Machine Learning to Detect Fungal Infections in Stored Pome Fruits via Mass Spectrometry Data: Industry, Economic, and Social Implications

Razia Sulthana Abdul Kareem [1,*], Nageena K. Frost [1], Charles A. I. Goodall [2], Timothy Tilford [1], and Ana Paula Palacios [1]

[1] School of Computing and Mathematical Sciences, Faculty of Engineering and Science, University of Greenwich, Old Royal Naval College, London, United Kingdom
[2] Faculty of Engineering and Science, University of Greenwich, Chatham Maritime, Chatham, United Kingdom
Email: razia.sulthana@greenwich.ac.uk (R.S.A.K.); n.k.frost@greenwich.ac.uk (N.K.F.); c.a.i.goodall@greenwich.ac.uk (C.A.I.G.); t.tilford@greenwich.ac.uk (T.T.); a.palacios@greenwich.ac.uk (A.P.P.)
*Corresponding author

Abstract*—Pome fruits, notably apples and pears, experience decay during storage due to fungal infections. The timely discernment of these infections is imperative to avert the deterioration of these fruits within warehouse confines. In an experimental setup, two distinct apple cultivars, Braeburn and Gala, were inoculated with fungi Monilinia laxa, Neonectria ditissima, and Botrytis cinerea. As the infection progresses, the apples release chemical volatile components, which are measured using mass spectrometry in both positive and negative ion modes, recording mass-charge ratios ranging from m/z 30 to m/z 900 with a 0.3 Dalton difference between each measurement. The dataset is then partitioned into 24 sets of three-dimensional data, encompassing attributes related to two types of apples, three types of fungi, and two types of ions. They are analyzed using various machine learning algorithms, including Logistic Regression, Support Vector Machines (SVM), XGBoost, Random Forest, and four distinct customised Neural Networks, to classify infected and uninfected apples. The outcomes from the different machine learning algorithms across the 12 combinations of Apple-Fungi-Ion are recorded, revealing that certain algorithms excel in different combinations. The performance metrics namely True Positive, True Negative, False Positive, False Negative, Accuracy are closely analysed and the algorithms that produce the highest and second-highest accuracy are highlighted. Upon thorough analysis of the 12 combinations, it is observed that Logistic Regression and SVM with a linear kernel achieve the highest accuracy in approximately 11 combinations. Specifically, Logistic Regression achieves a precision of 98% for Braeburn apples, while SVM attains 99% accuracy for Gala apples. This research project has a triple impact on industry, economy, and society. On an industrial level, the precision and early predictions of the proposed work can effectively safeguard large quantities of apples in storage bins. Economically, it has the potential to avert substantial monetary losses. Societally, it plays a crucial role in determining the ideal timing to release fruits to the market for consumption without jeopardizing human health.

***Keywords*—pome fruits, apples, fungal infection, mass spectrometry, machine learning algorithms, neural networks**

I. Introduction

Among the plethora of available fruits, apples emerge as the preeminent choice for individuals spanning all age groups, proving beneficial to intestinal health amidst various illnesses. Apples are convenient for on-the-go consumption at work or during travel and the judicious interplay between the price and the size of apples renders them reasonably priced and affordable [1]. Numerous countries globally have implemented legal regulations governing the cultivation, processing, packaging, and fresh delivery of diverse apple cultivars. In 2011, the United Kingdom established a standard to deliver fresh apples to consumers, emphasizing the preservation of apple quality by ensuring they are undamaged, clean, and free from pests and infections [2].

The fungal pathogens can infect apples trees in orchards causing the fruit to drop prematurely. Often these pathogens colonize to the interior of apple or on the surface and travel form the orchard to postharvest storage. These postharvest fungal infections cause significant loss compared to preharvest, up to 20% in United States [3] ranging from 30–40% in developing countries to as high as 60% in severe cases [4]. The spread and severity of the infection varies every year based on different factors such as the grower, apple variety, packaging design, storage temperature and other attributes. Conventional microbiological techniques [5] that are applied to detect the infection are either time consuming or involve destroying the fruit to study the infection. While experimental [6] and statistical models [7] also exist to identify infections, there is a need for an efficient AI model that can swiftly predict both the occurrence and severity of infections.

While extensive research has been conducted in predicting postharvest fruit fungal infections, particularly

Manuscript received June 20, 2024; revised July 2, 2024; accepted July 30, 2024; published October 23, 2024.

in apples, numerous challenges persist in accurately identifying these infections: Difficulty in detecting infections within apple cores due to the absence of obvious external symptoms; Microbiological studies conducted in labs often involve destructive methods; Visual inspection of fruit is convenient but prone to inconsistency and subjectivity; The application of RGB and hyperspectral imaging on apple images is utilized. While RGB images effectively capture shape, color, and texture for infection detection, they lack accuracy in revealing internal structures and type of infections [8]. Hyperspectral images offer detailed insights into internal structures, providing both spatial and spectral characteristics of objects. However, they require expensive imaging devices and skilled operators.

Nevertheless, subjecting apples to prolonged storage exerts minimal influence on their nutritional constituents, thereby preserving both the phytochemical composition and antioxidant attributes [9]. Hence the primary focus of the research is directed towards protecting apples from fungal infection, with an emphasis on early identification if such infections occur.

In this research, we explore the primary findings of mass spectrometry results conducted on two distinct apple cultivars, Gala and Braeburn during the post-harvest storage. These apples were inoculated with Monilinia laxa, Neonectria ditissima, and Botrytis cinerea. We then employ various machine learning models to precisely identify the infected and uninfected samples.

II. Literature Review

The pathogen Monilinia laxa, classified under the genus Monilia, primarily affects apple. This pathogen is responsible for causing brown rot disease in apples [10]. Apple canker or European canker [11] is one of most disease that affects apple across UK and Europe and is cause by Neonectria ditissima previously known as N. galligena that causes the apples to rot and affects the barks of apple trees. The pathogen Botrytis cinerea (B. Cinerea) [8] causes gray mold disease in pome fruits, affecting over 200 crop species globally. However, the most significant impact of gray mold occurs during the postharvest period in apples, with losses reaching levels as high as 20–50%. Surveys conducted in the Pacific Northwest have identified gray mold as the second most significant postharvest infection.

The aforementioned three fungus decay apples even in cold storage, with these fungi capable of growing at temperatures below 0 degrees Celsius. While low temperatures slow down fruit spoilage, they do not completely stop it, as evidenced by prior research. Resnicow and Botvin [12] investigates three types of fungi, and the literature review focuses on these fungal species.

Dutot *et al.* [3] described examines postharvest apple pathogens in the Okanagan region with the aim of preventing decay of apples in packinghouses. This research is a simulation-based study conducted using general observations of fungal pathogens, particularly P. Expansum and Botrytis cinerea and the simulated model is created using theoretical inquiries from data repositories, focusing on aspects like infection spread rate, pattern, and the influence of various parameters on disease incidence. A significant issue associated with solely relying on simulated models is that packinghouses are entirely sealed following apple storage and are opened only during the distribution of apples.

In the research conducted by Lennox *et al.* [13], the population density of Bortrytis cinerea on pear pome fruit is measured. They found a strong linear correlation coefficient of 0.761 between the presence of the pathogen on the fruit surface and in the surrounding air, and 0.765 between decaying fruit and air. It suggests that the pathogen's presence in soil and litter is less harmful compared to its presence on the fruit surface, which is a primary cause of decay. Consequently, the proposed method entails performing swab tests on apple surfaces to detect the presence of infections.

In numerous research articles, the Raman spectroscopy technique has been employed to detect fungal pores [14–16]. However, due to its weak signal generation in many fungi, dynamic surface-enhanced Raman spectroscopy was utilized in [17] along with machine learning algorithms for infection detection. For the Bortrytis cinerea fungi, the accuracy rates of prediction were RF 90.55%, KNN 93.88%, LeNet5 90.56%, ZFNet 99.44%, and Inception 98.89%.

Haghbin *et al.* [18] focused on detecting Botrytis cinerea contamination in kiwifruit samples using an electronic nose (e-nose) system. The e-nose comprised thirteen Metal Oxide Semiconductor (MOS) gas sensors, each extracting six features, resulting in a total of 78 features. Correlation-based Feature Selection (CFS) and principal component analysis (PCA) were applied to optimize feature selection process. Subsequently, they applied various machine learning classification techniques, including Multilayer Perceptron Neural Network (MLPNN), Linear Discriminant Analysis (LDA), Bayesian Network (BN), Naive Bayes (NB), Radial Basis Function Neural Network (RBFNN), Support Vector Machine (SVM) [19, 20], and Decision Tree (DT). Of all the techniques, RBFNN achieved an accuracy of 98.9%. The high accuracy attained by machine learning algorithms underscores their efficacy in precisely identifying infected kiwis. This positive outcome is encouraging to apply machine learning algorithms for mass spectrometry results obtained for the proposed work.

N. ditissima is observed to impact apple trees, leading to pre-harvest blossom-end rot or post-harvest infections. Weber [21] examines the occurrence of canker lesions in trees, the severity of fruit infections, and the timing of infection spread during spring and summer. It is suggested that the pathogen may have gained entry to the fruit stem while still on the tree leading to rot during storage and spreading within storage areas. Gelain *et al.* [22] conducted on N. ditissima infection, colonization, and reproduction in two apple fruit cultivars: Gala and Eva. It was determined that a minimum of 25 days was required for sporulation to occur on fruit subjected to incubation temperatures ranging from 16.8–21.7 °C. Therefore, early prediction is crucial for timely fruit preservation.

The literature extensively covers a variety of machine learning and deep learning algorithms for tasks like identifying rot on apple surfaces through image processing [23]. In contrast, there is a consistent focus on utilizing statistical methods and visualization techniques when analyzing mass spectrometer datasets, with minimal incorporation of ML algorithms. The restricted usage of ML algorithms in this scenario can be ascribed to the intrinsic high-dimensional complexity of mass spectrometer data, requiring expertise not only in the domain but also in farming, apple production, and spectrometer data analysis. Therefore, the current research is collaboratively conducted by experts from the three fields to explore the predictive accuracy concerning three common post-harvest fungal pathogens during early storage in pome fruits, especially apples, utilizing mass spectrometer datasets. Given the No Free Lunch Theorem's [24] assertion that no single optimization algorithm universally excels, indicating algorithm effectiveness varies across problem domains, we intend to apply various ML algorithms to the mass spectrometer dataset. This approach allows us to evaluate their performance and identify the most suitable ones for our objectives and dataset.

III. Dataset

The mass spectrometry experiment was performed on Radian-ASAP manufactured by Waters Corporation direct analysis system. The Atmospheric Pressure Solids Analysis Probe (ASAP) ionisation methods were used to profile volatolome changes of apples infected with Monilinia laxa, Neonectria ditissima and Botrytis cinerea during pathogenesis.

Gala and Braeburn fruit were collected and stored under a controlled atmosphere at the Natural Resources Institute's Produce Quality Centre at East Malling, Kent belonging to University of Greenwich. Fruits were washed, residues removed and were then inoculated using sequenced strains of Monilinia laxa, Neonectria ditissima, and Botrytis cinerea cultured on Potato Dextrose Agar (PDA). A picture of the apples in the rack and their storage is shown in Fig. 1. Inoculated apples were separated into sample-sets comprising 10 apples per set, then placed into sterilised net sacking and stored in sterilised crates (Fig. 1). The inoculation process is carried our every week with crates of inoculated fruit stored under ambient conditions to maximise rate of disease progression until fruit rot had progressed to non-viability. Subsequently sampling is done on every inoculated sample (near the infected area which is the inoculated location and the uninfected part) in intervals and spectrometry values are recorded.

The mass spectrometry data, focusing on both positive and negative ionization, is segregated for Braeburn and Gala apples and encompasses information related to three different types of fungal infections. Table I gives a detailed analysis on the number of records in each of the combinations. The data appears to be evenly spread between negative and positive ions as well as between infected and uninfected samples. Our experimental analysis involves utilizing these data values to apply machine learning algorithms.

a) b)

Fig. 1. Apples stored for experimental purpose. a) Collected Apples in rack. b) Controlled storage.

TABLE I. Tabulation of Number of Record's in the Dataset Involving 2900 Columns with 12 Different Apple-Fungi-Ion Combination

Apple	Fungi	+ve ion	–ve ion	File Name	Infected	Number of records
Braeburn	Bortrytis cinerea	✓		BBB+ve	✓	50
Braeburn	Bortrytis cinerea	✓			×	45
Braeburn	Monilinia laxa	✓		BBM+ve	✓	33
Braeburn	Monilinia laxa	✓			×	27
Braeburn	Neonectria ditissima	✓		BBN+ve	✓	36
Braeburn	Neonectria ditissima	✓			×	45
Gala	Bortrytis cinerea	✓		GAB+ve	✓	38
Gala	Bortrytis cinerea	✓			×	30
Gala	Monilinia laxa	✓		GAM+ve	✓	36
Gala	Monilinia laxa	✓			×	32
Gala	Neonectria ditissima	✓		GAN+ve	✓	30
Gala	Neonectria ditissima	✓			×	36
Braeburn	Bortrytis cinerea		✓	BBB-ve	✓	50
Braeburn	Bortrytis cinerea		✓		×	50
Braeburn	Monilinia laxa		✓	BBM-ve	✓	30
Braeburn	Monilinia laxa		✓		×	33
Braeburn	Neonectria ditissima		✓	BBN-ve	✓	37
Braeburn	Neonectria ditissima		✓		×	41
Gala	Bortrytis cinerea		✓	GAB-ve	✓	38
Gala	Bortrytis cinerea		✓		×	30
Gala	Monilinia laxa		✓	GAM-ve	✓	39
Gala	Monilinia laxa		✓		×	42
Gala	Neonectria ditissima		✓	GAN-ve	✓	37
Gala	Neonectria ditissima		✓		×	31

IV. Result and Discussion

A. Exploratory Analysis on the Data

Since there is limited existing literature on applying different machine learning algorithms to analyze apple fungal infections and classify them, the initial approach involved conducting a thorough Exploratory Data Analysis (EDA) on the mass spectrometry dataset. A boxplot image on the BBB+ve infected 50 samples is shown in Fig. 2. It distinctly illustrates the values within the interquartile range for each sample and the dispersion of the outliers.

From the boxplot of the BBB+ve ion, it's found that only 1% of the data points falls under outlier. The lower quartile, upper quartile and the median value of almost all the samples falls under the similar range.

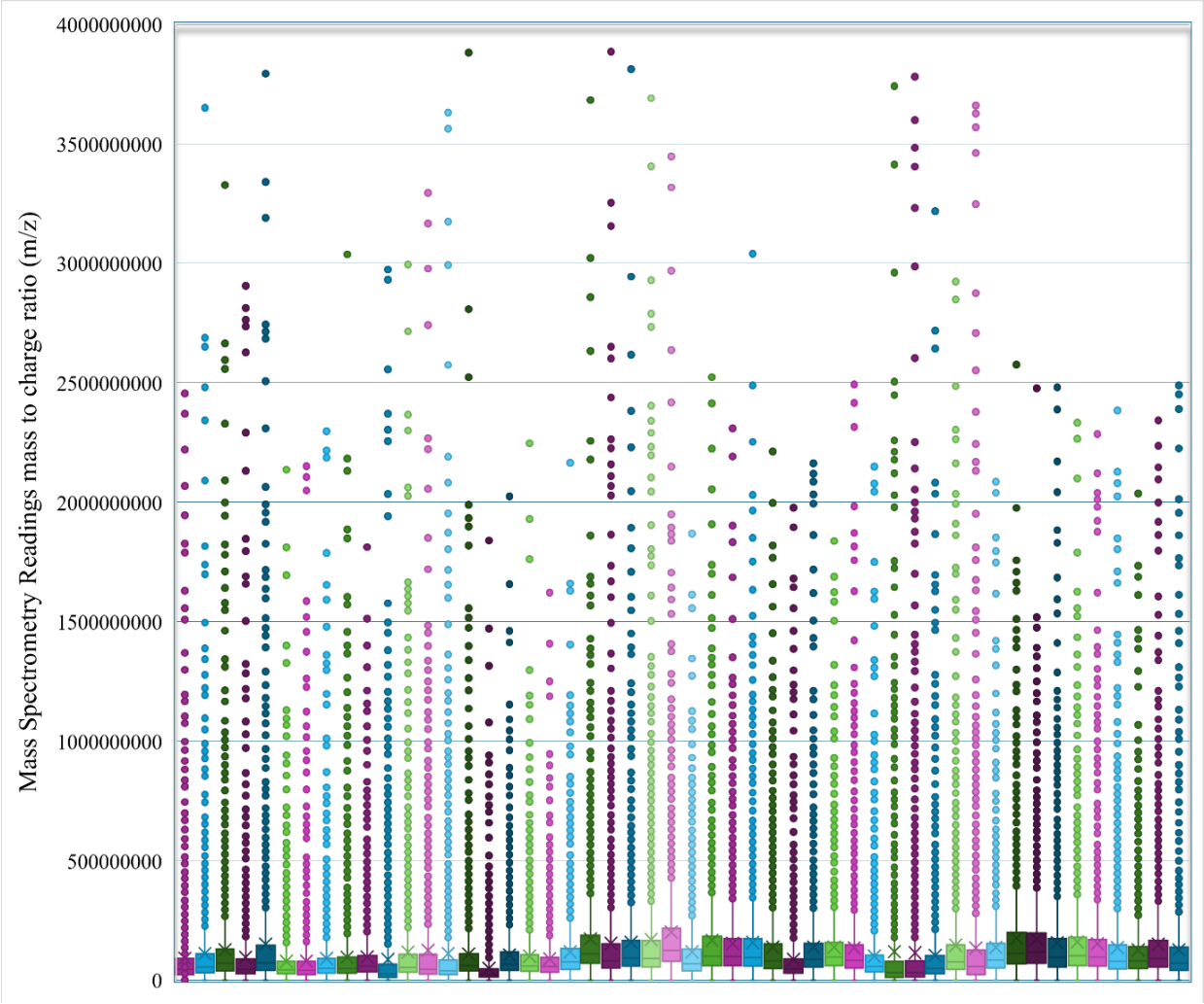

Fig. 2. Exploratory data analysis-box plot on the BBB+ve infected 50 samples.

From the analysis recorded during the literature study, supervised algorithms like Support Vector Machine (SVM), K-Nearest Neighbor Classification (KNN), Artificial Neural Network (ANN), Decision Tree (DT), and Naive Bayes (NB) [25] are predominantly used for classification problems and have produced good classification results. With this understanding, five distinct Machine Learning algorithms were employed in this study including Logistic Regression, SVM-XGBoost Gridsearch with Linear Kernel, SVM-XGBoost Gridsearch with Polynomial Kernel, XGBoost, and Random Forest. Additionally, four customised Neural Network algorithms were designed, leveraging both Adaptive Moment Estimation (ADAM) and Stochastic Gradient Descent (SGD) classifiers. The Neural Network hyperparameters were fine-tuned, and their performance was subsequently assessed and recorded. The true positive, true negative, false positive, false negative, and accuracy metrics were documented for each of the 12 combinations involving Apple-Fungi-Ion in Table II. The 12 combinations include BBB+ve, BBM+ve, BBN+ve, GAB+ve, GAM+ve, GAN+ve, BBB-ve, BBM-ve, BBN-ve, GAB-ve, GAM-ve, GAN-ve.

TABLE II. PERFORMANCE ANALYSIS OF THE ML ALGORITHMS ON THE 12 COMBINATIONS OF APPLE-FUNGI-ION. THE HIGHEST ACCURACY FOR EACH OF THE APPLE-FUNGI-ION ARE IN RED AND THE SECOND HIGHEST ARE IN BLUE

Model	BBB+ve Train:71, Test:24					BBM+ve Train:45, Test:15					BBN+ve Train:60, Test:21				
	TP	FP	FN	TN	Accuracy	TP	FP	FN	TN	Accuracy	TP	FP	FN	TN	Accuracy
Logistic Regression	13	0	1	10	**95.8**	6	2	0	7	**86.6**	8	0	0	13	**100**
SVM–Linear	12	1	1	10	**91.6**	7	1	0	7	**93.3**	8	0	0	13	**100**
SVM–Polynomial	13	0	8	3	66.6	6	2	0	7	**86.6**	5	3	0	13	85.7
XGBoost	11	2	2	9	83.3	5	3	0	7	80	6	2	0	13	90.4
Random Forest	11	2	3	8	79.1	6	2	2	5	73.3	6	2	0	13	90.4
NN model 1	10	3	3	8	75	6	2	2	5	73.3	5	3	0	13	85.7
NN model 2	10	3	3	8	75	6	2	2	5	73.3	8	0	1	12	**95.2**
NN model 3	2	7	3	7	47.3	3	0	3	6	75	2	2	1	12	82.3
NN model 4	7	2	3	7	73.6	2	1	4	5	58.3	3	1	1	12	88.2

Model	GAB+ve Train:71, Test:24					GAM+ve Train:45, Test:15					GAN+ve Train:60, Test:21				
	TP	FP	FN	TN	Accuracy	TP	FP	FN	TN	Accuracy	TP	FP	FN	TN	Accuracy
Logistic Regression	6	1	0	10	**94.1**	7	0	0	10	**100**	4	1	0	12	**94.1**
SVM–Linear	6	1	0	10	**94.1**	7	0	0	10	**100**	4	1	0	12	**94.1**
SVM–Polynomial	6	1	7	3	52.9	7	0	5	5	70.5	5	0	5	7	70.5
XGBoost	6	1	3	7	76.4	7	0	3	7	82.3	4	1	2	10	82.3
Random Forest	6	1	2	8	82.3	7	0	2	8	**88.2**	5	0	0	12	**100**
NN model 1	6	1	1	9	88.2	7	0	2	8	**88.2**	5	0	1	11	**94.1**
NN model 2	6	1	0	10	**94.1**	6	1	1	9	**88.2**	4	1	0	12	**94.1**
NN model 3	7	0	2	5	85.7	5	2	0	7	85.7	7	0	2	5	85.7
NN model 4	7	0	0	7	**100**	6	1	1	6	85.7	7	0	2	5	85.7
Model	**BBB-ve Train:71, Test:24**					**BBM-ve Train:45, Test:15**					**BBN-ve Train:60, Test:21**				
	TP	FP	FN	TN	Accuracy	TP	FP	FN	TN	Accuracy	TP	FP	FN	TN	Accuracy
Logistic Regression	12	1	1	11	**92**	4	3	1	8	**75**	7	1	2	10	**85**
SVM–Linear	12	1	1	11	**92**	4	3	1	8	**75**	7	1	2	10	**85**
SVM–Polynomial	8	5	1	11	76	3	4	1	8	68.75	8	0	2	10	**90**
XGBoost	11	2	0	12	**92**	2	5	1	8	62.5	6	2	1	11	**85**
Random Forest	10	3	1	11	84	2	5	1	8	62.5	7	1	2	10	**85**
NN model 1	6	7	0	12	72	3	4	3	6	56.25	7	1	3	9	80
NN model 2	8	5	0	12	80	3	4	3	6	56.25	7	1	3	9	80
NN model 3	6	3	1	10	80	2	4	1	6	61.5	5	0	3	8	81.25
NN model 4	6	3	1	10	80	3	3	1	6	69.2	5	0	3	8	81.25
Model	**GAB-ve Train:51, Test:17**					**GAM-ve Train:60, Test:21**					**GAN-ve Train:51, Test:17**				
	TP	FP	FN	TN	Accuracy	TP	FP	FN	TN	Accuracy	TP	FP	FN	TN	Accuracy
Logistic Regression	4	3	2	8	70.5	8	0	0	13	**100**	7	0	0	10	**100**
SVM–Linear	4	3	2	8	70.5	8	0	0	13	**100**	7	0	0	10	**100**
SVM–Polynomial	6	1	6	4	58.8	8	0	3	10	85.7	7	0	2	8	88.2
XGBoost	4	3	3	7	64.7	8	0	1	12	**95.2**	7	0	0	10	**100**
Random Forest	4	3	3	7	64.7	8	0	0	13	**100**	7	0	0	10	**100**
NN model 1	3	4	2	8	64.7	7	1	0	13	**95.2**	5	2	0	10	88.2
NN model 2	4	3	1	9	**76.4**	7	1	0	13	**95.2**	7	0	1	9	**94.1**
NN model 3	5	2	3	4	64.2	4	0	4	9	76.4	7	0	0	7	**100**
NN model 4	5	2	2	5	**71.4**	4	0	4	9	76.4	7	0	1	6	92.8

B. ML Model 1: Logistic Regression

The model takes as input a set of 2,900 attributes derived from mass spectrometry data, represented as $x_1, x_2, \ldots, x_{\{2900\}}$ with the response variable indicating a binary classification of {"infected"(0), "uninfected"(1)}. Applying logistic regression to classify the dataset yielded outstanding results, achieving the highest and second-highest accuracy scores in 11 combinations of Apple-Fungi-Ion (Fig. 3).

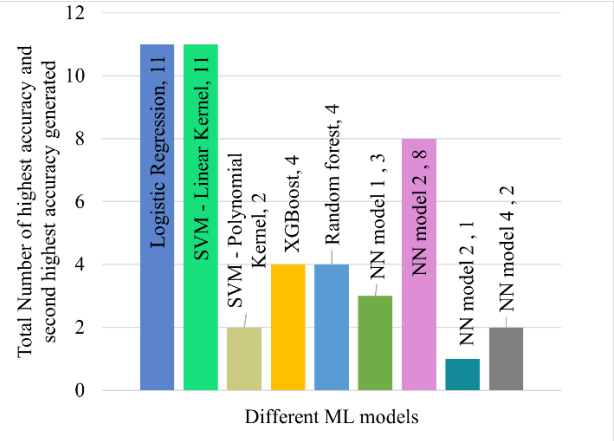

Fig. 3. The total number of highest accuracy and second highest accuracy generated by each of the ML models.

The reason behind the highest performance of the logistic model is that it follows a Sigmoid prediction curve represented by $predicted\_value = \frac{1}{1+e^{\{-x\}}}$. The error function of the logistic regression is given by

$$\frac{1}{2\times no\_of\_records} \times \sum_{i=1}^{no\_of\_records} (predicted\_value - actual\_value)^2 \quad (1)$$

The logistic regression error value is calculated in two iterations as given below

$$Error(predicted\_value - actual\_value) = \begin{cases} -log(predicted\_value), & actual\_value = 1 \\ -log(1 - predicted\_value), & actual\ value = 0 \end{cases} \quad (2)$$

On applying Eq. (2) in Eq. (1), and by compressing the conditions in Eq. (2), the differentiated final error function of the logistic regression would be

$$= \frac{-1}{m} \sum_{i=1}^{no\_of\_records} actual\_value \times \log(prediction\_value) + (1 - actual\_value) \log(1 - prediction\_value) \quad (3)$$

From the sample EDA boxplot in Fig. 2, outliers are seen in the spectrometry recordings which may disturb the behavior of the model causing it to be overfitting. Hence to combat the effect of overfitting caused by high variance in the data, regularisation is applied on them. The selection of regularisation over pruning stems from the fact that pruning features could result in data loss, potentially compromising the accuracy of predicting the extent of fungal infection. Hence the logistic function is further optimized applying the L1 regularisation (Eq. (4)) and L2 regularisation (Eq. (5)) on it. L1 regularisation considers the absolute values of the weights, while L2 regularisation involves the squares of the weights. Upon experimenting with both L1 and L2 regularisation, we observe minimal deviations in the results obtained.

$$Error\ from\ Eq.(3) + \frac{\lambda}{2\times no\_of\_records} \sum_{j=1}^{n} |W_j| \quad (4)$$

$$Error\ from\ Eq.(3) + \frac{\lambda}{2 \times no\_of\_records} \sum_{j=1}^{n} |W_j|^2 \quad (5)$$

One significant aspect of the dataset that greatly contributed to logistic regression achieving outstanding outcomes is its minimal number of outliers, possibly only a few among the 2,900 samples. This could be attributed to the dataset's origin from real experiments, reducing the probability of outlier occurrences, and any outliers present were addressed during the regularisation process. Moreover, the cross-entropy loss function for logistic regression was consistently minimal across the 11 models, indicating its suitability for the dataset's nature. The monotonicity of logistic regression aligned well with the dataset characteristics. Fig. 4(A) illustrates that logistic regression yielded excellent results for all combinations of Apple-Fungi-Ion except GAB-ve.

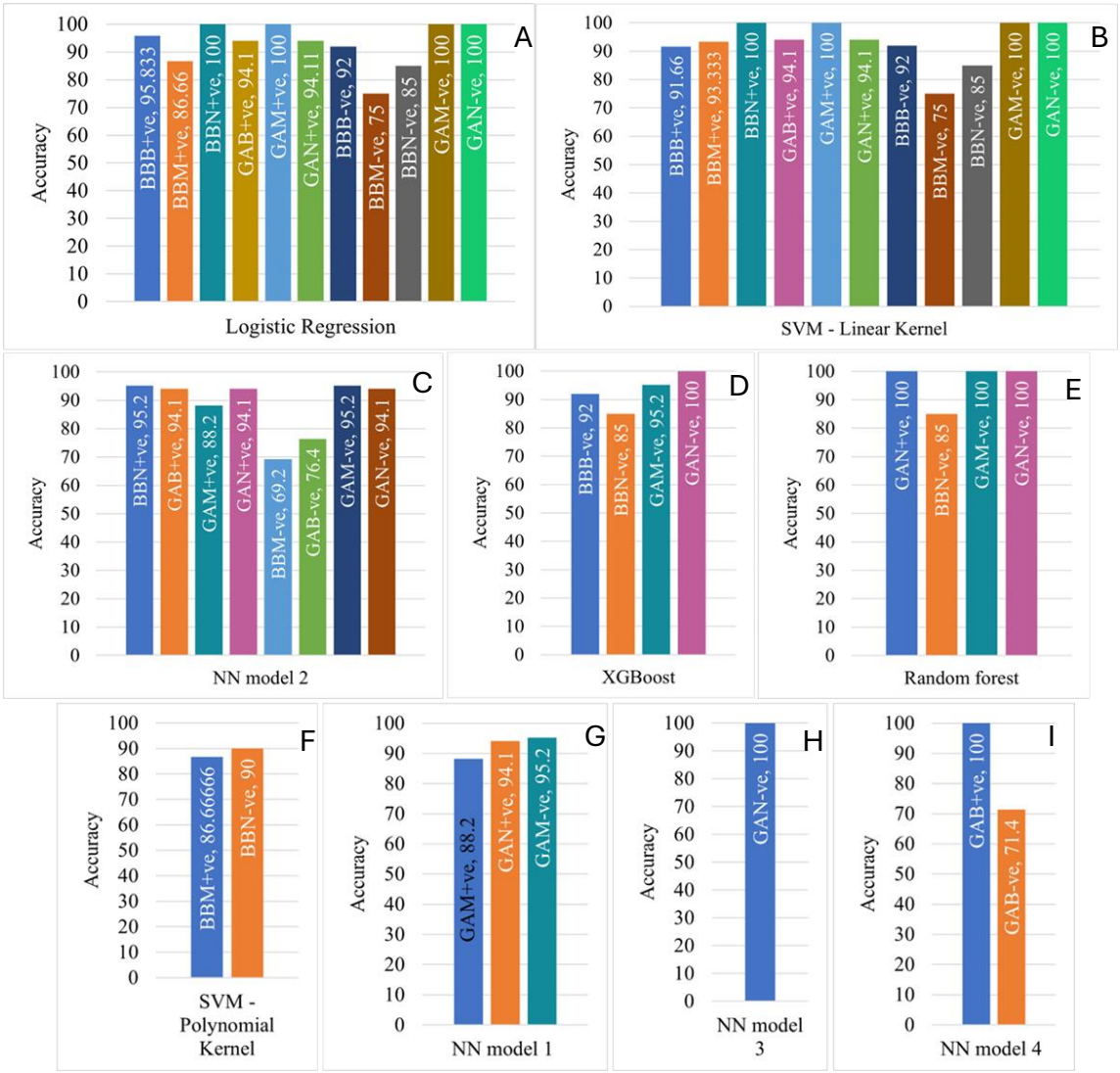

Fig. 4. Comparison of the performance of each machine learning model in terms of achieving the highest and second highest.

C. SVM–Linear and Polynomial regression

Support Vector Machines (SVM) is a supervised learning algorithm used for classification and regression tasks, known as Support Vector Classification (SVC) and Support Vector Regression (SVR) respectively. It's particularly suitable for smaller datasets due to longer processing times. SVM aims to find an optimal hyperplane that effectively separates different classes, especially in binary classification scenarios. SVM can employ various kernels such as linear, polynomial, and Radial Basis Function (RBF). Typically, the linear kernel performs well for linear data, while the polynomial kernel is suitable for nonlinear data. The RBF kernel is preferred for more complex datasets.

In general, Fig. 4(B) illustrates that the SVM with a linear kernel achieves high accuracy across all models except GAB-ve. Interestingly, the results of SVM with kernels closely resemble those of the logistic regression model. Fig. 4(F) depicts the outcomes of SVM with a polynomial kernel, indicating highest performance for only two models. This suggests that the data exhibits predominantly linear characteristics with slight non-linear deviations, despite having many dimensions in the labeled

dataset. The initial linear equation is represented by $y = w_0 + w_1 \times x_1 + w_2 \times x_2 + \cdots + w_m \times x_m$, where m is the dimensions. The mass spectrometry dataset has 2900 attributes, hence the input equation would be $y = w_0 + w_1 \times x_1 + w_2 \times x_2 + \cdots + w_{\{2900\}} \times x_{\{2900\}}$. This equation is hence reduced to a vector form as given below:

$$y = w_0 + \sum_{i=1}^{2900} w_i \times x_i,\ y = w_0 + W^T \times X \text{ and } y = b + W^T \times X \quad (6)$$

where W and X are weight and input vectors.

SVM constructs three hyperplanes to create a decision surface that effectively separates the data. These hyperplanes are represented by the equations:

$$y = w_0 + W^T \times X = 0, y^+ = w_0 + W^T \times X = 1$$

$$y^- = w_0 + W^T \times X = -1 \quad (7)$$

The goal is to find the optimal hyperplane that best separates the data, which involves solving a specific optimization problem as given in equation below:

$$min\ \varphi(w) = \frac{1}{2} ||\ w\ ||^2,\ where,\ \phi: R^n \rightarrow\ R^n \quad (8)$$

As per the above equation, a mapping is done from features in high dimensional feature space to low dimension, where these points can become linearly separable. Similarly for the polynomial kernel in SVM, the mapping function is represented by $k(x_i, x_i) = \left(x_j^T \times x_i + t\right)^d$, where d is the higher order power of the polynomial term. Based on the above equations, the SVM kernel classifies the input data into two classes.

D. XGBoost and Random Forest

XGBoost and Random Forest are selected due to their respective strengths: Random Forest effectively captures dataset behavior with numerous features, while XGBoost optimally tunes hyperparameters to capture relationships within the data.

Ensemble model:

Let there be *n* data samples with 2900 features and $y_i = \{0,1\}$.

Given that $D = \{(X_i, y_i)\}, (|D|) = N, X_i = \{1,2900\}$, $y_i = \{0,1\}$ The ensemble gradient boosting model uses B different functions to predict the output. The predicted output would be

$$predicted\_value(y\_i) = \sum_{b=1}^{B} func_{b(X_i)}, func_{b(X_i)} \in FUNC \quad (9)$$

where $FUNC = \{func(x) = W_j\}$ and B different regression functions or Classification and Regression Trees (CART) are used. Each func_b corresponds to each individual tree, with weight w_j and w is the weight of the leaf nodes in the tree, if the tree has T leaf nodes. Each of the regression decision tree derives a continuous score for the leaves and sum of the values of the leaf values are added to obtain the predicted result. The $FUNC$ are learned in an attempt by minimising the regularized objective:

$$L = \sum_{b=1}^{B} l(predicted\_value, actual\_value) + \frac{1}{2} \lambda \|w\|^2 \quad (10)$$

The function l applied is a convex loss function representing the difference between $predicted\_value$ and $actual\_value$. The complexity of the model is decided by the $weight\_factor$. At the $(t)^{th}$ iteration the value of the regularisation objective depends on the value of $(t-1)^{th}$ regularisation objective denoted by

$$L = \sum_{b=1}^{B} l(predicted\_value, actual\_value)^{t-1}) + \frac{1}{2} \lambda \|w\|^2 \quad (11)$$

This approach is a greedy method that incorporates the previous input values with the current one to enhance the model's performance.

Further down the model, the second layer optimization is done in XGBoost to optimize the results. In the proposed case the optimal results are the weights are calculated in reduced optimization steps using the following weight adjustment:

$$W = \frac{\sum_i d_{1i}}{\sum_i d_{2i} + \lambda} \quad (12)$$

The i in the above equation spans across all the leaf nodes, where d_{1i} and d_{2i} refers to the first order and second order differentiated values of the loss function l. Therefore, the optimization function simplifies to

$$L = \frac{-1}{2} \sum_{j=1}^{no\_of\_trees} \frac{\sum_i d_{1i}}{\sum_i d_{2i} + \lambda} \quad (13)$$

The values produced by the above equation measures the impurity of the tree. The XGBoost algorithm proposed in this solution is run until it converges to reach a minimal score but carefully stopped at a stage to prevent overfitting. The XGBoost algorithm is designed with a greedy approach to handle both local and global optimization efficiently, capable of processing sparse input data and executing parallel operations simultaneously.

The Random Forest algorithm however is not affected by loss function hence there is no option to minimize the loss as we could do in the XGBoost. However,

$$predicted\_value = \sum_{i=1}^{X} \left(\frac{1}{no_{of_{trees}}} \sum_{j=1}^{no_{of_{trees}}} W\right) \times actual\_value \quad (14)$$

where X is the number of attributes in the dataset. The random forest is trained by passing each of the data item as input and the weights are adjusted. Subsequently, the entropy (impurity value) is measured. The training process iterates until the entropy reaches a minimal value and at the point of convergence the algorithm is stopped. On average both the XGBoost and Random Forest were trained by generating 500-600 tress for each potential Apple-Fungi-Ion combination.

Both XGBoost and Random Forest demonstrated exceptional performance, achieving the highest and second-highest scores for four Apple-Fungi-Ion combinations and attaining a perfect 100% score on GAN-ve (Fig. 4(D) and Fig. 4(E)).

E. Neural Network Model

The neural network model architecture for the proposed system is illustrated in Fig. 5. It consists of four layers, with Rectified Linear Unit (ReLU) activation functions applied to the three hidden layers and a Sigmoid activation function used for the output layer. The model takes 2900

features as input and generates a binary output indicating whether the specified sample is infected or uninfected.

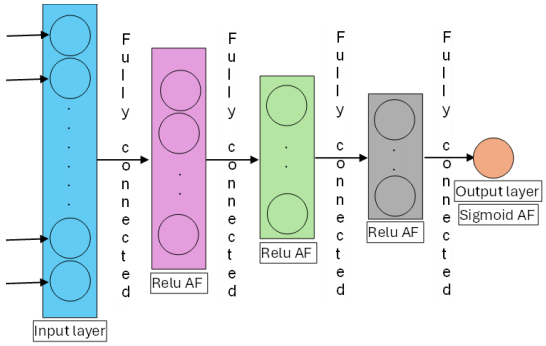

Fig. 5. Architecture of the Neural Network model.

Algorithm 1. Structure of the proposed Neural Network Algorithm

1. The input layer in the neural network has 2900 neurons denoted by, $X = \{x_1, x_2, x_3, x_4, \ldots\ldots, x_{\{2900\}}\}$. The output O_X from the neurons in the input layer is passed to all the neurons in the hidden layer 1 forming a fully connected network.

2. The hidden layer-1 has 2900 neurons. The input to each neuron in the hidden layer-1 is given by $I_{h1} = \sum_{i=1}^{2900} O_{Ii}$, where O_i is the output from the *ith* neuron in the input layer. Subsequently, the neurons in hidden layer-1 processes the inputs by computing the sum of all inputs, applies the ReLU activation function and generates an output as given below:

$$O_{h1} = \begin{cases} I_{h1}, & I_{h1} > 0 \\ 0, & I_{h1} \leq 0 \end{cases}$$

3. The output of the 2900 neurons in the hidden layer-1 is passed to the neurons in the hidden layer-2 forming a fully connected network. The hidden layer-2 has 500 neurons and input to each neuron is the sum of all inputs from the hidden layer-1, $I_{h2} = \sum_{i=1}^{2900} O_{h1}$, applies the RELU activation function and generates an output as given below:

$$O_{h2} = \begin{cases} I_{h2}, & I_{h2} > 0 \\ 0, & I_{h2} \leq 0 \end{cases}$$

4. The output of the 500 neurons in the hidden layer-2 is passed to the neurons in the hidden layer-3 forming a fully connected neural network. The hidden layer-3 has 10 neurons and the input to each neuron is the sum of all the inputs from the hidden layer-2, $I_{h3} = \sum_{i=1}^{500} O_{h2}$ applies the RELU activation function and generates an output as given below:

$$O_{h3} = \begin{cases} I_{h3}, & I_{h3} > 0 \\ 0, & I_{h3} \leq 0 \end{cases}$$

5. The output of the 10 neurons in the hidden layer-3 is passed to the neuron in the output layer. The output layer has one neuron and the input to the neuron is the sum of all the inputs from the hidden layer-3, $I_O = \sum_{i=1}^{10} O_{h3}$, applies the SIGMOID activation function and generates an output as given below:

$$O_o = \frac{1}{(1 + e^{-I_o})^2}$$

6. The output neuron generates a binary output corresponding to infected or uninfected apples.

7. Following each iteration, the error is computed, and subsequently, backpropagation is performed to adjust the weights. The parameters and hyperparameters of the neural network, including optimizers and epochs, are systematically varied. Four unique neural network models are developed, and their individual results are recorded.

Neural Network Model-1 and Neural Network Model-2 utilize the ADAM optimizer and undergo training for 10 epochs and 100 epochs, respectively, on the training dataset for each Apple-Fungi-Ion combination. After training, each combination is evaluated using the corresponding test dataset. Model-1 successfully achieved the highest and second-highest scores for three combinations of Apple-Fungi-Ion (Fig. 4(G)). Model-2 achieved the highest and second-highest scores for eight combinations of Apple-Fungi-Ion (Fig. 4(C)). The unique property of the ADAM optimiser is that it corrects the bias error as it combines the effects of two different optimisers AdaGrad and RMSProp and hence converges early to reach the global minima. The Adam optimiser is unique for the reason that it captures the first moment and the second moment, the mean and the variance of the gradients, enabling faster convergence. Let θ be the parameter of the model, l be the learning rate and the cost calculated at time/iteration i is given by

$$\theta = \theta - l \times grad_{cost\_i} \quad (15)$$

The objective is to navigate θ in the direction to minimize the cost function. The step width of the navigation is decided by l. The two main internal parameters of the Adam are momentum which speeds up the gradient and the step size. The version of the Adam that works on the mean and the variance are

$$mean_i = hp_1 \times mean_{\{i-1\}} + (1 - hp_1) \times grad_{cost\_i}$$

$$variance_i = hp_2 \times variance_{\{i-1\}} + (1 - hp_2) \times \left(grad_{cost\_i}\right)^2 \quad (16)$$

The hp_1 and hp_2 signifies the hyperparameters, $grad_{cost\_i}$ represents the gradient of the cost function.

The parameter θ is adjusted as per the following Eq. (17):

$$\theta = \theta - \left(l \times \frac{mean_i}{\sqrt{\{variance\_i + min_{const}\}}}\right) \quad (17)$$

where min_{const} represents a minimal value of constant aimed at preventing the denominator from approaching zero. Adam makes a warm start and keeps updating the value of $mean_i$ and $variance_i$ until N step which leads to convergence as shown below:

$$\widehat{mean}_i = \frac{mean_i}{1 - hp_{1_i}} \quad (18)$$

$$\widehat{variance}_i = \frac{variance_i}{1 - hp_{2_i}} \quad (19)$$

Further down, the θ value is adjusted based on the new capped mean and variance.

$$\theta = \theta - \left(l \times \frac{\widehat{mean}_i}{\sqrt{\{\widehat{variance}\_i + min_{const}\}}}\right) \quad (20)$$

By the 10^{th} epoch the model reaches a convergence value.

The Neural Network Model 3 is constructed with the ADAM optimizer and undergoes training over 100 epochs on the training datasets. This model is designed to be a 3-layer NN model with 6 neurons in each of the hidden layer and 1 neuron in the output layer. It is validated on a

separate validation dataset and subsequently tested on a test dataset. Despite this extensive training, the model achieved the highest and second-highest scores for one combination of Apple-Fungi-Ion. Unfortunately, the performance wasn't better than the other models (Fig. 4(H)), achieving outstanding results only for the GAN-negative combination.

Neural Network Model 4 is built using the SGD optimizer and trained over 10 epochs. This model achieved the highest and second-highest scores for two combinations of Apple-Fungi-Ion (Fig. 4(I)). The θ in SGD optimizer is calculated by:

$$\theta_{i+1} = \theta_i - l \times grad_{loss\_i} \tag{21}$$

Asper the above equation, the training parameter θ is updated every iteration where $grad_{loss\_i}$ represents the gradient of the loss function. In SGD, the weight and the bias are updated using the below formulae:

$$Weight_{\{i+1\}} = Weight_i - l \times grad_{loss\_i} \tag{22}$$

$$Bias_{\{i+1\}} = Bias_i - l \times grad_{\{loss\_i\}} \tag{23}$$

The $grad_{\{loss\_i\}}$ is the differentiated value of the loss calculated at the iteration i by applying

$$grad_{\{loss\_i\}} = \frac{\partial Loss}{\partial Weight\_i} \tag{24}$$

SGD updates the weight and bias and other parameters ater processing each record, resulting in a slower convergence to the global minima where the path becomes rough rather smooth. Comparing both optimizers, ADAM and SGD, ADAM showcased superior performance due to its quicker narrowing down to the global minima compared to SGD, and its effectiveness in optimization during training.

Based on the comprehensive study and analysis conducted, it has been demonstrated that out of the nine models developed, both Logistic Regression and SVM with linear kernels consistently yielded the best results for around 11 combinations of Apple-Fungi-Ion.

V. Conclusion

This research is a significant contribution to the field of food safety following the sustainable practices, with potential implications for the global scientific community and the apple industry.

Non-destructive approach to detection of fungal infections: The developed method utilizes mass spectrometry data and machine learning algorithms to detect fungal infections in apples without causing physical damage to the fruit. This non-destructive approach is significant to the apple industry, as it saves a lot of apples from being cut for testing.

Comprehensive analysis: This research conducted a thorough analysis by evaluating the performance of nine different supervised machine learning models, including logistic regression, Support Vector Machines (SVMs), XGBoost, Random Forest, and four customised neural networks. This comprehensive approach learns the linear nature of the mass spectrometry dataset as evidenced by the excellent performance of logistic regression and SVM with linear kernels for accurately classifying infected and uninfected apples across various combinations of apple varieties and fungal pathogens. The logistic regression and SVM models achieved over 90% accuracy across 11 different Apple-Fungi-Ion combinations demonstrating the robustness and reliability of the developed method, which is essential for practical implementation in the apple industry. This approach being rapid and produces quick results, offers time-saving benefits in the fast-paced agricultural and food industry.

Building on the initial extensive results, this study is set to evolve into a proposed funded project aimed at uniting academia, industry professionals, and stakeholders and is designed to benefit the apple industry and consumers, with the aim of providing significant economic and societal advantages to a wide range of people.

The research acknowledges the potential for future iterations of the project, such as quantifying the extent of infection and determining the apple's edibility duration or shelf life. Further down, it can also be extended to other stone fruits like pears or berry fruits like strawberry, raspberry, etc., thus opening avenues for further exploration and refinement.

Overall, this research also demonstrates the power of interdisciplinary collaboration between analytical techniques, machine learning, and agricultural sciences. By providing a reliable, non-destructive, and cost-effective method for detecting fungal infections in apples, this study has the potential to significantly impact the global scientific community, particularly in the areas of food security, agricultural productivity, and sustainable practices.

Conflict of Interest

The authors declare no conflict of interest

Author Contributions

Razia Sulthana Abdul Kareem conducted the research, analyzed the dataset, and wrote the paper. Charles A. I. Goodall provided the mass spectrometry dataset and provided detailed information on the data collection and apple fungal infections, while Timothy Tilford initiated the research collaboration. Along with Nageena Frost and Ana Paula Palacios, we all participated in numerous discussions, shared opinions, and contributed to the revisions leading to the final version. All authors had approved the final version.

References

[1] F. R. Harker, F. A. Gunson, and S. R. Jaeger, "The case for fruit quality: An interpretive review of consumer attitudes, and preferences for apples," *Postharvest Biology and Technology*, vol. 28, no. 3, pp. 333–347, 2003.

[2] E. Commission, "Commission implementing regulation (eu) no 543/2011 of 7 June 2011, laying down detailed rules for the application of council regulation (ec) no 1234/2007 in respect of the fruit and vegetables and processed fruit and vegetables sectors," *Off. J. Eur. Union.*, vol. 157, pp. 1–163, 2011.

[3] M. Dutot, L. Nelson, and R. Tyson, "Predicting the spread of postharvest disease in stored fruit, with application to apples," *Postharvest Biology and Technology*, vol. 85, pp. 45–56, 2013.
[4] J. Kohl, M. Wenneker, B. Groenenboom-de Haas, R. Anbergen, H. G. van de Geijn, C. L. van der Plas, F. Pinto, and P. Kastelein, "Dynamics of post-harvest pathogens *Neofabraea* spp. and *Cadophora* spp. in plant residues in Dutch apple and pear orchards," *Plant Pathology*, vol. 67, no. 6, pp. 1264–1277, 2018.
[5] L. Gao, Q. Zhang, X. Sun, L. Jiang, R. Zhang, G. Sun, Y. Zha, and A. R. Biggs, "Etiology of moldy core, core browning, and core rot of Fuji apple in China," *Plant Disease*, vol. 97, no. 4, pp. 510–516, 2013.
[6] P. Pieczywek, J. Cybulska, M. Szymanska-Chargot, A. Siedliska, A. Zdunek, A. Nosalewicz, P. Baranowski, and A. Kurenda, "Early detection of fungal infection of stored apple fruit with optical sensors—Comparison of bio speckle, hyperspectral imaging and chlorophyll fluorescence," *Food Control*, vol. 85, pp. 327–338, 2018.
[7] R. A. Spotts, M. Serdani, K. M. Wallis, M. Walter, T. Harris-Virgin, K. Spotts, D. Sugar, C. L. Xiao, and A. Qu, "At-harvest prediction of grey mould risk in pear fruit in long-term cold storage," *Crop Protection*, vol. 28, no. 5, pp. 414–420, 2009.
[8] J. Blasco, N. Aleixos, and E. Molto, "Computer vision detection of peel defects in citrus by means of a region oriented segmentation algorithm," *Journal of Food Engineering*, vol. 81, no. 3, pp. 535–543, 2007
[9] J. Boyer and R. H. Liu, "Apple phytochemicals and their health benefits," *Nutrition Journal*, vol. 3, pp. 1–15, 2004.
[10] F. Derikvand, E. Bazgir, M. E. Jarroudi, M. Darvishnia, H. M. Najafgholi, S.-E. Laasli, and R. Lahlali, "Unleashing the potential of bacterial isolates from apple tree rhizosphere for biocontrol of *Monilinia laxa*: A promising approach for combatting brown rot disease," *Journal of Fungi*, vol. 9, no. 8, 828, 2023. https://doi.org/10.3390/jof9080828
[11] L. Garkava-Gustavsson, A. Zborowska, J. Sehic, M. Rur, H. Nybom, J. Englund, M. Lateur, E. van de Weg, and A. Holefors, "Screening of apple cultivars for resistance to European canker, *Neonectria ditissima*," *Acta Horticulturae*, vol. 976, pp. 529–536, 2013.
[12] K. Resnicow and G. Botvin, "Effects decay?" *Preventive Medicine*, vol. 22, pp. 484–490, 1993.
[13] C. L. Lennox, R. A. Spotts, and L. A. Cervantes, "Populations of botrytis cinerea and penicillium spp. on pear fruit, and in orchards and packinghouses, and their relationship to postharvest decay," *Plant Disease*, vol. 87, no. 6, pp. 639–644, 2003.
[14] N. E. Dina, A. M. R. Gherman, V. Chis, C. Sarbu, A. Wieser, D. Bauer, and C. Haisch, "Characterization of clinically relevant fungi via sers fingerprinting assisted by novel chemometric models," *Analytical Chemistry*, vol. 90, no. 4, pp. 2484–2492, 2018.
[15] Z. Guo, M. Wang, A. O. Barimah, Q. Chen, H. Li, J. Shi, H. R. E. Seedi, and X. Zou, "Label-free surface enhanced Raman scattering spectroscopy for discrimination and detection of dominant apple spoilage fungus," *International Journal of Food Microbiology*, vol. 338, 108990, 2021.
[16] O. Zukovskaja, S. Kloß, M. G. Blango, O. Ryabchykov, O. Kniemeyer, A. A. Brakhage, T. W. Bocklitz, D. Cialla-May, K. Weber, and J. Popp, "UV-Raman spectroscopic identification of fungal spores important for respiratory diseases," *Analytical Chemistry*, vol. 90, no. 15, pp. 8912–8918, 2018.
[17] J. Wang, R. Zhu, Y. Wu, L. Tang, C. Wang, M. Qiu, L. Zheng, P. Li, and S. Weng, "Dynamic surface-enhanced Raman spectroscopy and positively charged probes for rapid detection and accurate identification of fungal spores in infected apples via deep learning methods," *Food Control*, vol. 157, 110151, 2024.
[18] N. Haghbin, A. Bakhshipour, S. Mousanejad, and H. Zareiforoush, "Monitoring botrytis cinerea infection in kiwifruit using electronic nose and machine learning techniques," *Food and Bioprocess Technology*, vol. 16, no. 4, pp. 749–767, 2023.
[19] A. R. Sulthana and A. Jaithunbi, "Varying combination of feature extraction and modified support vector machines based prediction of myocardial infarction," *Evolving Systems*, vol. 13, no. 6, pp. 777–794, 2022.
[20] S. A. Razia and R. Pranav, "Predicting the import and export of commodities using support vector regression and long short-term predition models," *International Journal of Computing and Digital Systems*, vol. 11, no. 1, pp. 635–648, 2022.
[21] R. Weber, "Biology and control of the apple canker fungus *Neonectria ditissima* (syn. *N. galligena*) from a northwestern European perspective," *Erwerbs-Obstbau*, vol. 56, no. 3, pp. 95–107, 2014.
[22] J. Gelain, S. A. M. Alves, R. R. Moreira, and L. L. M. D. Mio, "*Neonectria ditissima* physiological traits and susceptibility of 'gala' and 'eva' detached apple fruit," *Tropical Plant Pathology*, vol. 45, pp. 25–33, 2020.
[23] M. Turkoglu, D. Hanbay, and A. Sengur, "Multi-model LSTM-based convolutional neural networks for detection of apple diseases and pests," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 7, pp. 3335–3345, 2022.
[24] S. P. Adam, S.-A. N. Alexandropoulos, P. M. Pardalos, and M. N. Vrahatis, "No free lunch theorem: A review," *Approximation and optimization: Algorithms, Complexity and Applications*, pp. 57–82, 2019.
[25] R. Sulthana, A. Jaithunbi, and P. Sunraja, "Application of machine learning algorithms in predicting the heart disease in patients," in *Proc. 2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, IEEE, 2023, pp. 1–4.