MAKERERE UNIVERSITY

COLLEGE OF ENGINEERING, DESIGN, ART AND TECHNOLOGY

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Non-Destructive and Rapid Prediction of Protein and Fat Content in Black Soldier Fly Larvae Using Hyperspectral Imaging and Machine Learning

by

Immaculate Babirye 21/U/1048

Jovia Namulinda Nabwire 21/U/1352

Main Supervisor: Mr. Frank Ssemakula

Co-Supervisor: Dr. Roseline Akol

A Final Year Project Report Submitted in Partial Fulfilment of the Requirements for the Award of the Degree of Bachelor of Science in Electrical Engineering of Makerere University

June 2025

Declaration

I declare that this report is our original work, submitted in partial fulfillment of the requirements for the Bachelor of Science in Electrical Engineering degree at Makerere University. This document has not been submitted to any other institution. Additionally, no section of this report has been reproduced improperly or without permission. Citations, quotations, and references to other authors' work or sources of information have been appropriately acknowledged throughout.

Approval

This is to certify that, under the supervision of Mr.Frank Ssemakula and Dr.Roseline Akol of the Department of Computer and Electrical engineering, and in partial fulfilment of the requirements for the award of a Bachelor's of science in Electrical Engineering.

Main Supervisor

Mr.Frank Ssemakula

Signature: _____ Date: _____

Co-Supervisor

Dr.Roseline Akol

Signature: _____ Date: _____

Acknowledgement

I would like to sincerely acknowledge the invaluable support and contributions of the FLYgene Project at Makerere University, the Centre for Quantitative Genetics and Genomics at Aarhus University, funded by the Ministry of Foreign Affairs of Denmark under grant 21-09-AU, and the National Agricultural Research Organisation (NARO). Their generous support, resources, and expertise were instrumental in the successful completion of this research. I am deeply grateful for the opportunity to benefit from their guidance, collaboration, and commitment to advancing scientific research and innovation.

Abstract

The accurate determination of protein and fat content in Black Soldier Fly Larvae (BSFL) is essential for their utilization in sustainable animal feed production. Traditional methods such as Kjeldahl and Solvent extraction for assessing these nutritional parameters are time-consuming, destructive, and require extensive chemical analysis. This study presents a non-destructive, rapid, and reliable machine learning model for predicting BSFL protein and fat content using Hyperspectral Imaging (HSI). Hyperspectral camera was used to capture a broad range of spectral information from BSFL enabling detailed analysis of chemical composition without physical sample destruction. A dataset of 200 hyperspectral images of BSFL samples was collected using XIMEA IMEC camera and preprocessed to extract spectral features. These features were then used to train and validate machine learning models such as the Partial Least Squares Regression (PLSR) and the Support Vector Machine Regression (SVMR). The models' performance was then evaluated using metrics like the coefficient of determination (R^2) and Root Mean Square Error (RMSE). Results demonstrated that the developed models accurately predicted BSFL protein and fat content, with the SVMR yielding the best prediction, characterized by R^2 values closer to 1 compared to the PLSR. This approach provides a rapid, non-destructive alternative to traditional chemical analyses, making it suitable for real-time monitoring and quality control in BSFL production. The study emphasizes the potential of integrating hyperspectral imaging and machine learning for advancing precision nutrition analysis in the insect protein industry.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Networks
BSF	Black Soldier Fly
BSFL	Black Soldier Fly Larvae
HBO	boric acid
HSI	Hyperspectral Imaging
HSO	sulfuric acid
ISTE	In Situ Transesterification
MAE	Mean Absolute Error
ML	Machine Learning
MSE	Mean Squared Error
PCA	Principal Component Analysis
PLSR	Partial Least Squares Regression
RMSE	Root Mean Square Error
ROI	Region of Interest
SDK	Software Development Kit
SG	Savitzky–Golay
SVM	Support Vector Machine
SVMR	Support Vector Machine Regression

1. Introduction

As the global population rises from 7.7 billion in 2019 to 9.7 billion by 2050, there is increasing focus on alternative protein and fat sources which offer both nutritional benefits and environmental sustainability [1] particularly in the context of animal feed. Black Soldier Fly Larvae (BSFL) (*Hermetia illucens*) have emerged as a promising candidate for protein and fat-rich feed, primarily due to their ability to convert organic waste into valuable biomass efficiently. According to the available literature, BSFL has been adopted as a feed additive for different animals and poultry.

The nutritional content of BSFL, especially protein and fat, varies depending on factors such as diet, growth stage, and environmental conditions. Accurate and rapid assessment of these nutritional components is crucial for optimizing BSFL farming operations and ensuring the quality of the larvae used for animal feed. Currently, this assessment is being done using the soxhlet and kjeldahl methods. However, these have their own short comings. The Soxhlet extraction method presents multiple challenges for accurately quantifying fatty acids in microalgae. It often yields incomplete extractions, particularly with more resistant fatty acids, resulting in lower yields. Solvent choice also affects Soxhlet's reliability, with polar solvents providing higher yields than non-polar ones. Additionally, Soxhlet lacks the acidic conditions necessary to release certain fatty acids fully, often requiring pre-treatment with acid hydrolysis [12], which complicates the procedure.

On the other hand, the Kjeldahl method for protein assessment has several limitations, including a lengthy, labor-intensive process that involves multiple digestion,

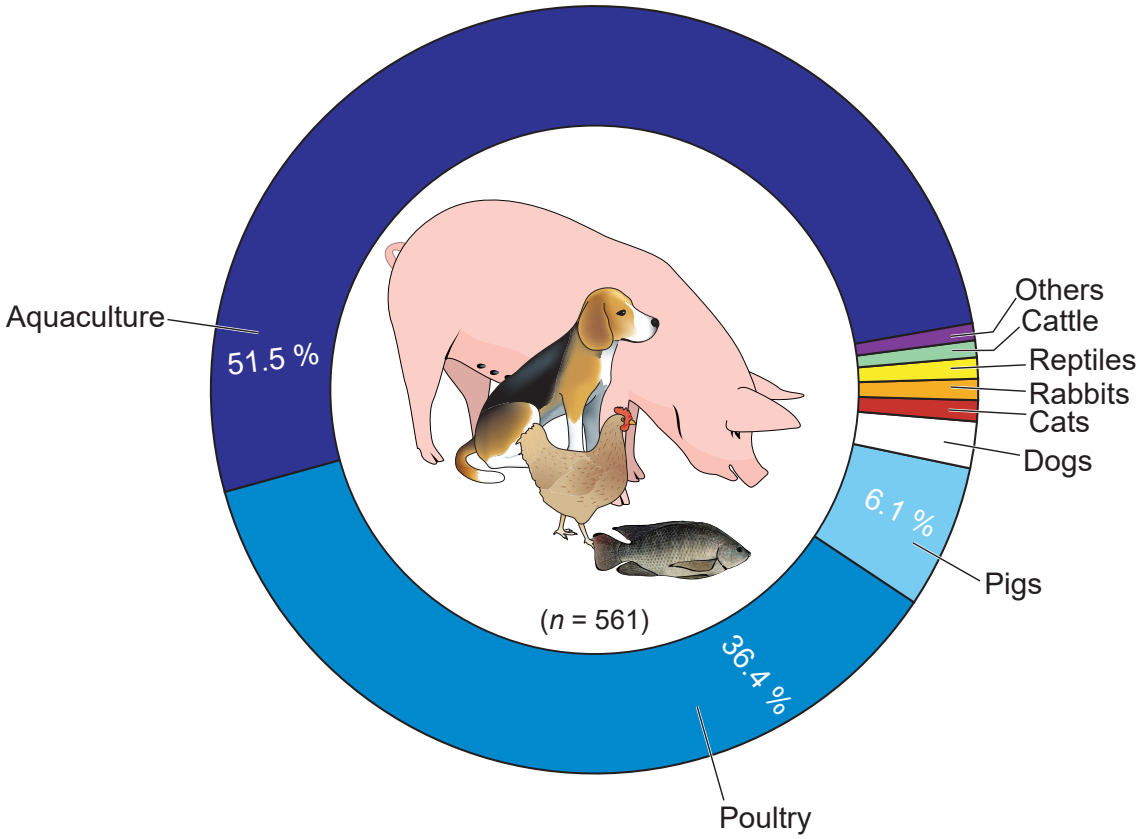

Figure 1.1: Adoption of Black Soldier Fly Larvae as feed additive

cooling, and distillation steps, making it unsuitable for high-throughput testing. It uses hazardous chemicals like concentrated sulfuric acid and sodium hydroxide, posing safety and environmental risks. Additionally, the method measures total nitrogen, leading to potential inaccuracies in protein estimation due to interference from non-protein nitrogen sources. The manual handling of samples and reagents also introduces variability, which can affect reproducibility. These challenges make

Kjeldahl less efficient [13] than automated alternatives.

HSI, a non-invasive and highly precise technology, has shown promise in the agricultural and food industries for predicting the composition of various biological materials [2]. By capturing data across a wide range of wavelengths, HSI can detect subtle chemical and structural differences in biological samples. Recent studies have explored its potential in predicting protein, fat, and moisture content in insects and other food sources. In this context, machine learning models will be employed to interpret hyperspectral data and make accurate predictions about fat and protein content in BSFL, offering an efficient, real-time solution to a growing industry need.

1.1 Background

Food quality and safety are pressing issues that continue to garner significant social attention. Recently, there has been an increasing demand for real-time information regarding food quality, prompting a shift away from traditional manual sensory testing and chemical analysis methods [4], which can be time-consuming and destructive. This transition highlights the strong potential of non-destructive testing techniques within the food supply chain.

With advancements in computer science and spectroscopic technologies, Machine Learning (ML) and Hyperspectral Imaging (HSI) have emerged as highly effective methods for evaluating the sensory characteristics and quality attributes of food products. These innovative techniques enable rapid and non-destructive assessments, making them invaluable for ensuring food safety and quality. By integrating ML with HSI, our study aims to harness these tools to predict protein and fat content in BSF larvae, offering a more efficient and reliable approach to food quality evaluation.

1.2 Problem Statement

Traditional methods for analyzing the nutritional content of BSFL such as the Kjeldahl and Soxhlet often require destructive sampling, expensive lab equipment, and highly trained personnel. These drawbacks hinder real-time quality control and large-scale monitoring [10].

1.3 Justification

A machine learning approach coupled with hyper-spectral imaging provides a non-destructive, rapid, and reliable solution for determining protein and fat levels in black soldier fly larvae. This reduces the need for a lab environment, use of chemicals, highly trained personnel, and also reduces prediction time. It addresses the critical needs of the BSF industry by offering faster, accurate, and scalable solutions. The use of this approach not only enhances the economic viability of BSF production but also supports sustainability goals through the reduction of chemical waste and optimization of resources.

1.4 Objectives of the Study

1.4.1 Main Objective

To develop a non-destructive and rapid machine learning model that accurately predicts the protein and fat content in BSF larvae using hyperspectral imaging data.

1.4.2 Specific Objectives

1. To obtain data from BSF larvae samples using XMEA xiQ Hyperspectral camera.
2. To obtain reference data for protein and fat using the standard chemical meth-

ods, Soxhlet and Kjeldahl.

3. To design and train a machine learning model using the cleaned data for the prediction.
4. To evaluate the model performance and accuracy in predicting the protein and fat content using the Root Mean Squared Error and the Coefficient of determination.

1.5 Scope

This study focuses on developing a machine learning-based regression model to predict the protein and fat content of black soldier fly larvae using hyperspectral imaging data. It aims to streamline the analysis process, making it faster and more efficient compared to traditional laboratory methods.

1.6 Significance of the Study

The study holds several key significances; it aims to significantly reduce the time required to determine protein and fat content compared to traditional methods. Additionally, by excluding the need for a lab environment and chemicals, the study offers a more efficient and eco-friendly approach. The model also supports selective breeding by providing accurate predictions, aiding in the enhancement of desired traits. Moreover, it is non-destructive, helping preserve the species while ensuring sustainable research practices.

2. Literature Review

This section contains a review of the Black Soldier Fly (*Hermetia illucens*) and its significance as a sustainable protein and fat source. It also explores the role of hyperspectral imaging (HSI) and Machine Learning (ML) techniques in enhancing non-destructive assessments while determining the protein and fat content in BSFL.

2.1 Black Soldier Fly

The Black Soldier Fly (BSF), scientifically known as Hermetia illucens, is originally from the Americas but has adapted well to subtropical and tropical regions worldwide. Unlike other flies, BSF doesn't pose any health risks to humans; it isn't a pest and doesn't carry diseases. The adult flies focus only on drinking water and steer clear of humans—they don't bite or sting.

What makes BSF particularly interesting is how easy they are to rear, even on a large scale. They don't need a lot of space or specialized equipment, making them suitable for various farming setups. At the pre-pupae stage, they even have the ability to move themselves out of the growth medium, making the harvesting process more efficient [10].

The life cycle of the Black Soldier Fly (BSF) involves a complete metamorphosis, transitioning through several distinct stages: egg, larva, pupa, and adult. It all starts when a female BSF lays between 200 to 800 eggs, usually in clusters near decaying organic matter. These eggs hatch within 2 to 4 days, releasing larvae that immediately begin feeding on organic waste. During this stage, the larvae grow rapidly,

shedding their outer layer through several molts as they progress through multiple larval stages. This phase of consuming organic material is crucial for their development, allowing them to store energy for the next stage.

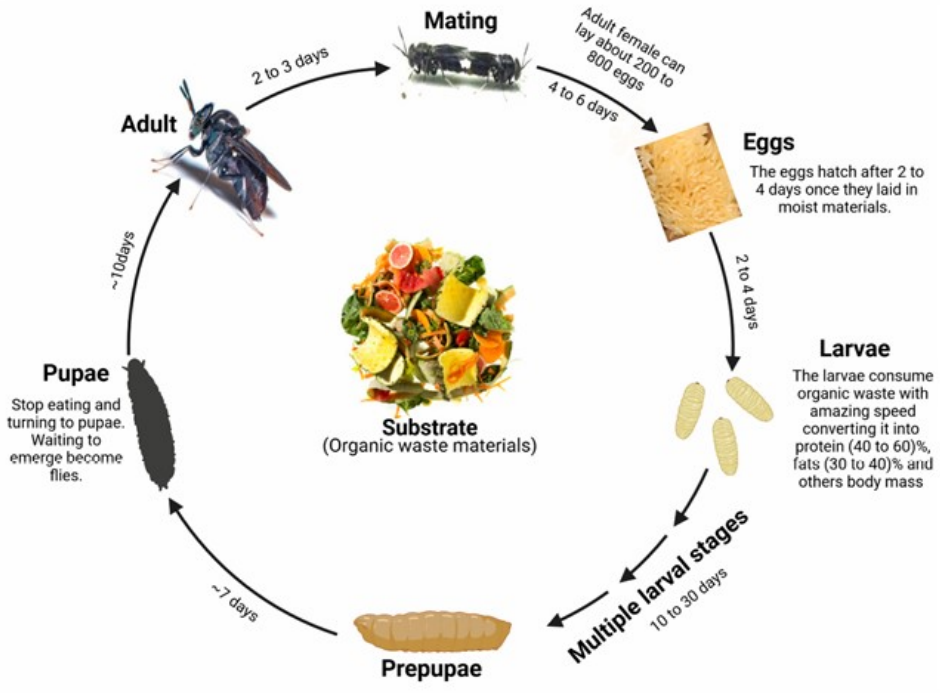

Figure 2.1: The life cycle of a black soldier fly (BSF) [15]

After a period of rapid growth, the larvae enter the pupal stage, where they transform into a resting phase. The pupal stage typically lasts around 10 days, during which the larvae develop into their final form. Once fully developed, adult flies emerge from the pupae, ready to continue the cycle. Within 2 to 3 days, the adult flies begin to mate, restarting the process with a new generation. The entire cycle showcases the

BSF's adaptability and efficiency, making it an effective decomposer and a valuable component in waste management and sustainable practices.

BSFL present a sustainable alternative to traditional feed sources like soybean meal and fishmeal, which are often associated with environmental issues such as deforestation and overfishing. These conventional ingredients can also represent a significant portion of poultry production costs. In comparison, BSFL can thrive on various organic wastes, helping to reduce environmental impact while delivering a high-quality protein source. This dual benefit—providing essential nutrients while reducing waste—has sparked significant interest in the animal feed industry, offering both economic and ecological advantages.

BSFL, have gained attention as a valuable ingredient in broiler chicken diets due to their rich protein content, high nutritional value, and eco-friendly benefits. These larvae excel at converting organic waste into a nutrient-dense biomass filled with proteins and lipids, making them both practical and sustainable. Depending on their growth stage and diet, BSFL can contain between 32% to 53% protein [15], providing essential amino acids that support muscle development and growth in chickens and fat levels ranging from 18% to 33% [15]. The notable fat content of BSFL significantly influences poultry feed formulation, necessitating adjustments in nutrient balance and energy density.

Existing Methods for Determining Protein and Fat content in BSFL

1. Fat Content Analysis

   The fat content of the sample is analyzed using a gravimetric method [10].

This process involves a direct extraction using a Soxhlet apparatus. First, 2 grams of the sample are placed into prepared hulls lined with cotton, which are then sealed with cotton at the top. The hulls are dried in an oven at 105°C for about 10-15 minutes [16]. After drying, the hulls are placed inside the Soxhlet apparatus, which is connected to a 300ml [3] flask containing dried and weighed boiling stones. Hexane is added to the apparatus until the hulls are fully submerged, and the setup is equipped with a condenser and water bath. The extraction process takes place over 3 hours.

Following extraction, the hexane solvent is distilled off, leaving behind the fat residue. This residue is further dried in an oven at 105°C [16]to ensure all solvent traces are removed. After drying, the flask containing the fat residue is cooled to room temperature in a desiccator before being weighed. The drying process is repeated until a consistent weight was achieved, ensuring accurate measurement of the fat content.

2. Protein Content Analysis

   Protein content analysis is conducted using a titrimetric method. To begin, about 1 gram of the sample is prepared, with 0.5 grams weighed out and placed into a 300 ml Kjeldahl tube. The sample is then mixed with 1 gram of a selenium catalyst and 12 ml of concentrated sulfuric acid (HSO) [10] before being subjected to digestion at 420°C for 1 hour using a Kjeldahl apparatus.

   Once digestion is complete, the sample was allowed to cool to room temperature. The digested sample was then transferred to a distillation unit, where 50 ml of distilled water and an excess of 40% sodium hydroxide (NaOH) were

added. A 250 ml Erlenmeyer flask containing 25 ml of 4% boric acid (HBO) [11] is set up as a reservoir for the distillation. The sample is distilled for 4 minutes, during which the solution in the reservoir changes color from red to green, indicating the collection of ammonia.

The collected distillate is then titrated with a 0.2 N hydrochloric acid (HCl) [10] solution until the color shifts from green back to red, marking the endpoint. This titration process allows for athe ccurate determination of the protein content in the sample.

2.2 Hyperspectral Imaging(HSI)

HSI is a powerful technology with applications across various fields, ranging from scientific research to practical uses in industries. It is particularly effective in analyzing and interpreting complex visual data, making it invaluable in areas such as astronomy, medicine, forensics, military operations, security, and environmental studies. This method allows experts—whether researchers, medical professionals, or scientists—to gain deeper insights that can lead to improved diagnostics, enhanced research outcomes, and more effective evidence collection.

Unlike traditional imaging methods, hyperspectral imaging captures a wide range of wavelengths across the electromagnetic spectrum. This enables the identification and characterization of materials based on their unique spectral properties, providing data that goes beyond what is possible with conventional color or multispectral imaging techniques [7]. By analyzing these wavelengths, researchers can extract detailed information about the materials and surfaces being studied.

One of the key advantages of hyperspectral imaging is its ability to provide detailed pattern recognition, including tasks such as classification and anomaly detection. The imaging process generates a three-dimensional data cube, with axes representing spatial dimensions (x and y) and the spectral dimension, which corresponds to hundreds or even thousands of wavelengths. Each pixel in this cube contains brightness values for each wavelength, offering a rich dataset for analysis.

The primary goal of hyperspectral imaging is not merely to produce visually appealing images but to capture the unique spectral signature of each object. This signature allows for the precise identification of materials, as each substance has a distinct way of reflecting or absorbing light across various wavelengths. By capturing this spectral information, hyperspectral imaging enables detailed analysis and identification that is crucial for applications where precision and specificity are key.

A hyperspectral imaging system in line-scan mode generally consists of four basic components: illumination, an imaging spectrograph, a camera, and a zoom lens. Figure 2:2 shows the schematic diagram and hardware components of the hyperspectral imaging system is used for measuring the spatial distribution [5] of diffuse reflectance.

2.3 Machine Learning for Prediction

Machine Learning (ML) is a branch of Artificial Intelligence (AI) that uses algorithms to perform tasks by recognizing patterns within data, rather than relying on direct programming for each specific action. The process involves setting a clear goal—such as predicting a numerical value, assessing [6] the model's performance against that goal, and then iteratively refining the model through repeated experimentation.

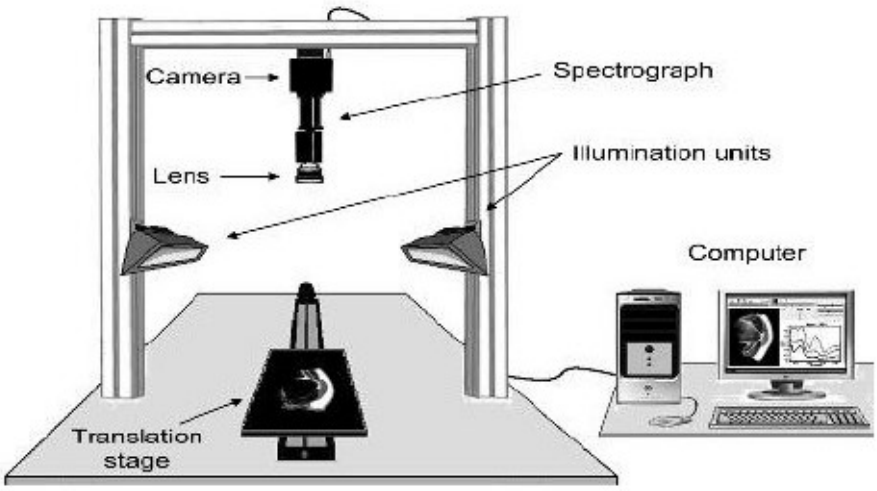

Figure 2.2: Configuration of the hyperspectral imaging system [5]

Advantages of Machine Learning

Machine Learning (ML) offers several practical advantages that make it an appealing choice for many applications. One of the key benefits is that the computer essentially learns on its own how to accomplish tasks, eliminating the need for extensive programming or manual instructions. This approach not only saves valuable time and effort but also enhances the system's ability to adapt and tackle a variety of problems.

Moreover, by leveraging existing data to predict outcomes, ML algorithms can significantly reduce the time and costs [6] associated with experimental verification. Instead of conducting lengthy experiments to validate results, organizations can rely on ML predictions to inform their decisions more quickly and efficiently.

Additionally, the tools needed to carry out ML experiments are quite minimal. Typically, all that's required is a computer, a virtual environment for development, and a relevant dataset. This accessibility allows researchers and practitioners to easily implement and experiment with ML techniques, driving innovation and insights across diverse fields.

2.3.1 Application of HSI and Machine Learning

The application of HSI combined with ML models for BSFL is a relatively new but growing field. The combination of HSI with machine learning (ML) techniques has enabled the development of predictive models [14] that can analyze complex datasets and accurately estimate the nutritional content of food and biological materials [2]. Commonly used algorithms include Partial Least Squares Regression (PLSR), Support Vector Machines (SVM), and Artificial Neural Networks (ANN), which are capable of extracting patterns from hyperspectral data. [1] In their study used two portable NIR spectrometers (900–1700 nm and 1350–2562 nm) and chemometric models—Partial Least Square Regression (PLSR) and Support Vector Machine Regression (SVMR)—to predict protein and lipid content in black soldier fly (BSF) larvae flour. Spectral data was analyzed with principal component analysis (PCA), while regression models were applied for prediction. Both models performed well for protein content, with spectrometer 2's broader range showing slightly better accuracy. For lipid content, SVMR outperformed PLSR, with spectrometer 2 achieving high predictive accuracy (RMSEP = 3.51% and RPD = 4.32). Variable selection methods further refined model performance by focusing on relevant wavelengths.

2.4 Related studies

Tirado and his team in their study used two portable NIR spectrometers (900–1700 nm and 1350–2562 nm) and chemometric models Partial Least Square Regression (PLSR) and Support Vector Machine Regression (SVMR) to predict protein and lipid content in BSF larvae flour. [1] Spectral data was analyzed with principal component analysis (PCA), while regression models were applied for prediction. Both models performed well for protein content, with spectrometer 2's broader range showing slightly better accuracy. For lipid content, SVMR outperformed PLSR, with spectrometer 2 achieving high predictive accuracy (RMSEP = 3.51% and RPD = 4.32). Variable selection methods further refined model performance by focusing on relevant wavelengths. [12] This study compares Soxhlet extraction with in situ transesterification (ISTE) for quantifying fatty acids (FA) in microalgae to assess biodiesel potential. Soxhlet extraction was found to inconsistently measure FA content, sometimes significantly overestimating due to non-saponifiable lipid content. ISTE, which combines extraction and transesterification in a single step, yielded more consistent and accurate FA results across 18 microalgae species. Acidic conditions in ISTE were crucial for complete FA recovery, as pre-treating samples with acid improved Soxhlet results.

3. Methodology

This chapter outlines the tools and procedures used to predict protein and fat content in Black Soldier Fly larvae using hyperspectral imaging and machine learning. MATLAB R2024b and Hypertools v3.0 were used for image processing, visualization, and model development. Samples were collected from ten different substrates, imaged using an IMEC hyperspectral camera, and preprocessed using Savitzky–Golay smoothing. Reference protein and fat values were obtained using the Kjeldahl and Soxhlet methods and used together with ontained data for model training and validation.

3.1 Tools Used

3.1.1 MATLAB R2024b

In this work, MATLAB R2024b was the central platform for all image processing and machine learning tasks. After incorporating Hypertools into MATLAB, the hyperspectral images of Black Soldier Fly Larvae were processed to extract the relevant spectral data. The extracted spectra were then smoothened using the Savitzky-Golay filter, a widely used technique for enhancing signal quality by reducing noise while preserving the shape and features of the spectral data. This preprocessing was crucial for ensuring the reliability of subsequent analyses. The clean, smooth spectral data served as input for the development of machine learning models within MATLAB. Specifically, Partial Least Squares Regression (PLSR) and Support Vector Machine Regression (SVMR) models were designed and trained to predict the protein and fat content of the larvae. MATLAB's advanced computational capabilities, along with its robust machine learning toolboxes, enabled efficient model development, training,

and evaluation, ensuring the accuracy and reproducibility of the results.

3.1.2 Hypertools v3.0

Hypertools v3.0 was used exclusively for the visualization of spectral data. Its primary role in this study was to generate spectral graphs directly from the raw hyperspectral files (.raw files). These spectral graphs provided an initial visual assessment of the spectral profiles for each sample, allowing for quick identification of trends, anomalies, or inconsistencies in the raw data. This visualization step was essential for confirming the quality of the acquired spectral data before any further processing or analysis in MATLAB. After obtaining and reviewing the spectral graphs with Hypertools, all subsequent data handling—including preprocessing, smoothening, and machine learning—was conducted within the MATLAB environment.

3.2 Data Collection

3.2.1 Obtaining Samples

Ten different substrates were carefully selected based on their availability. These substrates included brewery, food waste, soya, maize bran, jackfruit cowdung poultry, Irish, Irish cowdung poultry, soya + coconut + palm cake, Banana + cowdung poultry and wheat. These substrates were expected to provide varied nutritional compositions, influencing larval growth and composition.

At the start of the feeding experiment, 31.2 grams of six-day-old larvae were introduced to 3 kilograms of each substrate. Over the next nine days, the larvae were closely monitored, with daily observations made to track their growth progress. The controlled conditions ensured that the larvae had an optimal environment for feeding and development, such that they could be harvested successfully on the nineth day.

During the harvesting process, sieves of varying sizes were used to carefully sepa-

Figure 3.1: Substrate selection and sample distribution

rate the larvae from the substrate, with each substrate having replicas. The different samples from the different substrates were then packaged and taken to the National Agricultural Research Organisation(NARO) in Namulonge.

To achieve the required sample size of 200 samples, duplicates were prepared for each substrate. From these duplicates, representative samples were selected for IMEC hyperspectral imaging.

Table 3.1: Distribution of samples per substrate and replica

No.	Substrate	1	2	3	4	5	6	Total
1	Wheat	4	4	2	4	4	2	20
2	Soya	2	2	4	4	4	4	20
3	Food waste	5	5	5	5			20
4	Maize bran	4	4	4	4	2	2	20
5	Irish	4	4	4	4	2	2	20
6	Brewery	5	5	5	5			20
7	Irish cowdung poultry	8	4	8				20
8	Jackfruit cowdung poultry	4	8	8				20
9	Banana + cowdung poultry	6	6	8				20
10	Soya + coconut + palm cake	8	8	4				20
Overall Total								**200**

3.2.2 Image acquisition

Before the scanning and imaging process, the Black Soldier Fly Larvae (BSFL) from each substrate underwent a thorough cold-water bath to remove any residual surface fat and substrate particles. This step ensured that the imaging process was not affected by external contaminants, providing more accurate spectral data.

Once cleaned, the larvae samples were carefully scanned and imaged using the IMEC hyperspectral camera, capturing detailed spectral information for further analysis and image processing.

Below were the steps that were taken to capture the hyperspectral images using the XIMEA IMEC camera:

(a) Setting up the image acquisition system.

(b) Connecting the IMEC camera to the computer using USB 3.0.

(c) Dragging the calibration file of the camera to the software.

(d) In the Acquire tab; setting the integration time to 4ms.

(e) Leaving the demosaicing as default and edge detection off.

(f) Selecting the lens of the camera as predefined, that is, 25mm.

(g) Setting the F number as 2.8 on the camera and the software.

(h) Obtaining the dark reference.

(i) Obtaining the white reference.

(j) Removing the demosaicing from the default.

(k) Capturing the image.

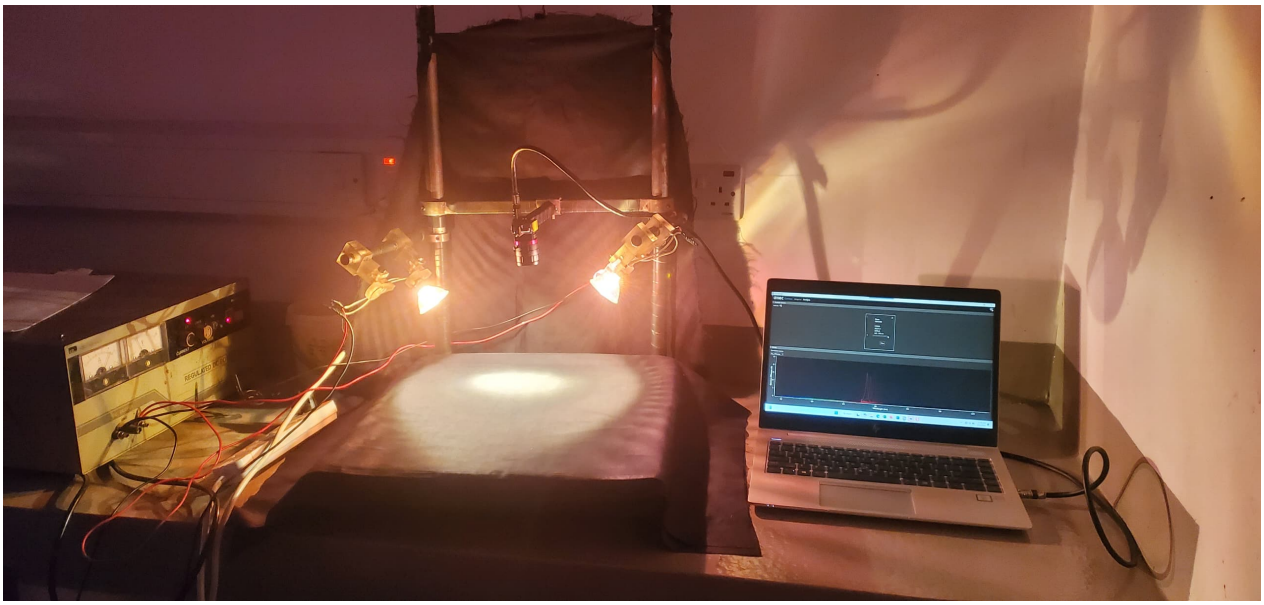

Figure 3.2: Image acquistion setup

3.2.3 Image processing

For image processing, HyperTools v3.0 incorporated in MATLAB R2024b was used to extract and analyze spectral data from the hyperspectral images captured by the IMEC XiQ Hyperspectral camera.

The image processing steps began with wavelength selection, where a total of 24 spectral bands were identified. The Region of Interest (ROI) extraction was carried out, isolating specific areas of the larvae samples for precise spectral analysis while ensuring that the background was not captured to prevent errors. Each sample scanned had about 80 to 150 larvae, and therefore, the spectral data obtained for a given image was averaged, and this was done for all the images.

The output was an CSV file containing the spectral data for all 200 samples across the spectral range(669nm to 939nm) for all the 24 bands, and it was used to obtain spectral graphs for the different samples.

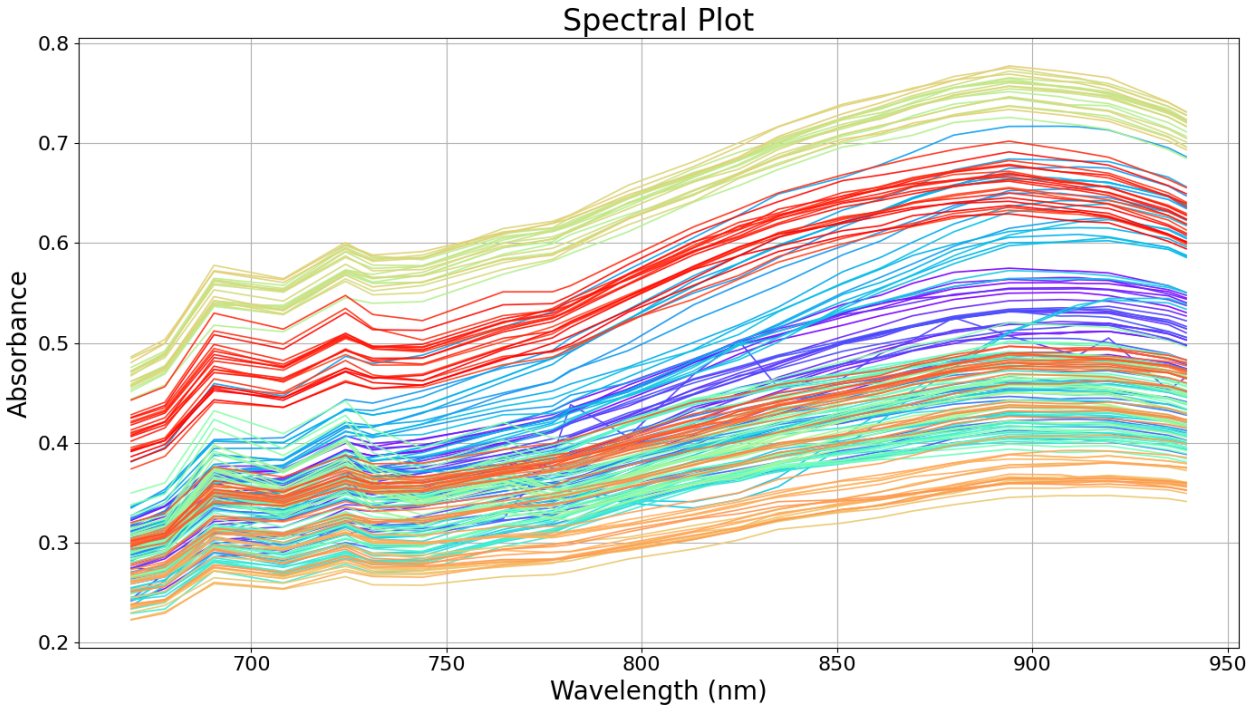

Figure 3.3: Spectral graph for the 200 samples

3.2.4 Spectral Data Smoothing: Savitzky–Golay Method

As part of the spectral data preprocessing pipeline, the Savitzky–Golay (SG) smoothing filter was implemented using MATLAB R2024 to reduce high-frequency noise and enhance the interpretability of the measured signals. The raw spectral dataset consisted of 200 samples, each with 24 spectral bands recorded across a defined wavelength range. For each sample, the wavelength values were extracted as the independent variable, while the corresponding absorbance values were treated as the dependent variable. The SG filter was applied to each spectrum individually using a third-order polynomial and a window length of five, selected to preserve the shape and height of spectral peaks while minimizing distortion. This choice of parameters ensured a balance between smoothness and fidelity to the original signal. The filtering process retained local spectral features while attenuating random fluctuations that could otherwise obscure important patterns, especially in multivariate analysis or derivative-based methods. A comparative visualization was produced for each sample, displaying both the original and smoothed spectra—highlighting the effectiveness of the filter in clarifying spectral trends. Finally, the smoothed data for all 200 samples were compiled into a structured Excel file, maintaining processed values for each spectral band. This step was critical in preparing the data for subsequent modeling and analysis.

The blue line is the original spectrum and the red line is smoothed spectrum for a given sample. This spectral smoothing was done for all the 200 samples and the output was an excel file containing smoothened spectral data across the spectral range for all the 24 bands.

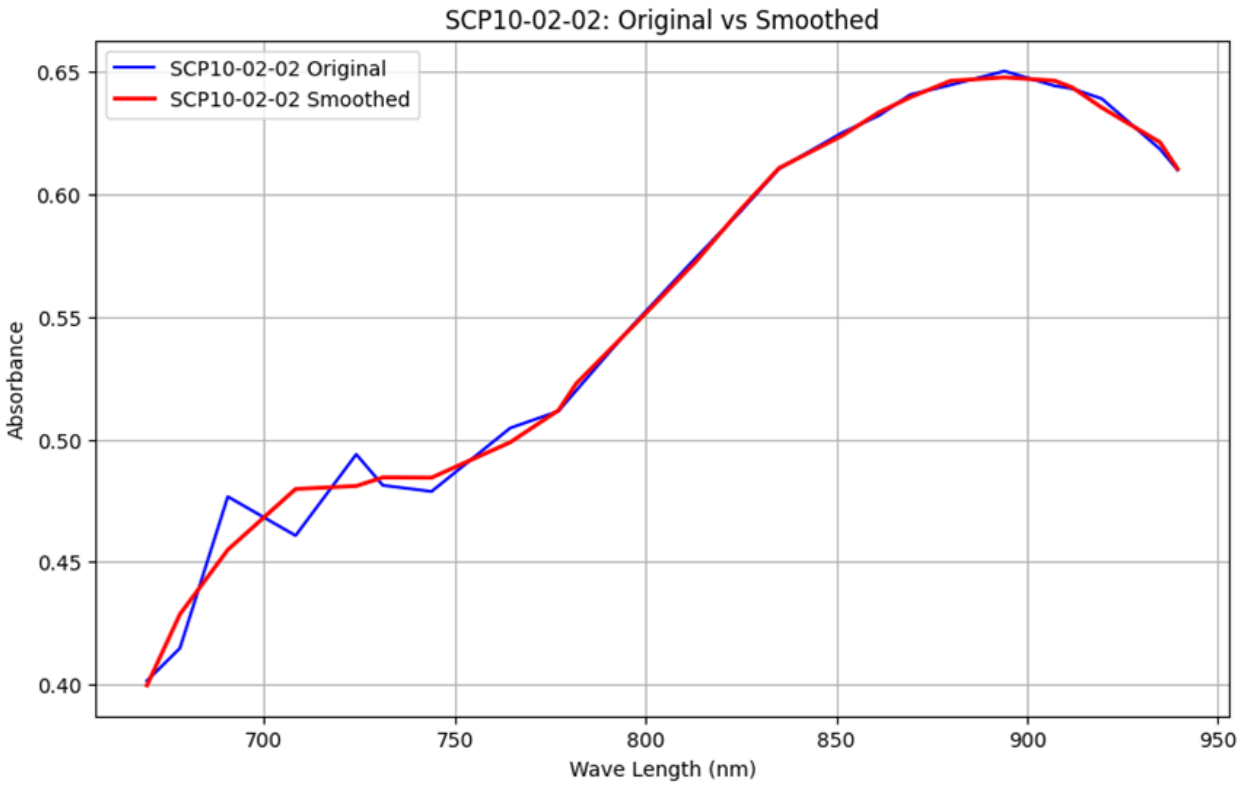

Figure 3.4: Original VS Smoothened Spectrum

3.3 Collection of Reference Data for Protein and Fat Using Standard Chemical Methods (Kjeldahl and Soxhlet)

To develop accurate machine learning models for predicting protein and fat content using hyperspectral imaging, it was essential to obtain reliable reference values. These reference values were determined using standard chemical analysis methods: the Kjeldahl method for protein and the Soxhlet extraction method for crude fat content.

3.3.1 Protein Analysis by Kjeldahl Method

Protein content was determined by measuring the total nitrogen content in dried BSF larvae samples using the micro-Kjeldahl technique, in accordance with AOAC Method 981.10 (Latimer, 2016) [9]. The procedure comprised three main steps:

Digestion: Approximately 1.0 g of dried sample (in triplicate) was accurately weighed and transferred into a Kjeldahl flask. To each flask, 30 mL of concentrated sulfuric acid (95–97%), 0.4 g of copper sulfate (catalyst), and 3.5 g of potassium sulfate (boiling point elevating agent) were added. The mixture was heated slowly in a fume hood to minimize frothing and then maintained at 400°C for 2.5 to 3 hours until the digest turned an iridescent blue, indicating complete digestion. The digest was allowed to cool and diluted to 100 mL with distilled water.

Distillation: A 10 mL aliquot of the digested solution was mixed with 10 mL of 40% sodium hydroxide (NaOH) and placed into a distillation apparatus. The liberated ammonia was captured in a receiving flask containing 10 mL of 4% boric acid solution along with 3 drops of a mixed indicator (methyl red and bromocresol green). The distillation continued until 25–30 mL of distillate was collected.

Titration: The collected distillate was titrated against 0.1 M hydrochloric acid (HCl) until a color change indicated the endpoint. The volume of HCl used was recorded and the percentage of nitrogen was calculated using equation 3.1:

$$\%\text{Nitrogen} = \left[\frac{\text{mL (titre – B)} \times M_{\text{HCl}} \times \text{Dilution Factor} \times 14.007}{\text{Sample mass (mg)} \times 10}\right] \times 100 \quad (3.1)$$

Where:

- M_{HCl} = Molarity of hydrochloric acid
- B = Volume of acid used in blank titration

Finally, protein content was calculated using the nitrogen-to-protein conversion factor:

$$\%\text{Protein} = \%\text{Nitrogen} \times 6.25 \quad (3.2)$$

3.3.2 Fat Content Analysis by Soxhlet Extraction

Crude lipid content was determined using the Soxhlet extraction method, following the protocol described by the Association of German Agricultural Investigation and Research Institutions (VDLUFA, 1976) [8], with minor modifications.

Approximately 2.0 g of homogenized, dried BSF larvae sample was weighed into cellulose extraction thimbles (26 × 60 mm). These were inserted into the Soxhlet extractor (Foss Soxhlet extraction unit, Denmark). Extraction was performed using 50 mL of petroleum ether (boiling point: 40–60 °C) as the solvent. The system was heated to 60 °C, allowing the ether to cycle through the thimble for 30 minutes.

After extraction, the solvent was evaporated using a heating block at 50 °C under a fume hood. The fat residue was left to cool in a desiccator for 30 minutes and then weighed. The mass difference before and after extraction was used to calculate

the crude fat content gravimetrically, using equation 3.3

$$\%\text{Fat} = \left(\frac{(\text{Crucible + oil}) \text{ - empty crucible}}{\text{Sample weight}} \right) \times 100 \tag{3.3}$$

3.3.3 Purpose of Reference Data

The protein and fat content values obtained through the Kjeldahl and Soxhlet methods served as the Ground truth data for model training and validation. These standard chemical analyses provide reliable, laboratory-grade benchmarks against which the predictions of hyperspectral imaging and machine learning models were evaluated.

3.4 Prediction Modeling

Following the preprocessing and alignment of the spectral data with corresponding reference values, machine learning models were developed to predict the protein and fat content in Black Soldier Fly (BSF) larvae. The hyperspectral image data, containing spectral reflectance values across multiple wavelengths, was utilized as input features, while the experimentally determined protein and fat concentrations were used as the target variables for supervised learning.

A variety of regression algorithms were explored to assess their predictive capabilities. These included:

- **Partial Least Squares Regression (PLSR)** – a robust method commonly used in chemometrics that handles multicollinearity and reduces dimensionality by projecting predictors and responses to a new space.

- **Support Vector Machine Regression (SVMR)** – a powerful technique

that uses kernel functions to capture non-linear relationships between spectral data and biochemical properties.

The dataset was divided into training and testing subsets using a 70:30 ratio. The training set (70%) was used to fit the models and learn underlying patterns between the spectral features and the target variables. The remaining 30% was held out as a test set to evaluate the generalizability and predictive performance of the models on unseen data.

3.5 Evaluation and Prediction

Model performance was quantitatively evaluated using multiple statistical metrics to ensure comprehensive assessment of prediction accuracy and reliability. The key evaluation criteria included:

- **Root Mean Squared Error (RMSE)** – measures how well the predicted dat matches the actual data.

- **R-squared ($\mathbf{R^2}$)** – represents the proportion of variance in the target variable explained by the model, with values closer to 1 indicating better predictive power.

4. Results and Discussion

This chapter presents the findings of the study on the prediction of protein and fat content in Black Soldier Fly (BSF) larvae using hyperspectral imaging combined with machine learning models. The spectral data acquired from the hyperspectral camera was used as input to build predictive models, while the corresponding reference values for protein and fat content were obtained through standard chemical methods that is Kjeldahl and Soxhlet techniques, respectively.

Regression algorithms such as Partial Least Squares Regression (PLSR) and Support Vector Machine Regression (SVMR), were trained and evaluated to compare their predictive capabilities. The performance of these models was assessed based on statistical metrics such as the coefficient of determination (R^2), Root Mean Squared Error (RMSE). The results are discussed with respect to the accuracy and robustness of each model in predicting the biochemical composition of the larvae.

4.1 Protein Prediction

The performance of Support Vector Machine Regression (SVMR) and Partial Least Squares Regression (PLSR) models was evaluated for predicting protein content in BSF larvae using hyperspectral data. Both models were trained using the same preprocessed spectral features, while the corresponding reference values for protein content were obtained via the Kjeldahl method.

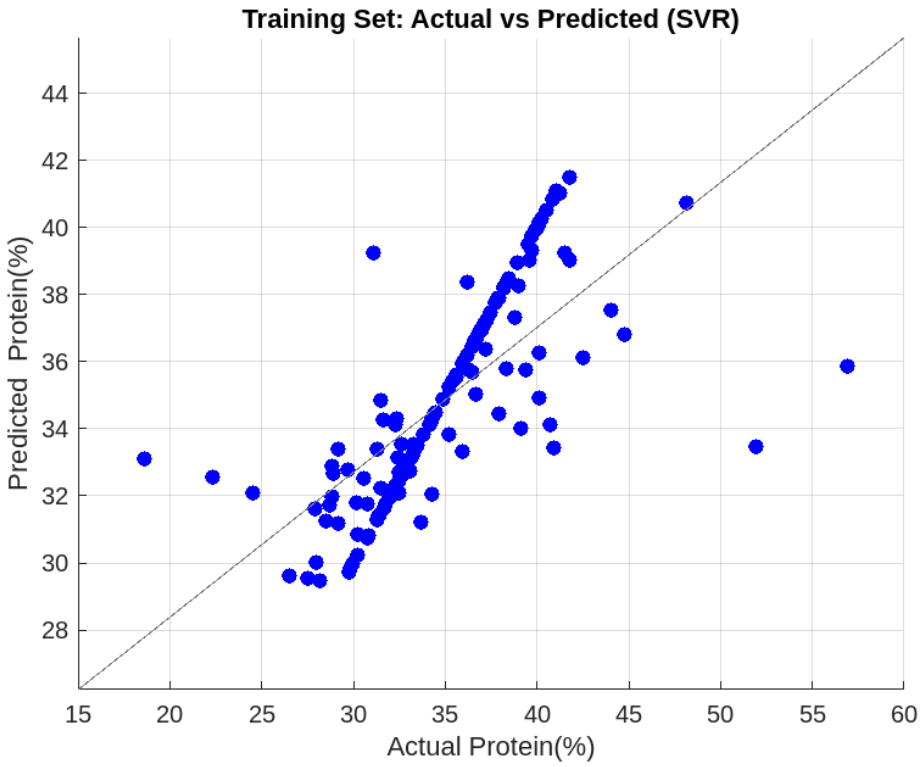

Figure 4.1: Training Set: Predicted Vs Actual (SVMR)

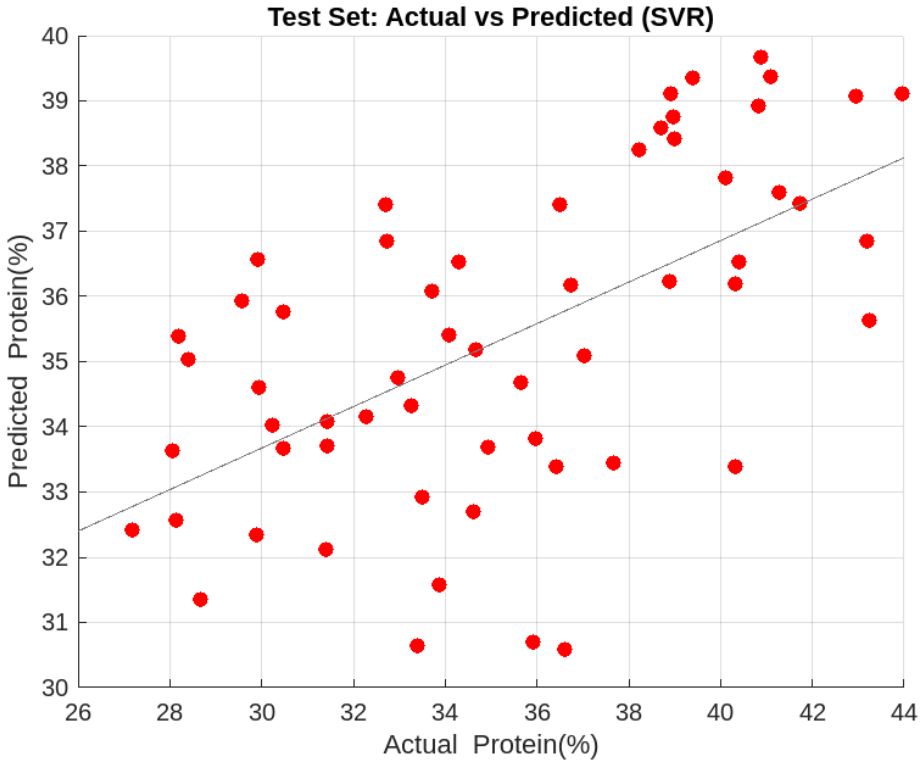

Figure 4.2: Test Set: Predicted Vs Actual (SVMR)

SVMR demonstrated moderate predictive performance across both the training and testing datasets. On the training set, it achieved an R^2 value of 0.5013 and a Root Mean Squared Error (RMSE) of 3.6408, indicating a fair fit to the data. For the test set, the model obtained an R^2 value of 0.3449 and an RMSE of 3.6992, suggesting reasonable generalization with relatively consistent prediction error across datasets, though there is still room for improvement in accuracy.

The PLSR model also demonstrated reasonable performance, though it was less accurate than SVMR. On the training set, PLSR achieved an R^2 value of 0.2293 with an RMSE of 4.5939. On the test set, the R^2 value slightly decreased to 0.2176,

accompanied by a lower RMSE of 3.7828. This modest performance and lower training R^2 suggest that PLSR may be more susceptible to underfitting in this context compared to SVMR.

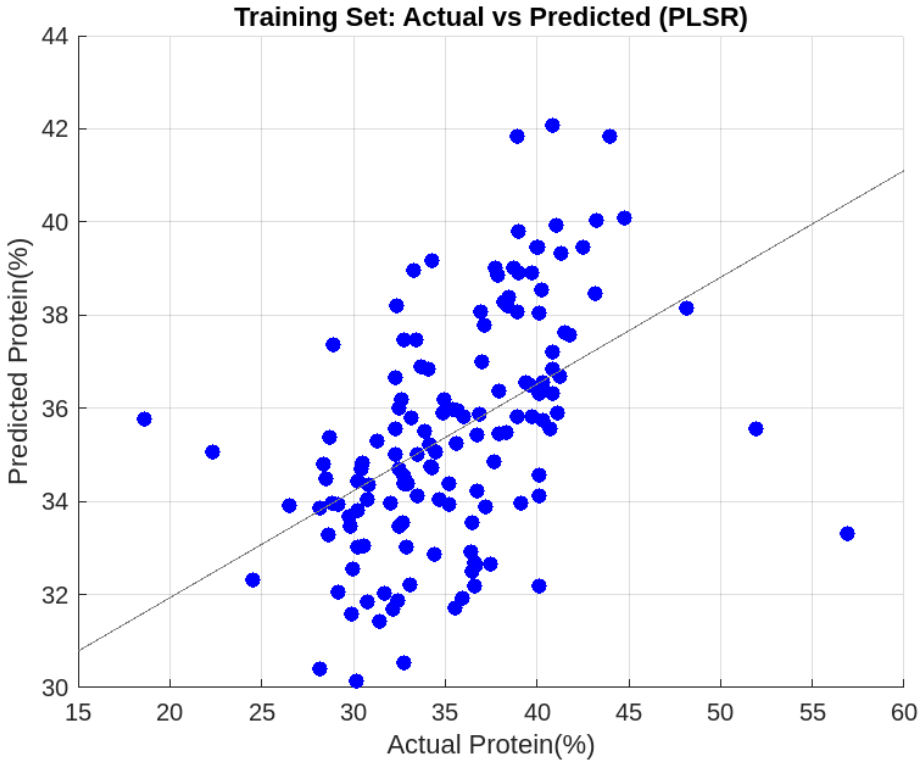

Figure 4.3: Training Set: Predicted Vs Actual (SVMR)

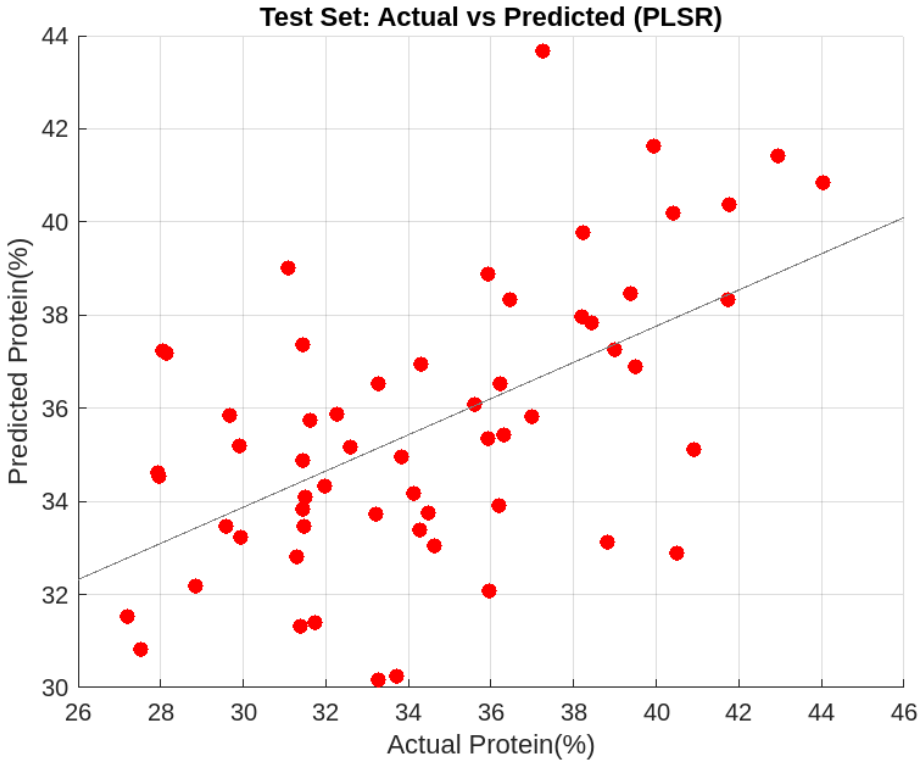

Figure 4.4: Test Set: Predicted Vs Actual (SVMR)

Overall, SVMR outperformed PLSR in predicting protein content, particularly in its ability to generalize to unseen test data. This makes SVMR a more suitable model for capturing the relationship between spectral data and protein content.

4.2 Fat Prediction

For fat content prediction, both SVMR and PLSR models were assessed against reference values derived using the Soxhlet extraction method. Unlike the case of protein prediction, the results revealed a stronger non-linear relationship between the hyperspectral data and fat content, which favored the SVMR model.

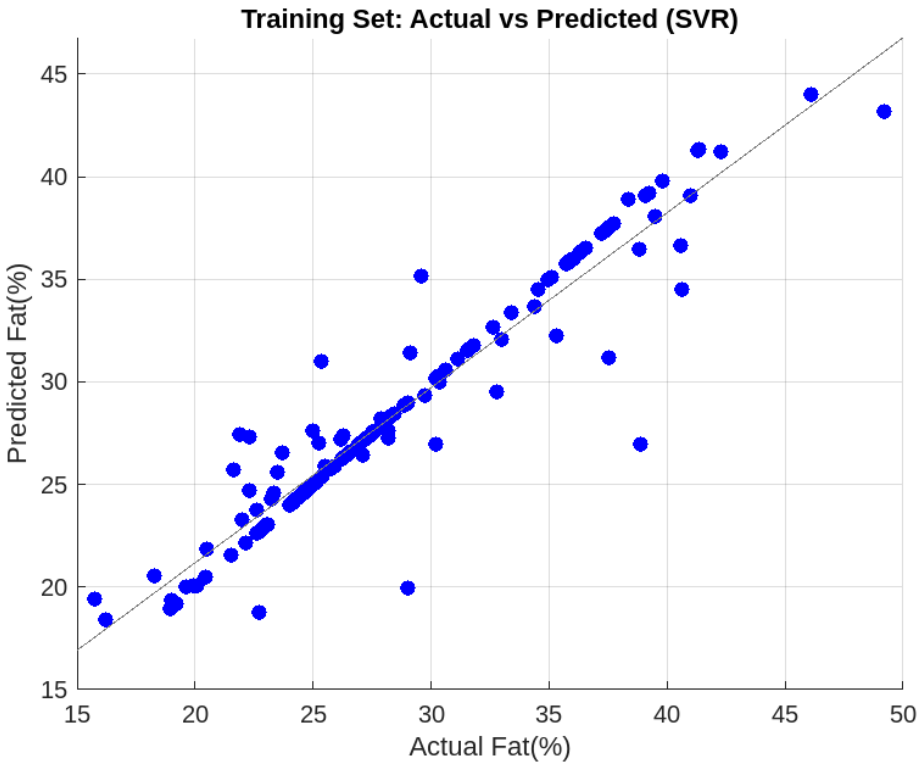

Figure 4.5: Training Set: Predicted Vs Actual (SVMR)

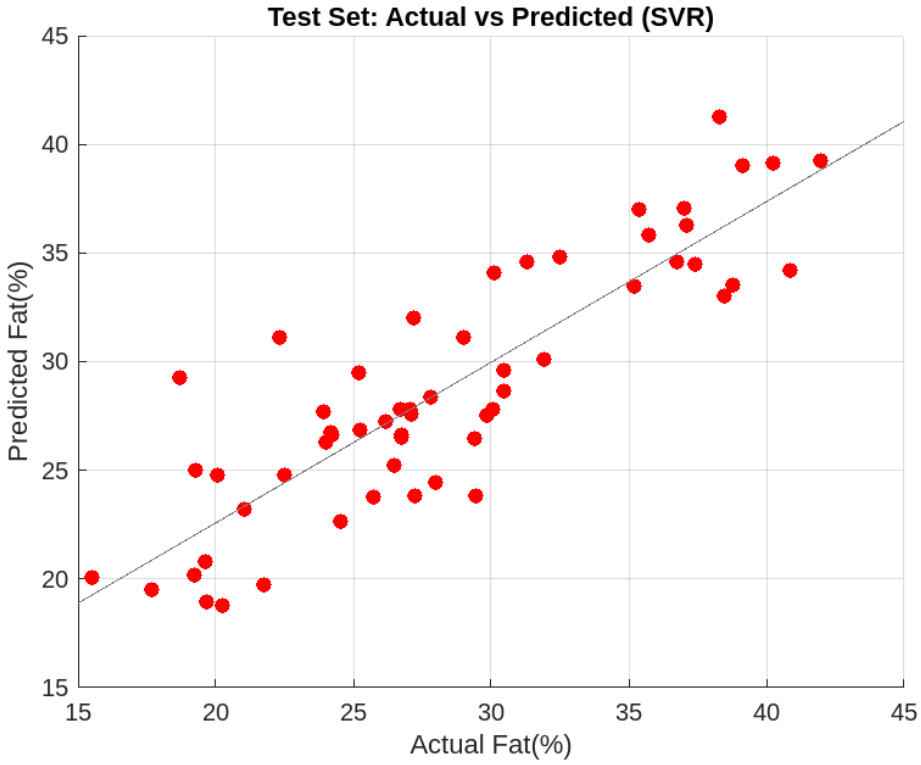

Figure 4.6: Test Set: Predicted Vs Actual (SVMR)

SVMR showed superior performance, with an R^2 of 0.8258 and RMSE of 2.7612 on the training set, and an R^2 of 0.7690 with an RMSE of 3.2156 on the test set. Its ability to model complex non-linear relationships likely contributed to its better performance in capturing spectral variations related to fat composition.

PLSR, while still acceptable, showed slightly lower predictive capability for fat content. It recorded an R^2 of 0.6051 and RMSE of 4.1576 on the training set, and an R^2 of 0.84 and RMSE of 4.2270 on the test set. These results indicate that while PLSR is competent, it may not fully capture the non-linear dependencies between spectral features and fat content.

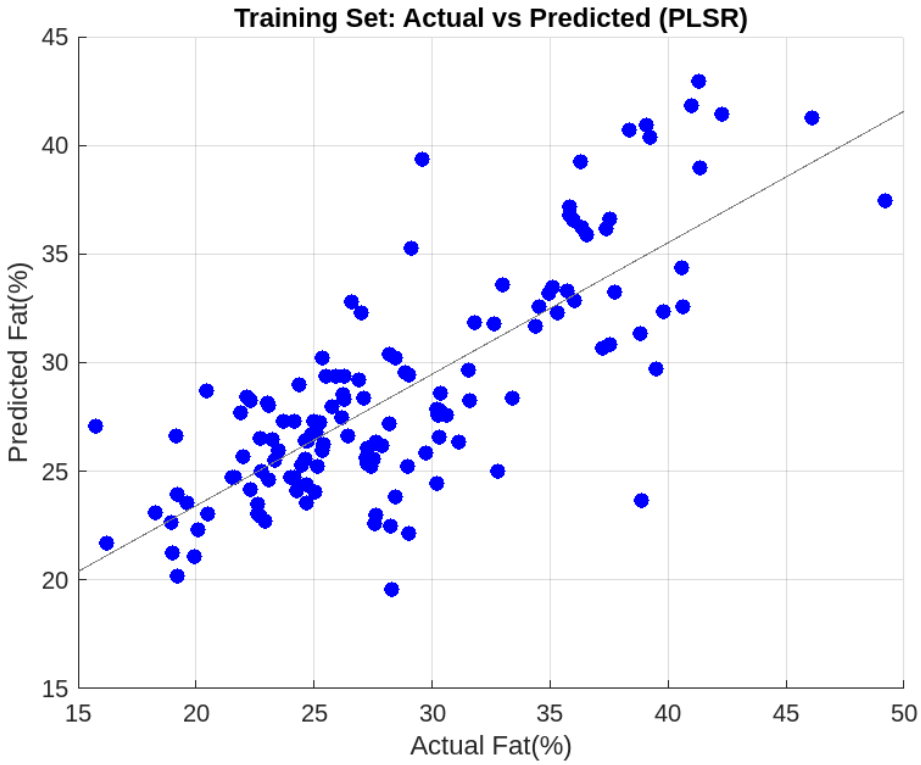

Figure 4.7: Training Set: Predicted Vs Actual (PLSR)

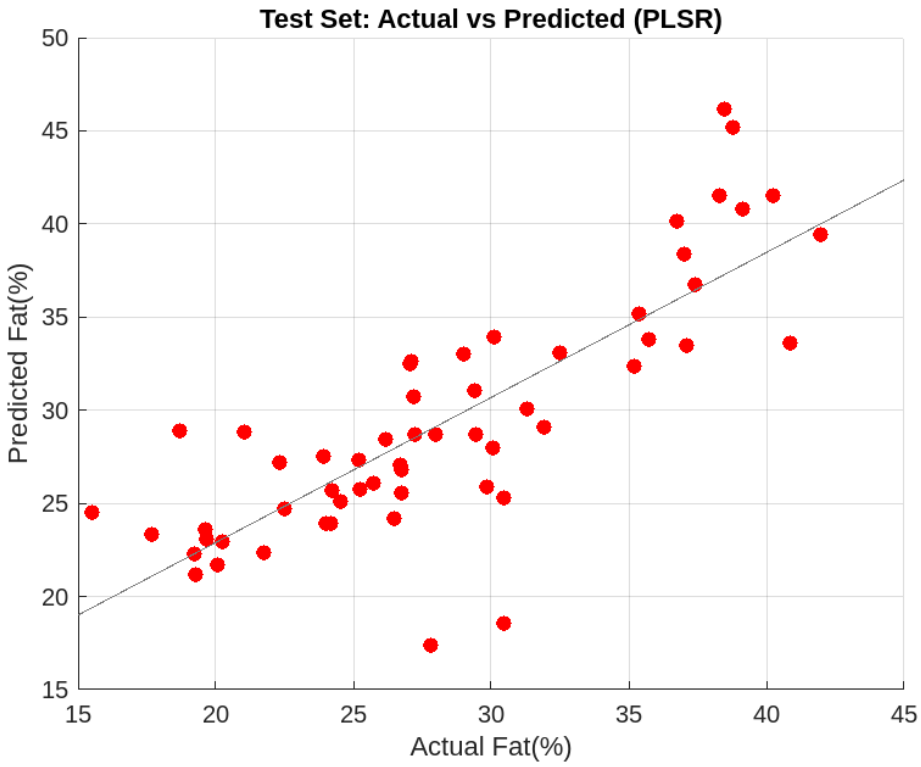

Figure 4.8: Test Set: Predicted Vs Actual (PLSR)

Therefore, SVMR again outperformed PLSR in predicting fat content, reinforcing the value of non-linear regression methods for biochemical attributes with complex spectral signatures.

Summary of Model Performance

Table 4.1: Comparison of SVMR and PLSR Models Performance Metrics on Training and Test Sets

Model	**Dataset**	**R^2**	**RMSE**
PLSR	Protein (Train)	0.2293	4.5939
PLSR	Protein (Test)	0.2176	3.7828
SVMR	Protein (Train)	0.5013	3.6408
SVMR	Protein (Test)	0.3449	3.6992
PLSR	Fat (Train)	0.6051	4.1576
PLSR	Fat (Test)	0.6008	4.2270
SVMR	Fat (Train)	0.8258	2.7612
SVMR	Fat (Test)	0.7690	3.2156

These findings demonstrate that SVMR consistently outperformed PLSR in predicting both protein and fat content, indicating its superior ability to model the underlying spectral–biochemical relationships. This highlights the importance of employing advanced, non-linear regression methods like SVMR for accurate estimation in hyperspectral imaging-based biochemical analysis.

5. Conclusion

This study demonstrated the effectiveness of Hyperspectral Imaging (HSI) combined with machine learning for the rapid, non-destructive prediction of protein and fat content in Black Soldier Fly Larvae (BSFL). Two regression models—Partial Least Squares Regression (PLSR) and Support Vector Machine Regression (SVMR)—were developed and rigorously evaluated.

The results revealed that while both models achieved high predictive accuracy, the SVMR model consistently outperformed PLSR in both protein and fat prediction tasks. This was evidenced by higher coefficients of determination (R^2) and lower Root Mean Square Error (RMSE) values for SVMR across both training and test datasets (see Table 4.1 in the report). The superior performance of SVMR underscores its suitability for handling the complex, nonlinear relationships inherent in hyperspectral data.

By leveraging machine learning with HSI, the study provides a robust, scalable, and environmentally friendly alternative to traditional chemical methods such as Kjeldahl and Soxhlet, which are destructive, time-consuming, and resource-intensive. The integration of these advanced techniques enables real-time, high-throughput quality assessment, supporting both operational efficiency and sustainability in the insect protein industry.

6. Recomendations

1. Prioritize SVMR for Industrial Deployment. Given its superior predictive performance, SVMR should be adopted as the primary machine learning model for routine protein and fat analysis in BSFL production facilities.
2. Expand and Diversify the Dataset. Future work should include a broader range of BSFL samples—encompassing different diets, growth stages, and environmental conditions—to further enhance the robustness and generalizability of the machine learning models.
3. Integrate HSI-SVMR Systems into Automated Pipelines. To maximize efficiency and minimize human error, HSI data acquisition and SVMR-based prediction should be incorporated into automated quality control workflows within industrial settings.
4. Broaden Analytical Scope. Investigate the application of HSI and SVMR for additional quality parameters such as moisture content, amino acid profiles, and potential contaminants, providing a comprehensive quality assessment platform.
5. Evaluate Environmental and Economic Benefits. Conduct detailed assessments of the environmental and economic impacts of replacing chemical analysis with HSI-SVMR methods, including reductions in hazardous waste, energy consumption, and operational costs.

Bibliography

[1] JP Cruz-Tirado, Matheus Silva dos Santos Vieira, José Manuel Amigo, Raúl Siche, and Douglas Fernandes Barbin. Prediction of protein and lipid content in black soldier fly (hermetia illucens l.) larvae flour using portable nir spectrometers and chemometrics. *Food Control*, 153:109969, 2023.

[2] Arthur Ricardo de Sousa Vitoria, Arlindo Rodrigues Galvao Filho, Clarimar Jose Coelho, Raylane Pereira Gomes, and Lilian Carla Carneiro. Bacteria gram staining differentiation using hyperspectral imaging and machine learning. In *2023 13th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, pages 1–5. IEEE, 2023.

[3] S Ishak, A Kamari, SNM Yusoff, and ALA Halim. Optimisation of biodiesel production of black soldier fly larvae rearing on restaurant kitchen waste. In *Journal of Physics: Conference Series*, volume 1097, page 012052. IOP Publishing, 2018.

[4] Zhilong Kang, Yuchen Zhao, Lei Chen, Yanju Guo, Qingshuang Mu, and Shenyi Wang. Advances in machine learning and hyperspectral imaging in the food supply chain. *Food Engineering Reviews*, 14(4):596–616, 2022.

[5] R Khodabakhshian, B Emadi, M Khojastehpour, and MR Golzarian. Combination of conventional imaging and spectroscopy methods for food quality evaluation. In *4th International Workshop on Computer Science and Engineering, United Arab Emirates, Dubai*, 2014.

[6] Daniel Kirk, Esther Kok, Michele Tufano, Bedir Tekinerdogan, Edith JM Feskens, and Guido Camps. Machine learning in nutrition research. *Advances in Nutrition*, 13(6):2573–2589, 2022.

[7] Shalom Hai Kobi, Mor David, Isaac August, and Dima Bykhovsky. Hyperspectral image prediction using a linear model in different illumination conditions. In *2023 13th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, pages 1–4. IEEE, 2023.

[8] Hartmut Kolbe. Prüfung der vdlufa-bilanzierungsmethode für humus durch langjährige dauerversuche. *Archives of Agronomy & Soil Science*, 51(2), 2005.

[9] Isaac Lee, Quanyin Gao, Wei Liu, Darshan Patel, Malissa Smith, Hong You, Tushar Patel, Karine Aylozyan, Silva Babajanian, Teddy Collins, et al. Determination of aloin a, aloin b, and aloe-emodin in raw materials and finished products using hplc multi-laboratory validation study, aoac 2016.09, final action status. *Journal of AOAC International*, page qsae070, 2024.

[10] L Loho and D Lo. Proximate and fatty acid analysis of black soldier fly larvae (hermetia illucens). In *IOP Conference Series: Earth and Environmental Science*, volume 1169, page 012082. IOP Publishing, 2023.

[11] Rosemary Mwende Matheka. *Investigating the Performance of Black Soldier Fly Larvae (Hermetia illucens) in Fecal Matter Co-digestion for Optimum Protein*. PhD thesis, JKUAT-IEET, 2022.

[12] Jesse McNichol, Karen M MacDougall, Jeremy E Melanson, and Patrick J McGinn. Suitability of soxhlet extraction to quantify microalgal fatty acids as determined by comparison with in situ transesterification. *Lipids*, 47:195–207, 2012.

[13] Purificación Sáez-Plaza, María José Navas, Sławomir Wybraniec, Tadeusz Michałowski, and Agustín García Asuero. An overview of the kjeldahl method of nitrogen determination. part ii. sample preparation, working scale, instrumental finish, and quality control. *Critical Reviews in Analytical Chemistry*, 43(4):224–272, 2013.

[14] Dhritiman Saha and Annamalai Manickavasagan. Machine learning techniques for analysis of hyperspectral images to determine quality of food products: A review. *Current Research in Food Science*, 4:28–44, 2021.

[15] Md Salahuddin, Ahmed AA Abdel-Wareth, Kohzy Hiramatsu, Jeffery K Tomberlin, Daylan Luza, and Jayant Lohakare. Flight toward sustainability in poultry nutrition with black soldier fly larvae. *Animals*, 14(3):510, 2024.

[16] Lijun Wu, Zhijun Li, Jianhang Li, Haifeng Fang, Xiaopeng Qiu, Pengrui Fu, and Tong Zhu. Lipid extraction of black soldier fly larva using aqueous enzymatic and soxhlet method. In *E3S Web of Conferences*, volume 406, page 03013. EDP Sciences, 2023.