and β-conglycinin, are the most important. Glycinin is like a cluster of six smaller proteins joined together, and it has a molecular weight of 300–380 kDa. β-conglycinin, on the other hand, is made up of three parts with a molecular weight of 150–200 kDa (Liu, 1997; W. Wang et al., 2008).

The amount of glycinin in soybean protein affects how firm tofu turns out, while the amount of β-conglycinin determines its springiness. When we make tofu, heating and coagulation make glycinin form a stable structure because of the way its parts interact with each other. β-conglycinin does not form as strong a structure because it relies on a different kind of interaction called hydrogen bonding. So, tofu made with more glycinin generally has better texture.

To make a gel with glycinin, you need a higher concentration of it (around 1.03%) compared to β-conglycinin (about 0.479%) under the same conditions (Zhao et al., 2017). Different types of soybeans have varying amounts of these proteins, and the ratio between 11S and 7S (β-conglycinin and glycinin) can determine the right soybeans for making tofu.

Moreover, tofu has pretty much the same amino acids as soybeans (Roger Wang & Kow-Ching Chang, 1995). These amino acids include things like aspartic acid, glutamic acid, leucine, lysine, methionine, phenylalanine, threonine, tryptophan, and valine. Soybeans are rich in lysine and methionine, which are important in plant-based proteins. This makes soybeans a good source of "complete" protein because they have all the essential amino acids human bodies need (Zarkadas et al., 2007).

Around 16.9% to 17.5% of the total protein in tofu is made up of basic amino acids. These amino acids help create a stable structure in tofu through different types of bonds, like hydrogen bonds and ionic bonds. Methionine, a specific amino acid, makes up around 1.98% to 2.10% of the total protein in tofu.

Basic amino acids can also make bonds with calcium or magnesium ions, which further helps tofu keep its structure. These factors have a significant impact to the texture and quality of tofu.

2.1.4.2. Water uptake capacity

Water uptake capacity plays a crucial role in determining the quality of tofu. Several factors can influence the water-uptake capacity of tofu, including the composition of soybean proteins, processing conditions and coagulants used.

It is a pivotal parameter for tofu quality as variety that hold more water are of greater importance to tofu manufacturers because they yield a great volume and weight of tofu from a given quantity of beans (A. Yang & James, 2013).

The protein composition and the processing conditions can affect the water uptake capacity of tofu. It has been reported that the water holding capacity of tofu decreases with increasing coagulant concentrations, regardless of protein composition (A. Yang & James, 2013). The choice of coagulant used in tofu production impacts its water uptake capacity. The concentration of magnesium chloride in water-in-oil emulsions was found to impact the yield, water content, protein content, and hardness of tofu (Zhu et al., 2016). Similarly, the concentration of magnesium chloride in soymilk was shown to affect the consistency and protein content of tofu (Toda et al., 2003). Traditional salt coagulants, such as magnesium chloride (MgCl2), lead to a bitter taste and low water uptake capacity (H. Gao et al., 2021). However, there are studies that have shown that the combination of lactic acid bacteria and salt coagulants can improve the yield and water uptake capacity (Y. Wang et al., 2020). Moreover, the solid content of soymilk affects the water uptake capacity of tofu, with lower solid content resulting in higher water uptake capacity and softer texture (Rekha & Vijayalakshmi, 2013).

Furthermore, the water-to-bean ratio during tofu production can affect both the water uptake capacity and composition of tofu. It has been studied that increasing the amount of drained water during tofu production does not significantly reduce the retention of hydrophilic compounds, such as daidzin and genistin (Kao et al., 2004)

Additionally, the use of ozonated water has been explored as a method to preserve the quality of tofu. Ozone is an antimicrobial agent that is safe to be in contact with food. A study examined the effect of exposure time and replacement of ozonated water on the quality of tofu and found that ozonated water effectively reduced the total mesophilic aerobic bacteria (TMAB) count, maintained pH, and preserved protein levels (Karamah et al., 2021)

Overall, the water uptake capacity of tofu is influenced by various factors, including the protein composition of soybeans, processing conditions, choice of coagulant, and water-to-bean ratio. Understanding and optimizing these factors can help improve the water-holding capacity and overall quality of tofu.

2.1.4.3. Fat

Tofu is regarded as one of the top plant-based protein sources due to the high levels of useful lipids. Palmitic acid (16:0), stearic acid (18:0), oleic acid (18:1), linoleic acid (18:2), and linolenic acid (18:3) are the five fatty acids found in tofu (Y. Guo et al., 2018). Most studies with oleic acid suggest that multi unsaturated fatty acids are useful dietary replacement for lipogenic carbohydrates and saturated fats (Poudyal et al., 2013). However, oxidative deterioration of soybean oil generates rancidity and off flavors that would result in flaws in the texture, color, flavor, and odor of tofu. Due to its effect on the shelf life and quality features of final soy products, lipid oxidation poses a significant challenge for tofu manufacturing. Due to lipid oxidation, texture, color, flavor, and odor are considered production flaws in fatty foods. Lipid oxidation also degrades vitamins and amino acids in stored foods (Elias et al., 2008). Ali et al. (2021) reported tofu with less seed oil, and less total saturated fat is supposedly of higher quality.

2.1.4.4. Carbohydrate

Carbohydrate mainly includes insoluble fiber, soluble fiber, and simple sugar in soybean. Insoluble fiber as well as the protein and oils conjugated with insoluble fiber is removed during the filtration step of tofu process. Therefore, soluble fiber and simple sugar are the major components of carbohydrate. Soybeans contain 11 to 25% soluble carbohydrates, including 15 to 20 distinct sugar species (Obendorf et al., 2008). Sucrose, raffinose, and stachyose are the soluble sugars that are found in the greatest abundance in tofu. The presence of sugar in tofu makes it an appetizing sweet flavor, and sugar is also a desirable component in soybean seeds. Raffinose, stachyose, and verbascose, three galacto-oligosaccharides, are regarded as antinutritional factors since ingesting them causes flatulence and digestive disturbances in people and nonruminant animals (Liying et al., 2010).

2.2. Hyperspectral imaging (HSI)

Hyperspectral imaging (HSI) is a new technique combines spectroscopy with imaging to extract spectral and spatial information from an object. There are three types of image process: push broom scanners, whisk broom scanners, and snapshot hyperspectral imaging. Push broom scanners acquire image data by sweeping a sensor across a scene, typically in a linear or rectangular pattern. Whisk broom scanners use a rotating sensor to acquire image data in a circular or spiral pattern. Snapshot hyperspectral imaging uses a single exposure to acquire a wide range of spectral information from a scene, typically using a sensor with many narrow band filters. Each of these methods have their own advantages and disadvantages and are used for different types of applications (Jurado et al., 2021). All the scanning method will collect and assemble the data into hypercubes. It is a three-dimensional array, with the first two dimensions representing the spatial information (x and y coordinates) and the third dimension representing the spectral information (usually a specific wavelength or band). The hypercube is used for data processing and analysis, such as feature extraction, classification, and target detection. **Figure. 2.3** shows hypercubes and each spatial point of an item under study is made up of hundreds of sequential wavebands.

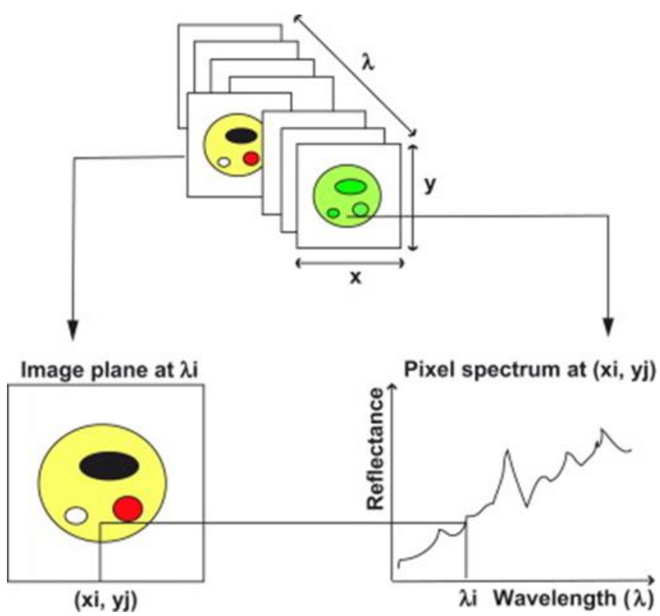


Figure 2.3 Schematic representation of hyperspectral imaging (HSI) hypercube showing the relationship between spectral and spatial dimensions, Adapted from Gowen et al., (2007)

The fundamental idea behind HSI is that chemical and physical information about food is shown by combining spectral and spatial data. The multispectral sensor of the HSI covering the spectrum from the visible to the longwave infrared (300 – 2500 nm). Visible spectrum (300 – 700 nm) is one of the first techniques used for non-destructive testing of food products, enabling rapid detection of mechanical damage and ripeness by extracting sensory data such as texture, color and morphology from RGB photographs of food products (Jain et al., 2014). The internal compositional changes of food products cannot be detected with this technology, though. From the standpoint of chemical composition, it is unable to validate the scientific validity of non-destructive testing methods.

The infrared spectrum, such as near infrared spectrum (700 – 2500 nm), could identify chemical bonds and estimate chemical composition of foods. The interaction of samples with light that might cause vibrational transitions in the molecules is the fundamental idea behind infrared rays. The secondary structural alterations of proteins are frequently studied using infrared spectroscopy. The C = O stretch vibration of the amide group is mostly represented by the amide I frequency bands (1600–1700 nm). The following describes the relationship between each sub-peak and secondary structure: Between 1648 and 1664 nm, -α helix absorbs signals, -sheet absorbs signals between 1615 and 1637 nm, β-turn absorbs signals between 1664 and 1681 nm, and the random coil signal peak is between 1637 and 1648 nm (Liao et al., 2022). The Near Infrared Spectrum is utilized to determine the approximate composition, amino acid profile, and fatty acid profile of grains and seeds such as wheat, soybean, and pea, among others (Osborne, 2006).

2.3. Machine learning is a powerful tool on image processing

2.3.1. Imaging processing

Image processing is required to convert raw data of imaging into pixel data to provide useful information. To retrieve the information, algorithms add value to the raw data. Getting the data is the initial stage in the analysis (Lee et al., 2014). Typically, traditional imaging techniques are based on simple statistical procedures, and experimental error and systematic bias are inescapable in the model (Russ, 2006). Image enhancement technology plays a significant role in image processing. In order to boost clarity and eliminate noise, image enhancement technology modifies a number of the properties of images. For example, the brightness, contrast, saturation, and hue of the image may be changed. Image enhancement is the process of selectively emphasizing the aspects of interest in an image, reducing the appearance of undesirable features, or both, to make the image more consistent with the visual response characteristics. The obtained image may occasionally be dim, low contrast, and noisy. The frequency domain method and the space domain method are two categories under which image enhancement falls. The first method provides signal augmentation based on the two-dimensional Fourier transform and treats the image as a two-dimensional signal. Local averaging and median filtering are two sample algorithms in the latter spatial domain technique (Q. Qiao, 2022).

2.3.2. Machine learning

Machine learning is a form of artificial intelligence in which the model is constructed using training data, sometimes referred to as sample data. Without being specifically trained to do so, this model can make decisions and predictions. Machine learning has recently risen to the top of the list of issues that are most frequently discussed. The digital images are processed using machine learning, and their classification and processing accuracy are further improved.

Image processing has made use of a variety of widely used algorithms. Convolutional neural networks (CNN), which are mostly used for classification and prediction, are a popular machine learning model. The input layer, hidden layer, and output layer are the three components that make up the CNN structure. Convolutional Neural Networks (CNNs) have several advantages when it comes to image processing: 1. When applied for tasks like image segmentation, object detection, and classification, the CNN algorithm can automatically learn and extract relevant characteristics from images. 2, Unlike traditional image processing methods, CNNs can accurately detect objects and features in images without consideration to their relative positioning.; 3, Compared to conventional image processing approaches, CNNs are more resilient since they can manage fluctuations in image size, rotation, and lighting.; 4, CNNs can handle large image datasets with high-dimensional data and can be trained on GPUs and cloud computing resources, which makes them more suitable to process big data Different facial expressions, such as sadness, anger, or enjoyment, can be distinguished using CNN. They are designed to work well under a variety of lighting situations and facial angle variations (Q. Qiao, 2022).

Extreme gradient boost (XGBoost) is a decision tree-based ensemble machine learning technique that makes use of the gradient boosting approach. Boosting is the fundamental concept of this algorithm. The weak decision trees serve as the foundation for numerous categorization models, which use "Gradient descent" to create a so-called stronger model after incorporating the output from the preceding model (M. Parsa, 2021). Some of the advantages of XGBoost for image processing include: 1, XGBoost can handle missing values, which is a common problem in image processing tasks, where parts of the image may be missing or obscured; 2, XGBoost can handle both categorical and numerical data, which makes it well-suited for image processing tasks that involve both types of data; 3, XGBoost is also efficient in terms of computational resources and fast in training and prediction; 4, XGBoost is designed to handle large datasets with high dimensionality, which is common in image processing tasks; 5, XGBoost is known for its high accuracy and ability to handle noisy and missing data, which makes it well-suited for image processing tasks. It is used in fingerprint localization task and in optimization of pollutant concentration (J. Li et al., 2022).

The k-nearest neighbors (k-NN) algorithm is a type of instance-based, or memory-based, supervised machine learning algorithm. It is used for both classification and regression tasks. The basic idea behind the k-NN algorithm is to find the k number of training examples that are closest to a new input, and then use those examples to make a prediction about the new input (G. Guo et al., 2003). The k-NN algorithm has several advantages when it comes to image processing: 1, the k-NN algorithm is simple to implement and understand, making it easy to use for image processing tasks such as classification and object recognition. 2, the k-NN algorithm can be used for a wide range of image processing tasks, including supervised and unsupervised learning, and can be easily adapted to different types of image data; 3, the k-NN algorithm is robust to noise and outliers in the data, making it a good choice for image processing tasks that may involve high levels of noise or unreliable data; 4, k-NN does not require to make assumptions about the underlying probability density function of the data, this is useful when working with image processing where the data could be affected by different lighting conditions, occlusions, etc., 5, K-NN is able to handle missing data and it's able to classify a data point with missing values by using the k nearest neighbors which have non-missing values. The k-NN algorithm is simple and easy to implement, but it can be computationally expensive when dealing with large datasets. It's also sensitive to the choice of the distance metric and k value, and it is not well suited for high-dimensional data. The kNN algorithm can be used for text categorization and for heart disease prediction with simplified health parameters of patients (Suryanegara et al., 2016). Random forest is a machine learning technique used to handle classification and regression issues. Using ensemble learning, a technique that combines several classifiers to solve challenging problems. The advantages of random forests are 1. Random forest is beneficial for working with distinct data subsets and excels with dimensionality. 2. It is highly tolerant to nonlinear data and outliers. Random forest has a way for balancing error in data sets with imbalanced class populations. 3. Each decision tree has a low degree of bias but a great degree of variance. 4. As the trees in random forest are averaged, the variance is also averaged, resulting in a model with low bias and moderate variation. Traditional random forest classifier can be used for image segmentation and can be an easier approach than U-net resulting in multiclass semantic segmentation (Genuer & Poggi, 2020).

3. UNRAVELING THE IMPACT OF DIFFERENT SOURCES ON SOYBEAN SEED ATTRIBUTES AND ASSOCIATED TOFU PROPERTIES

3.1. Abstract

This study investigates the effect of sourcing on soybean seed quality metrics and resultant tofu characteristics. 178 soybean varieties from diverse regions of the US and China were analyzed for protein, moisture, and other attributes. Tofu was produced and tested for yield, texture, and sensory qualities. Multivariate statistical tools elucidated relationships between seed origin and product parameters. The key findings including: 1, sourcing significantly impacted seed protein and moisture contents. US varieties exhibited higher protein while Chinese sources displayed higher moisture; 2, tofu yield, firmness, and gumminess were highest for US-sourced soybeans. Chinese varieties produced tofu with enhanced springiness, cohesiveness, and resilience; 3, Seed protein correlated positively with tofu protein but negatively with tofu yield and moisture. Higher seed sugars improved soy milk and tofu yield; 4, Moisture content strongly influenced tofu texture, with higher moisture decreasing firmness and chewiness. The study provides novel insights into tailoring soybean sourcing and tofu production practices to achieve desired sensory and textural qualities aligned with consumer preferences. It underscores the need to consider seed origin when optimizing tofu attributes.

3.2. Introduction

The assessment of soybean seeds and the quality of tofu is a pivotal aspect of soybean production and processing. Sourcing of soybeans are recognized as significant determinants that wield a considerable influence on seed quality and the subsequent attributes of tofu products. A multitude of factors, encompassing genetic makeup, environmental influences, and intrinsic seed characteristics, collectively shape the quality of soybean seeds, thus bearing a direct impact on the tofu derived from them.

Genetic factors, encapsulated within the genotype of soybean seeds, emerge as a key determinant in the quality spectrum. Extensive studies have revealed the presence of quantitative trait loci (QTL) associated with seed protein and oil concentrations, as well as seed size – all pivotal facets in defining soybean seed quality (Panthee et al., 2005). Furthermore, variations in sucrose and raffinose family oligosaccharides (RFOs) content within soybean seeds have been linked to genotype and growth sources. These intrinsic disparities in genotype and growing sources inherently contribute to the variations in soybean seed quality, consequently reverberating in the quality attributes of tofu derived from them (Kumar et al., 2010).

Environmental factors, among which sources stands prominent, cast their considerable influence on soybean seed quality. Research underscores that the content of sucrose and RFOs within soybean seeds can be markedly influenced by the sources (Kumar et al., 2010). Moreover, studies have delved into the spatial variance in soybean seed quality, highlighting the pronounced effect of sources on seed quality (Müller et al., 2018). A profound understanding of how source impacts soybean seed quality is imperative in the selection of suitable soybean genotypes for optimal tofu production.

Furthermore, it becomes evident that the quality of soybean seeds can fluctuate significantly across diverse regions and climatic conditions. Tropical regions characterized by elevated average temperatures are generally deemed less conducive for preserving soybean seed quality (Coradi et al., 2020). The deleterious impact of high temperatures on the physiological quality of soybean seeds necessitates meticulous evaluation of storage and cooling methodologies to safeguard seed quality. In addition, the presence of regional and temporal variations in soybean seed quality has been substantiated, with localized weather patterns emerging as influential factors (MacMillan & Gulden, 2020). The acknowledgment of these multifaceted variations in seed quality across divergent regions is of paramount importance to guarantee consistent tofu quality.

In summation, the comprehensive evaluation of soybean seed quality and its implications for tofu production mandates an intricate consideration of the sources from which soybeans are obtained. The confluence of factors such as genotype, environmental dynamics, and inherent seed characteristics collectively govern the quality of soybean seeds, thereby casting a profound impact on the resultant tofu quality. The profundity of understanding these multifaceted factors and their intricate interplay is indispensable in the direction of consistently delivering high-quality tofu products. Considering the mass factors at play, our study was carried out to comprehensively evaluate the quality of both soybeans and tofu, with a particular focus on their sources.

3.3. Materials and methods

3.3.1. Seeds and materials

One seventy-eight varieties of soybeans, harvested from North Dakota, Minnesota, California and different provinces in China, were generously provided by the different organizations throughout these countries. Magnesium chloride was purchased from the local market.

3.3.2. Water uptake capacity of soybean seeds

Six hundred grams of soybeans were soaked in 1500g of water for 16 hours at 4 ℃ and afterward, the excess water was drained, and the soaked soybeans were weighed to estimate the water uptake of the beans (Meng et al., 2016).

$$\text{Water uptake} = (W_S - W_D)/W_D \qquad \textbf{Equation (1)}$$

where W_S (kg) indicates the weight of soaked soybeans and W_D (kg) indicates the weight of dry soybeans.

3.3.3. Preparation of tofu

The tofu process was adapted from Meng et al. (2016) with modifications. Briefly, dry soybeans (W_0) were soaked for 16 hours following **Method 3.3.2**. The soaked soybeans were milled into slurries using a grinder machine. The soymilk was collected and weighed after the grinding of the soybeans. 1000 g of water was added to the soymilk to lower the solid content and speed up the boiling process. The soymilk was further mixed in a stained steel contained and placed in a water bath. The soymilk was kept until a temperature of 95^{o}C was reached and after that it was boiled for 5 minutes. The initially collected soy milk was weighed and recorded (W_1). Further, weight 3000g of soymilk and add 28.3 grams of liquid nigari (Magnesium chloride) and mix it properly. The bowl in which the soymilk was kept was covered with a cloth and let to rest for 5 minutes. After that, the curd was broken with the whisk and the mold prepared with the sheet cloth. The curd was poured into the mold and a heavy plate was placed on top of the mold for 5 minutes. With the mold under the tofu presser, the pressure was released to half for 5 minutes and kept for full pressure for 10 minutes. Then the presser was released and tofu cloth as well. The final weight (W_2) of the tofu was recorded and it was let to cool in water for 15 minutes and put for refrigeration.

The formula for calculating tofu yield is:

$$\text{Tofu yield (kg/kg soybean seeds)} = W2 \times (W1/11)/W0 \qquad \textbf{Equation (2)}$$

3.3.4. Chemical composition

Moisture was determined using AOAC official methods (AOAC, 1998). Crude protein content of the ground tofu samples was determined by nitrogen combustion method using a LECO FP428 nitrogen analyzer (LECO Corporation, St Joseph, MN, USA), and calculated with a nitrogen conversion factor of 6.25 (Q. Wang et al., 2022). Whole soybean seeds were scanned using an at-line NIRS analyzer, DA 7250 (PerkinElmer Health Sciences Canada Inc., Winnipeg, Canada). The DA 7250 belongs to the family of diode array spectrometers, and it analyzes several components in samples within 6 s. The wavelength range is from 950 nm to 1650 nm with an interval of 5 nm. Every sample (around 50 g) was scanned thrice in the rotating sample tray and the average spectrum was used for analysis (Hang et al., 2022).

3.3.5. Evaluation of tofu texture

The quality of the tofu was analyzed by a Texture analyzer using a Stable Micro System, model TA-XT2 (Texture Technologies Corp., White Plains, NY, USA). The cylinder-shaped samples (25 mm diameter) were obtained by vertically cutting the tofu using a cylindrical cutter with triplicates. The samples were pressed twice using a metal disc probe (60 mm diameter) to simulate a mouth bite. The Texture Analyzer recorded the hardness, springiness, and cohesiveness of the tofu (Beléia et al., 2005)

3.3.6. Clustering of soybean seeds

An unsupervised pattern recognition technique, hierarchical clustering analysis (HCA), was used to cluster the soybean seeds and the tofu quality based on sources. Differences between the sources was clustered based on the composition of soybean seeds and parameters of tofu, such as protein content, moisture content, seed water uptake rate, tofu yield, firmness, springiness, gumminess, chewiness, resilience, and cohesiveness. The data were standardized and processed with the Ward method. This method provides the clustering of samples based on the tofu quality (M. Xu et al., 2019). Overall, soybean seeds and tofu quality parameters were sorted into 6 classes.

3.3.7. Principal component analysis

The matrix of compositional data of soybean seeds and tofu quality parameters was analyzed by principal components analysis (PCA) using statistical package JMP® Pro 15.0.0 (SAS Institute Inc.), to indicate possible clustering between distinct attributes (chemical components and tofu quality parameters) and distinct clusters of samples.

3.3.8. Statistical analysis

The tofu quality analysis was performed in triplicate. The data was further subjected to analysis of variance followed by Duncans Multiple Range Test with SPSS Statistics 24 (IBM). Differences at $p<0.05$ were considered significant.

HCA was performed on JMP® Pro 15.0.0 (SAS Institute Inc.).

3.4. Results and discussion

3.4.1. Sources of soybean seeds

Table 3.1 Origins of soybean seeds for comparative evaluation of seeds from the United States and China

Country	Province/State	No. of soybean varieties
United States	North Dakota	98
United States	Minnesota	19
United States	California	2
China	Qinghai	1
China	Heilongjiang	12
China	Hubei	16
China	Shanxi	1
China	Hebei	1
China	Shandong	10
China	Hainan	1
China	Guizhou	1
China	Anhui	1
China	Beijing	4
China	Guangxi	1
China	Jiangxi	1
China	Zhejiang	5
China	Guangdong	2
China	Sichuan	1
China	Shanghai	1

In the United States, diverse states contribute to the soybean varieties, with significant numbers originating from North Dakota, Minnesota, and California. Specifically, in North Dakota, the soybean seeds were procured from the plant sciences department at NDSU, Fargo, USA. Similarly, in Minnesota, the soybean seeds were acquired from the University of Minnesota, Minnesota, USA. Soybean seeds were collected from 16 different provinces in China, covering latitudes from 20° to 48° and longitudes from 96° to 128°

3.4.2. Characteristics of soybean seeds collected from different sources in the United States and China

The study involved an examination of 178 soybean varieties to assess their protein and moisture content. Utilizing hierarchical cluster analysis (HCA), as depicted in **Figure 3.1**, the soybeans were categorized into six distinct clusters, the details of which are presented in Table 3.1 and 3.2.

The results derived from HCA highlighted the division of soybean seeds into six clusters, primarily based on the similarity of their protein and moisture attributes. Cluster i exhibited a protein content ranging from 25.61% to 30.77%, with a relatively less coefficient of variation (C.V.) at 3.69%. Conversely, Cluster ii displayed a protein content spanning from 23.49% to 26.72%, accompanied by a slightly lower C.V. of 3.72 when compared to Cluster i.

Moving on to Cluster iii, this group demonstrated a protein content within the range of 26.32% to 29.74%, with a C.V. of 3.41, signifying similarity with Cluster i and a lack of statistical significance between the two. In contrast, Cluster iv presented a notably higher protein range, fluctuating between 30.41% and 32.63%, and a relatively lower C.V. of 2.68, distinctly setting it apart from its predecessors.

Cluster v showcased a protein content ranging from 29.33% to 31.91%, accompanied by a C.V. of 2.47, which, although lower than the previous cluster, remained significantly distinct from the other clusters. Lastly, Cluster vi exhibited the highest protein content range, extending from 31.42% to 37.06%, and the highest C.V. at 4.37, distinctly elevating it above all other clusters in terms of variability.

It's worth noting that all clusters exhibited significant differences from one another, except for Cluster i and Cluster iii, which displayed similar protein and moisture characteristics.

Cluster vi encompassed soybeans originating from diverse regions within North Dakota and Minnesota, all of which hailed from the United States. Interestingly, this cluster exhibited a remarkable high range in protein %. Such findings might suggest that soybeans cultivated in the United States tend to possess higher protein levels compared to their counterparts in China. This distinction could bear significant implications to produce tofu, particularly concerning its textural attributes, potentially yielding a firmer tofu product.

On the other hand, Cluster ii presented the lower range of protein %, with all samples originating exclusively from China. This observed variation could likely be attributed to the distinct sources where these soybeans were grown. The sources factor appears to play a pivotal role, giving rise to discernible differences in protein content among clusters derived from various sources.

In contrast, the moisture content within the clusters exhibited notable variability. Cluster i displayed a moisture percentage ranging from 1.34% to 4.27%, with a particularly high coefficient of variation (C.V.) at 21.54%. This high C.V. underscored the considerable diversity in moisture levels within this cluster.

Cluster ii, on the other hand, demonstrated a moisture range of 3.07% to 4.60% and a comparatively lower C.V. at 12.75, indicating higher moisture content when compared to Cluster i. Cluster iii exhibited a wider moisture range, spanning from 4.71% to 9.43%, accompanied by a C.V. of 15.30, setting it apart significantly from the preceding clusters.

Cluster iv presented a more constrained moisture range, fluctuating between 4.24% and 6.52%, with the lowest C.V. in the dataset at 11.81. This low C.V. highlighted a relatively consistent moisture content within this cluster. Meanwhile, Cluster v showcased a moisture range of 3.19% to 4.84%, with a C.V. of 12.18.

Lastly, Cluster vi with a range spanning from 1.12% to 4.66% and an exceptionally high C.V. at 31.13, signifying the greatest degree of moisture variability across all clusters.

In summary, each cluster displayed distinct moisture behaviors, with significant differences observed between Cluster iii and the preceding clusters. Notably, Cluster ii and v, as well as Cluster i and vi, did not exhibit significant differences in their moisture content.
The moisture content patterns observed in Cluster i and vi displayed striking similarities. This similarity may be attributed to the fact that most samples within these clusters originated from the United States, with a few hailing from Chinese provinces. On the contrary, Cluster ii and v exhibited a more diverse composition, comprising samples from both the United States and China. This intriguing divergence in moisture content could potentially be influenced by a multitude of factors.

It stands to reason that moisture content within these clusters is subject to considerable variability. This variability could account for the observed analogous behavior between samples from the United States and China, suggesting that moisture content is indeed influenced by a complex interplay of factors that transcend sources. According to Song et al. (2016), the levels of both crude protein and water-soluble protein exhibited notable variations across distinct sources. in China. Specifically, these variations were prominent in the North Spring Planting Region (NSR), encompassing both NESR (North East Spring Region) and NWSR (North West Spring Region), as well as in HHHR (Highland Highland Region) and SMCR (Southern Mountainous Region).

Interestingly, there was a discernible trend of increasing protein content as one moved from the northern regions, characterized by higher latitudes, towards the southern regions, which lie at lower latitudes. This study shows that various factors such as different source distribution, climate factors, seed coat color, and seed size that can influence protein content in soybean seeds. The studies highlight the variability in protein content among different soybean cultivars and sources, providing insights into the differences between soybeans from China and the United States.

Table 3.2 Evaluation of soybean seed protein and moisture across various soybean classes

	Cluster	Min	Max	Mean	C.V
Protein (%)	i	25.61	30.77	28.39±1.05d	3.69
	ii	23.49	26.72	25.18±0.94e	3.72
	iii	26.32	29.74	28.30±0.96d	3.41
	iv	30.41	32.63	31.51±0.84b	2.68
	v	29.33	31.91	30.34±0.75c	2.47
	vi	31.42	37.06	33.47±1.46a	4.37
Moisture (%)	i	1.34	4.27	2.98±0.64d	21.54
	ii	3.07	4.60	4.03±0.51c	12.75
	iii	4.71	9.43	5.87±0.90a	15.30
	iv	4.24	6.52	5.30±0.63b	11.81
	v	3.19	4.84	3.94±0.48c	12.18
	vi	1.12	4.66	2.97±0.92d	31.13

Different letters indicate statistically significant difference within columns ($p<0.05$).

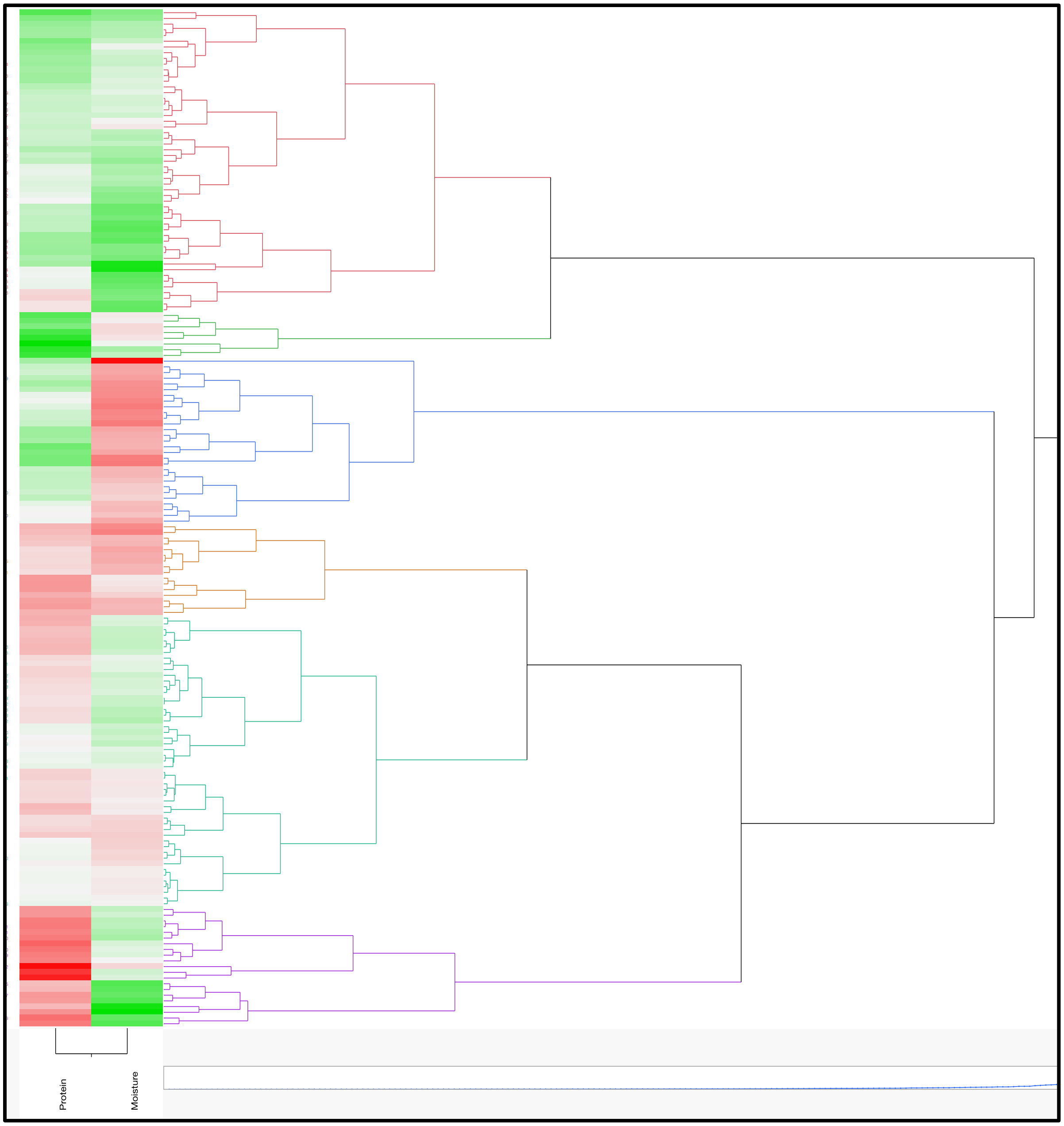


Figure 3.1 Clustering of soybean seeds considering their protein and moisture based on their sources using hierarchical clustering analysis (HCA). The color of Cluster i, ii, iii, iv, v and vi are indicated with red, green, blue, orange, dark green and purple respectively

3.4.3. Characteristics of tofu collected from different sources in the United States and China

The study involved an examination of tofu prepared from 178 soybean varieties to assess the tofu quality parameters influenced by the sources. Utilizing hierarchical cluster analysis

(HCA), as depicted in **Figure 3.2**, the tofu quality parameters were categorized into six distinct clusters, the details of which are presented in Table 3.3.

The results derived from HCA highlighted the division of tofu into six clusters, primarily based on the similarity of their water uptake, tofu yield, moisture, soymilk yield, brix, protein, firmness, gumminess, chewiness, springiness, cohesiveness, and resilience.

In examining water uptake across different clusters, several intriguing patterns emerged. Cluster 1 exhibited a water uptake range spanning from 1275.20 to 1501.40 kg/kg of soybean, accompanied by a relatively moderate coefficient of variation (C.V.) at 3.25. Conversely, Cluster 2 displayed a narrower water uptake range, falling between 1329.00 and 1486.20 kg/kg of soybean, with a notably low C.V. at 2.77. This low C.V. indicated a tighter concentration of data points and a higher degree of similarity in water uptake within this cluster. Furthermore, Cluster 2 exhibited analogous behavior to Cluster 1.

Cluster 3 showcased a water uptake range of 1317.00 to 1439.60 kg/kg of soybean, with an impressively low C.V. of 2.63, marking it as one of the clusters with the lowest variability. This behavior closely resembled that of Cluster 2. Cluster 4, while mirroring the patterns seen in Cluster 3, displayed a wider water uptake capacity range, stretching from 1193.80 to 1437.60 kg/kg of soybean, and featured a higher C.V. at 3.26.

Moving on to Cluster 5, it demonstrated a water uptake range spanning from 1351.00 to 1483.60 kg/kg of soybean, with the lowest C.V. among all clusters at 2.22. This cluster shared certain characteristics with Cluster 1.

Finally, Cluster 6 exhibited a broad water uptake range extending from 1137.00 to 1438.00 kg/kg of soybeans, coupled with the highest C.V. at 5.75. Interestingly, this cluster shared similarities with Clusters 2, 3, and 4 in terms of water uptake behavior. Cluster 1 showed the highest range and most of the soybeans in this cluster belonged to China and a few of them to specifically North Dakota, United States. Whereas Cluster 6 had the lowest water uptake and all of them belonged to the United States.

When examining the tofu yield of these clusters, notable variations emerged. Cluster 1 displayed a Tofu yield range from 557.00 to 762.60 kg/kg of soybeans, accompanied by a C.V. of 7.06. In contrast, Cluster 2 exhibited a narrower Tofu yield range, spanning from 527.40 to 623.00 kg/kg of soybeans, featuring the lowest C.V. in the dataset at 4.79. Notably, this cluster demonstrated different behavior compared to Cluster 1.

Meanwhile, Cluster 3 showcased a Tofu yield range of 515.20 to 631.60 kg/kg of soybeans, with a C.V. of 5.62, and its behavior closely resembled that of Cluster 2. However, Cluster 4 featured a wider Tofu yield range, fluctuating between 450.20 and 609.80 kg/kg of soybeans, and a high C.V. at 7.78, making it one of the clusters with the highest variability and significantly different from the previous clusters.

Cluster 5 displayed a Tofu yield range of 553.40 to 726.40 kg/kg of soybeans, along with a C.V. of 6.53, the second highest in the dataset. Its behavior bore similarities to Cluster 1. Finally, Cluster 6 exhibited a Tofu yield range from 492.00 to 599.40 kg/kg of soybeans, accompanied by a C.V. of 5.93. Cluster 1 had the highest range and had samples mostly from China and a few of them from United States, and Cluster 4 had the lowest range, and all the samples were from the United States.

In summary, each cluster demonstrated distinct Tofu yield characteristics, with certain clusters displaying greater variability than others. Clusters 1 and 5, as well as Cluster 2 and 3, shared similarities in their Tofu yield, while the remaining clusters exhibited significant differences from one another. These findings shed light on the Tofu production potential of the soybean varieties within each cluster.

Now, when we delve into Soymilk yield, we uncover distinctive trends among the clusters. Cluster 1 exhibited a Soymilk yield ranging from 2970.40 to 3664.80, with a C.V. of 4.06. Cluster 2, on the other hand, depicted a narrower range, spanning from 3043.00 to 3460.00, with a lower C.V. at 2.97. Despite the reduced variability and range, it still shared similarities in behavior with Cluster 1.

In contrast, Cluster 3 showcased a Soymilk yield range extending from 2405.00 to 3317.20, accompanied by the highest C.V. in the dataset at 9.73, signifying significant differences from its predecessors. Cluster 4 displayed a range from 2165.60 to 3279.80, featuring a C.V. of 9.42, which pointed to a different behavior compared to previous datasets. Notably, the minimum value within this range was the lowest among all the clusters.

Cluster 5 exhibited a Soymilk yield range of 2593.00 to 3370.40, along with a C.V. of 4.35. Interestingly, it showed statistically significant similarities to Cluster 3. Lastly, Cluster 6 portrayed a range spanning from 2280.00 to 3093.20, featuring a C.V. of 8.80. Remarkably, the maximum value in this dataset was the least among all the datasets, signifying its significant divergence from the remaining datasets. Cluster 1 had the highest soymilk yield and most of the samples were from China, whereas the Cluster 4 had the lowest range, and all the samples were from the United States and Cluster 6 also showed similar behavior to Cluster 4, wherein its samples were all of them from the United States.

In the Brix levels of Soymilk, we observe an intriguing pattern. Cluster 1 exhibited a Brix range from 7.10 to 8.50, coupled with a C.V. of 4.38. Cluster 2 displayed similar behavior to Cluster 1, with a range of 6.80 to 8.50 and a C.V. of 4.39.

In contrast, Cluster 3 showcased a Brix range spanning from 6.70 to 8.20, featuring a C.V. of 6.31, the second highest in the dataset. This cluster exhibited behavior distinct from its predecessors. Cluster 4 presented one of the narrowest Brix ranges, fluctuating between 5.90 and 7.60, with a C.V. of 6.00, significantly differing from the previous clusters.

Cluster 5 displayed a Brix range of 6.40 to 7.70, accompanied by a C.V. of 4.79, sharing similarities in behavior with Cluster 3. In contrast, Cluster 6 portrayed the smallest Brix range in the entire dataset, ranging from 5.60 to 7.30, despite having one of the highest C.V.s at 6.91. This cluster exhibited significant differences from the other clusters.

It's noteworthy that Cluster 1, primarily comprised of soybean samples from China, displayed the highest Brix levels. Conversely, Cluster 6, consisting entirely of samples from the United States, exhibited the lowest Brix range.

In summary, each cluster displayed distinct Brix characteristics in Soymilk, with varying ranges and degrees of variability. These findings provide valuable insights into the Brix levels of Soymilk produced from the soybean varieties within each cluster.

Examining firmness, we uncover diverse patterns among the clusters. Cluster 1 exhibited a wide range, spanning from 3946.09 to 21559.55, along with a high C.V. of 30.86. In contrast, Cluster 2 displayed a narrower range, from 15142.30 to 28594.87, featuring a C.V. of 16.13, signifying significant differences from its predecessor. Cluster 3 showcased a firmness range of 8115.42 to 21754.72, with a C.V. of 22.07, and it exhibited behavior akin to Cluster 1.

Cluster 4 presented the highest firmness range, fluctuating between 11849.00 and 44620.59, accompanied by a C.V. of 34.69, distinguishing it significantly from the other clusters in the dataset. Cluster 5 displayed a firmness range of 3725.04 to 17735.30, with a C.V. of 38.59, and its behavior was reminiscent of Cluster 1.

Lastly, Cluster 6 portrayed a range from 5021.51 to 23089.98, featuring the highest C.V. in the dataset at 40.87. Interestingly, the lowest firmness was observed in Cluster 1, which included samples from various provinces in China and North Dakota, United States. On the other hand, the highest firmness was associated with the United States tofu, with Cluster 4 predominantly comprising varieties from North Dakota and Minnesota.

An investigation into the springiness of tofu unveiled distinct patterns among the clusters. Cluster I exhibited a springiness range of 0.86 to 1.09, with a coefficient of variation (C.V.) at 4.42. Conversely, Cluster 2 displayed a slightly wider range, spanning from 0.95 to 1.27, and featured a higher C.V. at 7.16. Interestingly, Cluster 2 demonstrated behavior reminiscent of Cluster 1.

In stark contrast, Cluster 3 showcased the most extensive springiness range, extending from 1.26 to 2.35, coupled with the highest C.V. in the dataset at 22.66. This cluster exhibited significant differences from all the other datasets. Cluster 4 presented a springiness range of 0.91 to 1.53, with a C.V. of 15.27, the second highest in the dataset, and it shared similarities in behavior with Cluster 2.

Cluster 5 displayed a narrower springiness range, spanning from 0.83 to 1.00, featuring the lowest C.V. at 4.23, and exhibited similarities to Cluster 2. In contrast, Cluster 6 showcased the smallest springiness range within the dataset, ranging from 0.81 to 0.98, with a C.V. of 5.04. This cluster shared similarities with Clusters 1, 2, and 4.

Notably, Cluster 3, comprised of samples from various provinces in China and different states in the US, including Minnesota, California, and North Dakota, exhibited the highest springiness range. In contrast, Cluster 6, consisting solely of samples from the United States, displayed the lowest springiness range.

In the examination of tofu cohesiveness, distinct trends emerged among the clusters. Cluster 1 exhibited a cohesiveness range of 0.60 to 0.81, accompanied by a coefficient of variation (C.V.) of 6.25. Conversely, Cluster 2 displayed a cohesiveness range of 0.63 to 0.78, featuring a similar C.V. of 6.25, thus mirroring Cluster 1 in variability.

Cluster 3, in contrast, showcased a cohesiveness range spanning from 0.71 to 0.85, with the lowest C.V. in the dataset at 5.32. This cluster stood out significantly from the other datasets. Cluster 4 presented a cohesiveness range of 0.57 to 0.82, featuring the highest C.V. in the entire dataset at 10.34.

Cluster 5 displayed a cohesiveness range of 0.51 to 0.65, with a C.V. of 6.10, and exhibited significant differences from the rest of the dataset. In comparison, Cluster 6 demonstrated a cohesiveness range from 0.46 to 0.66, with the second-highest C.V. at 8.84, significantly differing from the other datasets.

Notably, Cluster 6 had the lowest cohesiveness, with samples predominantly originating from the United States. In contrast, Cluster III displayed the highest cohesiveness, followed by Cluster 2. Cluster 3 mainly consisted of samples from China, with a mixture of samples from the United States, while Cluster 2 exclusively represented Chinese provinces.

In summary, each cluster exhibited distinct cohesiveness characteristics in tofu, marked by varying ranges and degrees of variability. These findings provide valuable insights into the cohesiveness attributes of tofu produced from the soybean varieties within each cluster, with different sources of origin playing an important role.

In the comprehensive analysis of tofu resilience, intriguing observations were made across the clusters. Cluster 1 exhibited a resilience range spanning from 0.16 to 0.32, accompanied by a notable coefficient of variation (C.V.) of 13.18. Cluster 2 displayed a narrower resilience range, fluctuating between 0.21 and 0.31, featuring a C.V. of 10.04, thereby sharing similarities with Cluster 1. Notably, Cluster 3 showcased a resilience range of 0.23 to 0.34, accompanied by a C.V. of 11.03, which was one of the highest C.V. values, and exhibited behavior akin to Cluster 2.

Conversely, Cluster 4 presented a resilience range from 0.15 to 0.29, featuring the highest C.V. in the entire dataset at 16.82. This cluster stood out significantly from the rest of the dataset. Cluster 5 displayed a resilience range spanning from 0.13 to 0.23, with a C.V. of 13.28, marking it as significantly different from the remaining data. Cluster 6 portrayed the narrowest resilience range, fluctuating between 0.11 and 0.19, with a C.V. of 15.72, significantly differing from the rest of the dataset.

It's noteworthy that Cluster 6 exhibited the lowest resilience range and predominantly included samples from the United States. In contrast, Cluster 3 and Cluster 2 displayed higher ranges of resilience, with most of the varieties in these clusters originating from various Chinese provinces.

In the comprehensive examination of tofu chewiness, a complex picture emerged across the clusters. Cluster 1 exhibited a wide chewiness range, spanning from 2418.44 to 13779.49, accompanied by a substantial coefficient of variation (C.V.) of 31.91. Cluster 2 displayed a narrower chewiness range, fluctuating between 11619.74 and 21492.02, featuring one of the lowest C.V. values in the dataset at 18.07. Notably, Cluster 2 demonstrated behavior akin to Cluster 1.

On the other hand, Cluster 3 showcased a significantly broader chewiness range, extending from 8062.45 to 37698.82, accompanied by a high C.V. of 40.48, marking it as significantly different from the entire dataset. Cluster 4 presented an even wider chewiness range, ranging from 6393.78 to 45702.32, with the highest C.V. in the dataset at 49.45, while sharing similarities with Cluster 2.

Cluster 5 displayed a chewiness range spanning from 1937.28 to 10189.28, featuring the second-largest C.V. at 41.64, and exhibited behavior similar to Cluster 1. Cluster 5 portrayed a range of chewiness from 2256.64 to 11154.95, with a C.V. of 41.10, showcasing similar behavior to Clusters 1 and 5.

It's noteworthy that Cluster 5 exhibited the narrowest chewiness range, comprising various soybean seed varieties from North Dakota. Conversely, the highest chewiness range was observed in Cluster 4, followed by Cluster 3, with Cluster 4 primarily consisting of samples from the United States, while Cluster 3 comprised samples from Chinese provinces.

In the analysis of tofu gumminess, a nuanced panorama emerged across the clusters. Cluster 1 exhibited a gumminess range spanning from 2500.78 to 14136.09, marked by a considerable coefficient of variation (C.V.) at 32.57. In contrast, Cluster 2 displayed a notably narrower gumminess range, fluctuating between 11725.25 and 21814.25, featuring the lowest C.V. in the entire dataset at 16.48. This cluster demonstrated distinct behavior from the previous cluster.

Cluster 3 showcased a gumminess range extending from 6358.28 to 18407.85, with a C.V. of 24.83, signifying differences from the preceding data. In stark contrast, Cluster 4 presented the widest gumminess range, ranging from 7180.66 to 32983.29, and featured a C.V. of 36.87, while sharing similarities with Cluster 2.

Cluster 5 displayed a gumminess range spanning from 2013.39 to 10901.09, accompanied by the highest C.V. at 42.54, marking it as significantly different from the previous data. Cluster 6 portrayed a range of gumminess from 2329.11 to 11454.01, with a C.V. of 38.69, showcasing behavior akin to Clusters 1 and 5.

Notably, Cluster 5 exhibited the lowest gumminess and comprised varieties exclusively from North Dakota, United States. Conversely, Cluster 4 featured the highest gumminess and included varieties from Minnesota and North Dakota, in combination with a variety from China.

In the assessment of tofu protein content, a diverse array of patterns emerged among the clusters. Cluster 1 exhibited a protein content range spanning from 39.18 to 58.98, featuring a coefficient of variation (C.V.) of 7.55. Cluster 2 displayed a notably narrower protein content range, fluctuating between 50.12 and 59.41, accompanied by the lowest C.V. in the dataset at 3.94, mirroring characteristics of Cluster 1.

Conversely, Cluster 3 showcased a protein content range extending from 34.94 to 60.21, marked by the highest C.V. among the entire dataset at 13.97 while resembling Cluster 1. Cluster 4 presented the widest protein content range, ranging from 41.25 to 67.10, with a C.V. of 11.90, ranking as the third-largest C.V. in the dataset and showing similarity to Cluster 2. Cluster 5 exhibited a protein content range spanning from 42.29 to 60.61, which was the second highest range, with a C.V. of 8.11 like Cluster 2 and 3. Cluster 6 portrayed a range of protein content from 36.69 to 55.24, with the second-highest C.V. at 13.87 which was significantly different to other readings.

Notably, Cluster 4 featured the highest protein content, with most samples in this cluster originating from the United States. Cluster 5 had second highest protein and was also entirely composed of samples from the United States. In contrast, the lowest protein content was observed in Cluster 1, where samples had mixed origins from both China and the United States.

In the analysis of tofu moisture content, distinct patterns emerged within each cluster. Cluster 1 exhibited a moisture content range spanning from 72.61 to 78.92, featuring a relatively low coefficient of variation (C.V.) of 2.03. In stark contrast, Cluster 2 displayed a notably narrower moisture content range, fluctuating between 69.64 and 73.48, with the lowest C.V. in the entire dataset at 1.50. This cluster demonstrated significant differences from its predecessor.

Cluster 3 showcased a moisture content range extending from 71.35 to 75.34, marked by a C.V. of 1.70 and signifying significant differences from the previous dataset. Cluster 4 presented a range of moisture content ranging from 69.36 to 74.94, with a C.V. of 1.78, while sharing similarities with Cluster 2. Cluster 5 exhibited a moisture content range spanning from 72.88 to 80.18, featuring the highest C.V. in the dataset at 2.55 and marked by significant differences from the preceding data.

Cluster 6 portrayed the widest moisture content range, extending from 73.16 to 78.17, which was the highest range in the dataset, accompanied by a C.V. of 2.09, the second-highest C.V.

Notably, Cluster 2 displayed the lowest moisture content range in the entire dataset, with most samples originating from China. In contrast, Cluster 6 exhibited the highest moisture content, with samples primarily hailing from North Dakota, United States.

Table 3.3 Evaluation of tofu quality across various soybean classes

	Cluster	Min	Max	Mean	C.V
Water uptake (kg/kg soybean)	1	1275.20	1501.40	1396.75±45.42ab	3.25
	2	1329.00	1486.20	1385.32±38.41cb	2.77
	3	1317.00	1439.60	1391.35±36.63bc	2.63
	4	1193.80	1437.60	1381.45±45.04bc	3.26
	5	1351.00	1483.60	1414.66±31.39a	2.22
	6	1137.00	1438.00	1364.44±78.46c	5.75
Tofu yield (kg/kg soybean)	1	557.00	762.60	631.13±44.59a	7.06
	2	527.40	623.00	580.44±27.77b	4.79
	3	515.20	631.60	576.48±32.38b	5.62
	4	450.20	609.80	519.18±40.39d	7.78
	5	553.40	726.40	624.74±40.82a	6.53
	6	492.00	599.40	545.32±32.33c	5.93
Soymilk yield	1	2970.40	3664.80	3290.45±133.66a	4.06
	2	3043.00	3460.00	3326.09±98.78a	2.97
	3	2405.00	3317.20	2993.52±291.25b	9.73
	4	2165.60	3279.80	2638.54±248.61d	9.42
	5	2593.00	3370.40	3074.66±133.64b	4.35
	6	2280.00	3093.20	2769.01±243.68c	8.80
Brix	1	7.10	8.50	7.79±0.34a	4.38
	2	6.80	8.50	7.89±0.35a	4.39
	3	6.70	8.20	7.40±0.47b	6.31
	4	5.90	7.60	6.95±0.42c	6.00
	5	6.40	7.70	7.23±0.35b	4.79
	6	5.60	7.30	6.28±0.43d	6.91
Firmness (g force)	1	3946.09	21559.55	11970.40±3694.24cd	30.86
	2	15142.30	28594.87	20914.85±3373.28b	16.13
	3	8115.42	21754.72	14611.57±3225.11c	22.07
	4	11849.00	44620.59	23860.06±8277.85a	34.69

Different letters indicate statistically significant difference within columns ($p<0.05$).

Table 3.3 Evaluation of tofu quality across various soybean classes (Continued)

	Cluster	Min	Max	Mean	C.V
	5	3725.04	17735.30	9393.72±3624.82d	38.59
	6	5021.51	23089.98	11680.89±4774.01cd	40.87
Springiness	1	0.86	1.09	0.95±0.04c	4.42
	2	0.95	1.27	1.00±0.07cb	7.16
	3	1.26	2.35	1.68±0.38a	22.66
	4	0.91	1.53	1.04±0.16b	15.27
	5	0.83	1.00	0.94±0.04c	4.23
	6	0.81	0.98	0.93±0.05c	5.04
Cohesiveness	1	0.60	0.81	0.68±0.04b	6.25
	2	0.63	0.78	0.70±0.04b	5.39
	3	0.71	0.85	0.77±0.04a	5.32
	4	0.57	0.82	0.68±0.07b	10.34
	5	0.51	0.65	0.58±0.04c	6.10
	6	0.46	0.66	0.55±0.05d	8.84
Resilience	1	0.16	0.32	0.24±0.03b	13.18
	2	0.21	0.31	0.26±0.03ab	10.04
	3	0.23	0.34	0.27±0.03a	11.03
	4	0.15	0.29	0.21±0.04c	16.82
	5	0.13	0.23	0.18±0.02d	13.28
	6	0.11	0.19	0.15±0.02e	15.72
Chewiness	1	2418.44	13779.49	7366.47±2350.40c	31.91
	2	11619.74	21492.02	14634.54±2644.83b	18.07
	3	8062.45	37698.82	19831.71±8027.94a	40.48
	4	6393.78	45702.32	16600.13±8208.06b	49.45
	5	1937.28	10189.28	4921.54±2049.43c	41.64
	6	2256.64	11154.95	5543.59±2278.63c	41.1
Gumminess	1	2500.78	14136.09	7683.23±2502.80c	32.57
	2	11725.25	21814.25	14574.93±2401.82a	16.48
	3	6358.28	18407.85	11380.47±2825.29b	24.83
	4	7180.66	32983.29	15554.06±5734.04a	36.87
	5	2013.39	10901.09	5264.49±2239.62d	42.54
	6	2329.11	11454.01	5935.39±2296.18cd	38.69

Different letters indicate statistically significant difference within columns (p<0.05).

Table 3.3 Evaluation of tofu quality across various soybean classes (Continued)

	Cluster	Min	Max	Mean	C.V
Protein	1	39.18	58.98	52.31±3.95ab	7.55
	2	50.12	59.41	54.18±2.14a	3.94
	3	34.94	60.21	49.53±6.92b	13.97
	4	41.25	67.10	54.69±6.51a	11.90
	5	42.29	60.61	52.18±4.23ab	8.11
	6	36.69	55.24	46.11±6.39c	13.87
Moisture	1	72.61	78.92	74.94±1.52b	2.03
	2	69.64	73.48	72.12±1.08d	1.50
	3	71.35	75.34	73.55±1.25c	1.70
	4	69.36	74.94	72.12±1.28d	1.78
	5	72.88	80.18	76.29±1.94a	2.55
	6	73.16	78.17	75.43±1.58ab	2.09

Different letters indicate statistically significant difference within columns (p<0.05).

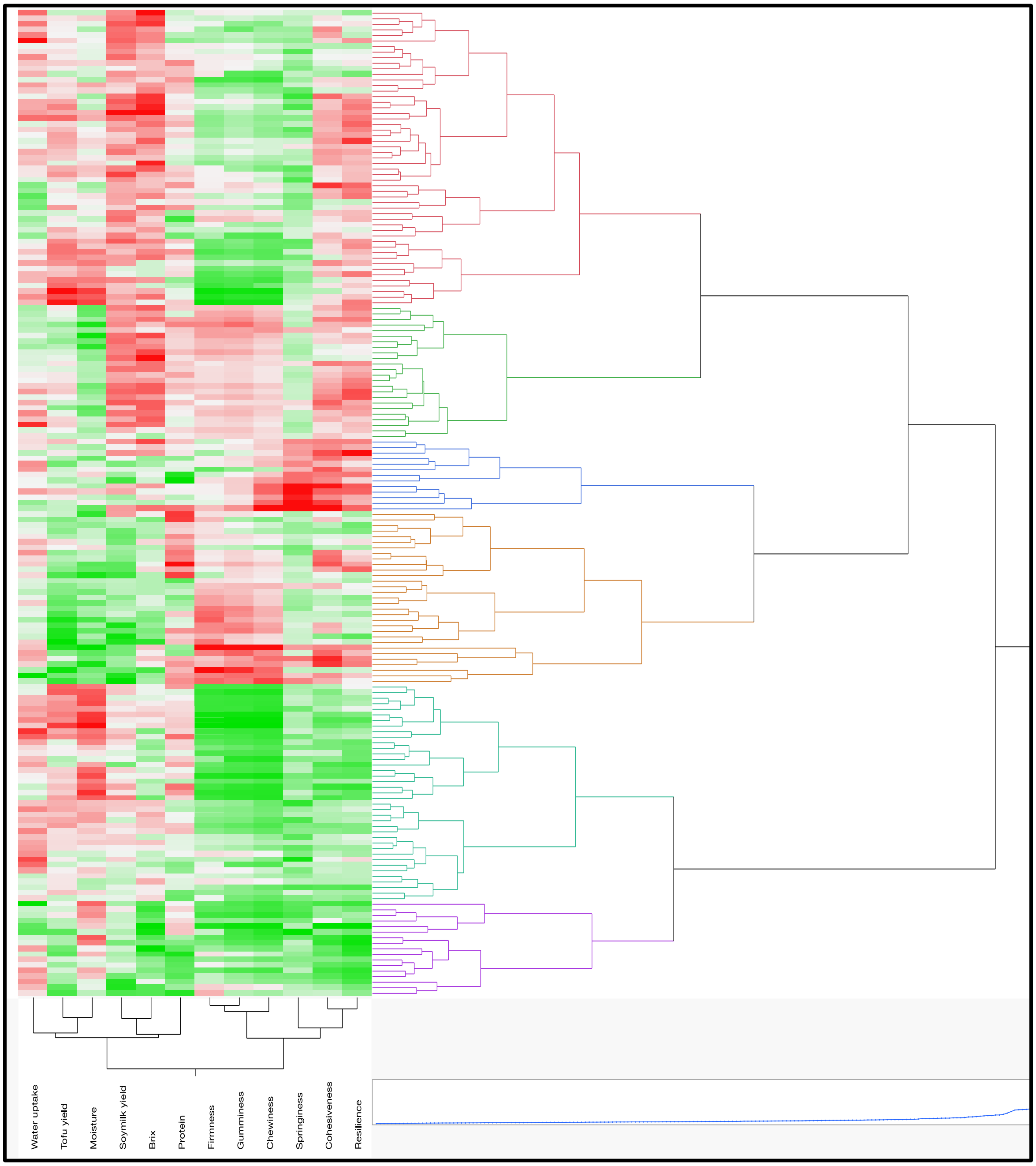


Figure 3.2 Clustering of tofu quality based on their sources using hierarchical clustering analysis (HCA). The color of Cluster 1, 2, 3, 4, 5 and 6 are indicated with red, green, blue, orange, dark green and purple respectively

3.4.4. Application of principal component analysis to assess the impact of seed origin on tofu quality parameters

Principal Component Analysis (PCA) is a powerful statistical tool which can assess the influence of sources on the quality parameters of various food products. In **Figure 3.3**, the PCA results depict the behavior of different samples from distinct clusters concerning their tofu quality. The measured values are used to construct a multidimensional dataset, which is then projected onto a biplot. A biplot is a scatter plot that illustrates the relationship between observed data and dependent variables, represented in terms of principal components.

PC1, PC2, and PC3 collectively account for 39.89%, 24.53%, and 9.76% of the total variance in tofu quality parameters. Variables and samples are situated separately within the four quadrants, revealing the presence of six clusters. In the first biplot, which illustrates the relationship between Component 1 and 2, springiness, chewiness, gumminess, and firmness align with the direction of PC1, indicating a positive correlation with PC1. Clusters 2, 3, and 4 fall within this region. Conversely, cohesiveness, resilience, protein, brix, soymilk yield, water uptake, and tofu yield align with PC2, displaying a positive correlation with PC2. Clusters 1 and 5 are positioned here. The plot reveals that PC2 is characterized by low protein and water uptake, while other components are high, whereas PC1 is characterized by low springiness.

When examining the relationship between Component 3 and 1, cohesiveness, chewiness, resilience, gumminess, and firmness exhibit a positive correlation with PC1. Cluster 3 is aligned with this direction. In contrast, moisture, tofu yield, and water uptake display a negative correlation with PC1, and clusters 1, 5, and 6 are positioned in this region. Springiness shows a positive correlation with PC3, and cluster 3 aligns with this direction. Soymilk yield, brix, and protein exhibit a negative correlation with PC3, and cluster 2 is situated here. PC1 is characterized by low resilience, while the remaining parameters have high loadings, and PC3 is marked by low brix, with the other parameters having high loadings.

Finally, regarding the relationship between Component 2 and 3, PC2 displays a positive correlation with cohesiveness, tofu yield, resilience, brix, soymilk yield, and clusters 1 and 3 are situated in this region. PC3 shows a positive correlation with springiness, water uptake, moisture, chewiness, and cluster 3 lies here. Conversely, PC3 is negatively correlated with gumminess, firmness, and protein, with clusters 2, 4, 5, and 6 positioned in this region. PC2 exhibits lower tofu yield and is loaded with the remaining parameters, while PC3 is characterized by low gumminess and chewiness, with the other parameters having high loadings.

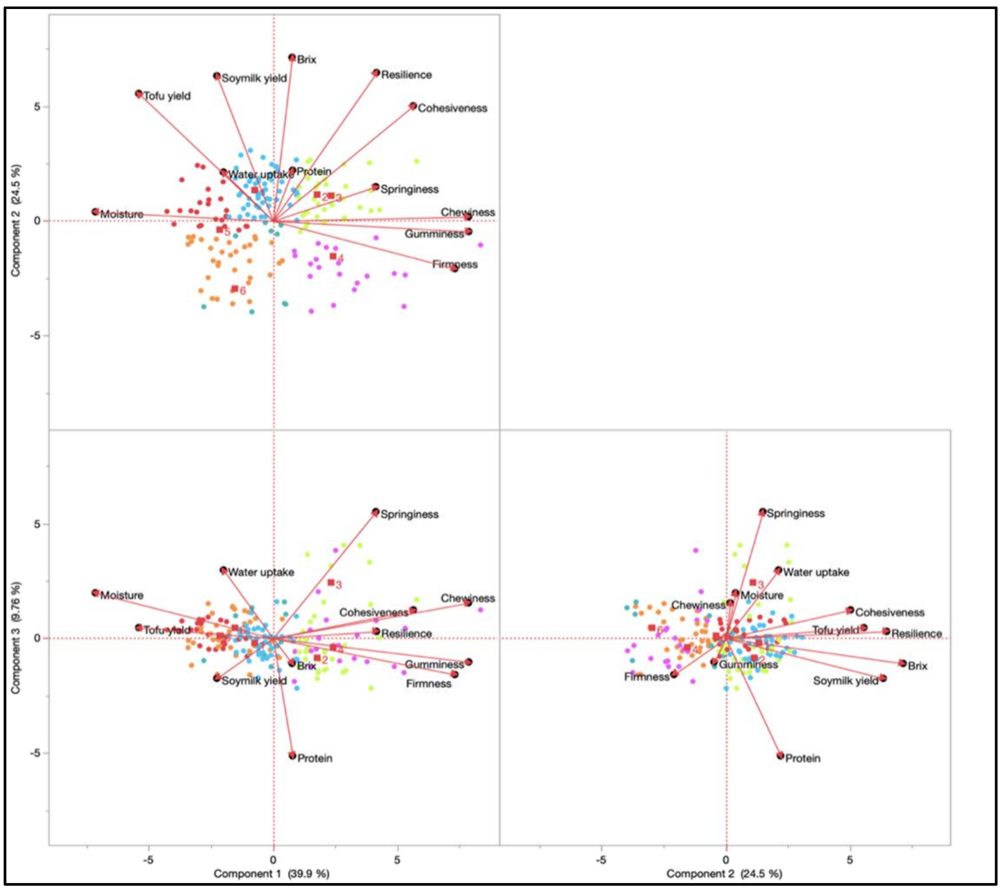


Figure 3.3 Clustering of soybean seeds based on tofu quality and characteristics of tofu from United States and China using principal component analysis (PCA)

3.4.5. Correlation coefficients between soybean seed characteristics and tofu quality parameters derived from soybean seeds from United States and China

In our comprehensive investigation, we have identified significant correlations among various tofu quality parameters and soybean seed characteristics, considering the sources of the soybean varieties. The results reveal intricate relationships that provide valuable insights into the interplay of these factors in shaping tofu quality represented in Figure 3.4.

Firstly, we observed a strong positive correlation between brix and tofu yield, as well as soymilk yield, p value <0.0001 indicating that higher brix levels are associated with increased soymilk and tofu yields. This suggests that the sweetness level, represented by brix, plays a pivotal role in influencing the yield of tofu and its associated soymilk.

Furthermore, our analysis unveiled a robust positive correlation between chewiness and firmness, p value <0.0001 emphasizing the close association between these textural attributes. Chewiness also exhibited noteworthy positive correlations with cohesiveness and gumminess, p value <0.0001 highlighting the co-dependency of these parameters in defining the tactile qualities of tofu. This suggests that alterations in one of these attributes may inherently impact the others, necessitating careful consideration during tofu production.

Conversely, we noted a negative relationship between firmness and tofu yield, p value <0.0001 indicating that firmer tofu tends to result in lower yields. Similarly, gumminess displayed a negative correlation with tofu yield, p value <0.0001 implying that excessively gummy textures may adversely affect tofu production, potentially leading to reduced yield.

Moreover, tofu moisture content exhibited a negative relationship with firmness, gumminess, and chewiness, p value <0.0001. This finding suggests that higher moisture content in tofu tends to yield softer, less gummy, and less chewy textures. These textural variations are essential factors to consider when striving to achieve desired sensory attributes in tofu products.

In summary, our study has elucidated intricate correlations among soybean seed characteristics and tofu quality parameters. These findings offer valuable insights into the multifaceted factors that contribute to the sensory and textural properties of tofu, with consideration for the sources of the soybean varieties.

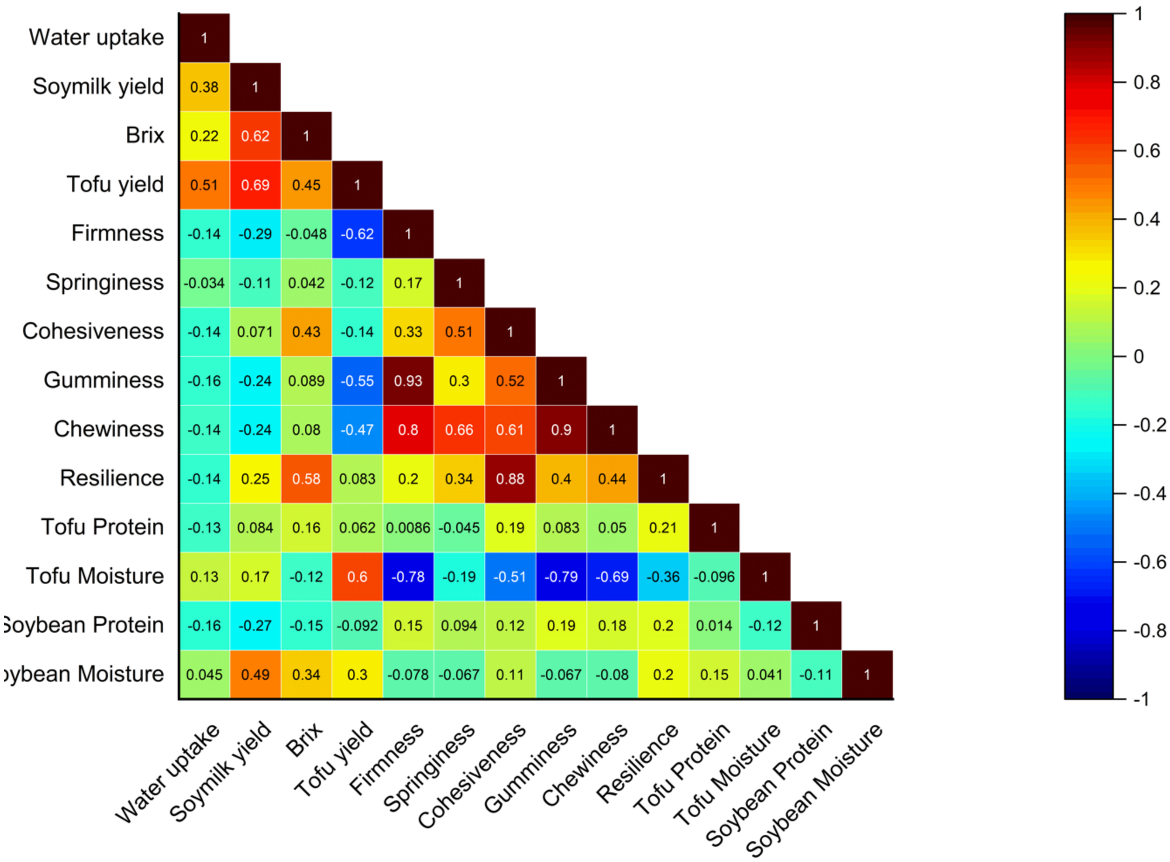


Figure 3.4 Pearson's correlation analysis of soybean seed characteristics and tofu quality across diverse clusters

3.5. Conclusion

In conclusion, this study has provided a comprehensive analysis of the effect of sources on the protein and moisture content of soybean seeds, as well as on various quality parameters of tofu produced from these soybeans. Through the application of hierarchical cluster analysis (HCA) and principal component analysis (PCA), we have gained valuable insights into the relationships and variations within the dataset.

The analysis of tofu quality parameters further highlighted the influence of sources. Variations in water uptake, tofu yield, soymilk yield, brix, firmness, gumminess, chewiness, springiness, cohesiveness, and resilience were observed among the clusters. These variations provided insights into the textural and sensory attributes of tofu, with implications for consumer preferences and industrial tofu production.

Furthermore, our analysis revealed that moisture content in tofu had a significant impact on textural attributes, with higher moisture content resulting in softer and less chewy textures. This information is crucial for producers seeking to tailor tofu products to meet specific consumer preferences.

In summary, this study has provided a comprehensive understanding of how source influences soybean seed characteristics and, subsequently, tofu quality parameters. These insights have practical implications for the soybean and tofu industries, offering opportunities for product optimization and market differentiation based on the sourcing. Additionally, the study underscores the importance of considering both soybean seed characteristics and tofu quality attributes when aiming to produce tofu that aligns with consumer expectations and preferences.

4. PREDICTING TOFU QUALITY FROM SOYBEAN SEEDS USING HYPERSPECTRAL IMAGING AND MACHINE LEARNING

4.1. Abstract

Assessing the quality of soybean seeds for tofu production traditionally requires the actual creation of tofu, a process that demands considerable time and effort. This study addresses this issue by employing machine learning to predict tofu quality from Hyperspectral Imaging (HSI) images of soybean seeds. Two hundred varieties of soybean seeds scanned with HSI have been classified into four categories based on their qualities of tofu products. Upon comparison, XGBoost was employed to pinpoint ten critical HSI wavelengths that show a potential correlation with the protein, carbohydrate, and oil contents in the soybean seeds. Subsequently, a Convolutional Neural Network model was formulated to forecast tofu quality, basing its predictions on HSI data of soybean varieties. Remarkably, the model successfully segregated the soybeans into four unique classes, demonstrating a predictive accuracy that varied between 96% and 99%. This research amalgamates cutting-edge technologies to revolutionize the conventional assessment of soybean seeds.

4.2. Introduction

Soybeans are a significant nutritional source worldwide, offering a complete protein profile containing all essential amino acids, dietary fiber, vitamins, minerals, and essential fatty acids. The beans are utilized in various products, including soy sauce, miso, natto, tempeh, sufu, kinako, soy milk, tofu, abura-age, and yuba (Fukushima, 2009). Tofu, in particular, is a traditional Asian food consumed in East-Asian countries for centuries and has gained popularity in Western countries due to the rising trend of plant-based food (Ali et al., 2021).

Tofu can be chemically described as a protein gel primarily composed of water, proteins, fats and carbohydrates. Tofu production involves adding a coagulating agent to soy milk and pressing the resulting curd into a block. Two traditional coagulants for tofu are calcium sulfate and magnesium chloride, which resulted in and marinated tofu, respectively. In addition to the coagulants, the quality of tofu is closely linked to the protein, fat, and carbohydrate content. It is noteworthy that while the protein content of soybean seeds does not significantly correlate with tofu yield, the protein content in soymilk does relate to water holding and tofu yield (Lim et al., 1990b). This suggests that protein quality, including protein subunit and amino acid composition, impacts tofu quality more than the protein content of soybeans (Stanojevic et al., 2011b). Protein quality also influences tofu textures such as hardness, cohesiveness, and springiness. Additionally, fats and carbohydrates can affect tofu quality through their interaction with proteins.

Traditional methods for evaluating tofu quality assess yield, texture, and sensory attributes (Poysa et al., 2006). However, these methods have shortcomings. They are labor-intensive, lack comparability due to variations in tofu processing parameters, and take a substantial amount of time, making them unfit for modern, rapid production capacities (Kurasch et al., 2018). Therefore, there is an urgent need for a swift, efficient, standardized method for evaluating soybean quality concerning tofu products.

Hyperspectral imaging (HSI) is a non-destructive testing method that combines imaging and spectral information, providing detailed spectral responses of target features (T. Gao et al., 2021; Kandpal et al., 2015; Kucha et al., 2021; Medus et al., 2021). It has been used in various studies to predict seed quality, analyze the chemical composition, such as protein, fat, and carbohydrate, and functionality of seeds through spectral information. Each chemical component has a unique spectral signature that can be detected using HSI (Erkinbaev et al., 2017) . Those chemical components are considered key factors for determining the tofu qualities. Several studies have shown HSI's potential as a rapid method for the evaluation of seed quality. For example, Squeo et al. (2022) developed a method using NIR- HIS to perform rapid, accurate and nondestructive quality control of TVP, as these parameters strongly influence the nutritional and textural properties of plant based meat analogues. Yang et al. (2018) used Raman hyperspectral imaging to detect the chemical compositions in maize seeds for online quality control. da Silva Medeiros et al. (2022) explored NIR HIS and successfully predicted the oil and erucic acid content in seeds.

Analyzing HSI data is challenging due to its complexity and high dimensionality, making it difficult to extract meaningful information using traditional statistical methods (Iqbal et al., 2014). Machine learning, however, can efficiently and accurately analyze high-dimensional data by learning patterns and relationships in data automatically (T. Gao et al., 2021). Within the context of hyperspectral imaging, machine learning can perform tasks such as classification, feature selection, regression, and anomaly detection. This research developed a machine learning model to predict tofu quality based on HSI data of soybean seeds. The objectives were to cluster soybean seeds based on corresponding tofu quality, select featured wavelengths scanned by HSI, and develop a predictive machine learning model using soybean HSI image data and corresponding tofu quality. This research has the potential to enhance the efficiency and accuracy of tofu quality evaluation, reduce waste and cost, and assist in soybean breeding and tofu manufacturing.

4.3. Materials and methods

4.3.1. Seeds and materials

Two hundred varieties of soybeans, harvested from North Dakota, Missouri, Minnesota, Illinois, and Ohio, were generously provided by the Agricultural Utilization Research Institute (Crookston, MN). Calcium sulfate was purchased from the local market.

4.3.2. Water uptake capacity of soybean seeds

Same as 3.3.2

4.3.3. Preparation of tofu

The tofu process was adapted from Meng et al. (2016) with modifications. Briefly, dry soybeans (W_0) were soaked following Method 4.3.2. The soaked soybeans were milled into slurries using a grinder hopper assembled on the automatic soymilk machine. Ten liters of water were added during the grinding procedure. The steam cooking (95 ℃) began automatically and lasted for five minutes. The milk exited via the catch pipe, while the okara exited through the pressure relief valve. The initially collected soy milk was weighed and recorded (W_1). In a pan, 11 kilograms of soy milk were weighed to make curd. The soy milk was cooled to 82 ℃ and placed in a pan. Then, 35 g of calcium sulfate was evenly dispersed in the soy milk. After 12 minutes, the curds were broken up with an edge scraper and whipped. After setting for one minute, the curds were poured into the drain pan with a mesh cloth at the bottom. After five minutes, the curds were wrapped with the mesh cloth and moved to the assembled air presser. The initial pressure and second pressure were added to the curds for 5 minutes and 15 minutes, respectively. The prepared tofu was soaked in cool water for 15 minutes, and the final weight (W_2) of tofu was recorded.

The formula for calculating tofu yield is calculated same as Equation 1 & 2.

4.3.4. Evaluation of tofu texture

The quality of the tofu was analyzed by a Texture analyzer using a Stable Micro System, model TA-XT2 (Texture Technologies Corp., White Plains, NY, USA). The cylinder-shaped samples (25 mm diameter) were obtained by vertically cutting the tofu using a cylindrical cutter with triplicates. The samples were pressed twice using a metal disc probe (60 mm diameter) to simulate a mouth bite. The Texture Analyzer recorded the hardness, springiness, and cohesiveness of the tofu (Beléia et al., 2005).

4.3.5. Clustering of soybean seeds

An unsupervised pattern recognition technique, hierarchical clustering analysis (HCA), was used in order to cluster the soybean seeds based on the tofu quality. Soybean seeds were clustered based on the qualities of soybean seeds and tofu, such as seed water uptake rate, tofu yield, firmness, springiness, and cohesiveness. The data were standardized and processed with the Ward method. This method provides not only the clustering of samples based on the tofu quality but is also an important source of knowledge with which to create cross-validation groups used in machine learning (M. Xu et al., 2019). Overall, soybean seeds were sorted into four clusters.

4.3.6. Hyperspectral scanning of soybean seeds

The hyperspectral data were recorded in the laboratory using the camera (Specim FX17, Specim, Oulu, Finland). The sensor is a push-broom type that captures hyperspectral cube data in the range of 900-1700 nm, with a spectral resolution of 8 nm, and has the ability to record 224 bands. To record the data, the researchers used Specim's LabScanner 40×20 platform, which features a halogen light source, a camera mount, and a 400×200 mm translation sample stage (**Figure 4.1**). To minimize external light interference, the data was recorded in a dark room with only the halogen bulbs of the platform as the light source. The researchers captured white and dark reference calibration images with each individual image, where the white reference was captured using a Teflon bar with > 95% reflectance, and the dark reference was captured by closing the sensor shutter. The data recording software used was Lumo Scanner. The kernels were placed in a petri dish to minimize the inertia generated by the translation stage.

To mitigate the effects of illumination changes and dark current in the sensor, the researchers calibrated the reflectance of the hyperspectral image using the formula below:

$$R = \frac{I - I_B}{I_W - I_B}$$

Equation (3)

Where $\boldsymbol{R}$ is the hyperspectral image after the reflectance calibration**,** $\boldsymbol{I}$ is the original hyperspectral image, $\boldsymbol{I_w}$ is the white reference hyperspectral image of the diffuse reflection whiteboard with 99% reflectance, and $\boldsymbol{I_B}$ is the dark reference hyperspectral image when the lens is covered (Feng, Makino, Oshita, & García Martín, 2018; He et al., 2022).

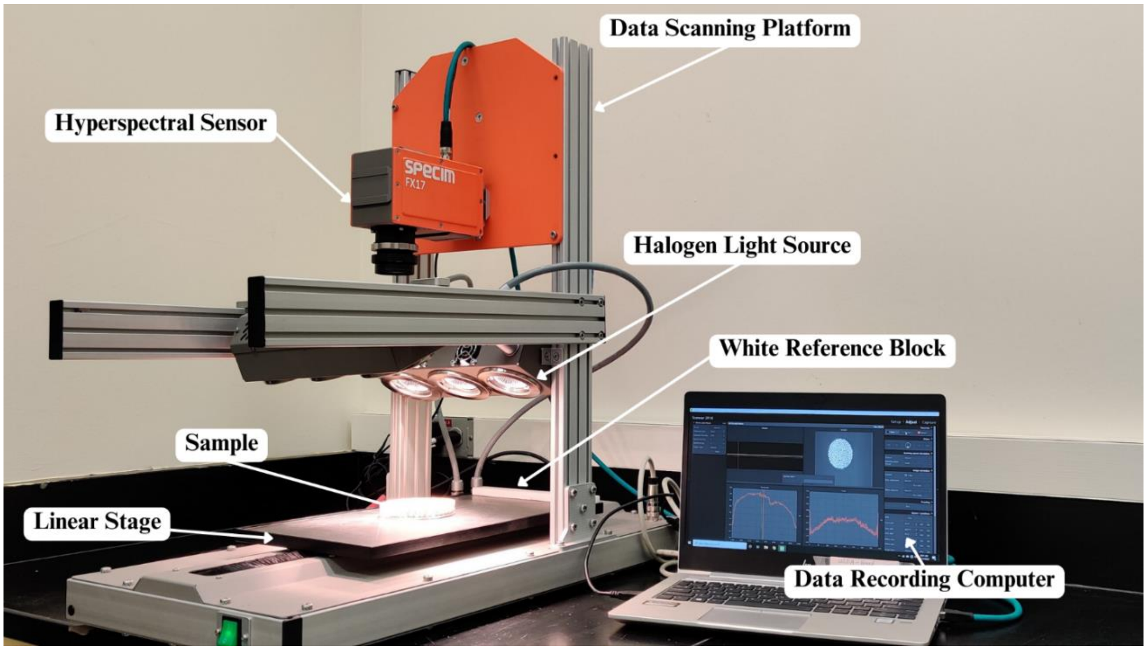


Figure 4.1 The hyperspectral imaging (HSI) system

4.3.7. Hyperspectral imaging (HSI) image processing

The HSI images were imported into MATLAB 2022a (The MathWorks, Natick, Massachusetts) and stored in a 3-D array. Each pixel value was normalized along the band axis. To increase the amount of data available for classification, each image was subdivided into 64×64 sub-pixel images. The selection of the 64×64 sub-pixel regions was carried out with the requirement that at least 30% of the pixels represented soybean seeds. Data augmentation was performed by rotating these images to 90, 180, and 270 degrees, as well as vertically and horizontally flipping the images. After these post-processing steps, a total of 25,000 images were generated for each class of soybean seeds. The processed images were saved in one document in a CSV format.

4.3.8. Feature selection of HSI

The feature selection method was adapted from Yang et al. (2021) with modifications.

4.3.8.1. Data segregation

After image processing, the dataset (CSV file) was randomly shuffled using a uniform distribution to ensure robust cross-validation later in the process. The dataset was divided into variables (X) and labels (y). The features, stored in X, consisted of columns 1 to 224 from the dataset. The labels, stored in y, consisted of column 225, with a subtraction of 1 applied to adjust for zero-based indexing. The dataset was further split into a training set (80% of data) and a testing set (20% of data) with a random seed of 0 for reproducibility.

4.3.8.2. Model training and evaluation

In this study, three distinct machine learning algorithms were utilized: Support Vector Machine (SVM), Extreme Gradient Boost (XGBoost), and Random Forest (RF), each chosen due to their unique characteristics. The SVM algorithm, configured with a linear kernel and a cost parameter set to 1, was selected for its effectiveness in high-dimensional spaces, a characteristic that makes it highly suitable for our hyperspectral data set. The XGBoost algorithm, applied both with all available features and with a subset of features selected based on their importance scores, was chosen due to its robustness to overfitting and its ability to handle a large number of features, making it ideal for feature importance analysis in our study. Finally, the RF algorithm, implemented using an ensemble of 250 decision trees, was selected for its inherent feature selection mechanism and ability to handle non-linear relationships, characteristics that are highly beneficial when dealing with complex hyperspectral data (Su et al., 2021).

The performance of the three algorithms was compared based on eight key parameters: calibration accuracy, prediction accuracy, correlation coefficients of calibration (rc), correlation coefficients of prediction (rp), coefficients of determination of calibration (Rc), coefficients of determination of prediction (Rp), root mean square error of calibration (RMSEC), and root mean square error of prediction (RMSEP). Following the detailed comparative analysis, the most efficient model was selected. In this chosen model, the ten most influential wavelengths were identified, and their respective importance scores were recorded. This step facilitated a deeper understanding of the spectral characteristics that significantly contributed to the performance of our most efficient model.

4.3.9. Model establishment using convolutional neural network (CNN)

CNN is a highly effective feed-forward network. CNN is advantageous for handling transformations such as titling, scaling, translation, and others. The CNN framework consists of two major components: the convolutional layer, which extracts features, and the pooling layer, which reduces the input data size. Using a variety of filters, the convolutional layer can extract the deep features. Using maximum or mean combinations, the pooling layer drastically reduces the number of parameters. By combining with one or more fully connected layers, the CNN outputs the highly refined features of an image

Based on the findings from the feature selection process (as outlined in Method 4.3.8.2), twenty-five thousands of 64×64×10 dimensional images were selected from each class to train the convolutional neural network (CNN). The network is composed of two convolutional layers with 32 and 64 filters each, both having a 2×2 filter size and a stride of 2. These layers were subsequently followed by a batch normalization layer, global max pooling, and four fully connected layers. The classifier was trained using 30 epochs and a randomly selected subset of 80% of the images. The remaining 20% of images were reserved as a test dataset for evaluating the performance model (Lv, Ming, Chen, & Wang, 2018). Ultimately, a predictive machine learning library was developed to facilitate future predictions based on this CNN model.

4.3.10. External validation of predictive machine learning model

The external validation of the predictive machine learning library was conducted using four untested soybean samples. These samples underwent the hyperspectral scanning process as outlined in Method 4.3.6 and the image processing procedure detailed in Method 4.3.7. The processed images were subsequently classified by the predictive machine learning library, developed in Method 4.3.9, into one of the categories defined in Method 4.3.5. Simultaneously, these four untested soybean samples were subjected to the tofu production process described in Method 4.3.3 and the tofu quality evaluation method presented in Method 4.3.4. The quality of the resulting tofu was then statistically compared to the tofu quality characteristics of the soybean category predicted by the machine learning library.

4.3.11. Statistical analysis

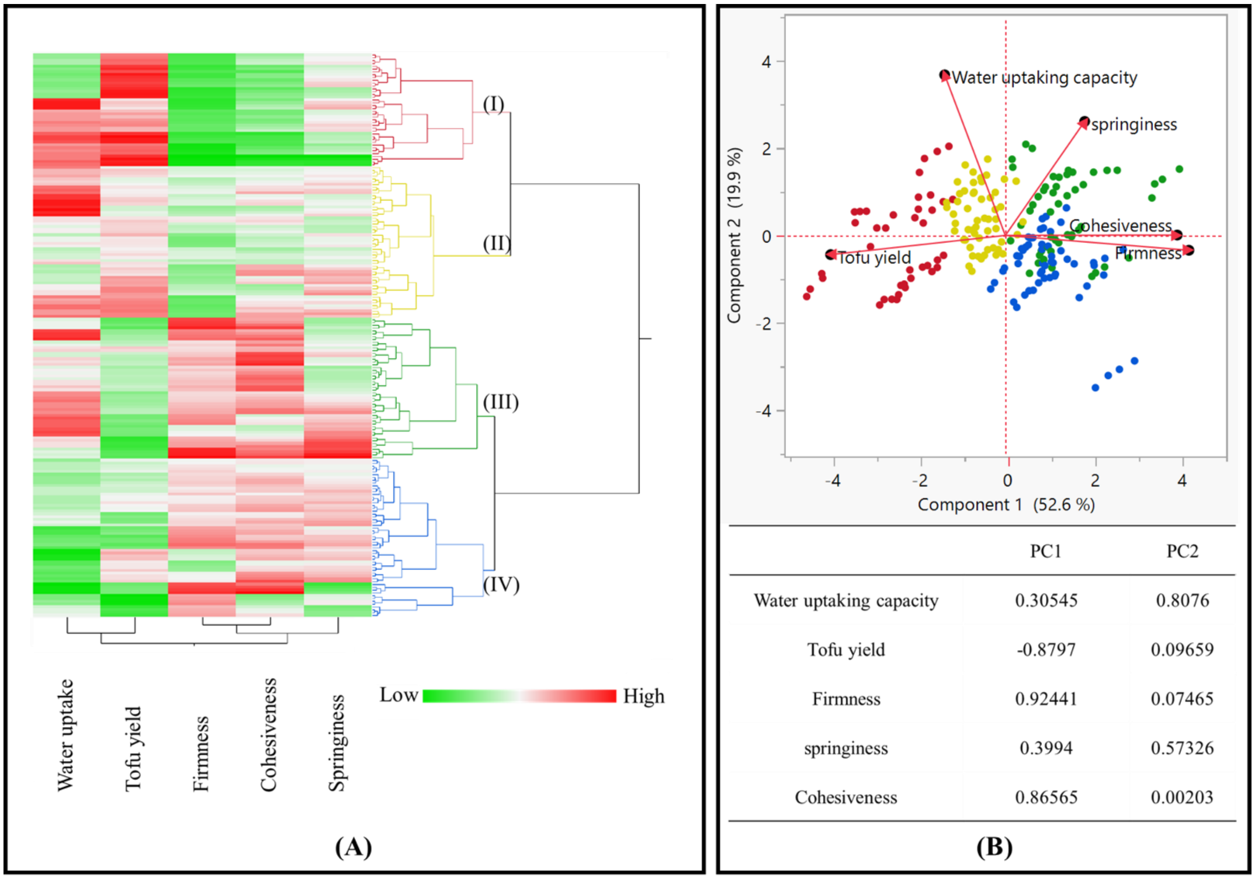
The tofu quality analysis was performed in triplicate. The data was further subjected to analysis of variance followed by Tukey's test with Statgraphics Plus 5.1 Software (Manugistics, Inc.). Differences at $p<0.05$ were considered significant.

HCA was performed on JMP® Pro 15.0.0 (SAS Institute Inc.). ENVI 5.3 (ITT Visual Information Solutions, Boulder, UT) was used to compute the spectral values of each pixel within the region of interest. MATLAB R2022a (The MathWorks, Natick, Massachusetts) was used for image processing. A 1D CNN model was constructed utilizing Python 3.8.3 and Jupyter Notebook. The CPU-based architecture of the 1D CNN model was programmed using the well-known deep learning framework Pytorch (https://pytorch.org/).

4.4. Results and discussion

4.4.1. Clustering of soybean based on tofu quality

Hierarchical clustering analysis (HCA) was utilized to sort tofu samples into different clusters. Soybean seeds were divided into four classes based on the similarity between each group regarding water uptake of soybean, yield, firmness, cohesiveness, and springiness of tofu (Figure 4.2A). PCA has also demonstrated similar results. The overall variance was explained by Principal Component 1 (PC1) and Principal Component 2 (PC2) by 72.5%, with 52.6% for PC1 and 19.9% for PC2, respectively (Figure 4.2B). Soybean seeds went from negative PC1 to positive PC1 following the group Class I, II, III, and IV. Class III and IV could not be well separated by PC1 alone; however, it is well separated by PC2. Class III was positive in PC2 while Class IV was negative in PC2. With the aid of Figure 4.2B, it was observed that soybeans in Class I and II exhibited a high water-uptake capacity and yielded a high amount of tofu. Conversely, soybeans in Class III and IV displayed higher firmness, cohesiveness, and springiness of tofu. Overall, Class I soybeans had the highest water uptake and tofu yield compared to the other classes. Class II had a lower water uptake capacity and tofu yield than Class I, but higher than Class III and IV. Additionally, the results indicated a positive correlation between tofu yield and water uptake capacity of soybean seeds. It is worth noting that Class III had a higher water-uptake capacity than Class IV, but both classes were characterized by higher values of tofu texture, such as firmness, cohesiveness, and springiness. The statistical data of Class I, II, III, and IV were listed in Table 4.1. The maximum tofu yield among the four classes was in Class I, about 3.6 kg/kg soybean seeds. The highest firmness and cohesiveness were found in Class III, which were 5.1 kg force and 0.67, respectively. The springiness had an insignificant difference ($p>0.05$) in the four classes.



	PC1	PC2
Water uptaking capacity	0.30545	0.8076
Tofu yield	-0.8797	0.09659
Firmness	0.92441	0.07465
springiness	0.3994	0.57326
Cohesiveness	0.86565	0.00203

Figure 4.2 Clustering of soybean seeds based on tofu quality using (a) hierarchical clustering analysis (HCA) and (b) principal component analysis (PCA) and the loading score of each component. The color of Class I, II, III, and IV are indicated with red, yellow, green, and blue, respectively.

In chemical terms, tofu is primarily a protein and water gel, with smaller amounts of fats, carbohydrates, and minerals. Soybean protein goes through denaturation, coagulation, and molding to hold water and soluble in the protein gel (Chen et al., 2023). The chemical composition of soybean, and processing conditions, are two major factors that affect the final quality of tofu. In this research, processing conditions have been fixed while different soybean varieties indicated the different chemical compositions are considered the only factors that affect the tofu quality.

Table 4.1 Evaluation of tofu and soybean seed quality across various soybean classes.

	Class	Min	Max	Mean
Water uptake (kg/kg soybean)	I	1.21	1.34	1.27 ± 0.05b
	II	1.23	1.34	1.28 ± 0.03b
	III	1.24	1.33	1.28 ± 0.03b
	IV	1.16	1.26	1.21 ± 0.03a
Tofu yield (kg/kg soybean)	I	2.98	3.98	3.59 ± 0.31c
	II	2.76	3.5	3.10 ± 0.21b
	III	2.06	3.11	2.50 ± 0.23a
	IV	1.87	3.26	2.62 ± 0.37a
Firmness (g force)	I	1357	2246	1959 ± 292a
	II	2255	3904	3082 ± 460b
	III	3539	7189	5092 ± 1079d
	IV	2627	6998	4483 ± 1012c
Springiness	I	0.93	0.98	0.96 ± 0.01a
	II	0.96	0.98	0.97 ± 0.01a
	III	0.95	1.00	0.97 ± 0.01a
	IV	0.94	0.98	0.97 ± 0.01a
Cohesiveness	I	0.4	0.58	0.50 ± 0.04a
	II	0.54	0.68	0.61 ± 0.04b
	III	0.56	0.75	0.67 ± 0.05d
	IV	0.51	0.77	0.64 ± 0.05c

Different letters indicate statistically significant difference within columns ($p<0.05$).

The sample number of classes I, II, III, IV is 40, 54, 50, and 56, respectively. Different letters indicate statistically significant differences ($p < 0.05$) by Tukey's The yield of tofu is intimately tied to the water-uptaking capability of the soybean seeds, a characteristic that denotes the ability of soybeans to hydrate during tofu production. Soybeans with superior water uptaking capabilities generally produce higher tofu yields compared to their less absorbent counterparts (Ali et al., 2021). This is because a higher water uptaking capacity suggests a greater water trapping capacity of the soybean protein. Ultimately, this leads to a higher yield of tofu since the weight of tofu is a sum of the weight of the solids and the absorbed water. Poysa and Woodrow

(2002) investigated ten soybean lines grown at three sources for two years. They found that higher water uptaking rate of the soybean seeds could result in the higher soymilk yield which was positively correlated with tofu yield per kilogram of soybeans.

Texture characteristics of tofu, including firmness, cohesiveness, and springiness, are fundamentally determined by the protein content and composition of the soybeans used in its production. Soybeans with a higher protein content typically produce tofu with enhanced firmness, cohesiveness, and springiness. However, it is noteworthy that as soybean seeds hydrate, the protein content becomes diluted, leading to a reduction in these texture attributes. This observation underpins the “Yield and Texture Trade-off Theory” that while a high-water uptaking capacity could lead to a high water content in the soymilk, resulting in a higher yield of tofu, it could simultaneously dilute the protein concentration in the soymilk. This dilution potentially diminishes tofu texture attributes such as firmness, cohesiveness, and springiness. Supporting this notion, Mujoo, Trinh, and Ng (2003) conducted a study on seven soybean varieties harvested from Michigan. Their research indicated that tofu firmness declined from 10.02 to 7.84 N as tofu yield increased from 2.93 to 3.43 kg/kg of soybeans, illustrating the balance between tofu yield and its textural attributes.

Contrarily, Class III soybeans serve as a counterexample to this theory, as their higher water-absorption capacity results in lower tofu yield and superior tofu texture. This implies that protein content is not the sole determinant of tofu quality and yield. Guan et al. (2021) underscored the influence of protein subunits on tofu yield and quality. To illustrate, soybeans with a lower 11S/7S ratio form a uniformly aggregated spherical gel, while beans with a higher 11S/7S ratio exhibit higher macroscopic phase separation, a coarser network structure, and larger pores (James & Yang, 2016). The role of amino acids in influencing tofu quality has also been reported. Coagulants such as calcium or magnesium salts are commonly used to bind the negatively charged amino acids together, forming a network-like structure (Ali et al., 2021; James & Yang, 2014). Given this information, it is plausible that the protein subunit composition and amino acid profile of Class III soybean seeds may vary significantly from those of Class IV.

In summary, soybean seeds from Class I, II, III, and IV each possess unique characteristics that influence the quality of tofu produced from them. The categorization of these soybean seeds provided valuable data for the application of supervised machine-learning techniques in the following research.

4.4.2. Hyperspectral imaging (HSI) of soybean seeds

4.4.2.1. Spectra of soybean HSI

The general trends of the HSI curves within the 900–1700 nm wavelength range were found to be quite similar (Figure 4.3A). However, the peak intensity of each soybean seed varied, ranging from 40 to 120. To better understand the relationship between the HSI data and tofu quality, the spectra were averaged and grouped into the four previously established classes of soybeans using HCA (Figure 4.3B). The intensity of the HSI spectra followed an overall order of Class I > II > III > IV, corresponding to the quality of tofu produced. These findings suggest that a predictive model could be established based on the HSI data and related parameters of tofu quality. However, with 224 wavelengths for each HSI curve, the dataset can be large, leading to potential computational complexity and noise in the predictive model (Ishida et al., 2018). As such, the following methods will explore ways to reduce the number of wavelengths in the dataset.

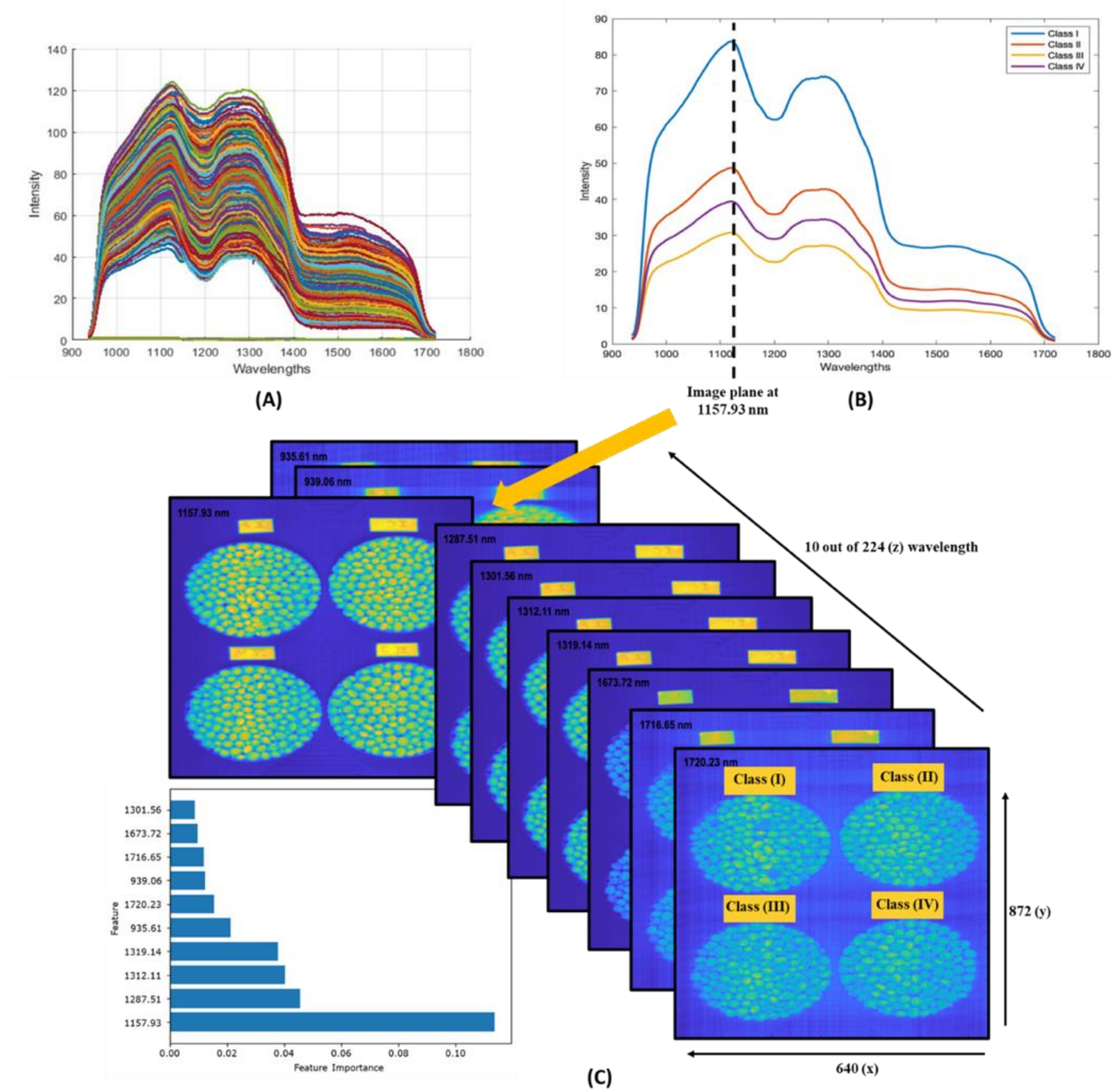


Figure 4.3 Hyperspectral imaging (HSI) profile of soybean seeds at the spectral range spanned from 900 to 1700 nm. (a) the HSI wavelength profile of all the soybeans; (b) the HSI wavelength profile of classified soybeans; (c) images of soybeans at ten featured wavelengths. The 10 featured wavelengths represented by the image planes were acquired by XGBoost with the feature importance listed.

Moreover, these observations imply that a predictive model could be built based on the HSI data and related parameters of tofu quality. However, as each HSI curve has 224 wavelengths, the dataset can be substantial, leading to potential computational complexity and noise in the predictive model (Loggenberg & Poona, 2022; Pal et al., 2020; Warner & Shank, 1997).

4.4.2.2. Selection of featured wavelengths

The HSI commonly includes neighboring bands that are highly correlated, causing problems of multicollinearity among closely positioned wavelength variables. To address this, featured wavelength selection is used to decrease data dimensionality and conserve storage space while preserving essential information. This strategy lessens collinearity issues, strengthens model resilience by reducing wavelength count, and potentially enhances model performance in accuracy and generalization.

Support Vector Machine (SVM), Extreme Gradient Boost (XGBoost), and Random Forest (RF) were widely applied in searching for the featured wavelength of HSI (Huang et al., 2022; Pal et al., 2020). Support Vector Machine (SVM) is capable of handling high-dimensional data effectively. This is particularly important in HSI where the number of features (wavelengths) can be very large. By using a linear kernel, the SVM is looking for a linear combination of featured wavelengths that best separates the classes. This makes the interpretation of the model simpler, as the weight given to each wavelength in the final model represents its importance (Huang et al., 2017). XGBoost introduces a regularization term on the basis of the gradient boosting algorithm, utilizes the second-order Taylor expansion for fitting residuals, and can be calculated in parallel, so it has the advantages of anti-overfitting and high computational efficiency (Liao et al., 2019). RF is a robust, scalable, and flexible algorithm that can handle complex and noisy data, identify the most informative bands, capture non-linear relationships, and reduce overfitting in HSI analysis (Qin et al., 2013).

In this study, both the XGBoost and RF algorithms showcased high calibration accuracy, exceeding 99%, as represented in **Table 4.2**. However, the SVM algorithm exhibited a significantly lower calibration accuracy of just 53.8%. Regarding prediction accuracy, despite its commendable performance, RF achieved a comparatively lower prediction accuracy of 56.4%, suggesting that the featured wavelengths selected by this method had limited predictive power. The XGBoost algorithm stood out with a good prediction accuracy of 99.5%, underscoring its superior capability in this context. These findings align with a similar study by Pal, Charan & Poriya that also reported the superior performance of the XGBoost algorithm for feature selection in their datasets, while the RF-based approach caused a drop in classification accuracy (Pal et al., 2020). Upon delving deeper into the XGBoost parameters, we found that the correlation coefficients of prediction (r_p) and coefficients of determination of prediction (R_p) were 85.5% and 76.3% respectively. Consequently, given its remarkable prediction accuracy of 99.5%, XGBoost was chosen as the optimal algorithm for selecting the featured wavelengths in this study.

Table 4.2 Performance of featured wavelength selected by different models.

Model	**Calibration accuracy**	$\mathbf{r_C}$	$\mathbf{R_C}$	**RMSEC**	**Prediction accuracy**	$\mathbf{r_P}$	$\mathbf{R_P}$	**RMSEP**
XGBoost	0.997	0.998	0.995	0.077	0.995	0.855	0.763	0.544
RF	1.000	1.000	1.000	0.000	0.564	-0.090	-0.090	1.163
SVM	0.538	0.135	0.135	1.037	0.534	0.122	0.122	1.049

Abbreviations: Support Vector Machine (SVM), Extreme Gradient Boost (XGBoost), and Random Forest (RF), correlation coefficients of calibration (rc), correlation coefficients of prediction (rp), coefficients of determination of calibration (Rc), coefficients of determination of prediction (Rp), root mean square error of calibration (RMSEC), and root mean square error of prediction (RMSEP).

It is known that the visible and near-infrared spectra (900-2500 nm) of soybean seeds mainly provide chemical information about the components such as protein, oil and water with the bands of O-H, N-H, and C-H groups (Sun et al., 2020; Teye, Anyidoho, Agbemafle, Sam-Amoah, & Elliott, 2020). The difference in the reflectance is due to the variation in the content and structure of protein, and oil, reflecting different varieties of soybean seeds moreover, it is also due to the physical properties of the light when interact with matter such as light scattering effects. Soybean seeds contain a lot of proteins, oils, and carbohydrates, but the chemical composition varies largely by the method of cultivation, temperature, sun, and rainfall (Song et al., 2016).

Ten featured wavelengths from XGBoost were 935.62, 939.08, 1157.93, 1287.51, 1301.50, 1312.11, 1319.14, 1673.72, 1716.65, and 1720.23 nm, respectively (Figure 4.2C). According to Table 4.3 (Curran, 1989), there were several compounds observed using the featured wavelengths like oil (935.62 nm, 939.08 nm), proteins (1157.93 nm, 1673.72 nm, 1716.65 nm, 1720.23 nm), water (1287.51 nm), cellulose (1287.51 nm), and lignin (1287.51 nm, 1673.72 nm, 1716.65 nm, 1720.23 nm). Those results covered the major three chemical components, including protein, oil, and carbohydrates, in the featured wavelengths. Another research suggested that absorption at 1187 nm (-CH), 1496 nm (-NH), 1674 nm (-CH), 1743 nm (-CH), 1980 nm (-NH), 2055 nm (-ROH/NH), and 2167 nm (-NH) increased as the protein content increased (Ingle et al., 2016). Therefore, the ten featured wavelengths were good indicators of protein quality of soybean seeds. The relationship between those wavelengths to the functionalities of soybean flour needs to be further studied.

Table 4.3 Featured wavelengths and the corresponding bonds.

Wavelength (nm)	**Bond Vibration**	**Chemicals**
935.61	C-H Stretch	Oil
939.06	C-H Stretch	Oil
1157.93	N-H Stretch	Protein
1287.51	O-H Bend, 1st Overtone	Water, Cellulose, Lignin
1673.72	C-H Stretch, 1st Overtone	Protein, Lignin, Nitrogen
1716.65	C-H Stretch, 1st Overtone	Protein, Lignin, Nitrogen
1720.23	C-H Stretch, 1st Overtone	Protein, Lignin, Nitrogen

The data is cited from (Curran, 1989)

4.4.3. Predicting tofu quality based on HSI with CNN model

4.4.3.1. Establishment of CNN model

Spectral data is rich in complex features, making it an ideal candidate for analysis using CNN. These models, characterized by their extensive architectures, offer advantages over traditional classifiers by extracting more abstract data features, leading to heightened performance levels. Although the training time for Convolutional Neural Networks (CNN) tends to be longer compared to other models, the trade-off is a superior performance, particularly in image classification tasks, where CNNs are considered one of the most effective algorithms (Zhou et al., 2019).

In this study, CNN was employed to develop a model based on ten selected spectral bands of interest. The essential parameters of this model are outlined in Table 4.4. A predictive model was established using the developed algorithm, which was subsequently verified by inputting random soybean seed images and evaluating the accuracy of its class predictions.

For this assessment, one hundred images were utilized from each class to determine the prediction percentages. As indicated in Table 4.5, Class I achieved a 98% prediction rate, Class II had a 99% prediction rate, and Class III also had a 99% prediction rate, while Class IV only attained a 96% prediction rate. These high prediction rates serve as a testament to the testing accuracy of model, demonstrating its efficacy and robustness in predicting soybean seed classifications.

While there have been no reported applications of CNN in predicting food processes, there is a growing body of research leveraging CNN for food quality and safety prediction. For instance, Yu et al. (2018) employed a deep learning model to analyze visible/near-infrared hyperspectral data from shrimps, aiming to predict their freshness. They used a Stacked Autoencoder (SAE) model to extract deep features from the samples, and then applied logistic regression to classify the freshness grade of shrimp based on these features. This novel approach yielded impressive results, with calibration and prediction set accuracies reaching 96.55% and 93.97% respectively, demonstrating the potential of deep learning methods in food quality assessment. Similar applications can be found in an illustrative study on the use of CNNs for HSI analysis. Qiu et al. (2018) explored the potential of CNNs to identify rice seed varieties. Significantly, the CNN model outperformed the SVM model in most scenarios, with an impressive total accuracy rate of 89.6%, showcasing the effectiveness of CNNs in analyzing spectral data. Our research demonstrates the promising application of CNNs in hyperspectral imaging for food product prediction, especially in predicting the tofu quality based on soybean seeds. The findings suggest that with the aid of rapid sample collection through hyperspectral imaging, CNNs, and HSI are a good combination to predict food quality based on the ingredients profile.

Table 4.4 Parameters of CNN architecture used in this work

Layer Type	Architecture Unit	Kernel Size	Output Size
C1	Rectified Linear Unit	3 × 3	32×32×32
C2	Rectified Linear Unit	3 × 3	16×16×64
Batch Normalization			16×16×64
Global Max Pooling			64
Dense			4
Output	SoftMax		

Table 4.5 Prediction of tofu quality with CNN based on 10 featured wavelengths

Class	No. of Samples Run	Prediction accuracy (%)
I	100	98
II	100	99
III	100	99
IV	100	96

4.4.3.2. Verification of CNN model

Four untested soybean seeds were employed to evaluate the quality of soybean seeds scanned with HSI. The resulting images were processed and fed into the Convolutional Neural Network (CNN) model, which classified the seeds into Class I, II, III, and IV. Tofu made from these soybeans was evaluated for quality using **Methods 4.3.4**, and the results were presented in **Figure 4.4**. In general, the CNN model accurately predicted the quality of soybeans in Class II and Class III, and most parameters for Class I and Class IV were also well predicted. Nonetheless, there are certain limitations to these results. Specifically, each quality parameter of tofu made from Class II soybeans fell within the interquartile range (IQR), while those of Class III tofu were situated between the lower and upper whiskers. These results are considered acceptable because the predicted tofu quality remains within the range of the training dataset.

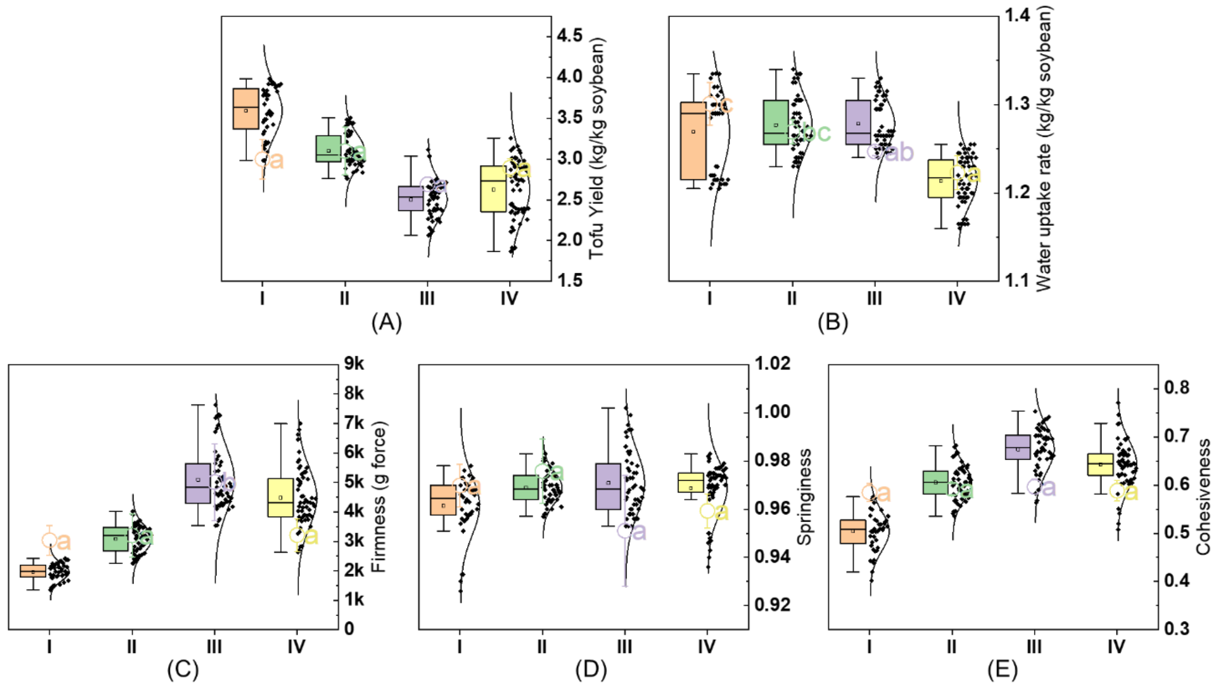


Figure 4.4 Verification of soybean seeds with tofu quality (a) tofu yield, (b) water uptaking capacity, (c) firmness, (d) springiness, and (e) cohesiveness. Note: The circle symbol indicated the mean value of tested tofu quality. The line of each box from top to bottom indicates upper whisker, upper quartile, median, lower quartile, and lower whisker. The black dots indicate the parameter values in the training dataset. Different letters indicate statistically significant differences ($p < 0.05$).

Although most quality parameters for Class I and Class IV also fell within the whisker range, there were outliers in the predicted results. The springiness of tofu (Figure 4.4D) made with predicted Class IV soybeans was lower than the lower whisker, which can be considered an outlier; this may be attributed to the presence of outliers in the training dataset itself. Conversely, the firmness of Class I tofu (Figure 4.4C) was higher than the upper whisker, indicating a model prediction outlier. In addition to these outliers, some predicted parameters were close to the whiskers, such as the water uptake capacity of Class I and Class III, and the tofu yield of Class IV. These results may be due to the bimodal distribution (Scheres, 2010; Xu et al., 2022) of the training dataset, as illustrated in Figure 4.4A&B.

Classifying soybeans into more categories using the training dataset might help reduce outliers and the bimodal distribution. However, increasing the number of categories may lead to fewer samples per category, potentially causing issues such as overfitting, high variance, or inappropriate model selection (Scheres, 2010). The optimal solution would be to collect additional data to enhance the performance of the machine learning model.

4.5. Conclusions

This study successfully determined ten featured wavelengths from Hyperspectral Imaging (HSI) data, spanning 200 soybean varieties, with the help of the XGBoost algorithm. These wavelengths potentially correlate with the protein, carbohydrate, and oil contents in the soybean seeds. However, further validation is needed to substantiate the relationship between these wavelengths and the respective chemical compositions.

A CNN model for predicting tofu quality has been successfully developed based on these ten featured wavelengths from the HSI data. This model, trained on data from 200 soybean varieties, is capable of classifying soybeans into four distinct classes using HSI images of individual seeds. The predictive accuracy for each class of soybeans impressively ranges from 96% to 99%.

The robustness of this model was further validated using untested soybean samples. These samples were accurately categorized into distinct classes, each representing a specific range of tofu quality parameters. Upon comparison, it was observed that the model accurately predicted the majority of tofu quality traits.

This research sets the groundwork for understanding the relationship between hyperspectral image, chemical composition, and tofu qualities. The feasibility of predicting tofu quality based on the hyperspectral image has been demonstrated; however, the machine learning prediction model requires further enhancements. It is recommended to collect more soybean samples to classify seeds into additional categories. This would equip the prediction model with a more comprehensive ability to accurately estimate tofu quality based on a diverse set of quality parameters.

5. OVERALL SUMMARY AND CONCLUSION

5.1. Conclusion

In conclusion, this comprehensive study delved into the intricate relationship between soybean seed characteristics, sources, and tofu quality parameters. Through the analysis of 178 soybean varieties, we identified distinct clusters based on protein and moisture content, shedding light on the impact of sourcing on these key attributes. These findings emphasized the significance of considering soybean origin when aiming to optimize tofu quality.

Furthermore, our exploration of tofu quality parameters using hierarchical cluster analysis and Principal Component Analysis offered valuable insights into the multifaceted nature of tofu quality. We found that various factors, including brix, texture, and moisture content, played pivotal roles in defining tofu quality. These findings provide essential guidance for producers and researchers seeking to tailor tofu to meet specific consumer preferences.

Additionally, this study pioneered the use of cutting-edge technologies such as Hyperspectral Imaging and machine learning to predict tofu quality directly from soybean seed characteristics. By successfully classifying soybean varieties into four distinct quality categories with an impressive accuracy ranging from 96% to 99%, our research represents a groundbreaking advancement in the field. This innovative approach has the potential to revolutionize the traditional assessment of soybean seeds for tofu production, offering a more efficient and resource-saving method.

In summary, our study not only deepened our understanding of the factors influencing tofu quality but also introduced a novel and accurate method for assessing soybean seed quality for tofu production. These findings have significant implications for the tofu industry, paving the way for improved quality control and consumer satisfaction.

5.2. Future research

In the future research, we can consider including the following aspects:

- Diverse Samples: Collecting an extensive and diverse dataset of soybeans from various sources worldwide, representing different growing conditions and climates, to ensure the model's robustness and applicability across different regions.
- Enhanced Hyperspectral Data Collection: Implementing an expanded data collection strategy for the hyperspectral imaging, ensuring a more comprehensive coverage of various chemical compositions within soybean seeds, thereby improving the training process of the machine learning model.
- Integration of Advanced Machine Learning Techniques: Exploring advanced machine learning methodologies beyond the conventional approaches, such as deep learning algorithms, recurrent neural networks, or generative adversarial networks, to further improve the accuracy and robustness of the predictive model for soybean seed chemical composition analysis.
- Comprehensive Validation and Testing Protocols: Designing rigorous validation and testing procedures to thoroughly assess the model's performance and reliability under diverse conditions, ensuring its efficacy in practical applications and commercial settings.
- Multi-dimensional Analysis: Integrating additional parameters or features, such as environmental factors, soil characteristics, and agricultural practices, to develop a more comprehensive understanding of the intricate relationship between soybean seed chemical composition and its environmental determinants.

- By incorporating the following suggestions into the future research, we can contribute to the advancement of hyperspectral imaging applications for soybean seed analysis and enhance the robustness and applicability of the machine learning model for comprehensive chemical composition detection.

REFERENCES

Ali, F., Tian, K., & Wang, Z. X. (2021). Modern techniques efficacy on tofu processing: A review. Trends in Food Science & Technology, 116, 766-785. https://doi.org/10.1016/j.tifs.2021.07.023

Anjum, S., Agnihotri, V., Rana, S., Pandey, A., & Pande, V. (2023). The Impact of Processing Methods and Conditions on Nutritional Properties of Soybean-Based Tofu: A Review. *Journal of Food Engineering and Technology*, *12*(1), 1-13. https://doi.org/10.32732/jfet.2023.12.1.1

Beleia, A., Prudencio-Ferreira, S. H., Yamashita, F., Sakamoto, T. M., & Ito, L. (2004). Sensory and instrumental texture analysis of cassava (*Manihot esculenta*, crantz) roots. *Journal of Texture Studies*, *35*(5), 542-553. https://doi.org/10.1111/j.1745-4603.2004.35505.x

Bi, C. H., Li, D., Wang, L. J., & Adhikari, B. (2013). Viscoelastic properties and fractal analysis of acid-induced SPI gels at different ionic strength. *Carbohydrate Polymers*, *92*(1), 98-105. https://doi.org/10.1016/j.carbpol.2012.08.081

Cai, T. D., Chang, K. C., Shih, M. C., Hou, H. J., & Ji, M. (1997). Comparison of bench and production scale methods for making soymilk and tofu from 13 soybean varieties. *Food research international*, *30*(9), 659-668. https://doi.org10.1016/s0963-9969(98)00032-5

Cai, T., & Chang, K. C. (1999). Processing effect on soybean storage proteins and their relationship with tofu quality. *Journal of Agricultural and Food chemistry*, *47*(2), 720-727. https://doi.org/10.1021/jf980571z

Chang, Y. Y., Li, D., Wang, L. J., Bi, C. H., & Adhikari, B. (2014). Effect of gums on the rheological characteristics and microstructure of acid-induced SPI-gum mixed gels. *Carbohydrate polymers*, *108*, 183-191. https://doi.org/10.1016/j.carbpol.2014.02.089

Chen, C. C., Hsieh, J. F., & Kuo, M. I. (2023). Insight into the processing, gelation and functional components of tofu: A review. *Processes*, *11*(1), 202. https://doi.org/10.3390/pr11010202

Colletti, A., Attrovio, A., Boffa, L., Mantegna, S., & Cravotto, G. (2020). Valorisation of by-products from soybean (Glycine max (L.) Merr.) processing. *Molecules*, *25*(9, 2129), 1- 33. https://doi.org/10.3390/molecules25092129

Coradi, P. C., Lima, R. E., Alves, C. Z., Teodoro, P. E., & Cândido, A. C. D. S. (2020). Evaluation of coatings for application in raffia big bags in conditioned storage of soybean cultivars in seed processing units. *PLoS One*, *15*(11), e0242522. https://doi.org/10.1371/journal.pone.0242522

Curran, P. J. (1989). Remote sensing of foliar chemistry. *Remote Sensing of Environment*, *30*(3), 271–278. https://doi.org/10.1016/0034-4257(89)90069-2

Enriko, I. K. A., Suryanegara, M., & Gunawan, D. (2016). Heart disease prediction system using k-Nearest neighbor algorithm with simplified patient's health parameters. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, *8*(12), 59-65. https://doi.org/jtec/article/view/1436/947

Fukushima, D. (2009). Recent progress of soybean protein foods: Chemistry, technology, and nutrition. *Food Reviews International*, *7*(3), 323–351. https://doi.org/10.1080/87559129109540915

Gao, H., Xu, J., Tan, M., Mu, D., Li, X., Zhao, Y., & Zheng, Z. (2021). Effect of high-intensity ultrasound soymilk pretreatment on the physicochemical properties of microbial transglutaminase-catalyzed tofu gel. *Journal of Food Science*, *86*(6), 2410-2420. https://doi.org/10.1111/1750-3841.15735

Gao, T., Chandran, A. K. N., Paul, P., Walia, H., & Yu, H. (2021). HyperSeed: An end-to-end method to process hyperspectral images of seeds. *Sensors 2021*, *21*(24), 8184. https://doi.org/10.3390/s21248184

Genuer, R., Poggi, J. M., Genuer, R., & Poggi, J. M. (2020). *Random forests* (pp. 33-55). Springer International Publishing. https://doi.org/10.1007/978-3-030-56485-8_3

Gowen, A. A., O'Donnell, C. P., Cullen, P. J., Downey, G., & Frias, J. M. (2007). Hyperspectral imaging–an emerging process analytical tool for food quality and safety control. *Trends in food science & technology*, *18*(12), 590-598. https://doi.org/10.1016/j.tifs.2007.06.001

Guan, X., Zhong, X., Lu, Y., Du, X., Jia, R., Li, H., & Zhang, M. (2021). Changes of soybean protein during tofu processing. *Foods*, *10*(7), 1594. https://doi.org/10.3390/foods10071594

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN model-based approach in classification. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003,* *Catania, Sicily, Italy, November 3-7, 2003. Proceedings* (pp. 986-996). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-39964-3_62

Hang, J., Shi, D., Neufeld, J., Bett, K. E., & House, J. D. (2022). Prediction of protein and amino acid contents in whole and ground lentils using near-infrared reflectance spectroscopy. *LWT*, *165*, 113669. https://doi.org/10.1016/j.lwt.2022.113669

Hartman, G. L., Pawlowski, M. L., Herman, T. K., & Eastburn, D. (2016). Organically grown soybean production in the USA: Constraints and management of pathogens and insect pests. *Agronomy*, *6*(1), 16. https://doi.org/10.3390/agronomy6010016

He, W., He, H., Wang, F., Wang, S., Li, R., Chang, J., & Li, C. (2022). Rapid and uninvasive characterization of bananas by hyperspectral imaging with extreme gradient boosting (XGBoost). *Analytical Letters*, *55*(4), 620–633. https://doi.org/10.1080/87559129109540915

Hou, H. J., & Chang, K. C. (2004). Storage conditions affect soybean color, chemical composition and tofu qualities. *Journal of Food Processing and Preservation*, *28*(6), 473-488. https://doi.org/10.1111/j.1745-4549.2004.24015.x

Huang, L., Liu, Y., Huang, W., Dong, Y., Ma, H., Wu, K., & Guo, A. (2022). Combining random forest and XGBoost methods in detecting early and mid-term winter wheat stripe rust using canopy level hyperspectral measurements. *Agriculture*, *12*(1), 74. https://doi.org/10.3390/agriculture12010074

Huang, L., Zhou, Y., Meng, L., Wu, D., & He, Y. (2017). Comparison of different CCD detectors and chemometrics for predicting total anthocyanin content and antioxidant activity of mulberry fruit using visible and near infrared hyperspectral imaging technique. *Food Chemistry*, *224*, 1–10. https://doi.org/10.1016/j.foodchem.2016.12.037

Hui, T., & Xing, G. (2022). Effect of transglutaminase pre-crosslinking treatment incorporated with glucono-δ-lactone on the physicochemical and digestive properties of tofu. *Polymers*, *14*(12), 2364. https://doi.org/10.3390/polym14122364

Ishida, T., Kurihara, J., Viray, F. A., Namuco, S. B., Paringit, E. C., Perez, G. J., Takahashi, Y., & Marciano, J. J. (2018). A novel approach for vegetation classification using UAV-based hyperspectral imaging. *Computers and Electronics in Agriculture*, *144*, 80–85. https://doi.org/10.1016/j.compag.2017.11.027

Jain, S., Paventhan, A., Chinnaiyan, V. K., Arnachalam, V., & Pradish, M. (2014, March). Survey on smart grid technologies-smart metering, IoT and EMS. In *2014 IEEE Students' Conference on Electrical, Electronics and Computer science* (pp. 1-6). IEEE. https://doi.org/10.1109/sceecs.2014.6804465

James, A. T., & Yang, A. (2014). Influence of globulin subunit composition of soybean proteins on silken tofu quality. 2. Absence of 11SA4 improves the effect of protein content on tofu hardness. *Crop and Pasture Science*, *65*(3), 268–273. https://doi.org/10.1071/CP13399

James, A. T., & Yang, A. (2016). Interactions of protein content and globulin subunit composition of soybean proteins in relation to tofu gel properties. *Food Chemistry*, *194*, 284-289. https://doi.org/10.1016/j.foodchem.2015.08.021

Joo, K. H., & Cavender, G. A. (2020). Investigation of tofu products coagulated with trimagnesium citrate as a novel alternative to nigari and gypsum: Comparison of physical properties and consumer preference. *LWT, 118,* 108819. https://doi.org/10.1016/j.lwt.2019.108819

Jurado, J. M., Pádua, L., Hruška, J., Feito, F. R., & Sousa, J. J. (2021). An efficient method for generating UAV-based hyperspectral mosaics using push-broom sensors. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 6515-6531. https://doi.org/10.1109/jstars.2021.3088945

Kao, F. J., Su, N. W., & Lee, M. H. (2004). Effect of water-to-bean ratio on the contents and compositions of isoflavones in tofu. *Journal of agricultural and food chemistry, 52*(8), 2277-2281. https://doi.org/10.1021/jf035410w

Karamah, E. F., Zahirsyah, M. Z., & Najeges, R. R. (2021). Preliminary Study of Ozonated Water to Preserve the Quality of Tofu: The Effect of Exposure Time and Replacement of Ozonated Water. *International Journal on Advanced Science, Engineering and Information Technology*, *11*(3), 962-967. https://doi.org/10.18517/ijaseit.11.3.13130

Kim, M. K., Kim, J. H., Nam, S. J., Ryu, S., & Kong, G. (2008). Dietary intake of soy protein and tofu in association with breast cancer risk based on a case-control study. *Nutrition and cancer, 60*(5), 568-576. https://doi.org/10.1080/01635580801966203

Kohyama, K., & Nishinari, K. (1993). Rheological studies on the gelation process of soybean 7 S and 11 S proteins in the presence of glucono-. delta.-lactone. *Journal of Agricultural and Food Chemistry, 41*(1), 8-14. https://doi.org/0021-8561/93/1441-0008

Kumar, V., Rani, A., Goyal, L., Dixit, A. K., Manjaya, J. G., Dev, J., & Swamy, M. (2010). Sucrose and raffinose family oligosaccharides (RFOs) in soybean seeds as influenced by genotype and growing location. *Journal of agricultural and food chemistry*, *58*(8), 5081-5085. https://doi.org/10.1021/jf903141s

Kurasch, A. K., Hahn, V., Miersch, M., Bachteler, K., & Würschum, T. (2018). Analysis of tofu-related traits by a bench-scale tofu production method and their relationship with agronomic traits in European soybean. *Plant Breeding*, *137*(3), 271-282. https://doi.org/10.1111/pbr.12581

Lee, J., Kao, H. A., & Yang, S. (2014). Service innovation and smart analytics for industry 4.0 and big data environment. *Procedia cirp*, *16*, 3-8. https://doi.org/10.1016/j.procir.2014.02.001

Li, J., An, X., Li, Q., Wang, C., Yu, H., Zhou, X., & Geng, Y. A. (2022). Application of XGBoost algorithm in the optimization of pollutant concentration. *Atmospheric Research*, *276*, 106238. https://doi.org/10.1016/j.atmosres.2022.106238

Li, M., Chen, F., Yang, B., Lai, S., Yang, H., Liu, K., ... & Deng, Y. (2015). Preparation of organic tofu using organic compatible magnesium chloride incorporated with polysaccharide coagulants. *Food Chemistry*, *167*, 168-174. https://doi.org/10.1016/j.foodchem.2014.06.102

Liao, X., Cao, N., Li, M., & Kang, X. (2019). Research on short-term load forecasting using XGBoost based on similar days. *2019 International Conference on Intelligent Transportation, Big Data and Smart City*, 675–678. https://doi.org/10.1109/icitbs.2019.00167

Liao, X., Wang, S., Li, Y., Michael Olajide, T., Zhai, X., Qian, J., Miao, S., & Huang, J. (2022). Effects of "nine steaming nine sun-drying" on proximate composition, protein structure and volatile compounds of black soybeans. *Food Research International*, *155*, 111070. https://doi.org/10.1016/j.foodres.2022.111070

Lim, B. T., Deman, J. M., Deman, L., & Buzzell, R. I. (1990). Yield and quality of tofu as affected by soybean and soymilk characteristics: Calcium sulfate coagulant. Journal of Food Science, 55(4), 1088–1092. https://doi.org/10.1111/j.1365-2621.1990.tb01605.x

Liying, Z., Li, D., Qiao, S., Johnson, E. W., Li, B., Thacker, P. A., & Han, I. K. (2003). Effects of stachyose on performance, diarrhoea incidence and intestinal bacteria in weanling pigs. *Archives of Animal Nutrition*, *57*(1), 1-10. https://doi.org/10.1080/0003942031000086662

Loggenberg, K., & Poona, N. (2022). A feature selection approach for terrestrial hyperspectral image analysis. *South African Journal of Geomatics*, *9*(2), 302–320. https://doi.org/10.4314/sajg.v9i2.20

Lusas, E. W., Erickson, D. R., & Nip, W. K. (Eds.). (1989). *Food uses of whole oil and protein seeds*. The American Oil Chemists Society.

Lv, X., Ming, D., Chen, Y. Y., & Wang, M. (2018). Very high resolution remote sensing image classification with SEEDS-CNN and scale effect analysis for superpixel CNN classification. *International Journal of Remote Sensing*, *40*(2), 506–531. https://doi.org/10.1080/01431161.2018.1513666

MacMillan, K. P., & Gulden, R. H. (2020). Effect of seeding date, environment and cultivar on soybean seed yield, yield components, and seed quality in the Northern Great Plains. *Agronomy journal*, *112*(3), 1666-1678. https://doi.org/10.1002/agj2.20185

Meng, S., Chang, S., Gillen, A. M., & Zhang, Y. (2016). Protein and quality analyses of accessions from the USDA soybean germplasm collection for tofu production. *Food Chemistry*, *213*, 31–39. https://doi.org/10.1016/j.foodchem.2016.06.046

Mujoo, R., Trinh, D. T., & Ng, P. K. W. (2003). Characterization of storage proteins in different soybean varieties and their relationship to tofu yield and texture. *Food Chemistry*, *82*(2), 265–273. https://doi.org/10.1016/S0308-8146(02)00547-2

Müller, D. H., Camili, E. C., Scaramuzza, W. L. M. P., & Albuquerque, M. C. D. F. (2018). Spatial variability in soybean seeds quality and in chemical attributes of the soil of a production field in the Brazilian Cerrado. *Journal of Seed Science*, *40*, 25-35. https://doi.org/10.1590/2317-1545v40n1182375

Muraro, M. A., Giampietro, P. G., & Galli, E. (2002). Soy formulas and nonbovine milk. *Annals of Allergy, Asthma & Immunology*, *89*(6), 97-101. https://doi.org/10.1016/s1081-1206(10)62132-1

Osborne, B. G. (2006). Near-infrared spectroscopy in food analysis. *Encyclopedia of analytical chemistry: applications, theory and instrumentation*. https://doi.org/10.1002/9780470027318.a1018

Pal, M., Charan, T. B., & Poriya, A. (2020). K-nearest neighbour-based feature selection using hyperspectral data. *Remote Sensing Letters*, *12*(2), 128–137. https://doi.org/10.1080/2150704x.2020.1864051

Pan, Z., & Tangratanavalee, W. (2003). Characteristics of soybeans as affected by soaking conditions. *LWT-Food Science and Technology*, *36*(1), 143-151. https://doi.org/10.1016/s0023-6438(02)00202-5

Pang, L., Men, S., Yan, L., & Xiao, J. (2020). Rapid vitality estimation and prediction of corn seeds based on spectra and images using deep learning and hyperspectral imaging techniques. *IEEE Access*, *8*, 123026–123036. https://doi.org/10.1109/access.2020.3006495

Panthee, D. R., Pantalone, V. R., West, D. R., Saxton, A. M., & Sams, C. E. (2005). Quantitative trait loci for seed protein and oil concentration, and seed size in soybean. *Crop Science*, *45*(5), 2015-2022. https://doi.org/cropsci2004.0720

Parsa, M. (2021). A data augmentation approach to XGboost-based mineral potential mapping: an example of carbonate-hosted ZnPb mineral systems of Western Iran. *Journal of Geochemical Exploration*, *228*, 106811. https://doi.org/10.1016/j.gexplo.2021.106811

Parsa, N., & Byrne, M. F. (2021). Artificial intelligence for identification and characterization of colonic polyps. *Therapeutic advances in gastrointestinal endoscopy*, *14*, 26317745211014698. https://doi.org/10.1177/26317745211014698

Peng, X., Ren, C., & Guo, S. (2016). Particle formation and gelation of soymilk: Effect of heat. *Trends in Food Science & Technology*, *54*, 138-147. https://doi.org/10.1016/j.tifs.2016.06.005

Poudyal, H., Kumar, S. A., Iyer, A., Waanders, J., Ward, L. C., & Brown, L. (2013). Responses to oleic, linoleic and α-linolenic acids in high-carbohydrate, high-fat diet-induced metabolic syndrome in rats. *The Journal of nutritional biochemistry*, *24*(7), 1381-1392. https://doi.org/10.1016/j.jnutbio.2012.11.006

Poysa, V., & Woodrow, L. (2002). Stability of soybean seed composition and its effect on soymilk and tofu yield and quality. *Food Research International*, *35*(4), 337–345. https://doi.org/10.1016/s0963-9969(01)00125-9

Poysa, V., Woodrow, L., & Yu, K. (2006). Effect of soy protein subunit composition on tofu quality. *Food Research International*, *39*(3), 309-317. https://doi.org/10.1016/j.foodres.2005.08.003

Qiao, Q. (2022). Image Processing Technology Based on Machine Learning. *IEEE Consumer Electronics Magazine*. https://doi.org/10.1109/mce.2022.3150659

Qin, Q., Wang, Q., Li, J., & Ge, S. S. (2013). Linear and nonlinear trading models with gradient boosted random forests and application to Singapore stock market. *Journal of Intelligent Learning Systems and Applications*, *2013*(01), 1–10. https://doi.org/10.4236/jilsa.2013.51001

Qiu, Z., Chen, J., Zhao, Y., Zhu, S., He, Y., & Zhang, C. (2018). Variety identification of single rice seed using hyperspectral imaging combined with convolutional neural network. *Applied Sciences*, *8*(2), 212. https://doi.org/10.3390/app8020212

Rekha, C. R., & Vijayalakshmi, G. (2013). Influence of processing parameters on the quality of soycurd (tofu). *Journal of Food Science and Technology*, *50*, 176-180. https://doi.org/10.1007/s13197-011-0245-z

Russ, J. C. (2006). *The image processing handbook*. CRC press. https://doi.org/10.1201/9780203881095

Saio, K., Koyama, E., & Watanabe, T. (1968). Protein-calcium-phytic acid relationships in soybean: Part II. effects of phytic acid on combination of calcium with soybean meal protein. *Agricultural and Biological Chemistry*, *32*(4), 448-452. https://doi.org/10.1080/00021369.1968.10859080

Schaefer, M. J., & Love, J. (1992). Relationships between soybean components and tofu texture 1. *Journal of Food Quality*, *15*(1), 53-66. https://doi.org/10.1111/j.1745-4557.1992.tb00975.x

Scheres, S. H. W. (2010). Classification of structural heterogeneity by maximum-likelihood methods. *Methods in Enzymology*, *482*(C), 295–320. https://doi.org/10.1016/s0076-6879(10)82012-9

Song, W., Yang, R., Wu, T., Wu, C., Sun, S., Zhang, S., Jiang, B., Tian, S., Liu, X., & Han, T. (2016). Analyzing the effects of climate factors on soybean protein, oil contents, and composition by extensive and high-density sampling in China. *Journal of Agricultural and Food Chemistry*, *64*(20), 4121–4130. https://doi.org/10.1021/acs.jafc.6b00008

Stanojevic, S. P., Barac, M. B., Pesic, M. B., & Vucelic-Radovic, B. V. (2011). Assessment of soy genotype and processing method on quality of soybean tofu. *Journal of Agricultural and Food Chemistry*, *59*(13), 7368–7376. https://doi.org/10.1021/jf2006672

Su, W. H., Yang, C., Dong, Y., Johnson, R., Page, R., Szinyei, T., Hirsch, C. D., & Steffenson, B. J. (2021). Hyperspectral imaging and improved feature variable selection for automated determination of deoxynivalenol in various genetic lines of barley kernels for resistance screening. *Food Chemistry*, *343*, 128507. https://doi.org/10.1016/j.foodchem.2020.128507

Sun, J., Wang, G., Zhang, H., Xia, L., Zhao, W., Guo, Y., & Sun, X. (2020). Detection of fat content in peanut kernels based on chemometrics and hyperspectral imaging technology. *Infrared Physics & Technology*, *105*, 103226. https://doi.org/10.1016/j.infrared.2020.103226

Teye, E., Anyidoho, E., Agbemafle, R., Sam-Amoah, L. K., & Elliott, C. (2020). Cocoa bean and cocoa bean products quality evaluation by NIR spectroscopy and chemometrics: A review. *Infrared Physics & Technology*, *104*, 103127. https://doi.org/10.1016/j.infrared.2019.103127

Ting, Ching-Hua, Feng-Jui Kuo, Cheng-Chang Lien, and Chung-Teh Sheng. (2009). "Use of ultrasound for characterising the gelation process in heat induced $CaSO_4 \cdot 2H_2O$ tofu curd." *Journal of Food Engineering* 93(1), 101-107. https://doi.org/10.1016/j.jfoodeng.2009.01.015

Toda, K., Ono, T., Kitamura, K., Hajika, M., Takahashi, K., & Nakamura, Y. (2003). Seed protein content and consistency of tofu prepared with different magnesium chloride concentrations in six Japanese soybean varieties. *Breeding science*, *53*(3), 217-223. https://doi.org/10.1270/jsbbs.53.217

Wang, C. C. R., & Chang, S. K. C. (1995). Physicochemical properties and tofu quality of soybean cultivar Proto. *Journal of Agricultural and Food Chemistry*, *43*(12), 3029-3034. https://doi.org/0021-8561/95/1443-3029

Wang, F., Meng, J., Sun, L., Weng, Z., Fang, Y., Tang, X., Zhao, T., & Shen, X. (2020). Study on the tofu quality evaluation method and the establishment of a model for suitable soybean varieties for Chinese traditional tofu processing. *LWT*, *117*, 108441. https://doi.org/10.1016/j.lwt.2019.108441

Wang, W., Bringe, N. A., Berhow, M. A., & Gonzalez de Mejia, E. (2008). β-Conglycinins among sources of bioactives in hydrolysates of different soybean varieties that inhibit leukemia cells in vitro. *Journal of agricultural and food chemistry*, *56*(11), 4012-4020. https://doi.org/10.1021/jf8002009

Wang, Y., Yang, X., & Li, L. (2020). A new style of fermented tofu by Lactobacillus casei combined with salt coagulant. *3 Biotech*, *10*, 1-8. https://doi.org/10.1007/s13205-019-2040-x

Warner, T. A., & Shank, M. C. (1997). Spatial autocorrelation analysis of hyperspectral imagery for feature selection. *Remote Sensing of Environment*, *60*(1), 58–70. https://doi.org/10.1016/s0034-4257(96)00138-1

Xu, M., Jin, Z., Lan, Y., Rao, J., & Chen, B. (2019). HS-SPME-GC-MS/olfactometry combined with chemometrics to assess the impact of germination on flavor attributes of chickpea, lentil, and yellow pea flours. *Food Chemistry*, *280*, 83-95. https://doi.org/10.1016/j.foodchem.2018.12.048

Xu, P., Tan, Q., Zhang, Y., Zha, X., Yang, S., & Yang, R. (2022). Maize seed classification and recognition based on machine vision and deep learning. *Agriculture*, *12*(2), 232. https://doi.org/10.3390/agriculture12020232

Yadav, D. N., Chauhan, G. S., Chauhan, O. P., Sharma, P., & Bajpai, A. (2003). Quality evaluation of curd prepared from milk-soymilk blends. *Journal of food science and technology (Mysore)*, *40*(4), 403-405. https://doi.org/15376075

Yang, G., Wang, Q., Liu, C., Wang, X., Fan, S., & Huang, W. (2018). Rapid and visual detection of the main chemical compositions in maize seeds based on Raman hyperspectral imaging.

Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, *200*, 186–194. https://doi.org/10.1016/j.saa.2018.04.026

Yang, L., Gao, H., Meng, L., Fu, X., Du, X., Wu, D., & Huang, L. (2021). Nondestructive measurement of pectin polysaccharides using hyperspectral imaging in mulberry fruit. *Food Chemistry*, *334*, 127614. https://doi.org/10.1016/j.foodchem.2020.127614

Yu, X., Tang, L., Wu, X., & Lu, H. (2018). Nondestructive freshness discriminating of shrimp using visible/near-infrared hyperspectral imaging technique and deep learning algorithm. *Food Analytical Methods*, *11*(3), 768–780. https://doi.org/10.1007/S12161-017-1050-8

Yuan, S., & Chang, S. K. C. (2007). Texture profile of tofu as affected by instron parameters and sample preparation, and correlations of instron hardness and springiness with sensory scores. *Journal of Food Science*, *72*(2), S136-S145. https://doi.org/10.1111/j.1750-3841.2006.00263.x

Zarkadas, C. G., Gagnon, C., Poysa, V., Khanizadeh, S., Cober, E. R., Chang, V., & Gleddie, S. (2007). Protein quality and identification of the storage protein subunits of tofu and null soybean genotypes, using amino acid analysis, one-and two-dimensional gel electrophoresis, and tandem mass spectrometry. *Food Research International*, *40*(1), 111-128. https://doi.org/10.1016/j.foodres.2006.08.005

Zhang, Q., Wang, C., Li, B., Li, L., Lin, D., Chen, H., Liu, Y., Li, S., Qin, W., Liu, J., Liu, W., & Yang, W. (2017). Research progress in tofu processing: From raw materials to processing conditions. *Critical Reviews in Food Science and Nutrition*, *58*(9), 1448-1467. https://doi.org/10.1080/10408398.2016.1263823

Zhang, T., Wei, W., Zhao, B., Wang, R., Li, M., Yang, L., Wang, J., & Sun, Q. (2018). A reliable methodology for determining seed viability by using hyperspectral data from two sides of wheat seeds. *Sensors, 18*(3), 813. https://doi.org/10.3390/S18030813

Zhao, H., Zhang, F., Hu, H., Liu, S., & Han, J. (2017). Experimental study on freezing of liquids under static magnetic field. *Chinese Journal of Chemical Engineering*, *25*(9), 1288-1293. https://doi.org/10.1016/j.cjche.2016.10.026

Zhou, L., Zhang, C., Liu, F., Qiu, Z., & He, Y. (2019). Application of deep learning in food: A Review. *Comprehensive Reviews in Food Science and Food Safety*, *18*(6), 1793–1811. https://doi.org/10.1111/1541-4337.12492

Zhu, Q., Wu, F., Saito, M., Tatsumi, E., & Yin, L. (2016). Effect of magnesium salt concentration in water-in-oil emulsions on the physical properties and microstructure of tofu. *Food Chemistry, 201*, 197-204. https://doi.org/10.1016/j.foodchem.2016.01.065