BRILL
WAGENINGEN
ACADEMIC

JOURNAL OF INSECTS AS FOOD AND FEED (2026) DOI:10.1163/23524588-BJA10450

Journal of Insects as Food and Feed
brill.com/jiff

RESEARCH ARTICLE

Utilising Fourier-transform infrared spectroscopy to determine harvest time of black soldier fly (Diptera: Stratiomyidae) larvae for large-scale monitoring

A.P. Lubrano[1]*, *A.P. Holman*[2], *D. Kurouski*[2] and *J.K. Tomberlin*[1,3]

[1]Department of Entomology, Texas A&M University, College Station, TX, USA; [2]Department of Biochemistry and Biophysics, Texas A&M University, College Station, TX USA; [3]Texas A&M AgriLife, Texas A&M University, College Station, TX, USA; *angelaplubrano@gmail.com

Received 12 May 2026 | Accepted 19 July 2026 | Published online 4 August 2026

Abstract

Black soldier fly (BSF) is an established gear in the global bioeconomy. Presently, the use of black soldier fly larvae (BSFL) has been emphasised with recycling organic side-streams to produce products of value; however, due to regulatory and other factors, the process has limited profitability in Global North nations. Consequently, research over the past decade has in part focused on methods for optimisation. The objective of this pilot study was to determine whether Fourier-transform infrared spectroscopy (FT-IR) could be used as an automated system for determining harvest time of individual pans of BSFL at a large scale. Pan material was collected from four independent rearing units containing approximately equal portions of BSFL fed a wet Gainesville diet over a 16-day growth period. FT-IR spectra were used to train and optimise chemometric and machine learning-based classification models, with emphasis on identifying the best-performing algorithm and hyperparameter configuration for distinguishing ready-to-harvest larvae from others. External validation demonstrated that FT-IR coupled with eXtreme gradient boosting discriminant analysis achieved greater than 85% balanced accuracy with precision and recall in classifying not ready vs ready-to-harvest larvae. Notably, the model exhibited a conservative bias toward 'not ready' classifications in cases of uncertainty, which is operationally favourable in industrial settings where premature harvesting can reduce yield and product quality. These results suggest that FT-IR-based chemometric monitoring, when integrated with machine learning, offers a viable pathway toward real-time, automated harvest decision-making in BSF production systems, ultimately supporting more efficient and economically sustainable bioresource utilisation.

Keywords

frass – FT-IR – *Hermetia illucens* – mass production

1 Introduction

The black soldier fly (BSF), *Hermetia illucens* (Linnaeus) (Diptera: Stratiomyidae), is a non-pest insect when properly managed, whose larvae exhibit great potential for sustainable waste management. As has been repeatedly demonstrated, black soldier fly larvae (BSFL) recycle a wide variety of organic waste streams by converting them into protein- and lipid-rich biomass throughout their life cycle (Sheppard *et al.*, 1994). In fact, resulting BSFL can consume a wide variety of organic matter such as manure, kitchen waste and fruits and vegetables

Published with license by Koninklijke Brill BV | DOI: 10.1163/23524588-BJA10450
© A.P. LUBRANO *ET AL.*, 2026 | ISSN: 2352-4588 (online)

(Nguyen *et al.*, 2015), which opens the possibilities to be used for recycling organic residues from various facets of society (e.g. home to agricultural waste streams).

BSFL have been successfully employed as a partial feed substitute for swine (Newton *et al.*, 1977), poultry (Zhang *et al.*, 2021) and pets (Jian *et al.*, 2022). They have also proven useful in aquaculture as supplemental fish feed, maintaining the growth of fish (St. Hilaire *et al.*, 2007a,b; Yamamoto *et al.*, 2022). Moreover, these larvae exhibit potential to promote health as a dietary ingredient for pets and fish (Mouithys-Mickalad *et al.*, 2020). As a feed additive, BSFL can improve immune responses in animals such as broilers, decreasing their mortality rates and enhancing their ability to fight infection (Zhang *et al.*, 2021). These examples shed light on the potential BSFL hold as a nourishment alternative in a variety of animals.

Relatively low waste is generated from this system. In fact, the resulting frass (i.e. residual remaining of the feedstock after digestion by the BSFL) has been shown to serve as a nutrient-rich soil amendment/fertiliser that is safe for plant growth (Chavez *et al.*, 2023, 2024, 2025) and promotes soil health (Jiang *et al.*, 2025). Hence, even the 'waste' that BSFL produce after consuming substrates can be reused to nourish vegetation, further supporting this insect's capability in sustainable waste management. These ecological and agricultural benefits highlight the key role that BSFL play in a circular economy. While all of these values have been demonstrated for the BSFL system, industrialisation in Global North nations or developed countries, has been challenging in part due to costs associated with the production of the insect (Tariq *et al.*, 2025).

Successful development of the BSF industry is partially driven by stable production of BSF, while reducing associated costs. Low variation (e.g. weight of larvae produced, bioconversion rate, survival, nutritional quality) allows for industry to project outputs and to arrange client engagements without concerns of not meeting expectations. In regions where mechanisation has been employed, greater efficiencies have been gained; however, variation still exists. For example, when working with a batch system where large numbers of pans are set up at the same time, they do not 'cure' at the same rate (Tomberlin *et al.*, unpublished data). In this example, 'curing' is in reference to the larvae growing to the targeted harvest weight. Consequently, variation in weight impacts total amounts generated of larvae as well as the bioconversion of waste streams to insect biomass and their own residual waste (i.e. frass).

Abiotic and biotic factors impact BSFL growth patterns and the bioconversion process, resulting in variation. Abiotic factors include nutrient profile of the substrate (Grossule *et al.*, 2025) and concentration, scale (e.g. pan dimensions) (Yang and Tomberlin, 2020), temperature (Myers *et al.*, 2008), pH (Ma *et al.*, 2018, Meneguz *et al.*, 2018) and moisture of diets (Holmes *et al.*, 2012). Biotic factors include fly population (Zhou *et al.*, 2013), density (Barragan-Fonseca *et al.*, 2018) and microbial communities (Yu *et al.*, 2024) associated with the feedstock.

In efforts to industrialise BSFL production and reduce variability across rearing systems, recent studies have emphasised automated sensing and monitoring technologies capable of improving process consistency. Deep-learning models enhanced with optical flow can autonomously recognise larval stages, supporting large-scale production with greater precision and sustainability while optimising resource management (Manduca *et al.*, 2025). Automation and Internet of Things (IoT)-based environmental controls that stabilise temperature, humidity and light intensity further demonstrate the transition toward digitally managed bioconversion facilities (Manduca *et al.*, 2025). Non-contact sensing platforms, including hyperspectral imaging, RGB cameras and thermal sensors have also been proposed for integration directly into production lines (Proietti *et al.*, 2022). However, many analytical procedures remain difficult to implement at scale and hyperspectral systems may require improved illumination to achieve reliable discrimination (Proietti *et al.*, 2022). Vibrational spectroscopy, particularly infrared (IR) methods, has emerged as a promising strategy for strengthening process control while aligning with circular-economy principles. Unlike destructive and energy-intensive traditional analyses, IR spectroscopy is rapid, non-destructive and reagent-free, making it suitable for industrial environments (Alagappan *et al.*, 2024). Attenuated total reflectance (ATR) mid-infrared (MIR) spectroscopy has demonstrated the ability to classify larvae according to waste-stream diet ($R^2 > 0.80$), supporting traceability and risk evaluation during production (Hoffman *et al.*, 2022). Presently, adoption remains limited due to a deficiency in training and deep understanding of spectroscopy and chemometrics, consequently leading to reliance on conventional measurements (Alagappan *et al.*, 2024). Building on these advances, the proceeding study addresses the need for scalable monitoring by applying IR spectroscopy directly to pan material (the combined matrix of BSFL substrate and frass), which may provide a more compre-

hensive indicator of BSFL pan health and help reduce batch-to-batch variation.

2 Materials and methods

Black soldier fly colony maintenance

The BSF colony operated in this study is located at the Forensic Laboratory for Investigative Entomological Sciences (F.L.I.E.S.) Facility at Texas A&M University in College Station, TX, USA. The standard diet used for rearing BSFL, the Gainesville diet, consists of 50% wheat bran, 30% alfalfa and 20% cornmeal at 70% moisture (Hogsette, 1992). Colony maintenance is modified from Sheppard *et al.* (2002) and follows the technique described by Tomberlin *et al.* (2023) in patent number US12490725B2. Adult BSF were kept in a 1 (W) × 1 (H) × 2 (L) m cage within a greenhouse set at approximately 27–30 °C and 70% RH (Cammack and Tomberlin, 2017). Adult BSF laid eggs in a plastic egg trap on top of a 5.7-litre Sterilite plastic container (Sterilite, Townsend, MA, USA) filled with 1 kg of wet Gainesville, approximately 500 BSFL (Cammack and Tomberlin, 2017) and a border of dry Gainesville to prevent larvae escape, all covered with a mesh lid. Eggs were collected within 24 hours of being laid and put in a 30 ml Dart plastic cup (Dart Container, Mason, MI, USA) that was placed into an 800 ml Ball mason jar (Ball, Westminster, CO, USA). The jar was covered with a paper towel and secured with a jar ring, then stored in a Rheem Environmental Chamber (Rheem Manufacturing, Atlanta, GA, USA) at 27 °C, 70% RH and 14:10 light/dark cycle until hatching. Upon hatching, 0.616 ml of neonates (approx. 13 000 BSF neonates) were placed directly on top of 220 g of wet Gainesville inside a 473 ml Uline plastic deli container (Uline, Pleasant Prairie, WI, USA), which was then topped with approx. 35 g of dry Gainesville. Four replicates of these containers were made and stored in the Rheem Environmental Chamber uncovered, with the same settings as previously mentioned. After three days, containers were covered with porous lids for breathability. After one month, each container was transferred to its own 27-litre Sterilite plastic pan (Sterilite) that contained 4 kg of wet Gainesville substrate for larvae to feed on. A border of 1 kg of dry Gainesville was applied to each pan to prevent larvae escape. The pan size and feed amount used in this study are comparable to those described as industrial-scale in previous BSFL production studies (Miranda *et al.*, 2020, Yang and Tomberlin, 2020). Each of the four pan replicates were labelled A, B, C or D and stored in the Rheem Environmental Chamber with the same settings. The date of assembly of pans is referred to as Day 0. Pans were kept in this incubator for the remainder of larval development until Day 16 when larvae were harvested. On Days 7-11, each pan was supplemented with 1 kg wet Gainesville per day, in between the hours of 10:00-13:00.

Sample collection

Pan material was obtained for FT-IR scanning on Days 1-16 at approximately 09:00 each morning. Every day, approximately 5 g of frass was shovelled from the top layer of each pan replicate (A, B, C and D), resulting in four primary samples collected each day. These primary samples of pan material were weighed on an Ohaus Scout Pro SP401 balance (Ohaus, Parsippany, NJ, USA). Then, the primary samples were placed in separate 473 ml Uline plastic deli containers and capped with a porous lid. They were labelled A, B, C or D, corresponding to the respective pan replicate from which the sample was taken. Samples of the Gainesville diet without larvae present were also obtained as a baseline reference for the input material.

Fourier-transform infrared (FT-IR) spectroscopy

The instrument used to collect spectral data of pan material each day, as well as food alone, was a Spectrum 100 IR spectrometer (PerkinElmer, Shelton, CT, USA). The wavenumber range was 800 to 4000 cm^{-1} with 10 cumulative scans at a resolution of 2 cm^{-1}. Every day, three subsamples of pan material were obtained from each primary sample for FT-IR analysis. In order to obtain and scan each subsample, a spoon was used to scoop enough pan material from the corresponding primary sample to completely cover the ATR crystal. A force of 70 N was applied using a flat anvil. With four primary samples (from the four pan replicates), this produced a total of 12 pan material subsamples analysed every day. Three FT-IR spectra were acquired per subsample, which generated a total of 36 spectra recorded per day. Spectra were visualised using PerkinElmer spectrum express software. This process was completed on Days 1-16.

Data analysis

All spectra were trimmed to the 880-1470 cm^{-1} range to reduce edge noise and water contribution so that only spectral regions unaffected by water were retained. Then, spectra were baseline-corrected using the adaptive smoothness least squares (ASLS) algorithm (λ = 1e7, p = 0.005), smoothed using 1st order Savitzky-Golay fil-

tering (window length = 6) and area-normalised before analysis, in that order. This process was done using NumPy, pandas, SciPy, pybaselines and scikit-learn in Python 3.13. A visual demonstration of our data before and after preprocessing, as well as a water signal for reference, is presented in Figure A1 in the Appendix.

Three machine learning models were tried in this experiment: partial least squares discriminant analysis (PLS-DA), eXtreme gradient boosting discriminant analysis (XGB-DA) and artificial neural network discriminant analysis (ANN-DA). These models were chosen due to their establishment as robust models throughout spectral machine learning, incorporating both flexible and non-flexible threshold decisions (Barton *et al.*, 2022, Holman *et al.*, 2024, Yang *et al.*, 2019). Model hyperparameters were automatically tuned via grid search 5-fold cross-validation where the model is trained on 4 folds (or 80% of the data), tested on the remaining fold (20%) and then re-trains the model 4 more times until all of the data has been seen during testing. Models were selected using the highest macro-averaged F1 score (Macro F1) from the mean performance of all folds following cross-validation. On the other hand, PLS-DA contains a single primary tuneable parameter, the number of latent variables (LVs). While increasing the number of LVs can improve apparent fit to the training data, excessive model complexity increases the risk of overfitting and leads to overly optimistic performance estimates that do not generalise to unseen data (Westerhuis *et al.*, 2008). Accordingly, model complexity was controlled by selecting the optimal number of LVs at the point where classification efficiency for the cross-validation and independent test sets converged, thereby balancing explanatory power and generalisability (Pomerantsev and Rodionova, 2025). Table A1 in the Appendix provides a short description of each model, as well as the optimisable parameters and the associated Python libraries used for their implementation.

Additionally, model-specific variable-importance analyses were performed to visualise which IR bands contributed most strongly to class discrimination. For PLS-DA models, variable contributions were assessed using Variable Importance in Projection (VIP) scores, which summarise the influence of each spectral variable on the latent structures that maximise covariance between predictors and class membership. For XGB-DA models, spectral importance was quantified using the mean decrease in impurity (Gini importance) aggregated across all trees, reflecting the extent to which each IR band reduced classification error through node splitting. For ANN-DA models, variable importance was evaluated using permutation importance, calculated as the decrease in classification performance observed when the values of each IR band were randomly permuted thirty times, thereby estimating the dependence of the trained network on individual spectral regions. Collectively, these complementary approaches provide interpretable insight into the spectral features driving model decisions while accounting for the differing mathematical structures of linear, tree-based and neural-network classifiers.

Accuracy, Macro F1, sensitivity, specificity, false positive rate (FPR) and false negative rate (FNR) were calculated for each model, where applicable, to comprehensively evaluate diagnostic performance. Accuracy was defined as the proportion of correctly classified spectra relative to the total number tested and provides a general measure of overall correctness; however, it can be misleading in the presence of class imbalance. To address this limitation, Macro F1 was employed as a primary performance metric. Macro F1 is calculated as the unweighted mean of class-specific F1 scores, where each F1 score represents the harmonic mean of precision and recall for a given class. As such, Macro F1 weights all classes equally and is widely regarded as a robust metric for imbalanced datasets commonly encountered in diagnostic applications (Powers, 2020).

Sensitivity (Equation 1) was calculated as the proportion of true positives correctly identified and reflects the assay's ability to detect the target condition when it is present. Specificity (Equation 2) represents the proportion of true negatives correctly classified and quantifies the assay's ability to correctly exclude non-target samples. The FPR (Equation 3) is calculated by subtracting specificity from 1 and represents the proportion of negative samples that were misclassified as positive. The FNR (Equation 4), on the other hand, is calculated by subtracting sensitivity from 1, representing the proportion of positive samples misclassified as negative.

$$\text{True Positive Rate (TPR)} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \quad (1)$$

$$\text{True Negative Rate (TNR)} = \frac{\text{True Negatives (TN)}}{\text{True Negatives (TN)} + \text{False Positives (FP)}} \quad (2)$$

$$\text{False Positive Rate (FPR)} = 1 - \text{TNR} \quad (3)$$

$$\text{False Negative Rate (FNR)} = 1 - \text{TPR} \quad (4)$$

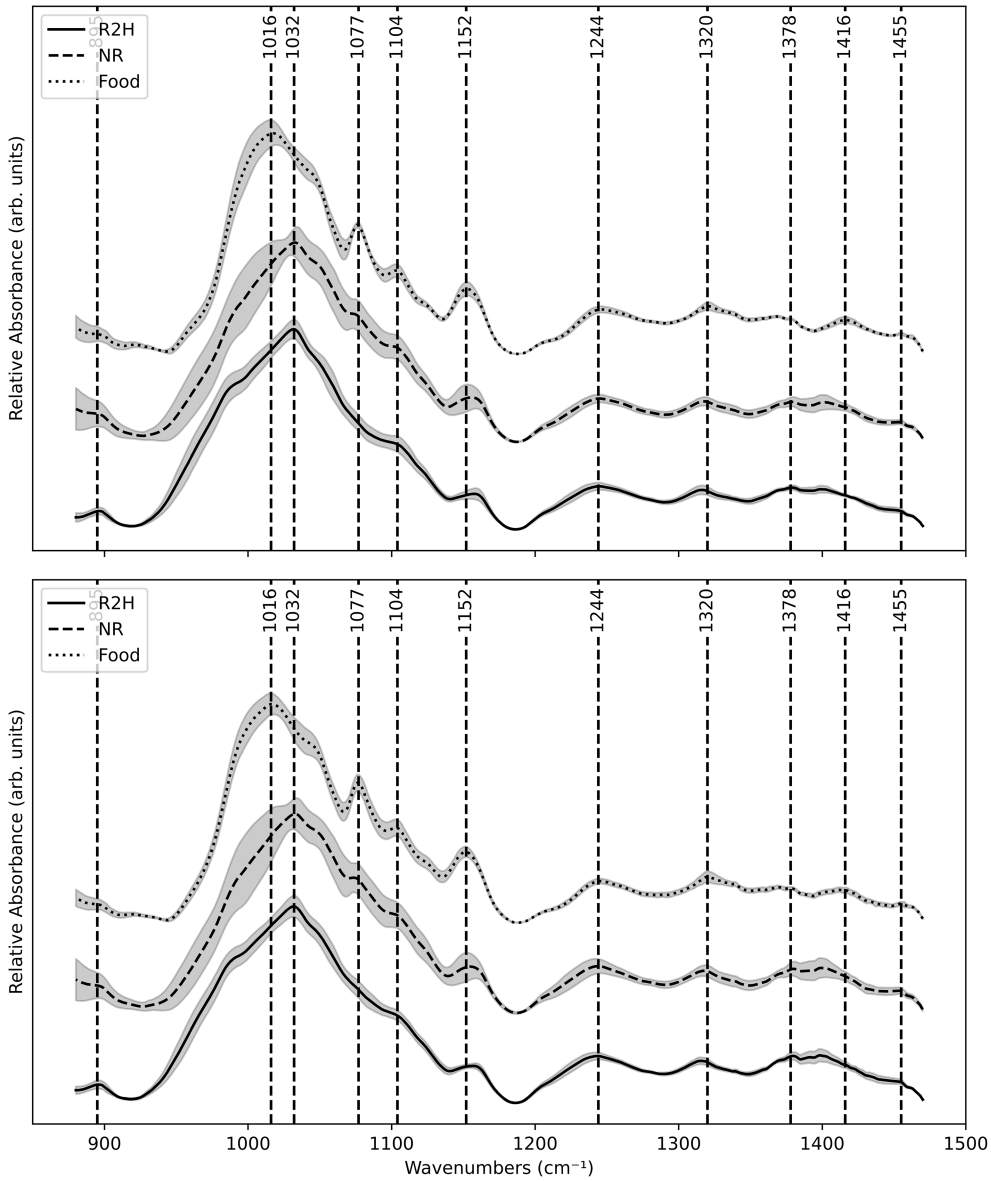

FIGURE 1 Mean and standard deviation FT-IR spectra for (top) train and (bottom) test datasets, separated by class (food, NR and R2H). Offset, relative FT-IR absorbance spectra of Gainesville diet (food), not ready (NR) samples and ready-to-harvest (R2H) samples are displayed following preprocessing.

3 Results

To evaluate the robustness and generalisability of each model, the dataset was split using an independent, pan-wise partitioning strategy. Specifically, spectra collected from pans A and C were used exclusively for model training, while spectra from pans B and D were reserved for testing. This design resulted in a 50:50 train-test split composed of fully independent experimental units, thereby preventing information leakage arising from shared sample preparation, instrumental conditions, or within-pan correlations. Independent hold-out evaluation of this type is widely recommended in chemometrics when replicate measurements are nested within higher-level sampling units, as it provides a conservative and realistic estimate of performance on unseen data (Varoquaux, 2018). Principal component analysis further revealed that the split did not alter representativeness of any class (Figure A2 in the Appendix).

Additionally, spectra from several food samples were included in training and testing (also split 50:50) to account for background spectral variation arising from both abiotic and biotic factors during sampling. Figure 1 displays mean and standard deviation IR spectra for food, ready-to-harvest (R2H) and not ready (NR) to harvest frass/pan samples for both train and test datasets. There are clear differences in the low-frequency region between approximately 1016-1152 cm^{-1}, where progres-

TABLE 1 Infrared absorptive band assignments

Wavenumber (cm^{-1})	Vibrational mode	Molecular assignments
895	C-O-C bending	B-glycosidic linkages (cellulose and other polysaccharides)
1016	C-O-H bending; C-C-H bending	Polysaccharides
1032	C-C stretching	Polysaccharides
1077	C-O stretching; C-C-O bending	Polysaccharides
1104	C-O stretching (Ring)	Polysaccharides
1152	C-O-C stretching	Glycosidic linkages (polysaccharides)
1244	C-N stretching; N-H bending	Proteins (Amide III)
1320	O = C-O bending; C-H wagging	Lignin and polysaccharides
1378	C-C bending	Lipids and polysaccharides
1416	C-O bending; C-H deformation	Polysaccharides and proteins
1455	C = C stretching; C-H deformation	Proteins, lignin, polysaccharides

Molecular assignments were adapted from Baum *et al.* (2017), Kong and Yu (2007), Mohan Rao (2019), Parihar *et al.* (2019), Robert *et al.* (2005) and Santos *et al.* (2015).

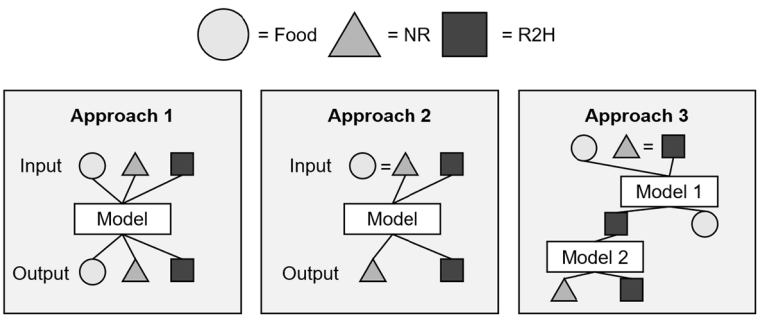

FIGURE 2 Summary of our three strategies for machine learning classification. In Approach 1, all three classes are utilised as labels during training. In Approach 2, food and NR spectra are umbrellaed under the NR label, so that only NR and R2H are used as labels during training. Approach 3 represents the most complex framework, where first spectra are declared frass or food, before being declared NR or R2H if classified as frass. The circle represents a food sample, the triangle represents a frass sample from a pan that is not yet ready for harvesting (NR) and the square represents a frass sample from a ready-to-harvest (R2H) pan.

sive attenuation and spectral broadening are evident as the food is digested and the larvae grow older. This region is commonly associated with collective vibrational modes arising from cellulose and other polysaccharides, presumably originating from the wheat bran and alfalfa cell wall carbohydrates (Hogsette, 1992) (Table 1).

Although distinct spectral features were apparent between the food and frass through visual examination, practical decision-making requires an objective and scalable classification framework. To quantitatively leverage these multivariate spectral differences, supervised machine-learning models were employed to classify pan-associated samples under three complementary approaches (Figure 2) with specific hyperparameters (Table 2). Approach 1 represents the ideal scenario, where food, NR and R2H spectra are easily distinguishable. Alternatively, Approach 2 obtains spectra acquired from food material and frass produced by early-stage larvae as NR, reflecting their shared biochemical characteristics and practical indistinguishability from non-harvestable material. Approach 3 employs a dual-phase classification pipeline: in Phase 1, samples are classified as food or frass and in Phase 2, spectra identified as frass are further classified as NR or R2H. Together, these strategies enable systematic evaluation of classification performance under progressively realistic sampling and decision-making conditions.

Under the first approach, none of the three classifiers achieved a test Macro F1 score exceeding 70% (Table 3) (variable-importance plots and confusion matrices for each model can be found in Figures A3 and A4 in the Appendix), unlike the other two approaches. The highest performance was obtained using PLS-DA, which reached a test Macro F1 of 67.77% with a relatively simple model comprising only two LVs (Table 2). ANN-DA

TABLE 2 Hyperparameters for each approach in this study

Model	Approach 1	Approach 2	Approach 3, Phase 1
PLS-DA	LVs = 2	LVs = 6	LVs = 2
XGB-DA	Learning Rate = 0.2	Learning Rate = 0.1	Learning Rate = 0.2
	Maximum Depth = 6	Maximum Depth = 1	Maximum Depth = 3
	No. Estimators = 200	No. Estimators = 500	No. Estimators = 50
ANN-DA	Hidden layer nodes = (1024, 512)	Hidden layer nodes = (512, 256)	Hidden layer nodes = (1024, 512)
	Learning rate = 0.005	Learning rate = 0.005	Learning rate = 0.001
	Dropout = 0.2	Dropout = 0.0	Dropout = 0.0
	L1 penalisation = 0.0	L1 penalisation = 0.0	L1 penalisation = 0.0

TABLE 3 Summary of Approach 1 model classification results

Model	CV/test	Macro F1	Accuracy	Food		NR		R2H	
				TPR	FPR	TPR	FPR	TPR	FPR
PLS-DA	CV	0.6606	0.6761	1.0	0.0938	0.5769	0.0476	0.9259	0.2727
	Test	0.6777	0.7201	0.9	0.125	0.6325	0.0357	1.0	0.1894
XGB-DA	CV	0.942	0.9623	0.9667	0.0139	0.9829	0.0952	0.8704	0.0
	Test	0.5543	0.7767	0.1	0.0417	0.9017	0.5714	0.6111	0.0417
ANN-DA	CV	0.7552	0.783	1.0	0.0625	0.7179	0.0357	0.9444	0.1818
	Test	0.5538	0.6981	0.2	0.0528	0.6923	0.2857	1.0	0.2159

TABLE 4 Summary of Approach 2 model classification results

Model	CV/Test	Macro F1	Accuracy	NR		R2H	
				TPR	FPR	TPR	FPR
PLS-DA	CV	0.7955	0.8553	0.8409	0.0741	0.9259	0.1591
	Test	0.761	0.8208	0.7955	0.0556	0.9444	0.2045
XGB-DA	CV	0.9612	0.978	0.9848	0.0556	0.9444	0.0152
	Test	0.8505	0.9214	0.9697	0.3148	0.6852	0.0303
ANN-DA	CV	0.7699	0.8239	0.7879	0.0	1.0	0.2121
	Test	0.7862	0.8396	0.8068	0.0	1.0	0.1932

and XGB-DA achieved test Macro F1 scores of approximately 55.38 and 55.43%, respectively.

Under the second approach, XGB-DA outperformed the other classifiers, achieving the highest test Macro F1 of 85.05% (Table 4). PLS-DA and ANN-DA exhibited comparable but slightly lower performance, with test Macro F1 scores of 76.1 and 78.62%, respectively. Analysis of model feature importance helps explain these differences (Figure A5 in the Appendix). For XGB-DA, the strongest contributors to accurate classification were spectral regions near 1016 and 1032 cm^{-1}. By contrast, ANN-DA emphasised the 1032 cm^{-1} region, while PLS-DA distributed importance across many spectral features rather than focusing on the most informative ones. This broader weighting likely reduced PLS-DA's overall performance and explains why ANN-DA performed slightly better. Upon closer inspection of XGB-DA, a lower FPR for R2H samples (approx. 3%) compared to the FNR (31.48%) is noteworthy (Table 4 and Figure 3) (remaining confusion matrices can be found in Figure A6 in the Appendix).

Our third and final approach, which implemented two separate model architectures within a single pipeline, did not improve Phase 1 classification of food versus frass, as no model surpassed the Macro F1 performance achieved under Approach 2 (Table 5). Under this framework, PLS-DA achieved a Macro F1 of 72.41%, while ANN-DA was 66.87% and XGB-DA was 48.87%, with corresponding variable-importance plots and confusion matrices shown in Figures A7 and A8 in the Appendix. Inspection of the ANN-DA and XGB-DA confusion matrices (Figure A8 in the Appendix) and associated error rates revealed substantial overfitting to frass-specific spectral features, as reflected by high false pos-

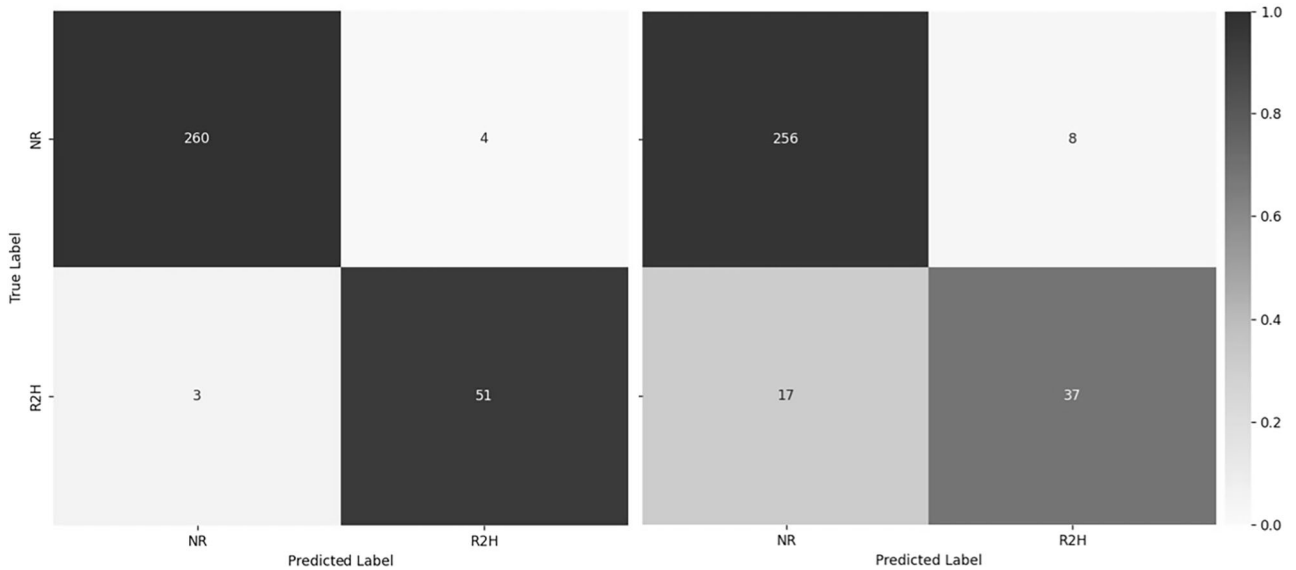

FIGURE 3 Confusion matrices for cross-validation (left) and testing (right) XGB-DA under Approach 2 framework. Rows represent true class labels and columns represent predicted class labels of not ready (NR) or ready-to-harvest (R2H) samples.

TABLE 5 Summary of Approach 3, Phase 1 model classification results

Model	CV/Test	Macro F1	Accuracy	Food		Frass	
				TPR	FPR	TPR	FPR
PLS-DA	CV	0.8407	0.9277	1.0	0.0799	0.9201	0.0
	Test	0.7241	0.8774	0.7	0.1042	0.8958	0.3
XGB-DA	CV	0.9816	0.9937	0.9667	0.0035	0.9965	0.0333
	Test	0.4887	0.8711	0.0333	0.0417	0.9583	0.9667
ANN-DA	CV	0.8685	0.9434	1.0	0.0625	0.9375	0.0
	Test	0.6687	0.8868	0.4	0.0625	0.9375	0.6

itive rates for frass (60.0% and 96.67%, respectively; Table 5). Given this overfitting behaviour and the added computational cost that would be required to introduce an additional downstream model to recover misclassified food spectra, this approach was not advanced to Phase 2 classification of NR versus R2H.

4 Discussion

In this pilot study, FT-IR spectroscopy was used to analyse the frass of BSFL as an indicator of time to harvest. After testing several strategies to classify spectra successfully, it was found that the binary classification system (Approach 2) yielded the most accuracy, sorting pan-material samples as either ready-to-harvest (R2H) or not ready (NR). We concluded that the biochemical variation between these two stages of larval development is distinct enough for FT-IR spectroscopy to differentiate between them.

There were three different approaches tested to classify pan-associated samples. Approach 1 was a multiclass system, Approach 2 employed binary classification and Approach 3 used dual-phase, binary classification. Approach 1 produced a suboptimal measurement of precision and recall. The non-linear models (XGB-DA and ANN-DA) performed more poorly relative to the linear model (PLS-DA), but all three were weak and inconsistent under this approach. These results suggest that treating food, NR and R2H samples as fully separable classes represents an unrealistic classification scenario. This outcome indicates substantial spectral overlap between food material and early-stage frass (as also evident from the confusion matrices, Figure A4B2 and A4C2 in the Appendix), reflecting the gradual and continuous nature of larval development in BSF. This finding motivated the adoption of more conservative and operationally representative classification strategies evaluated in subsequent approaches.

The second approach demonstrated substantially improved classification performance across all models.

Binary classification was the ideal strategy for determining when larvae are ready to harvest, indicating that the biochemical transition to this larval stage is more distinguishable. The nonlinear models, specifically XGB-DA, resulted in the highest Macro F1 scores, demonstrating their accuracy with precision and recall of spectra classification. This result points to the complexity of nutrient absorption by BSFL. They have an especially wide variety of digestive enzymes (Kim *et al*., 2011), which may provide some reasoning as to why nonlinear analysis better captures their digestion and developmental process.

The nonlinear models displayed strong levels of importance at specific peaks under Approach 2, particularly with the band near 1016 cm^{-1}, which is a region corresponding to polysaccharide hydroxyl and alkyl vibrations. Fiber is a significant nutritional factor in the Gainesville diet (Hogsette, 1992), explaining the detection of hydroxyl bending. This finding indicates that the 1016 cm^{-1} plays a key role in distinguishing NR from R2H BSFL, an observation that is also evident by visual inspection of the spectra in Figure 1. The observed trends are consistent with established principles of biochemical transformation during digestion, wherein dietary constituents undergo absorption and enzymatic modification and thus are not excreted in their original molecular forms (Karasov and Douglas, 2013). From an agricultural perspective, this suggests that a small number of chemically meaningful spectral markers can reliably differentiate larval rearing conditions.

Further analysis of the results from Approach 2 reflects the conservative decision tendency of XGB-DA, given the low FPR paired with the slightly higher FNR for R2H samples. In this instance, pans are more likely to be classified as NR rather than prematurely identified as R2H. From an operational standpoint, such behaviour is preferable, as it minimises the risk of harvesting underdeveloped material while allowing additional time for marginal cases to reach harvest readiness. While this conservative classification could cause a delay in production, it also helps avoid the harvesting of larvae that might not yet have reached their desired maturity level. Depending on the operational goals of the production system, this trade-off may be worth the risk, especially if achieving a target larval stage is prioritised over minor harvest delays.

Under Approach 3, the linear model performed better compared to the nonlinear model but was still poor compared to the precision and recall of Approach 2. The high FPR also demonstrated that dual-phase classification was unfavourable, further supporting that binary classification through our second approach was the superior method.

In the past, other researchers have also demonstrated IR spectroscopy's potential to improve the BSF industry. For example, Hoffman *et al*. (2022) successfully monitored developmental changes in BSFL using attenuated total reflectance (ATR) mid-infrared (MIR) spectroscopy. Their research analysed spectral differences in various BSFL sample types, including later-stage instars and frass, for the purpose of classifying larvae according to waste stream diet (Hoffman *et al*., 2022). The present study, however, focuses on using frass specifically as a biomarker of larval physiology and predicting harvest readiness. While these two studies differ in many ways, they both ultimately display the meaningful and diverse applications of spectral technology to BSF production.

The findings of this study demonstrate the potential for this rapid and non-destructive analytical technique to be applied in the mass production of BSFL. Implementing FT-IR does require an initial investment, but the direct cost of analysis is nearly zero. Based on our experience, one instrument can be used by a laboratory or a group of researchers to analyse samples and an undergraduate can independently operate the spectrometer with relatively little training.

Utilising such technology in the context of this research leads to the possibility of improved efficiency through automated harvest. However, it is important to address shortcomings to the present study. One limitation here is that other biological performance data, such as larval mass, survival rate and protein-fat content of larvae were not measured on harvest day. We exclusively analysed frass substrate, not insect body, to predict larval age. In a future study, other biological observations, like the ones mentioned, could be used as confirmation of larval maturity.

Additionally, future research could apply IR spectroscopy to other stages of larval development or further investigation of their digestion with different diets. The use of one standardised laboratory diet can be viewed as a limitation because the models were trained on frass substrate produced from digestion of only the Gainesville feed, so if other organic waste products were used, the classification accuracy may decrease. Therefore, additional sampling across a wider variety of feedstocks is the next step towards validating and advancing this research because it would allow for evaluation of the classification model's performance on diverse substrate diets that are relevant to sustainable waste management. Future studies should also include mul-

tiple life cycles and independent production batches to improve the robustness of the resulting models and further validate them for industrial adoption.

5 Conclusion

Through this pilot study, it was demonstrated that FT-IR could be used for real-time assessment of large-scale production of BSFL. This method eliminates the need for a batch system approach for mass production, which decreases losses due to pans either being under- or overdeveloped. Among the evaluated models, the binary classification framework that grouped NR and food-only samples as NR achieved the strongest performance for differentiating R2H vs NR larvae. In particular, XGB-DA yielded greater than 85% balanced accuracy and macro F1 score during external validation, while conservatively favouring NR classifications under uncertainty. Collectively, these findings support the integration of FT-IR spectroscopy with machine learning as a practical and scalable solution for improving harvest timing and operational efficiency in BSF production.

Acknowledgements

We acknowledge the Department of Entomology at Texas A&M University in College Station, Texas, USA for its support. This material is based upon work supported by United States Department of Agriculture, National Institute of Food and Agriculture awards 2023-38821-39977 and 2023-67019-39248, as well as the National Science Foundation under Award Nos 2052454, 2052565 and 2052788. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Special thanks to Dr Ezra Bailey, Dr Sienna McPeek and Dr Amber MacInnis for reviewing and editing this manuscript.

Conflict of interest

Ther authors declare no conflict of interest.

Funding

This material is based upon work supported by the United States Department of Agriculture, National Institute of Food and Agriculture awards 2023-38821-39961 and 2023-67019-39248 and the National Science Foundation under Award Nos. 2052454, 2052565, 2052788. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Alagappan, S., Kolobaric, A., Hoffman, L.C. and Cozzolino, D., 2024. Current and potential applications of vibrational spectroscopy as a tool in black soldier fly production and the circular economy. Applied Sciences 14: 7318. https://doi.org/10.3390/app14167318

Barragan-Fonseca, K.B., Dicke, M. and van Loon, J.J.A., 2018. Influence of larval density and dietary nutrient concentration on performance, body protein and fat contents of black soldier fly larvae (*Hermetia illucens*). Entomologia Experimentalis et Applicata 166: 761-770. https://doi.org/10.1111/eea.12716

Barton, B., Thomson, J., Lozano Diz, E. and Portela, R., 2022. Chemometrics for raman spectroscopy harmonization. Applied Spectroscopy 76: 1021-1041. https://doi.org/10.1177/00037028221094070

Baum, A., Dominiak, M., Vidal-Melgosa, S., Willats, W.G.T., Søndergaard, K.M., Hansen, P.W., Meyer, A.S. and Mikkelsen, J.D., 2017. Prediction of pectin yield and quality by FTIR and carbohydrate microarray analysis. Food and Bioprocess Technology 10: 143-154. https://doi.org/10.1007/s11947-016-1802-2

Cammack, J.A. and Tomberlin, J.K., 2017. The impact of diet protein and carbohydrate on select life-history traits of the black soldier fly *Hermetia illucens* (L.) (Diptera: Stratiomyidae). Insects 8: 56. https://doi.org/10.3390/insects8020056

Chavez, M.Y., Uchanski, M. and Tomberlin, J.K., 2023. Impacts of black soldier fly, (Diptera: Stratiomyidae) larval frass on tomato production. Journal of Economic Entomology 116: 1490-1495. https://doi.org/10.1093/jee/toad150

Chavez, M.Y., Uchanski, M. and Tomberlin, J.K., 2024. Impacts of black soldier fly, *Hermetia illucens*, larval frass on lettuce and arugula production. Frontiers in Sustainable Food Systems 8: 1399932. https://doi.org/10.3389/fsufs.2024.1399932

Chavez, M.Y., Villa Ignacio, A., Craver, J.K. and Bousselot, J., 2025. Investigating black soldier fly larval (*Hermetia illu-*

cens) frass applications as a partial peat replacement and liquid fertiliser in brassicaceae crop production. Agrochemicals 4: 8. https://doi.org/10.3390/agrochemicals4020008

Grossule, V., Henjak, M., Beggio, G. and Tomberlin, J.K., 2025. Biowaste treatment using black soldier fly larvae: effect of substrate macronutrients on process performance. Journal of Environmental Management 373: 123605. https://doi.org/10.1016/j.jenvman.2024.123605

Hoffman, L.C., Zhang, S., Alagappan, S., Wills, V., Yarger, O. and Cozzolino, D., 2022. Monitoring compositional changes in black soldier fly larvae (BSFL) sourced from different waste stream diets using attenuated total reflectance mid infrared spectroscopy and chemometrics. Molecules 27: 7500. https://doi.org/10.3390/molecules27217500

Hogsette, J.A., 1992. New diets for production of house flies and stable flies (Diptera: Muscidae) in the laboratory. Journal of Economic Entomology 85: 2291-2294. https://doi.org/10.1093/jee/85.6.2291

Holman, A.P., Pickett, D.N., Orr, A.E., Tarone, A.M. and Kurouski, D., 2024. A nondestructive technique for the sex identification of third instar *Cochliomyia macellaria* larvae. Journal of Forensic Sciences 69: 2075-2081. https://doi.org/10.1111/1556-4029.15619

Holmes, L.A., Vanlaerhoven, S.L. and Tomberlin, J.K., 2012. Relative humidity effects on the life history of *Hermetia illucens* (Diptera: Stratiomyidae). Environmental Entomology 41: 971-978. https://doi.org/10.1603/en12054

Jian, S., Zhang, L., Ding, N., Yang, K., Xin, Z., Hu, M., Zhou, Z., Zhao, Z., Deng, B. and Deng, J., 2022. Effects of black soldier fly larvae as protein or fat sources on apparent nutrient digestibility, fecal microbiota and metabolic profiles in beagle dogs. Frontiers in Microbiology 13: 1044986. https://doi.org/10.3389/fmicb.2022.1044986

Jiang, S.-Y., Shen, K.-W., Brandón, M.G., Lu, S.-B., Tomberlin, J.K., Tang, X.-T., Wang, H., Xiang, F.-M., Chen, X.-X. and Zhang, Z.-J., 2025. Using black soldier fly larval frass to restore soil health. Bioresource Technology 432: 132701. https://doi.org/10.1016/j.biortech.2025.132701

Karasov, W.H. and Douglas, A.E., 2013. Comparative digestive physiology. Comprehensive Physiology 3: 741-783. https://doi.org/10.1002/j.2040-4603.2013.tb00501.x

Kim, W., Bae, S., Park, K., Lee, S., Choi, Y., Han, S. and Koh, Y., 2011. Biochemical characterization of digestive enzymes in the black soldier fly, *Hermetia illucens* (Diptera: Stratiomyidae). Journal of Asia-Pacific Entomology 14: 11-14. https://doi.org/10.1016/j.aspen.2010.11.003

Kong, J. and Yu, S., 2007. Fourier transform infrared spectroscopic analysis of protein secondary structures. Acta Biochimica et Biophysica Sinica 39: 549-559. https://doi.org/10.1111/j.1745-7270.2007.00320.x

Ma, J., Lei, Y., ur Rehman, K., Yu, Z., Zhang, J., Li, W., Li, Q., Tomberlin, J.K. and Zheng, L., 2018. Dynamic effects of initial pH of substrate on biological growth and metamorphosis of black soldier fly (Diptera: Stratiomyidae). Environmental Entomology 47: 159-165. https://doi.org/10.1093/ee/nvx186

Manduca, G., Wilson, L.T., Stefanini, C. and Romano, D., 2025. Automated detection of larval stages of the black soldier fly (*Hermetia illucens* Linnaeus) through deep learning augmented with optical flow. Information Processing in Agriculture 12: 501-510. https://doi.org/10.1016/j.inpa.2025.05.001

Meneguz, M., Gasco, L. and Tomberlin, J.K., 2018. Impact of pH and feeding system on black soldier fly (*Hermetia illucens*, L; Diptera: Stratiomyidae) larval development. PLOS One 13: e0202591. https://doi.org/10.1371/journal.pone.0202591

Miranda, C.D., Cammack, J.A. and Tomberlin, J.K., 2020. Mass production of the black soldier fly, *Hermetia illucens* (L.), (Diptera: Stratiomyidae) reared on three manure types. Animals 10: 1243. https://doi.org/10.3390/ani10071243

Mohan Rao, U., Fofana, I., Kartheek, R., Yapi, K.M.L. and Jaya, T., 2019. Mineral oil and ester based oil/paper insulation decaying assessment by FTIR measurements. In: Proceedings of the 21st international symposium on high voltage engineering. Lecture Notes in Electrical Engineering 598, pp. 615-624. Available at https://doi.org/10.1007/978-3-030-31676-1_58

Mouithys-Mickalad, A., Schmitt, E., Dalim, M., Franck, T., Tome, N.M., van Spankeren, M., Serteyn, D. and Paul, A., 2020. Black soldier fly (*Hermetia illucens*) larvae protein derivatives: potential to promote animal health. Animals 10: 941. https://doi.org/10.3390/ani10060941

Myers, H.M., Tomberlin, J.K., Lambert, B.D. and Kattes, D., 2008. Development of black soldier fly (Diptera: Stratiomyidae) larvae fed dairy manure. Environmental Entomology 37: 11-15. https://doi.org/10.1093/ee/37.1.11

Newton, G.L., Booram, C.V., Barker, R.W. and Hale, O.M., 1977. Dried *Hermetia illucens* larvae meal as a supplement for swine. Journal of Animal Science 44: 395-400. https://doi.org/10.2527/jas1977.443395

Nguyen, T.T.X., Tomberlin, J.K. and Vanlaerhoven, S., 2015. Ability of black soldier fly (Diptera: Stratiomyidae) larvae to recycle food waste. Environmental Entomology 44: 406-410. https://doi.org/10.1093/ee/nvv002

Parihar, A., Vongsvivut, J. and Bhattacharya, S., 2019. Synchrotron-based infra-red spectroscopic insights on thermocatalytic conversion of cellulosic feedstock to levoglucosenone and furans. ACS Omega 4: 8747-8757. https://doi.org/10.1021/acsomega.8b03681

Pomerantsev, A.L. and Rodionova, O.Y., 2025. A new method for soft probabilistic discrimination of ranked classes. Microchemical Journal 212: 113329. https://doi.org/10.1016/j.microc.2025.113329

Powers, D.M., 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint. arXiv:2010.16061. https://doi.org/10.48550/arXiv.2010.16061

Proietti, M., Marini, A., Garinei, A., Rossi, G., Bianchi, F. and Marconi, M., 2022. Non-invasive measurements for characterization of Hermetia Illucens (BSF) life cycle in rearing plant. 2022 IEEE Workshop on Metrology for Agriculture and Forestry (MetroAgriFor). https://doi.org/10.1109/MetroAgriFor55389.2022.9964687

Robert, P., Marquis, M., Barron, C., Guillon, F. and Saulnier, L., 2005. FT-IR investigation of cell wall polysaccharides from cereal grains. Arabinoxylan infrared assignment. Journal of Agricultural and Food Chemistry 53: 7014-7018. https://doi.org/10.1021/jf051145y

Santos, J.I., Fillat, U., Martín-Sampedro, R., Ballesteros, I., Manzanares, P., Ballesteros, M., Eugenio, M.E. and Ibarra, D., 2015. Lignin-enriched fermentation residues from bioethanol production of fast-growing poplar and forage sorghum. BioResources 10: 5215-5232. http://hdl.handle.net/10261/290250

Sheppard, D.C., Newton, G.L., Thompson, S.A. and Savage, S., 1994. A value added manure management system using the black soldier fly. Bioresource Technology 50: 275-279. https://doi.org/10.1016/0960-8524(94)90102-3

Sheppard, D.C., Tomberlin, J.K., Joyce, J.A., Kiser, B.C. and Sumner, S.M., 2002. Rearing methods for the black soldier fly (Diptera: Stratiomyidae). Journal of Medical Entomology 39: 695-698. https://doi.org/10.1603/0022-2585-39.4.695

St-Hilaire, S., Sheppard, C., Tomberlin, J.K., Irving, S., Newton, L., McGuire, M.A., Mosley, E.E., Hardy, R.W. and Sealey, W., 2007a. Fly prepupae as a feedstuff for rainbow trout, *Oncorhynchus mykiss*. Journal of the World Aquaculture Society 38: 59-67. https://doi.org/10.1111/j.1749-7345.2006.00073.x

St-Hilaire, S., Cranfill, K., McGuire, M.A., Mosley, E.E., Tomberlin, J.K., Newton, L., Sealey, W., Sheppard, C. and Irving, S., 2007b. Fish offal recycling by the black soldier fly produces a foodstuff high in omega-3 fatty acids. Journal of the World Aquaculture Society 38: 309-313. https://doi.org/10.1111/j.1749-7345.2007.00101.x

Tariq, M.R., Liu, S., Wang, F., Wang, H., Mo, Q., Zhuang, Z., Zheng, C., Liang, Y., Liu, Y., ur Rehman, K., Helvaci, M., Qin, J. and Li, C., 2025. Black soldier fly: a keystone species for the future of sustainable waste management and nutritional resource development: a review. Insects 16: 750. https://doi.org/10.3390/insects16080750

Tomberlin, J.K., Yang, F. and Cammack, J.A., 2023. Google Patents. Long-term storage of black soldier fly larvae. The Texas A&M University System, College Station, TX. Available at https://patents.google.com/patent/US12490725B2/en

Varoquaux, G., 2018. Cross-validation failure: small sample sizes lead to large error bars. NeuroImage 180: 68-77. https://doi.org/10.1016/j.neuroimage.2017.06.061

Westerhuis, J.A., Hoefsloot, H.C.J., Smit, S., Vis, D.J., Smilde, A.K., van Velzen, E.J.J., van Duijnhoven, J.P.M. and van Dorsten, F.A., 2008. Assessment of PLSDA cross validation. Metabolomics 4: 81-89. https://doi.org/10.1007/s11306-007-0099-6

Yamamoto, F.Y., Suehs, B.A., Ellis, M., Bowles, P.R., Older, C.E., Hume, M.E., Bake, G.G., Cammack, J.A., Tomberlin, J.K. and Gatlin III, D.M., 2022. Dietary fishmeal replacement by black soldier fly larvae meals affected red drum (*Sciaenops ocellatus*) production performance and intestinal microbiota depending on what feed substrate the insect larvae were offered. Animal Feed Science and Technology 283: 115179. https://doi.org/10.1016/j.anifeedsci.2021.115179

Yang, F. and Tomberlin, J.K., 2020. Comparing selected life-history traits of black soldier fly (Diptera: Stratiomyidae) larvae produced in industrial and bench-top-sized containers. Journal of Insect Science 20: 1-6. https://doi.org/10.1093/jisesa/ieaa113

Yang, J., Xu, J., Zhang, X., Wu, C., Lin, T. and Ying, Y., 2019. Deep learning for vibrational spectral analysis: recent progress and a practical guide. Analytica Chimica Acta 1081: 6-17. https://doi.org/10.1016/j.aca.2019.06.012

Yu, Z., Xie, C., Zhang, Z., Huang, Z., Zhou, J. and Wang, C., 2024. Microbial fermentation and black soldier fly feeding to enhance maize straw degradation. Chemosphere 353: 141498. https://doi.org/10.1016/j.chemosphere.2024.141498

Zhang, Y., Yang, C.-Y., Li, C., Xu, Z., Peng, P., Xue, C., Tomberlin, J.K., Hu, W. and Cao, Y., 2021. Black soldier fly (*Hermetia illucens* L.) larval diet improves $CD8^+$ lymphocytes proliferation to eliminate chicken coronavirus at an early infection stage. Veterinary Microbiology 260: 109151. https://doi.org/10.1016/j.vetmic.2021.109151

Zhou, F., Tomberlin, J.K., Zheng, L., Yu, Z. and Zhang, J., 2013. Developmental and waste reduction plasticity of three black soldier fly strains (Diptera: Stratiomyidae) raised on different livestock manures. Journal of Medical Entomology 50: 1224-1230. https://doi.org/10.1603/ME13021

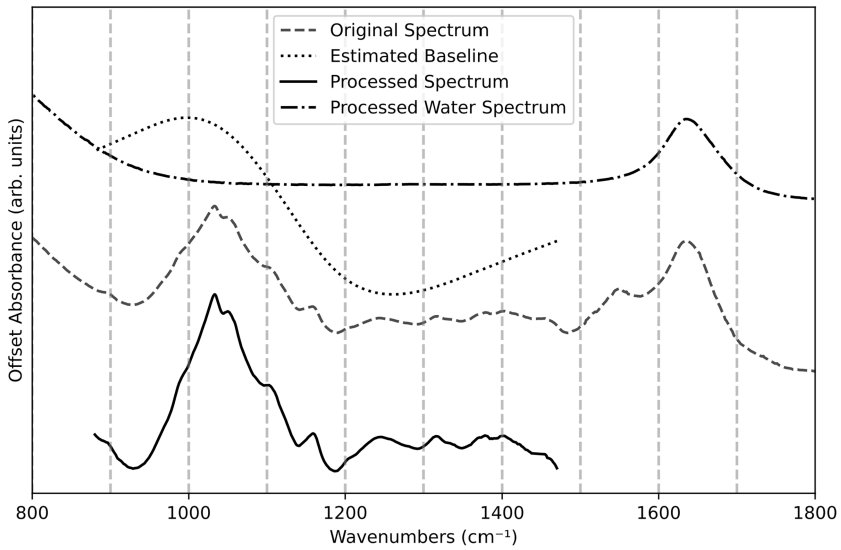

FIGURE A1 Pre-processing effects on FT-IR spectra before modelling. Offset absorbance is displayed for comparison of the original spectrum, estimated baseline and processed spectra without overlap.

TABLE A1 Machine learning models and relevant parameters and Python libraries

Model	Description	Tunable parameters	Python libraries
PLS-DA	A supervised latent-variable method that projects predictors into components maximising covariance between features and class membership.	LVs: 2, 3...20	sklearn
XGB-DA	An ensemble of decision trees built sequentially, where each new tree minimises the residual error of the preceding trees via gradient boosting.	Learning rate: [0.01, 0.1, 0.2] Maximum depth: [1, 3, 6, 10] No. of estimators: [50, 100, 200, 500]	xgboost
ANN-DA	A fully connected neural network that learns weighted connections between nodes by minimising prediction error through backpropagation.	Hidden layer nodes: [(1024, 5120], (512, 256)] Learning rate: [0.001, 0.005, 0.0001] Dropout: [0, 0.1, 0.2, 0.3] L1 Penalisation: [0, 0.01, 0.001]	Torch; torch.nn

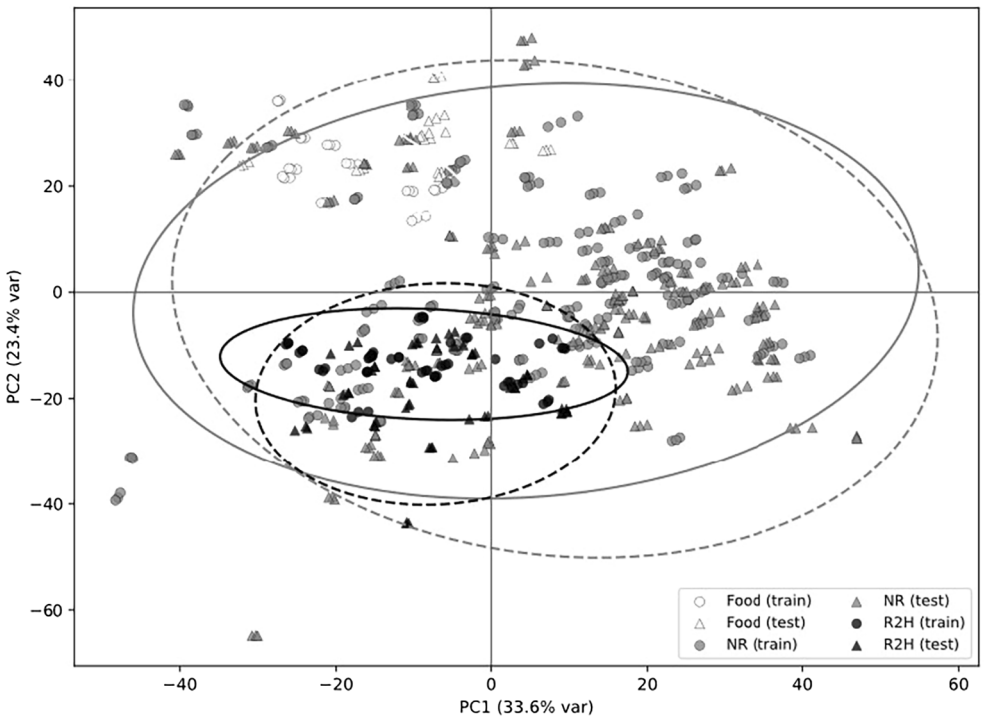

FIGURE A2 Principal component (PC) analysis PC1 and PC2 (approx. 57% model variance) score plot for training/CV and test data. The solid and dashed 95% confidence ellipses correspond to train and test datasets, respectively. Our independent train and test datasets share the same underlying distribution of discrimination variance indicating strong generalisation potential for modelling.

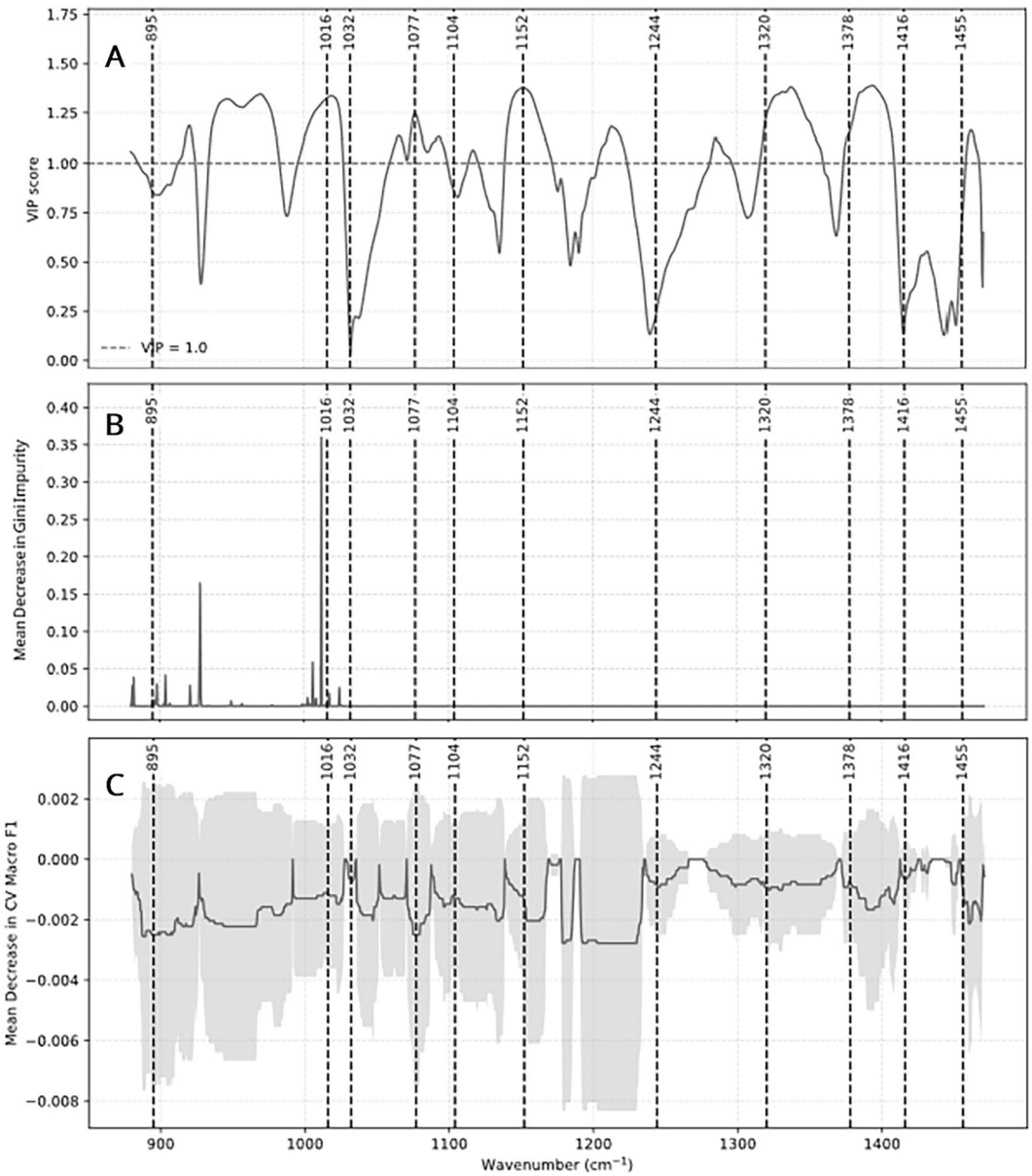

FIGURE A3 Importance measures across all features for Approach 1 (A) PLS-DA, (B) XGB-DA and (C) ANN-DA models. PLS-DA VIP plots the relative amount of variance per wavenumber across all wavenumbers and over all LVs, where values above 1 are considered significant contributors to class discrimination. XGB-DA importance plots the mean decrease in Gini impurity, a measure of the probability of misclassification on random data, normalised across all wavenumbers. ANN-DA importance plots the mean decrease in CV Macro F1 for each wavenumber by removing one at a time and re-evaluating the model each time across all wavenumbers.

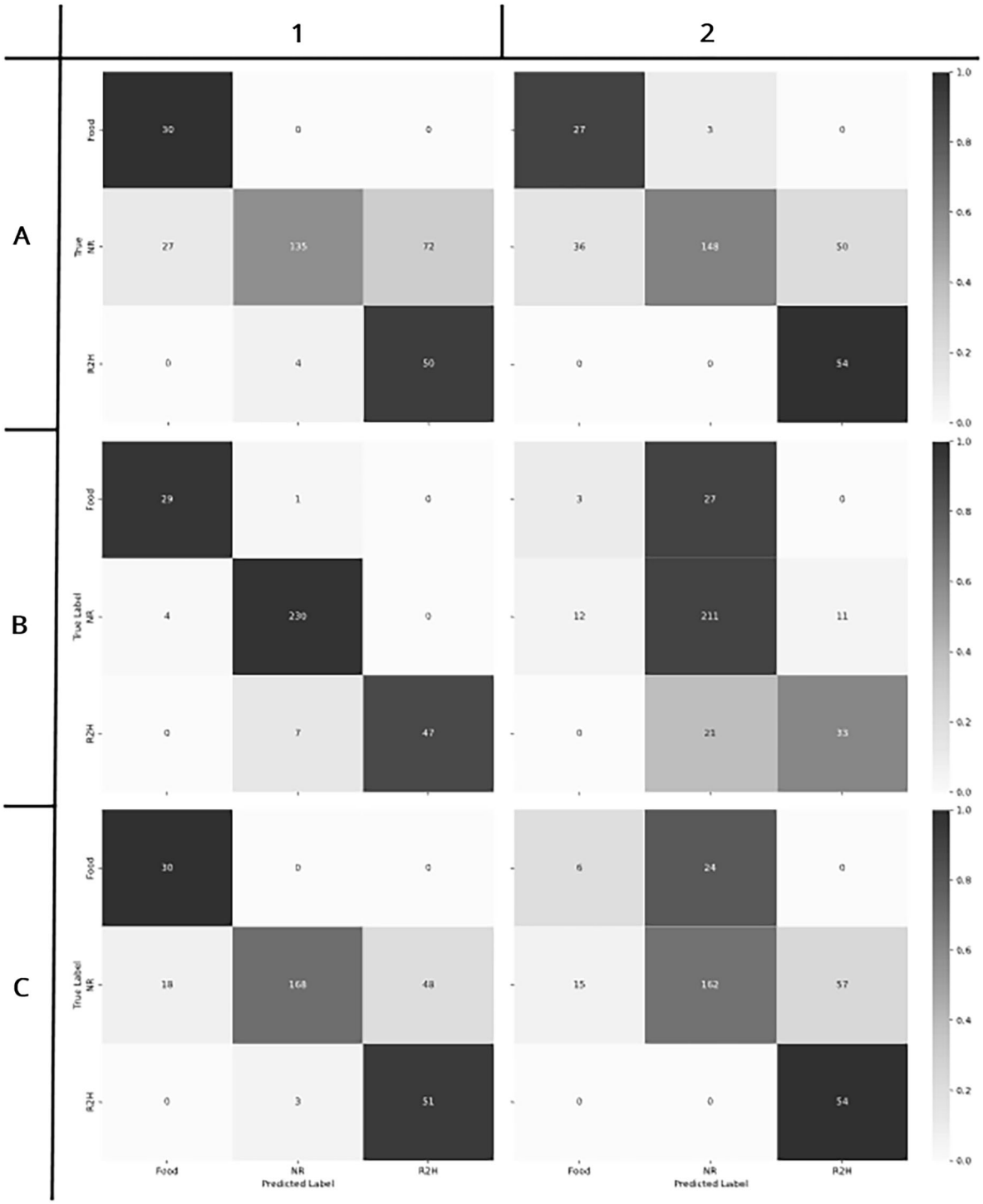

FIGURE A4 Confusion matrices for Approach 1 (A) PLS-DA, (B) XGB-DA and (C) ANN-DA models using the (1) train and (2) test datasets. Rows represent true class labels and columns represent predicted class labels of Gainesville diet (food), not ready (NR), or ready-to-harvest (R2H) samples.

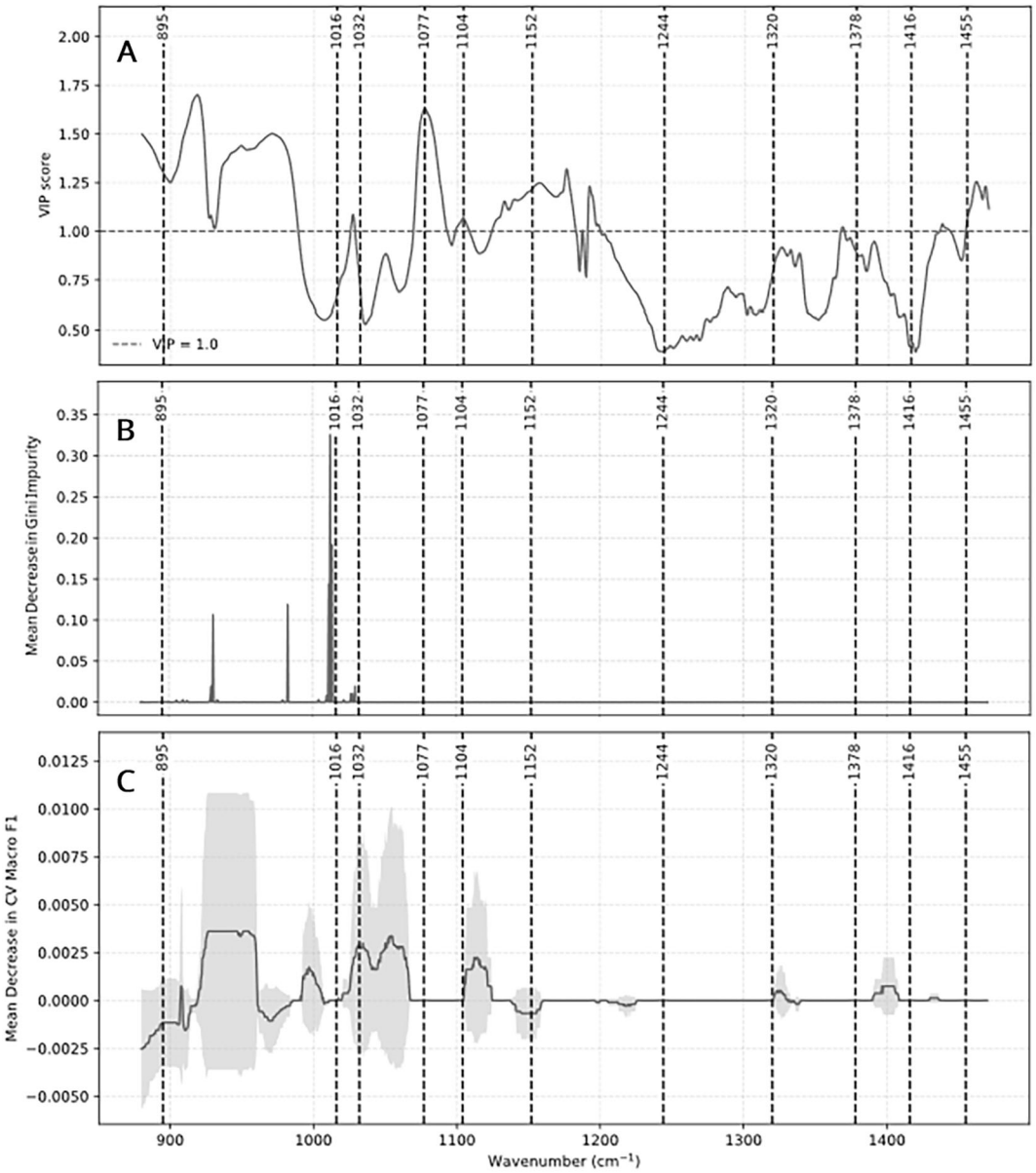

FIGURE A5 Importance measures across all features for Approach 2 (A) PLS-DA, (B) XGB-DA and (C) ANN-DA models. PLS-DA VIP plots the relative amount of variance per wavenumber across all wavenumbers and over all LVs, where values above 1 are considered significant contributors to class discrimination. XGB-DA importance plots the mean decrease in Gini impurity, a measure of the probability of misclassification on random data, normalised across all wavenumbers. ANN-DA importance plots the mean decrease in CV Macro F1 for each wavenumber by removing one at a time and re-evaluating the model each time across all wavenumbers.

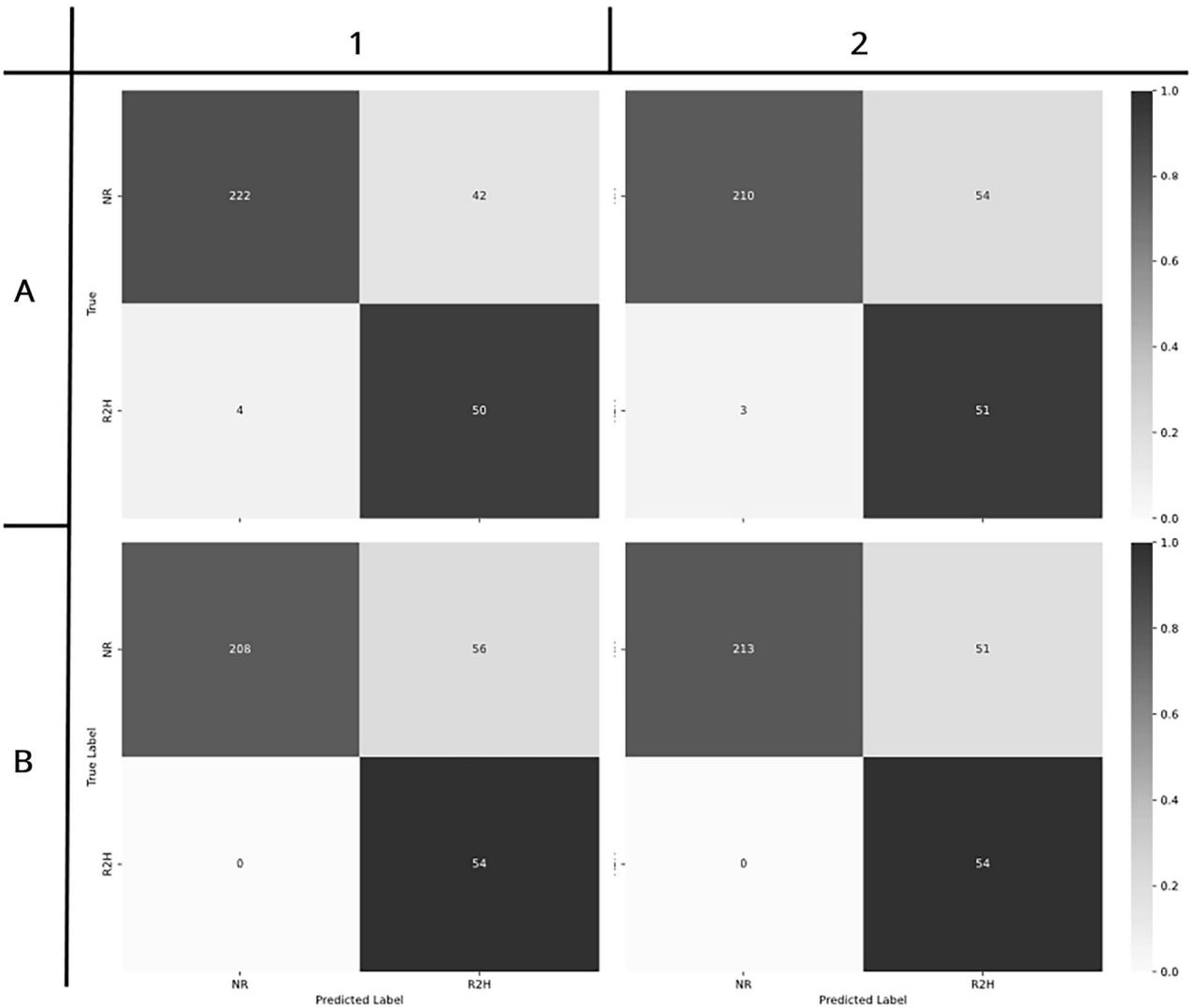

FIGURE A6 Confusion matrices for Approach 2 (A) PLS-DA and (B) ANN-DA models using the (1) train and (2) test datasets. Rows represent true class labels and columns represent predicted class labels of not ready (NR) or ready-to-harvest (R2H) samples.

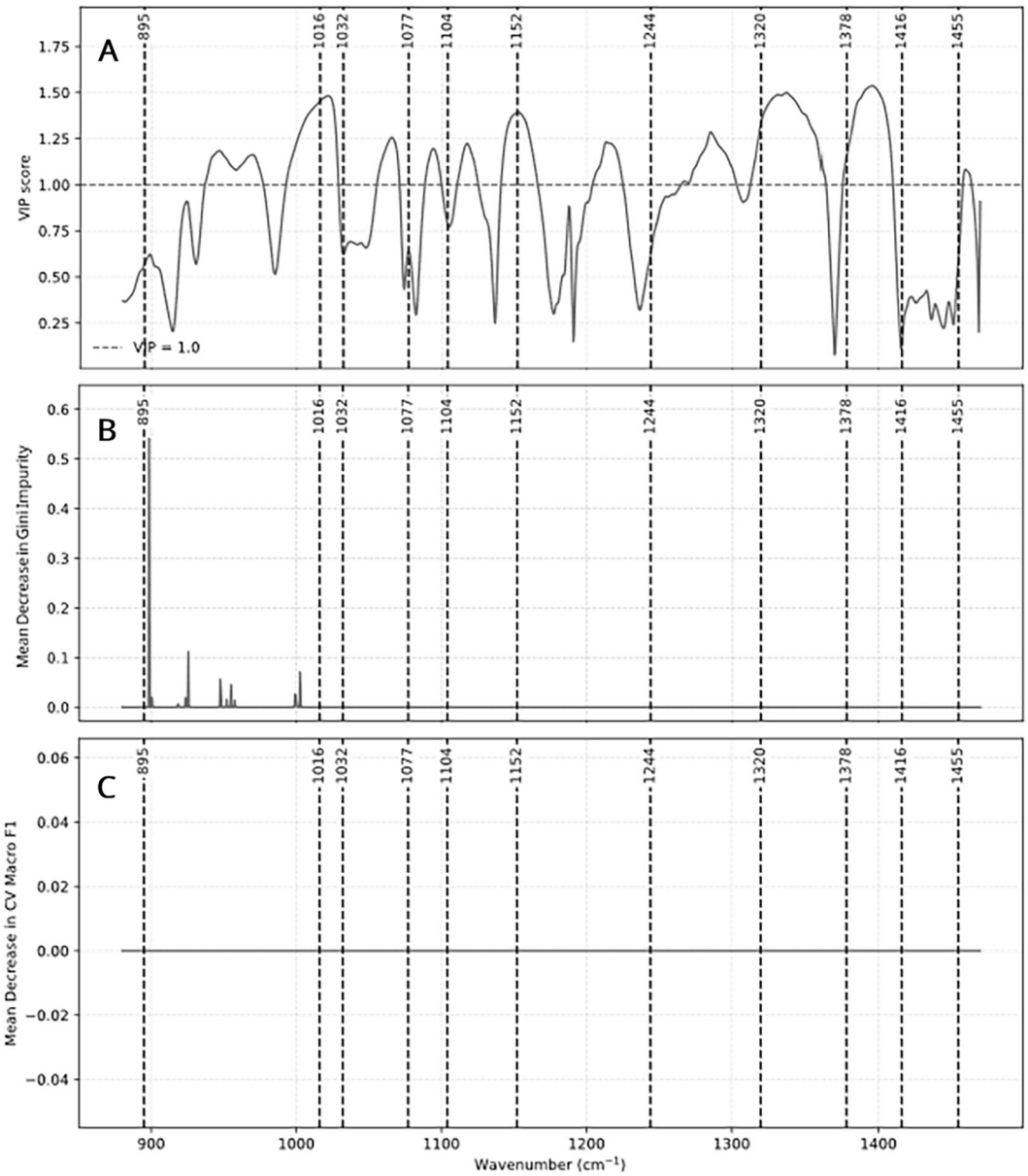

FIGURE A7 Importance measures across all features for Approach 3 (A) PLS-DA, (B) XGB-DA and (C) ANN-DA models. PLS-DA VIP plots the relative amount of variance per wavenumber across all wavenumbers and over all LVs, where values above 1 are considered significant contributors to class discrimination. XGB-DA importance plots the mean decrease in Gini impurity, a measure of the probability of misclassification on random data, normalised across all wavenumbers. ANN-DA importance plots the mean decrease in CV Macro F1 for each wavenumber by removing one at a time and re-evaluating the model each time across all wavenumbers.

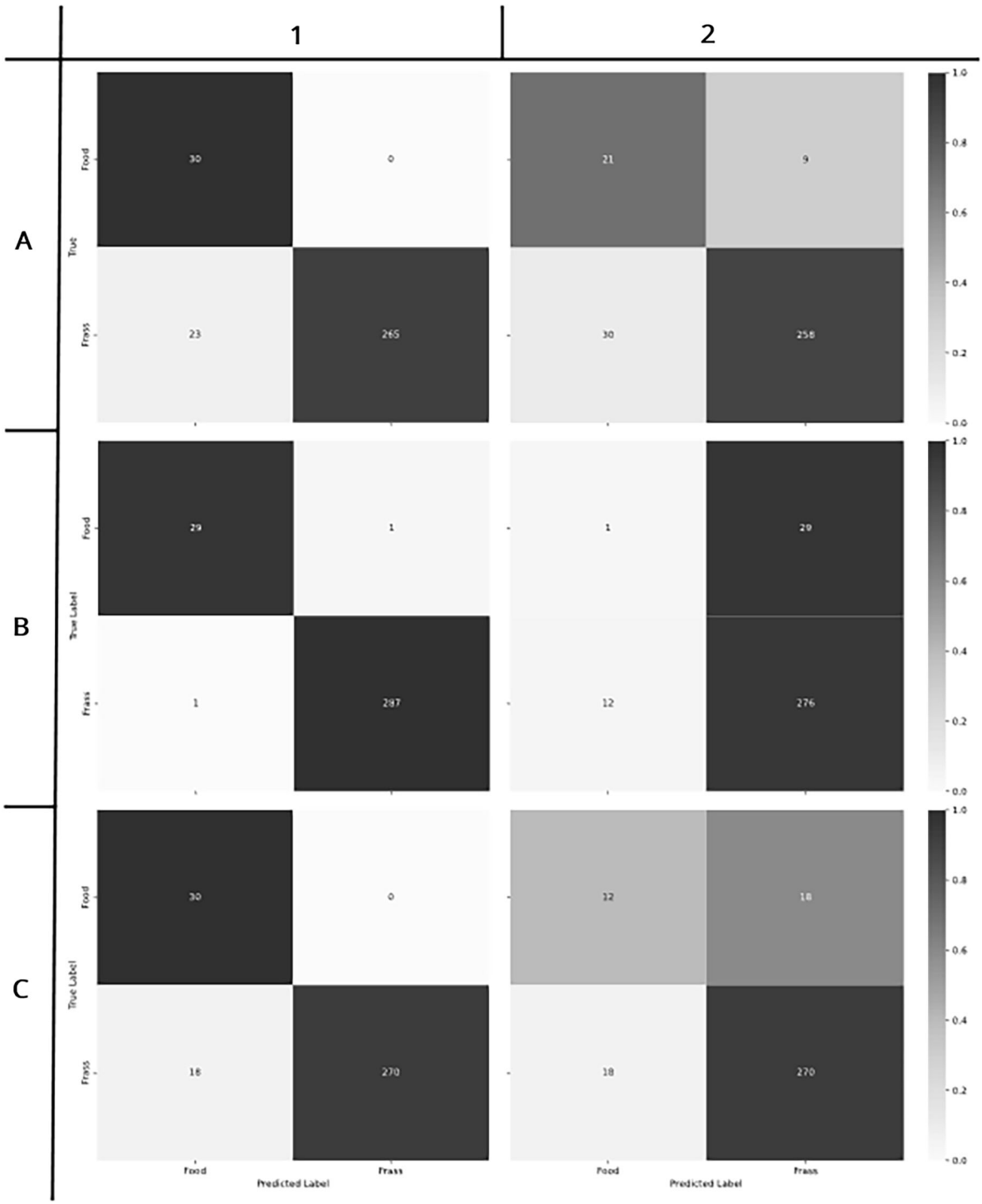

FIGURE A8 Confusion matrices for Approach 3 (A) PLS-DA, (B) XGB-DA and (C) ANN-DA models using the (1) train and (2) test datasets. Rows represent true class labels and columns represent predicted class labels of Gainesville diet (food) or pan material (frass) samples.