StrawSense

WFBR Trainee Project - Identifying *Botrytis* Related Volatiles In Strawberries

Miaomiao Yao, Charlotte Harbers and Elizaveta Genke

PUBLIC

WAGENINGEN
UNIVERSITY & RESEARCH

Identifying Botrytis Related Volatiles In Strawberries

WFBR Trainee Project

Authors: M. (Miaomiao) Yao PhD, C.J. (Charlotte) Harbers MSc, E. (Liza) Genke MSc

Institute: Wageningen Food & Biobased Research

This study was carried out by Wageningen Food & Biobased Research, funded by Wageningen Food & Biobased Research.

Wageningen Food & Biobased Research
Wageningen, December 2024

Public

Report 2642
DOI: 10.18174/684103

WAGENINGEN
UNIVERSITY & RESEARCH

Version: Final
Reviewer: dr. R.E. (Rob) Schouten
Approved by: dr.ir. H. (Henk) Wensink
Carried out by: Wageningen Food & Biobased Research
Funded by: Wageningen Food & Biobased Research
This report is: Public

PO box 17, 6700 AA Wageningen, The Netherlands, T + 31 (0)317 48 00 84, E info.wfbr@wur.nl, www.wur.eu/wfbr.

Contents

Summary

Botrytis cinerea (B. cinerea) is a common fungal pathogen of strawberries. It is highly adaptive and capable of surviving under cold storage conditions, posing a significant threat throughout the supply chain. The development of non-destructive, early detection methods are essential to timely identify infected (batches of) strawberries. In this regard, this study aimed to identify volatile organic compounds (VOCs) associated with Botrytis infection and develop machine learning models for predictive analysis.

To investigate this, strawberries of cultivar Favori were divided into two treatment groups: one control group inoculated with water and a test group inoculated with Botrytis. VOCs were analysed using thermal desorption gas chromatography mass spectrometry (TD-GC-MS), whilst infection was monitored through imaging. Partial Least Squares Discriminant Analysis (PLS-DA) was used to identify volatiles most effective at separating healthy from infected strawberries. Subsequently, random forest models were trained to predict both infection class (healthy vs. Botrytis) and infection rate based on the volatile profiles.

Among others, the PLS-DA identified VOCs 1-propanol, 2-methyl-, 1-butanol, 2-methyl-, and pentanoic acid, 4-methyl-, methyl ester as the most discriminatory biomarkers for Botrytis infection. The random forest classifier achieved an accuracy of 87% in predicting infection class, whilst the regressor explained 70% of the variance in infection rate (R^2=0.70). Both models used slightly different subsets of VOCs in their final configurations, reflecting differences in the optimal number and selection of features.

These findings demonstrate that volatile information can be used to effectively detect Botrytis infection in strawberries. However, the generalizability of the identified volatiles is constrained by the study's limitations, including the limited size of the dataset, the inclusion of a single cultivar and the feature selection method (recursive feature elimination). Future research should further investigate volatile variability within and across cultivars to ensure the robustness and applicability of these detection methods in the strawberry supply chain.

1 Introduction

Strawberries are one of the most perishable fruits, especially during the postharvest stage, even when they are apparently healthy at the time of harvest, they can undergo spoilage [1]. *Botrytis cinerea* (*B. cinerea*) is regarded as the most common and important fungus among the strawberry postharvest pathogens [2]. It is a highly adaptive fungal pathogen capable of infecting strawberries during both pre-harvest and post-harvest stages. Under optimal conditions, such as high humidity and moderate temperatures, Botrytis can cause up to 50% yield losses in strawberries [3]. Furthermore, the fungus can remain viable under cold storage conditions, making it a significant threat throughout the supply chain. These losses pose a challenge to the sustainability and economic stability of certain fresh produce markets, highlighting the urgent need for innovative approaches to minimize post-harvest waste.

In strawberry cultivation, the use of synthetic fungicides and bio fungicides plays an essential role, with up to 20 different fungicides commonly applied to control diseases like *B. cinerea* [4]. Without these fungicides, successful strawberry cultivation would be extremely challenging due to the susceptibility of the crop to various pathogens [5]. Although most fungicides, when properly applied, degrade quickly to levels below the maximum residue limit (MRL), these fungicides may leave residues on fruit if not applied correctly, thus causing health and environmental concerns [6]. Additionally, during the postharvest stage, fungicide use is strictly prohibited, leading to significant challenges in managing postharvest decay. Therefore, it is necessary and urgent to find alternative methods to control the postharvest decay in strawberries, and these methods should be safe, environmentally friendly, and sustainable. Such approaches, which may include non-synthetic fungicides or biocontrol strategies, align with global green and organic agriculture initiatives, meeting modern consumer expectations for healthier and more sustainable produce. In this context, volatile organic compounds (VOCs) can serve as indicators of Botrytis infection, providing a non-destructive method for early detection rather than being linked directly to fungicide usage.

VOCs are molecules emitted during fruit metabolic processes, and their profiles are often altered in response to physiological changes, including fungal infections [7, 8]. For instance, one study has shown that *B. allii* and *B. cepacia* inoculated onion bulbs emit different volatile metabolites from healthy onions and demonstrated that VOCs could potentially be used as biomarkers for onion postharvest disease detection during storage [9]. Another study proved that volatile compounds act as biomarkers for stem-end rot and green mould of citrus and can be used to discriminate the infected citrus fruit [10]. In strawberries, several Botrytis-related volatiles, including 3-methylbutanal, cis-4-decenal, 2-methyl-1-butanol, 2-methyl-1-propanol, 1-octen-3-one, 1-octen-3-ol, ethyl butanoate and 1-hexanol, have been identified as potential biomarkers for early detection of *B. cinerea* infections [11, 28]. In light of these findings, it is plausible to suggest that the detection and analysis of VOCs could serve as a powerful tool for the early identification of Botrytis decay. However, despite their potential, the application of VOC analysis in postharvest supply chain management remains largely underutilized.

VOCs are inherently complex in their composition, origin, synthesis, and breakdown, making them difficult to fully comprehend using traditional analytical approaches. This complexity presents a challenge for understanding their roles in postharvest quality and spoilage processes. Machine learning (ML) has shown remarkable potential in agricultural and food science, particularly in uncovering meaningful relationships within complex datasets that would be difficult to interpret [12]. For instance, combining VOC analysis (HS-GC-MS-IMS) with ML techniques including linear discriminant analysis (LDA), support vector machines (SVM), and the k-nearest-neighbor (kNN), can significantly enhance classification accuracy and highlight unique volatile profile of citrus juices [13]. In addition, ML also has been utilized to estimate the post fermentation year of pu-erh tea [14], and to evaluate food safety and quality in food processing [15]. For postharvest research, machine learning enables the collection and analysis of data from various fruit sources to identify the optimal preservation conditions for each fruit type. The data can encompass environmental variables such as temperature, humidity, light exposure, and other factors influencing spoilage rates. Additionally, chemical parameters, including pH levels, sugar content, and other compositional elements, can be incorporated, as they directly impact fruit longevity. By leveraging this comprehensive dataset, machine learning algorithms can develop predictive models that forecast the ideal preservation conditions for a specific fruit type, enhancing

storage and shelf-life management [16, 17]. In the context of postharvest disease detection, ML models such as Partial Least Squares-Discriminant Analysis (PLS-DA) and Random Forest (RF) have demonstrated effectiveness in classification and prediction tasks, particularly for high-dimensional data such as VOC profiles [21].

The aim of this study is to develop a predictive model for Botrytis cinerea infection in strawberries through the analysis of VOC data, utilizing PLS-DA and Random Forest techniques. By leveraging the discriminatory power of PLS-DA and the robustness of RF models, our approach seeks to identify key VOC biomarkers associated with Botrytis decay while providing accurate predictions for disease progression. This integration not only facilitates early detection of decay but also supports data-driven decision-making in the postharvest supply chain, ultimately reducing food waste and enhancing sustainability.

2 Methodology

2.1 Fruit material

Around 6 kg of one cultivar strawberry, namely "*Favori*", was collected from a local grower Fruitbedrijf van Beusichem-de Waal (beusichemdewaal.nl), Wielseweg 40, 4024 BK Eck en Wiel , the Netherlands, in July 2024. Only fully ripe fruits of comparable size, devoid of any diseases or physical injuries, were carefully chosen for the subsequent analysis.

2.2 Treatment

16 fruits were divided into 2 groups, and stably placed on the lid of the plastic box (Figure 1). For the Botrytis inoculated group 5 µL of the Botrytis inoculum (*Botrytis cinerea* B05.10, $1x10^7$ spores/mL in demineralized water) was pipetted on the centre of each strawberry tip, without damaging the exocarp. An equal amount of sterile water was used on a control group. Each treatment included 8 replicates with a single strawberry. The fruits were stored in a growing container (EPS box), containing a wet source (soaked tissue paper) and closed with a lid. The treated and control fruits were stored in two different EPS boxes The internal environment was kept at 100% relative humidity (RH), which was confirmed with an RH logger. The growing containers were stored at 10°C for 13 days. Measurements were performed on day 0, 3, 6, 8, 10 and 13.

2.3 Image analysis

On each measurement day, photographs of individual strawberries were taken using a controlled setup. The imaging setup consisted of an enclosed chamber illuminated by two LED strips to ensure consistent lighting. Images were captured with a smartphone camera (model: iPhone 14). The picture was taken from above as shown in Figure 1. The images were manually segmented using the software Labelme [19] to identify the strawberry surface area and the area of regions infected with Botrytis.

Figure 1 Example of segmentation of strawberry, in dark blue is total area of strawberry and in green are the areas affected by Botrytis (Source M. Yao, 2024)

The areas of segmented areas were computed using Python implementation of Shoelace formula for computing area of polygons:

$$Area = \frac{1}{2}\sum_{i=0}^{n-1}(x_i y_{i+1} - x_{i+1} y_i)$$

The target variable was then defined as the ratio of the total area labelled as covered by Botrytis mould to the total area of the strawberry.

$$y = \frac{\sum_{i=1}^{n} Area_{mould_i}}{Area_{strawberry}}$$

For training the classification model, an additional (binary) target variable was computed, being 'true' if any mould was detected (y > 0)

2.4 Analysis of volatile organic compounds (VOCs)

2.4.1 Sampling process

For the purpose of measuring the volatiles, single strawberries were placed on the lid of a plastic box (700 mL). Before measurement, strawberries were taken out of the EPS box to a room temperature environment. Medical air was flushed through the box for 2 minutes before sealing, and the samples were then allowed to accumulate for 50 minutes. Afterwards, a small hole was made in each box, and a needle was used to connect to the sampling channel of the desorption unit (Markes International, California, USA) for analysis.

2.4.2 Thermal desorption–gas chromatography–mass spectrometry (TD−GC−MS) measurements & analysis

The samples were loaded into the TD-100-xr thermal desorption unit (Markes International, California, USA) for analysis. The desorption of samples was performed according to the following settings: pre-desorption: Pre-purge time: 2 min, Split flow: 20 mL/min; Primary tube desorption: 260 °C for 7 min, Trap flow: 40 mL/min; Secondary trap desorption: Trap temperature: 25 – 300, Heating rate: 25 °C/s, Trap desorption time: 8 min, Split flow: 20 mL/min; TD split: 8:1 outlet split. The sample tube was desorbed entirely (splitless desorption) onto the cold trap held at 25 °C. The cold trap was then heated at a rate of 25 °C/s to 300 °C and held for 5 min. During trap desorption, samples were split using an 8:1 outlet split. Analysis was conducted using a gas chromatograph-spectrometer equipped with a DB-5MS capillary column. The GC oven was initially set at 35 °C for 4 min, followed by a ramp of 10 °C/min to 300 °C with a 2-minute isothermal period. The total run time was 32 min. The MS ion source temperature was 280 °C, and the interface temperature was 200 °C. Ions were acquired in the m/z range of 35–350. The peaks in the GC-MS chromatogram were integrated using Chromeleon software. Each peak was identified using the NIST17 library.

2.4.3 Pre-processing of VOC data

To account for the large variations in VOC values across several orders of magnitude, the data was first transformed using a logarithmic scale. This is a quite common approach to avoid that the model fitting is solely driven by the measured high concentrations. Missing VOC values were treated as zeros. After scaling, the dataset was split into two subsets: one for training and validating machine learning algorithms, and another for testing the models. The test set was created by setting aside four strawberries that were inoculated with Botrytis. These were chosen because the inoculation resulted in earlier (visible) infection, providing a more balanced representation of both healthy and Botrytis-infected samples in our test set. The min-max normalization technique (using the Python sklearn implementation) was then fitted to the training data and applied consistently to both the training and test datasets. See Figure 2 for the combined boxplot, created using the matplotlib library.

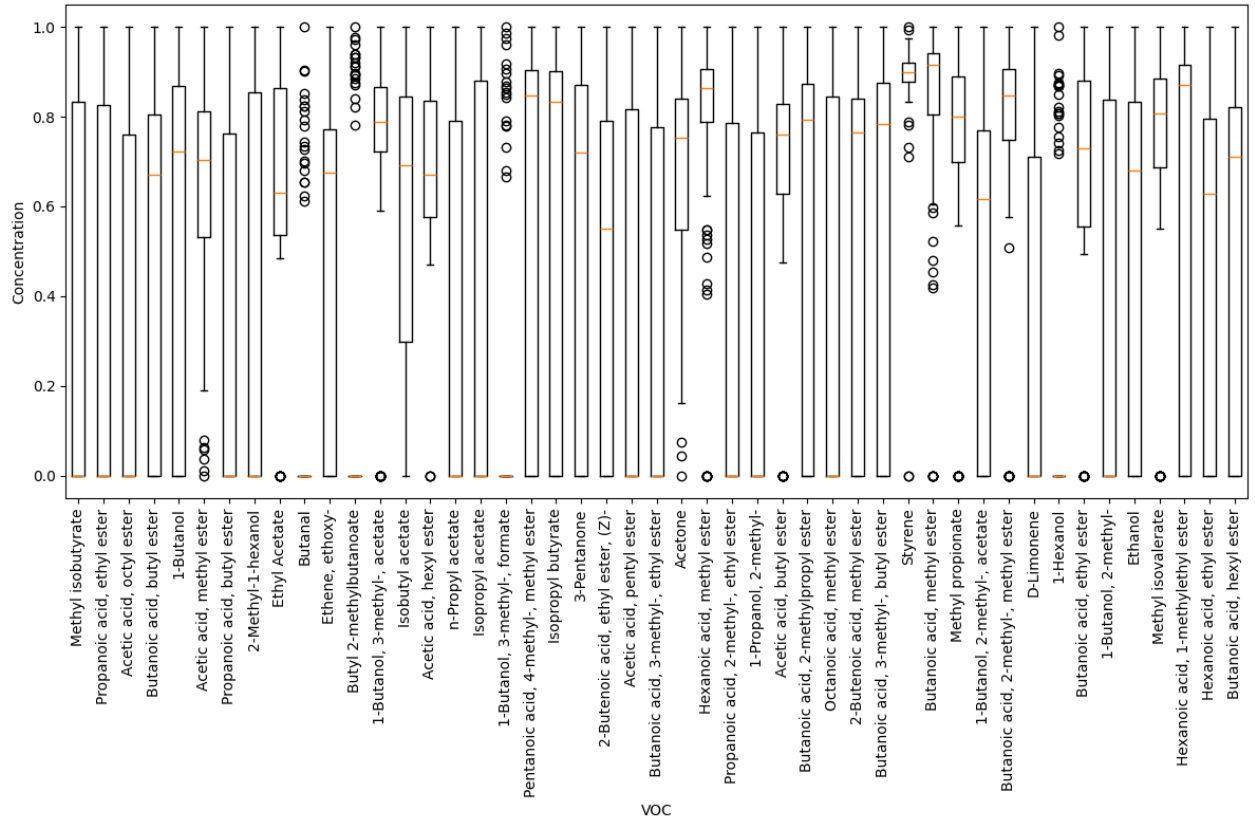

Figure 2 VOC data after applying logarithmic and min-max scaling

2.5 PLS-DA

In order to identify volatiles that were indicative of Botrytis, a Partial Least Squares Discriminant Analysis (PLS-DA) was executed. A PLS-DA is a supervised dimensionality reduction method that utilizes class labels to maximize the separation between different groups [20]. In this case, the PLS-DA aimed to identify latent components (i.e., linear combinations of the original features) that maximized the separation between the healthy and Botrytis infected strawberries. To this end, all strawberries that showed some degree of mould, were classified as infected. The extent to which features contributed to class separation, was reflected by the Variable Importance in Projection (VIP) score. Commonly, features that have a VIP score above 1 are deemed significant and therefore included in subsequent model building [21]. The PLS-DA was implemented in Python using the sklearn library, and was applied to the training dataset only. The model was built using two principal components. The features that had a VIP score above 1 were used in building the model (see next section). The visualizations of the PLS-DA results were created using the Matplotlib library.

2.6 Random Forest Classifier

The machine learning algorithm used to predict whether a strawberry was healthy or Botrytis infected was a random forest. A random forest is an ensemble of decision trees. In classification problems, these trees recursively split data on certain features to reduce node impurity (i.e., a measure of mixed classes within a node). By combining different trees that are slightly different from each other, a more robust model is created, which is less prone to overfitting [22]. Additionally, Recursive Feature Elimination (RFE) was implemented. This method takes an estimator (e.g. random forest) and initially trains the estimator on all features, after which it iteratively eliminates features that have a low importance score until the specified number of features is retained [23]. By using RFE, the goal was to improve the explainability of the final model and reduce overfitting. Note that the initial feature set consisted of those features having a VIP score above one. This essentially served as a pre-selection step, optimizing the RFE's runtime. As for the PLS-DA, this model was implemented using the sklearn package.

The optimization of hyperparameters was done using GridSearchCV, making use of a stratified 5-fold cross validation, ensuring that the class distribution in the training and validation set was approximately the same. Both the hyperparameters of the random forest and the RFE were optimized within a Pipeline structure. GridSearchCV was applied to the entire pipeline, ensuring that both the feature selection and random forest were tuned simultaneously. The final model for the random forest classifier was trained using n_estimators=50 and min_samples_leaf=10. The latter ensured rather shallow trees, which prevented overfitting for individual trees. Regarding the RFE, the required features being used in the random forest was set to 2. Also, the visualization of the random forest results were created using Matplotlib.

2.7 Random Forest Regression

A random forest regression model was selected to predict the degree of mould growth on strawberries. Random forest regression, like random forest classifier, is an ensemble learning method, it combines multiple decision trees to create more accurate and stable predictions by averaging the outputs of individual trees, thereby reducing overfitting and improving generalization. The implementation in Python sklearn was used. Like the random forest classifier, Recursive Feature Elimination (RFE) was employed for feature selection. The feature selection and parameter tuning process followed the same steps as those described for the random forest classifier. The model was trained with the following parameters: max_depth=4, min_samples_leaf=2, min_samples_split=4, and n_estimators=100, optimized using GridSearchCV with 5-fold cross-validation. The final model uses four types of volatile compounds.

3 Results

3.1 Change in *Botrytis cinerea* infection

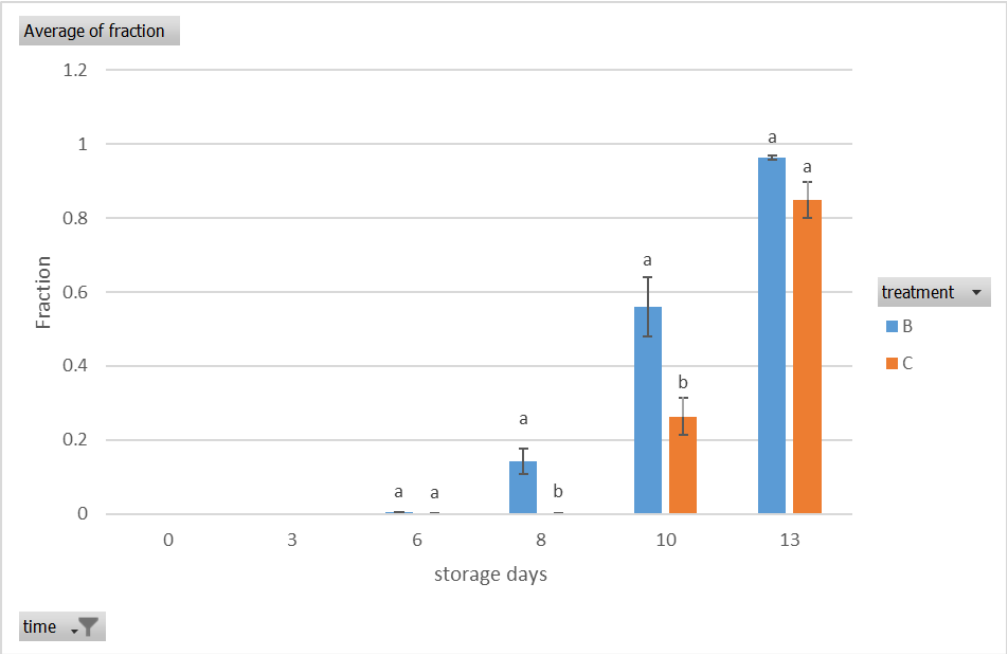

Figure 3 ***Changes in Botrytis infection during the storage time (day 0, 3, 6, 8, 10, 13) in the treatment (B) and control group (C). The letters indicate significant differences between treatment and control group within the same storage day. The significance is determined by t-tests.***

Figure 3 shows the average fraction of Botrytis lesion size on strawberries under Botrytis treatment (B) and control (C) conditions during the storage time. At the beginning of the storage (day 0-3), no lesions were observed in either treatment group (fraction = 0), with no significant differences between the Botrytis inoculated and control fruits. On day 6, Botrytis infection started to be visible in both groups regardless of treatment. This suggests that Botrytis growth occurs not only as a result of inoculation but also due to natural contamination or endogenous factors. By day 8, the fraction of Botrytis infection in the Botrytis group significantly increased compared to the control fruits ($p < 0.05$), indicating the onset of Botrytis infection under inoculation. This trend continued on day 10, where the lesion fraction in the Botrytis group was significantly higher than in the control group ($p < 0.05$), suggesting a more rapid progression of Botrytis infection in the inoculated samples. On day 13, both Botrytis and Control groups showed high levels of infection (fractions approaching 1.0), with no significant differences observed between treatments ($p > 0.05$). This indicates that by this stage, the infection had saturated in both groups regardless of treatment, as all strawberries displayed very poor quality.

3.2 PLS-DA

The scores plot of the PLS-DA shows clear clustering between the healthy and Botrytis infected strawberries in the training set (Figure 4a). This suggests that the volatile information is informative in detecting Botrytis in strawberries. Note that the cluster label is determined by the visible presence of mould (Botrytis), rather than by the sample's membership in either experimental group. Interestingly, looking at the scores plot of the Botrytis samples only, it can be observed that the samples also cluster by days (Figure 4b). Also, some outliers can be observed. For example, the two healthy samples in the lower part of Figure 4a (indicated by a circle) are closer to the Botrytis cluster than the healthy cluster. These samples refer to strawberries C5 and C6 on day 8, with 'C' indicating membership to the control group. Looking at the corresponding pictures in Figure 5, the strawberries indeed look free of mould, but show small spots that have a slightly deviating red colour. This could indicate a pre-phase of the mould. The corresponding VIP scores (> 1) of Figure 4a are presented in Table 1.

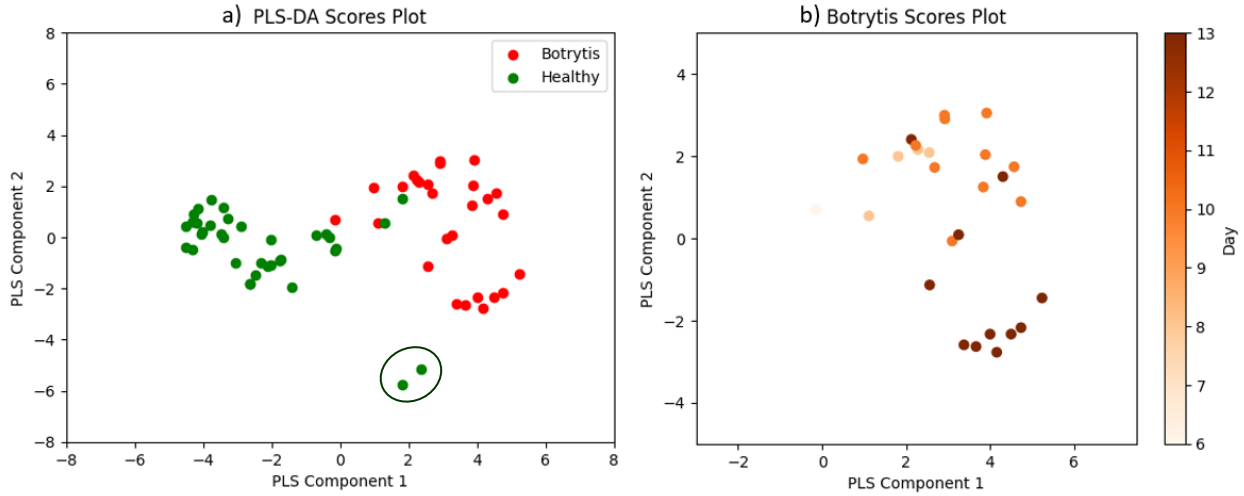

Figure 4 a) PLS-DA applied to training data, and b) Botrytis samples only, coloured by day

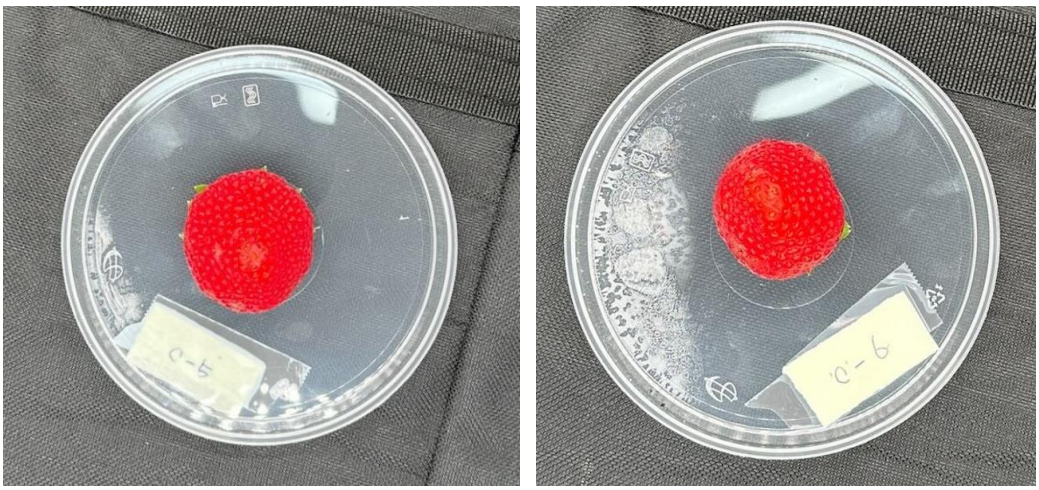

Figure 5 Samples C5 and C6 on day 8 (Source M. Yao, 2024)

Table 1 VIP scores > 1

Volatile	VIP Score
1-Propanol, 2-methyl-	2.33
D-Limonene	1.71
1-butanol, 2-methyl-	1.63
Pentanoic acid, 4-methyl-, methyl ester	1.59
Hexanoic acid, 1-methylethyl ester	1.51
Ethanol	1.49
Isopropyl butyrate	1.42
Isopropyl acetate	1.40
2-Methyl-1-hexanol	1.37
1-Butanol	1.34
Propanoic acid, 2-methyl-, ethyl ester	1.29
Butanoic acid, butyl ester	1.24
Methyl isobutyrate	1.20
Methyl isovalerate	1.13
Butanoic acid, methyl ester	1.13
Ethene, ethoxy-	1.09

3.3 Random Forest Classifier

The two volatiles included in the final model were 1-Propanol, 2-methyl and Pentanoic acid, 4-methyl, methyl ester (ranked 1st and 4th in Table 1). Note that because of the limited size of the dataset, the decision upon hyperparameters and selected volatiles proved to be somewhat dependent on the random seed, thus splits of the data. In the final model, the mean score on the training set(s) was 96.9% and the mean score on the cross validation set(s) was 95.1%. The performance on the test set was 87%, meaning that 20 out of the 23 samples were correctly classified. See Figure 6 for the confusion matrix. It can be observed that all healthy samples were correctly identified as healthy, while 9 out of the 12 Botrytis samples (i.e. those with visible mould) were accurately classified as being infected with Botrytis.

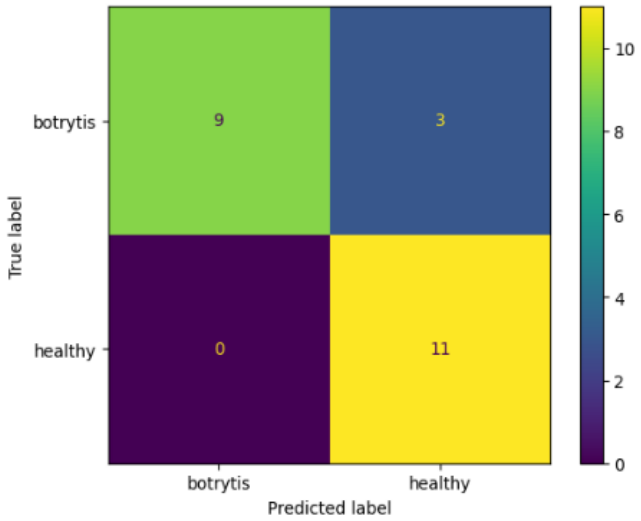

Figure 6 Confusion matrix for the random forest classifier on the test set

3.4 Random Forest Regressor

After recursive feature elimination the following four volatiles were chosen as the best predictors: 1-Propanol, 2-methyl, Butanoic acid, methyl ester, Acetic acid, butyl ester, and Butanoic acid, butyl ester. Using these volatiles the model was re-trained on training and validation data. The highest cross validation score (R^2-score) is 0.694. That resulted in root mean squared error (RMSE) of test set being equal to 0.137.

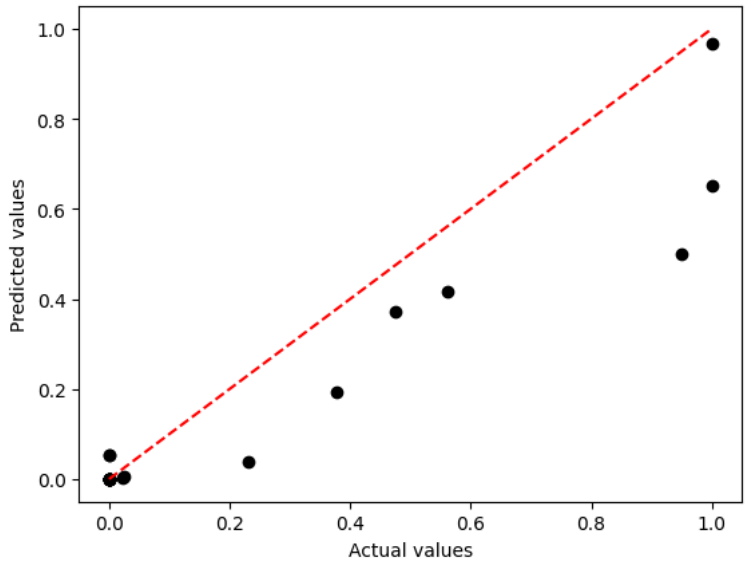

Figure 7 Predicted values based on volatiles vs. Actual values (results for training data set).

3.5 Change in concentration

Below the changes in VOC concentrations are reported for 1-Propanol, 2-methyl (also known as isobutanol); 1-Butanol, 2-methyl; and Pentanoic acid, 4-methyl-, methyl ester. Those VOCs are selected based on VIP-score and literature.

Figure 8 and Figure 9 illustrate the changes in the peak areas of 1-propanol, 2-methyl and 1-butanol, 2-methyl under Botrytis-inoculated (B) and control (C) throughout the storage period. From day 0 to day 3, the levels of both compounds remain negligible and show no significant differences between treatment groups ($p > 0.05$). By day 6, slight increases in both compounds are observed in the Botrytis-inoculated group, although these differences remain statistically insignificant ($p > 0.05$). This suggests that, early in storage, there is limited metabolic activity related to 1-propanol, 2-methyl and 1-butanol, 2-methyl production. On day 8, there is a marked elevation in both 1-propanol, 2-methyl and 1-butanol, 2-methyl production in the Botrytis-inoculated group compared to the control ($p < 0.05$), coinciding with the significant increase in Botrytis lesion size. This association implies that 1-propanol, 2-methyl and 1-butanol, 2-methyl may be linked to the metabolic processes activated by Botrytis infection. From day 10, 1-propanol, 2-methyl is showing elevated levels in both two groups. However, by day 13, the compound continues to increase in both groups, with significantly higher levels in the control group compared to the Botrytis-inoculated group ($p < 0.05$). At the end of storage, from 10 to day 13, 1-butanol, 2-methyl exhibits elevated levels in both the Botrytis-inoculated and control groups. While the compound's levels increase over time, no statistically significant differences are observed between the two groups ($p > 0.05$).

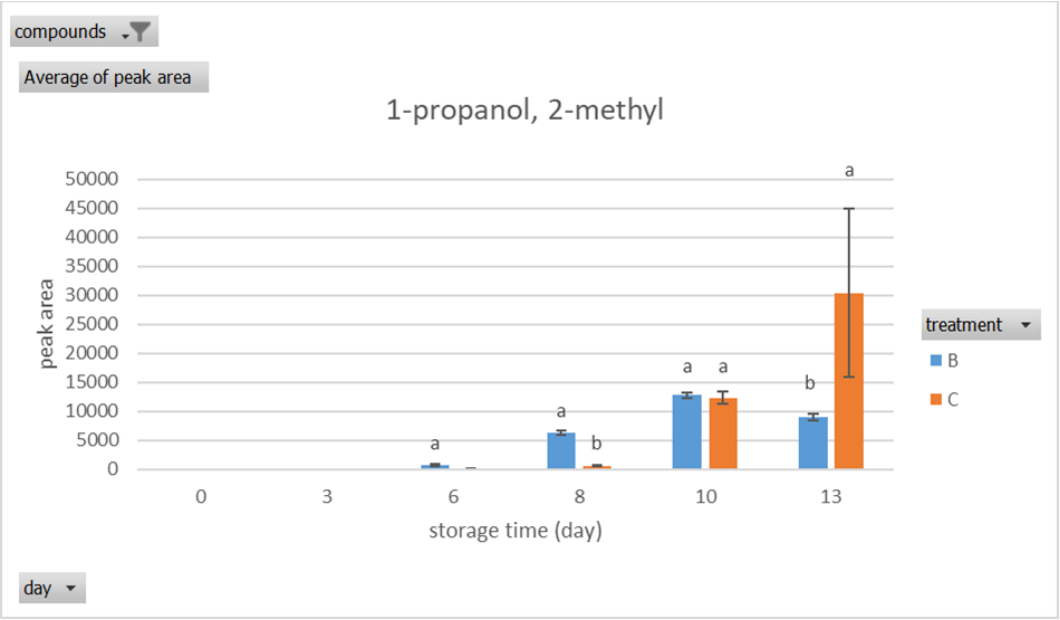

Figure 8 ***Changes in the concentration of 1-propanol, 2-methyl during the storage time (day 0, 3, 6, 8, 10, 13) in the treatment (B) and control group (C). The letters indicate significant differences between treatment and control group within the same storage day. The significance is determined by t-test.***

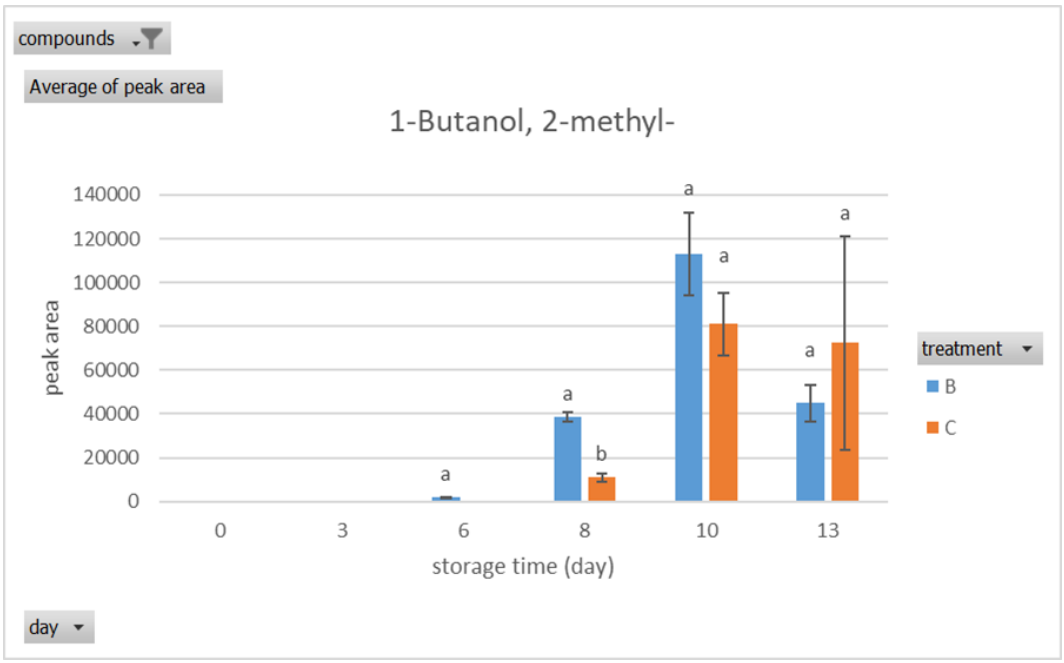

Figure 9 ***Changes in the concentration of 1-Butanol, 2-methyl- during the storage time (day 0, 3, 6, 8, 10, 13) in the treatment (B) and control group (C). The letters indicate significant differences between treatment and control group within the same storage day. The significance is determined by t-test.***

Figure 10 shows how the peak area of Pentanoic acid, 4-methyl-, methyl ester changes over time in strawberries under both Botrytis-inoculated (B) and control (C) conditions. From the start of storage, both groups have high concentrations of the ester, with no significant difference between them, consistent with the absence of Botrytis lesions seen in Figure 3. However, on day 6, a significant drop in the concentration of Pentanoic acid, 4-methyl-, methyl ester is observed in the Botrytis-inoculated group compared to the control ($p < 0.05$). This decrease corresponds with the first appearance of Botrytis lesions, as shown in Figure 3, suggesting that the ester may be metabolically degraded or transformed as infection begins to take hold. On day 8, the ester concentration in the Botrytis group stays lower than in the control, and both groups show a general decline. This period corresponds to the rapid spread of Botrytis lesions in the inoculated fruit, indicating that the ester continues to be depleted as the infection grows. By day 10, ester levels are minimal in both groups. At this point, Figure 3 shows that Botrytis lesions have extensively developed in the inoculated group, while the control group begins to show more noticeable infection. This indicates widespread metabolic breakdown associated with advanced decay. Until the end of storage, the concentration of Pentanoic acid, 4-methyl-, methyl ester is very low and nearly the same in both groups. This aligns with the extensive Botrytis infection in both groups, as observed in Figure 3. The low ester levels indicate that by this stage, the strawberries are severely decayed, and the ester is almost completely depleted.

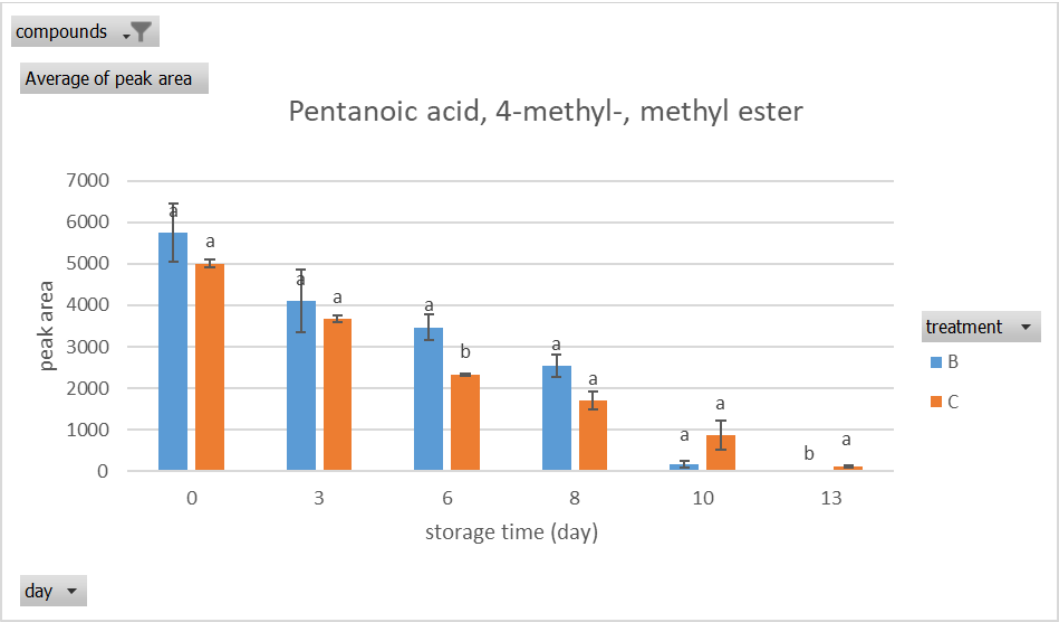

Figure 10 ***Changes in the concentration of Pentanoic acid, 4-methyl-, methyl ester during the storage time (day 0, 3, 6, 8, 10, 13) in the treatment (B) and control group (C). The letters indicate significant differences between treatment and control group within the same storage day. The significance is determined by t-test.***

4 Discussion

The findings of this study provide significant insights into the potential of volatile organic compounds (VOCs) for non-destructive, early detection of Botrytis infection in strawberries. Using thermal desorption gas chromatography-mass spectrometry (TD-GC-MS) and machine learning models, we demonstrated that VOC profiles can effectively differentiate between healthy and infected strawberries. Key biomarkers, including 1-propanol, 2-methyl- and 1-butanol, 2-methyl- as positive indicators, and Pentanoic acid, 4-methyl-, methyl ester as a negative indicator, exhibited significant relationships with lesion size, highlighting their relevance in assessing infection progression. However, several aspects warrant further discussion, particularly in the context of the biological significance of these biomarkers, the performance of the predictive models, and the practical implications and limitations of the study.

4.1 Potential biomarkers for strawberry Botrytis

The identification of 1-propanol, 2-methyl- and 1-butanol, 2-methyl- as positive biomarkers for Botrytis infection highlights their potential involvement in metabolic processes during fruit-pathogen interactions. Based on the findings from the literature [24, 25], these compounds are likely derived from both fungal and host metabolism, reflecting a dynamic interplay between the pathogen and the physiological responses of plant or fruit. Both compounds are products of branched-chain amino acid metabolism via the Ehrlich pathway, where valine and leucine serve as precursors. In our study, where Botrytis is visible, these 1-propanol, 2-methyl- and 1-butanol, 2-methyl- increase significantly, reflecting metabolic shifts due to Botrytis infection and tissue degradation. A previous study found that some higher alcohols such as 2-phenylethanol, 3-methylbutan-1-ol, 2-methylpropan-1-ol, and 2-methylbutan-1-ol were produced by *Botryosphaeria spp* [26]. The structural isomers of 1-butanol, 2-methyl-, 3-Methylbutan-1-ol, have been identified as volatile biomarkers in fungal pathogen spoilage of apples, particularly during the decay caused by pathogens like *Penicillium expansum* and *Botryosphaeria dothidea* [27]. Vandendriessche reported that compounds 1-propanol, 2-methyl- and 1-butanol, 2-methyl- could be potential biomarkers for Botrytis infection in strawberries, highlighting their association with fungal colonization and quality deterioration in Elsanta strawberries [11]. Although this study measured a different cultivar (Favori), similar potential Botrytis biomarkers were found.

Pentanoic acid, 4-methyl-, methyl ester has not been widely reported in the context of fruit systems, but its decline during Botrytis development and quality decay suggests it may play a role in the aroma profile of healthy strawberries. As an ester, it may contribute to fruity and sweet aroma characteristics, which are diminished as quality deteriorates. The observed reduction may result from metabolic shifts during fungal colonization, where precursors for ester synthesis are redirected to stress response pathways, or from enzymatic degradation by fungal esterase. This consistent decrease positions Pentanoic acid, 4-methyl-, methyl ester as a potential negative biomarker for infection and quality loss, highlighting its relevance to both sensory degradation and disease progression. Further research into its biosynthesis and role in fruit metabolism could provide valuable insights into its function during postharvest deterioration.

4.2 Size of the dataset

Many applications in bio-based and food fields face the challenge of limited data, which complicates the implementation of machine learning (ML) models that typically perform best with large datasets. In this study, the data volume was constrained by both the limited budget and the inherent challenges of the VOC measurement process. Given the limitations of the GC-MS measuring device, only 16 strawberries could be measured on any given day. A general guideline for ML algorithms suggests collecting at least 10 samples per feature to avoid overfitting. This constraint directly impacts the number of features that can be incorporated into the ML model, ultimately simplifying the intricate and complex VOC profiles exhibited by strawberries. As a result, the richness of these profiles may not be fully captured, highlighting the trade-off between dataset size and feature representation in ML-based analyses. Robustness of the model would be improved by repeating the experiment with different cultivars and performing a joint analysis.

4.3 Image analysis

Image analysis in this study was conducted manually by a single researcher, which introduces potential bias and limits scalability. This process could be improved by involving multiple individuals in labelling the images, reducing subjective bias and enabling the quantification of uncertainty in the target variable through the assessment of expert disagreement. A more advanced approach would involve developing a computer vision solution to automate image processing. Such a system would enhance reproducibility, minimize bias, and significantly increase scalability, making it a more robust and efficient method for image analysis. Another problem we have encountered in the study is that natural infection often grows from the calyx which is difficult to see on the pictures collected in this study. Involving 3D imaging techniques would solve this and make the scoring less prone to underestimating the degree of infection.

4.4 Machine learning choices

In this study, we aimed to make informed decisions in selecting model architectures, processing data, and training machine learning algorithms. To achieve this, we employed GridSearch for hyperparameter tuning, Partial Least Squares Discriminant Analysis (PLS-DA) for feature selection, and Recursive Feature Elimination (RFE) to further refine the feature set. We also implemented cross-validation to ensure robust model evaluation and set aside a subset of strawberries as a validation set to make the model less biased to individual strawberry VOC profiles. Random Forest models, both regressor and classifier, were trained on the processed data and yielded results that we deemed optimal for this study. However, given the ever-evolving landscape of machine learning solutions, some decisions were necessarily empirical rather than entirely data-driven. While our approach was systematic, it is possible that alternative algorithms or configurations could have yielded even better outcomes.

4.5 Potential falsification of the model

A dataset intended to validate the model was obtained from a different project related to Botrytis development in strawberries. Unfortunately, the collected data hint at a model falsification. The volatile profiles in the validation data differed somewhat from those observed in the initial experiment. This experiment was executed a little different than the experiment in the initial study. With hindsight we doubt the experimental design of the experiment that we used to validate the model. In view of budget and time limitations we were not able to repeat the validation experiment. Hence, the project ends with only a provisional conclusion and we recommend to execute a proper validation experiment later on.

4.6 Future work

While we believe that the results achieved in the study are quite promising we think that this work could have benefitted from the following:

- Inclusion of more data (more cultivars, batches from different harvest times)

The data used for developing the machine learning model(s) came from VOC measurements of one cultivar and the same batch, so collected on the same day. This has limited the applicability of this study to that same cultivar and potentially same time of harvest. In the future, the inclusion of different cultivars and different harvest times would be preferred. Another interesting thing to investigate is how storage temperature influences development of Botrytis and VOCs release related to that.

- Ageing vs. Botrytis

Related to the previous point, it is unclear in this study whether the selected volatiles are an indicator of ageing of postharvest strawberries or if they indicate Botrytis disease degree. There is a high correlation between time since harvest and degree of mould. Therefore we suspect that the VOCS that were most useful for predicting Botrytis infection were likely related to the ageing of strawberry. Also, note that ripening can trigger the development of Botrytis during postharvest period in strawberries [28].

- Understanding parameters influencing VOC expression

Keeping in mind that eventually this study could be useful for commercial purposes we expect that the measurement of VOCs would have to accommodate different measurement protocol setups. Therefore it would be good to build a model that would incorporate the temperature of the measured strawberries, the volume of measured strawberries, time of accumulation, and other factors that can influence the expression of VOCs.

- Predicting the onset of Botrytis infection prior to visual signs

It would also be beneficial to be able to predict the onset of Botrytis infection rather than the presence or the degree of mould. The most valuable insight for the supply chain would be an improved prediction of strawberry shelf-life. That would create the opportunity to better distribute strawberries, and avoid waste since customer acceptance of amount of mould present on strawberries is quite low. One study measured a maximum acceptable deterioration (MAD) limit of 13% [29].

5 Conclusion

This study underscores the role of VOCs in assessing Botrytis infection in strawberries through the integration of metabolomics and machine learning approaches. Using data from 'Favori' strawberries, we identified key descriptive biomarkers, including 1-propanol, 2-methyl-, 1-butanol, 2-methyl-, and pentanoic acid, 4-methyl-methyl ester, achieving high predictive performance with random forest models, notably an accuracy of 87% for infection classification and an R^2 of 0.70 for infection severity. These findings highlight the potential of VOC profiling for postharvest disease prediction and quality management. While challenges remain, such as limited dataset size and cultivar variability, this study validates VOC analysis as a powerful tool for improving strawberry supply chain monitoring.

Literature

1. Feliziani, E., & Romanazzi, G. (2016). Postharvest decay of strawberry fruit: Etiology, epidemiology, and disease management. Journal of Berry Research, 6(1), 47–63.
2. Romanazzi, G., & Feliziani, E. (2014). *Botrytis cinerea* postharvest decay: Control strategies. In D. R. H. Jones & S. F. Melzer (Eds.), Postharvest Decay Control (pp. 131–146).
3. Mertely, J. C., Oliveira, M. S., & Peres, N. A. (2018). Botrytis fruit rot or gray mould of strawberry. EDIS.
4. Abbey, J. A., Percival, D., Abbey, Lord, Asiedu, S. K., Prithiviraj, B., & Schilder, A. (2018). Biofungicides as an alternative to synthetic fungicide control of grey mould (*Botrytis cinerea*) – prospects and challenges. Biocontrol Science and Technology, 29(3), 207–228.
5. Wang, Z., Di, S., Qi, P., Xu, H., Zhao, H., & Wang, X. (2021). Dissipation, accumulation and risk assessment of fungicides after repeated spraying on greenhouse strawberry. Science of The Total Environment, 758, 144067.
6. Stensvand, A., & Christiansen, A. (2000). Investigation on fungicide residues in greenhouse-grown strawberries. Journal of Agricultural and Food Chemistry, 48(3), 917–920.
7. Prithiviraj, B., Vikram, A., Kushalappa, A. C., & Yaylayan, V. (2004). Volatile metabolite profiling for the discrimination of onion bulbs infected by *Erwinia carotovora ssp. carotovora, Fusarium oxysporum*, and Botrytis *allii.* European Journal of Plant Pathology, 110(4), 371–377.
8. Simon, J. E., Hetzroni, A., Bordelon, B., Miles, G. E., & Charles, D. J. (1996). Electronic sensing of aromatic volatiles for quality sorting of blueberries. Journal of Food Science, 61(5), 967–969.
9. Li, C., Schmidt, N. E., & Gitaitis, R. (2011). Detection of onion postharvest diseases by analyses of headspace volatiles using a gas sensor array and GC-MS. LWT - Food Science and Technology, 44(4), 1019–1025.
10. Wu, J., Cao, J., Chen, J., Huang, L., Wang, Y., Sun, C., & Sun, C. (2023). Detection and classification of volatile compounds emitted by three fungi-infected citrus fruits using gas chromatography-mass spectrometry. Food Chemistry, 412, 135524.
11. Vandendriessche, T., Keulemans, J., Geeraerd, A., Nicolai, B. M., & Hertog, M. L. A. T. M. (2012). Evaluation of fast volatile analysis for detection of *Botrytis cinerea* infections in strawberry. Food Microbiology, 32(2), 406–414.
12. Feng, Y., Wang, Y., Beykal, B., Qiao, M., Xiao, Z., & Luo, Y. (2024). A mechanistic review on machine learning-supported detection and analysis of volatile organic compounds for food quality and safety. Trends in Food Science & Technology, 143, 104297.
13. Brendel, R., Schwolow, S., Rohn, S., & Weller, P. (2021). Volatilomic Profiling of Citrus Juices by Dual-Detection HS-GC-MS-IMS and Machine Learning—An Alternative Authentication Approach. Journal of Agricultural and Food Chemistry, 69(5), 1727-1738.
14. Ku, K. M., Kim, J., Park, H. J., Liu, K. H., & Lee, C. H. (2010). Application of metabolomics in the analysis of manufacturing type of pu-erh tea and composition changes with different postfermentation years. Journal of Agricultural and Food Chemistry, 58(1), 345–352.
15. Zhu, L., Spachos, P., Pensini, E., & Plataniotis, K. N. (2021). Deep learning and machine vision for food processing: A survey. Current Research in Food Science, 4, 233–249.
16. Morin-Crini, N., Lichtfouse, E., Torri, G., & Crini, G. (2019). Applications of chitosan in food, pharmaceuticals, medicine, cosmetics, agriculture, textiles, pulp and paper, biotechnology, and environmental chemistry. Environmental Chemistry Letters, 17, 1667–1692.
17. Naser-Sadrabadi, A., Zare, H. R., & Benvidi, A. (2020). Photochemical deposition of palladium nanoparticles on TiO2 nanoparticles and their application for electrocatalytic measurement of nitrate ions in potato, onion, and cabbage using bipolar electrochemical method. Measurement, 166, 108222.
18. Lutz, É., & Coradi, P. C. (2022). Applications of new technologies for monitoring and predicting grain quality stored: Sensors, Internet of Things, and artificial intelligence. Measurement, 188, 110609.
19. K. Wada, "labelme: Image Polygonal Annotation with Python," 2024. [Online]. Available: https://github.com/wkentaro/labelme.
20. Ruiz-Perez, D., Guan, H., Madhivanan, P., Mathee, K., & Narasimhan, G. (2020). So you think you can PLS-DA?. BMC bioinformatics, 21, 1-10.

21. Do, E., Kim, M., Ko, D. Y., Lee, M., Lee, C., & Ku, K. M. (2024). Machine learning for storage duration based on volatile organic compounds emitted from 'Jukhyang' and 'Merry Queen' strawberries during post-harvest storage. Postharvest Biology and Technology, 211, 112808.
22. Müller, A. C., & Guido, S. (2016). Introduction to machine learning with Python: a guide for data scientists. O'Reilly Media, Inc.
23. RFE. (n.d.). Scikit-learn. https://scikit-learn.org/dev/modules/generated/sklearn.feature_selection.RFE.html
24. Weber, F. J., & de Bont, J. A. M. (1996). Adaptation mechanisms of microorganisms to the toxic effects of organic solvents on membranes. Biochimica et Biophysica Acta (BBA) - Reviews on Biomembranes, 1286(3), 225–245.
25. Schumacher, K., Asche, S., Heil, M., Mittelstädt, F., Dietrich, H., & Mosandl, A. (1998). Methyl-branched flavor compounds in fresh and processed apples. Journal of Agricultural and Food Chemistry, 46(11), 4496–4500.
26. Etschmann, M. M. W., Bluemke, W., Sell, D., & Schrader, J. (2002). Biotechnological production of 2-phenylethanol. Applied Microbiology and Biotechnology, 59(1), 1–8.
27. Kim, S. M., Lee, S. M., Seo, J. A., & Kim, Y. S. (2018). Changes in volatile compounds emitted by fungal pathogen spoilage of apples during decay. Postharvest Biology and Technology, 146, 51-59.
28. Li, H., Larsen, D. H., Cao, R., van de Peppel, A. C., Tikunov, Y. M., Marcelis, L. F. M., Woltering, E. J., van Kan, J. A. L., & Schouten, R. E. (2022). The association between the susceptibility to Botrytis cinerea and the levels of volatile and non-volatile metabolites in red ripe strawberry genotypes. Food Chemistry, 393, 133252.
29. C. Matar, S. Gaucel, N. Gontard, S. Guilbert, and V. Guillard, "A global visual method for measuring the deterioration of strawberries in MAP," MethodsX, vol. 5, pp. 944–949, Jan. 2018, doi: 10.1016/j.mex.2018.07.012.

Acknowledgements

We would like to thank Wout van Beusichem and and Louise van Beusichem from Fruitbedrijf van Beusichem-de Waal for generously providing the strawberries for the experiment, and for their enthusiasm. Additionally, we would also like to thank Leo Lukasse for his supervision in this project. Furthermore, we express our appreciation to Rob Schouten for reviewing and help with developing our ideas and Maxence Paillart, Marcel Staal, Jan Verschoor and Nicole Koenderink for supporting the experiment. We thank Jeroen Buil for guiding the machine learning modelling. Lastly, we would like to thank Henk Wensink for coordinating the trainee program, of which this project is part of.

Wageningen Food & Biobased Research
Bornse Weilanden 9
6708 WG Wageningen
The Netherlands
E info.wfbr@wur.nl
wur.eu/wfbr

Report 2642

The mission of Wageningen University & Research is "To explore the potential of nature to improve the quality of life". Under the banner Wageningen University & Research, Wageningen University and the specialised research institutes of the Wageningen Research Foundation have joined forces in contributing to finding solutions to important questions in the domain of healthy food and living environment. With its roughly 30 branches, 7,700 employees (7,000 fte), 2,500 PhD and EngD candidates, 13,100 students and over 150,000 participants to WUR's Life Long Learning, Wageningen University & Research is one of the leading organisations in its domain. The unique Wageningen approach lies in its integrated approach to issues and the collaboration between different disciplines.