A review on deriving useful parameters that may affect the harvest of an apple using Supervised Learning

Ruthvik Raja.M.V, Jayasaikiran.A, ENGG 6500, University of Guelph

***Abstract*—As the availability of skilled harvest labours is declining for harvesting crops, it is very important to find an alternate solution like inventing an automated transplanter. It saves time and meets our requirements with the current demand and supply. Whereas, in the current scenario, most people won't consider much about the canopy parameters that may affect the harvest of an apple. As a result, it is crucial to analyze the previous data for obtaining valuable insights to increase the efficiency of a harvesting system and provide guidance for the growers to classify whether an apple is harvested or un-harvested using the canopy parameters. This paper consists of a detailed compilation of the latest research works on specific methodologies that are used to find the key parameters that may affect the harvest of an apple in a vertically trained "Sci-fresh" and V trellis grown "Envy" trees.**

***Index Terms*—Sci-fresh, Envy, Supervised Learning, Principal Component Analysis, K-Nearest Neighbours.**

ABBREVIATIONS

ANN	Artificial Neural Networks
BBD	Branch Basal Diameter
BED	Branch End Diameter
BL	Branch Length
BMD	Branch Middle Diameter
FD	Fruit Density
FL	Fruit Load
FLoc	Fruit Location
FSM	Fruit Single Mass
Ha	Harvested
KNN	K-Nearest Neighbours
PCA	Principal Component Analysis
SBD	Shoot Basal Diameter
SI	Shoot Index
SL	Shoot Length
SVM	Support Vector Machine
Un	Un-Harvested

I. INTRODUCTION

THE annual apple production in the United States has increased from 2.9 to 3.6 billion kilograms,(Figure 1) while the annual production value has increased from US$ 2.3 billion to US$ 3.2 billion over the last ten years[1]. At the same time, there were fewer workers due to a variety of factors, including the United States of America's restrictive border policies. As a result, an alternate solution for ensuring continuous production of apples in both local and international markets with fewer workers has been discovered. To address the shortage of seasonal workers, mass mechanical harvesting was invented. Previous research [2,3] indicated that this would produce better results with higher efficiency and lower cost.

Modern apple tree systems have played an important role in achieving desired results with orchard mechanization methodologies such as mechanical harvesting [4]. Among all the methods used to produce apples in the North Pacific region of the United States, vertical and inclined V-trellis systems, as shown in Figure 2, are widely used orchard mechanization [5]. These are high-density systems with 3000 to 4500 trees per hectare. Since they are highly denser systems the trees have closely packed canopies and high exposure to the sunlight [6]. However, the goal of this research is to find the key parameters that may affect the harvest of an apple.

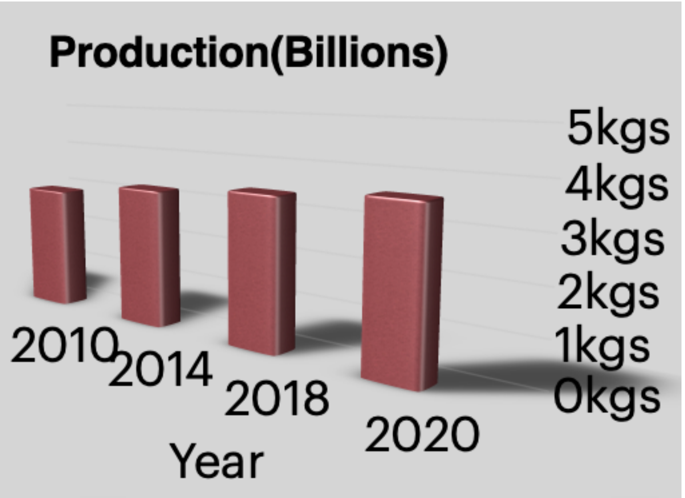


Fig. 1: Annual Production of apples in the last ten years

In most of the research and studies, the authors were concentrating only on the machine inputs but they did not bother much about the parameters that are affecting the growth of an apple- like branch diameter, interaction between the canopy and the machine etc[8,9]. Hence, to overcome this problem this study dives into the machine learning concepts like KNN classification, support vector machines, dimensionality reduction etc to define the key canopy(tree) parameters that may affect the harvest of an apple in a tree.

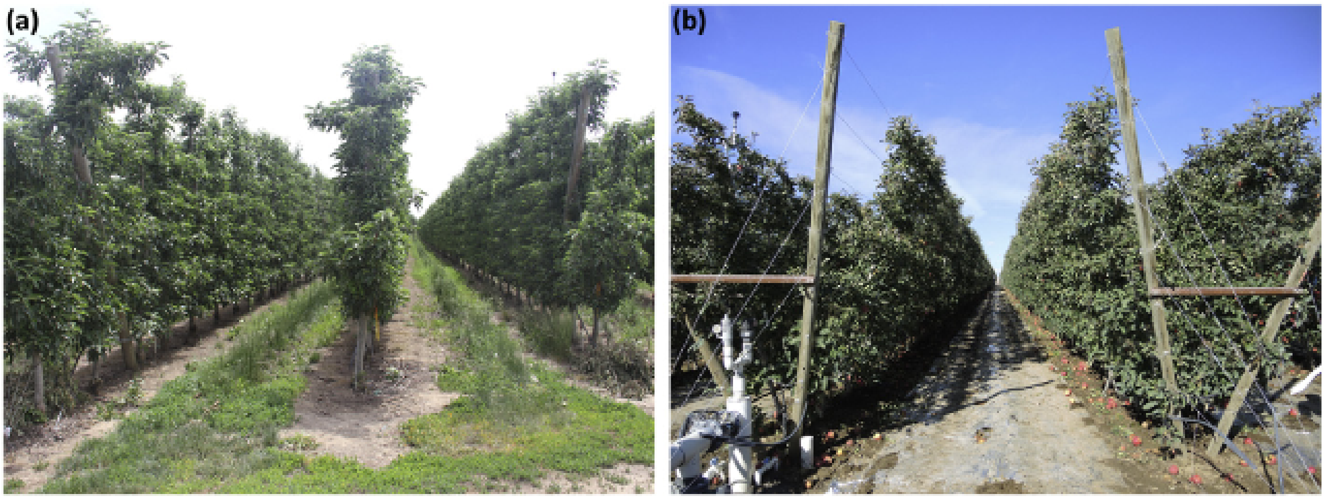


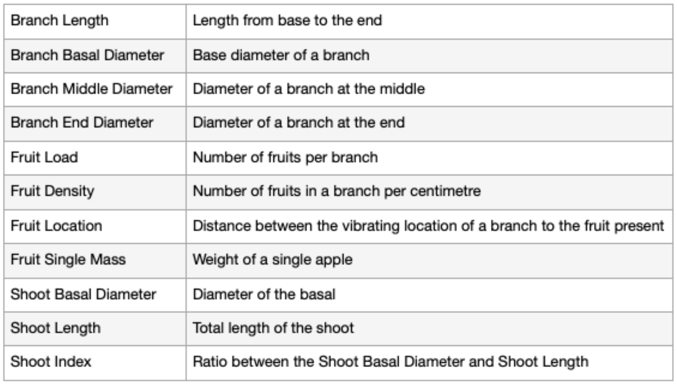
Fig. 2: (a) “Sci-fresh” and (b) “Envy” apple trees [9]

These algorithms are widely used for classification problems and this paper tells how these algorithms help us to classify an apple whether it is a mechanically harvested or mechanically unharvested apple based on the canopy parameters. It is also important to address the actual meaning of mechanically harvested and mechanically unharvested apple.

II. Canopy Parameters and Harvesting Trails

A. *Canopy Parameters*

The tree parameters can also be known as canopy parameters and various parameters are present for a particular tree, but this research considers only 11 parameters. The naming for both the orchard systems envy and sci-fresh are as follows(Figure 3):



Branch Length	Length from base to the end
Branch Basal Diameter	Base diameter of a branch
Branch Middle Diameter	Diameter of a branch at the middle
Branch End Diameter	Diameter of a branch at the end
Fruit Load	Number of fruits per branch
Fruit Density	Number of fruits in a branch per centimetre
Fruit Location	Distance between the vibrating location of a branch to the fruit present
Fruit Single Mass	Weight of a single apple
Shoot Basal Diameter	Diameter of the basal
Shoot Length	Total length of the shoot
Shoot Index	Ratio between the Shoot Basal Diameter and Shoot Length

Fig. 3: Definition of 11 Canopy Parameters

The optimal values for the above parameters are manually calculated in the field and shown in the Figure 4.

The apples were collected using an already developed vibratory machine [10], which is used to collect the fruits by shaking the tree at various locations for varying lengths of time. Shaking can be done for 2 to 5 seconds in the middle of the branches or at the base of the tree, which can be considered the trunk. Let us consider SB2, SB5, SM2, SM5, EB2, EM2, EM5 and EB5 where S denotes sci-fresh, E denotes Envy, M denotes Middle, B denotes Base and the numerical value represents the duration of vibration. The predictor variables, which are the input variables, were calculated for 2085 samples using the sci-fresh mechanism and 593 samples using the envy mechanism, with 2086 samples mechanically harvested and the rest mechanically un-harvested. 5).

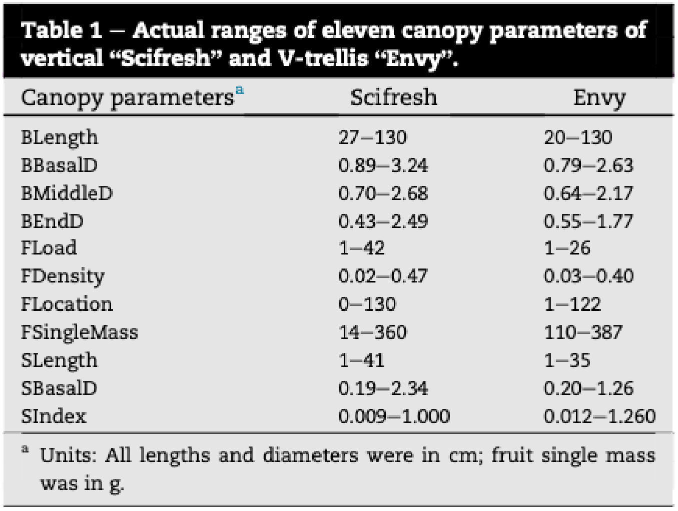


Table 1 – Actual ranges of eleven canopy parameters of vertical “Scifresh” and V-trellis “Envy”.

Canopy parameters[a]	Scifresh	Envy
BLength	27–130	20–130
BBasalD	0.89–3.24	0.79–2.63
BMiddleD	0.70–2.68	0.64–2.17
BEndD	0.43–2.49	0.55–1.77
FLoad	1–42	1–26
FDensity	0.02–0.47	0.03–0.40
FLocation	0–130	1–122
FSingleMass	14–360	110–387
SLength	1–41	1–35
SBasalD	0.19–2.34	0.20–1.26
SIndex	0.009–1.000	0.012–1.260

[a] Units: All lengths and diameters were in cm; fruit single mass was in g.

Fig. 4: Range values of 11 parameters for “sci-fresh” and “envy” [9]

B. Field characteristics and Methods

The data collection and validation were conducted in two apple orchards, one is vertically trained sci-fresh and the other one being V trellis architecture of Envy apple. This architecture is most commonly used in the united states due to its high productivity rate and high accessibility to the canopy parameters. Here the data is manually collected by the orchard workers during the harvesting seasons. 85% of the input data that is Sci-fresh(1772) and Envy(504) is used for training and validating the model whereas the remaining 15% of the input data that is Sci-fresh(313) and Envy(89) is used for testing the model.

[16, Heetal] developed a prototype shake and catch apple harvester which is used in conducting the field harvesting trails. The equipment consists of three major components a vibrating shaker, a four-wheel driving ground vehicle and a fruit catching machine. Vibration duration here is 2 seconds and 5 seconds and locations are branch base and middle portions(Figure 5).

Figure 6 tells the actual probability distribution function of the eleven canopy parameters that are manually calculated in the field in terms of Harvested and Un-harvested. From the figure some parameters like(Ex:FLoad) has showed a great difference in the distribution for harvested and un-harvested apples. Thereby, it is clear that the canopy parameter FLoad may have significant effect on the harvest that is harvested or un-harvested apple(ripe of un-ripe apple).

Whereas, some other parameters like(Ex:FLocation) is completely opposite to the above mentioned parameter because the probability density function for harvested and unharvested apples is getting overlapped. Hence it denotes that the canopy parameter FLocation has no impact on the output i.e harvest of an apple. Therefore to obtain the most relevant canopy parameters that may affect the harvest of an apple PCA algorithm i.e Principal Component Analysis is used and from the Figure 6 it is also clear that the canopy parameter SIndex is skewed towards one side that is left side so Logarithm can be applied to obtain a normal distribution for the canopy parameter.

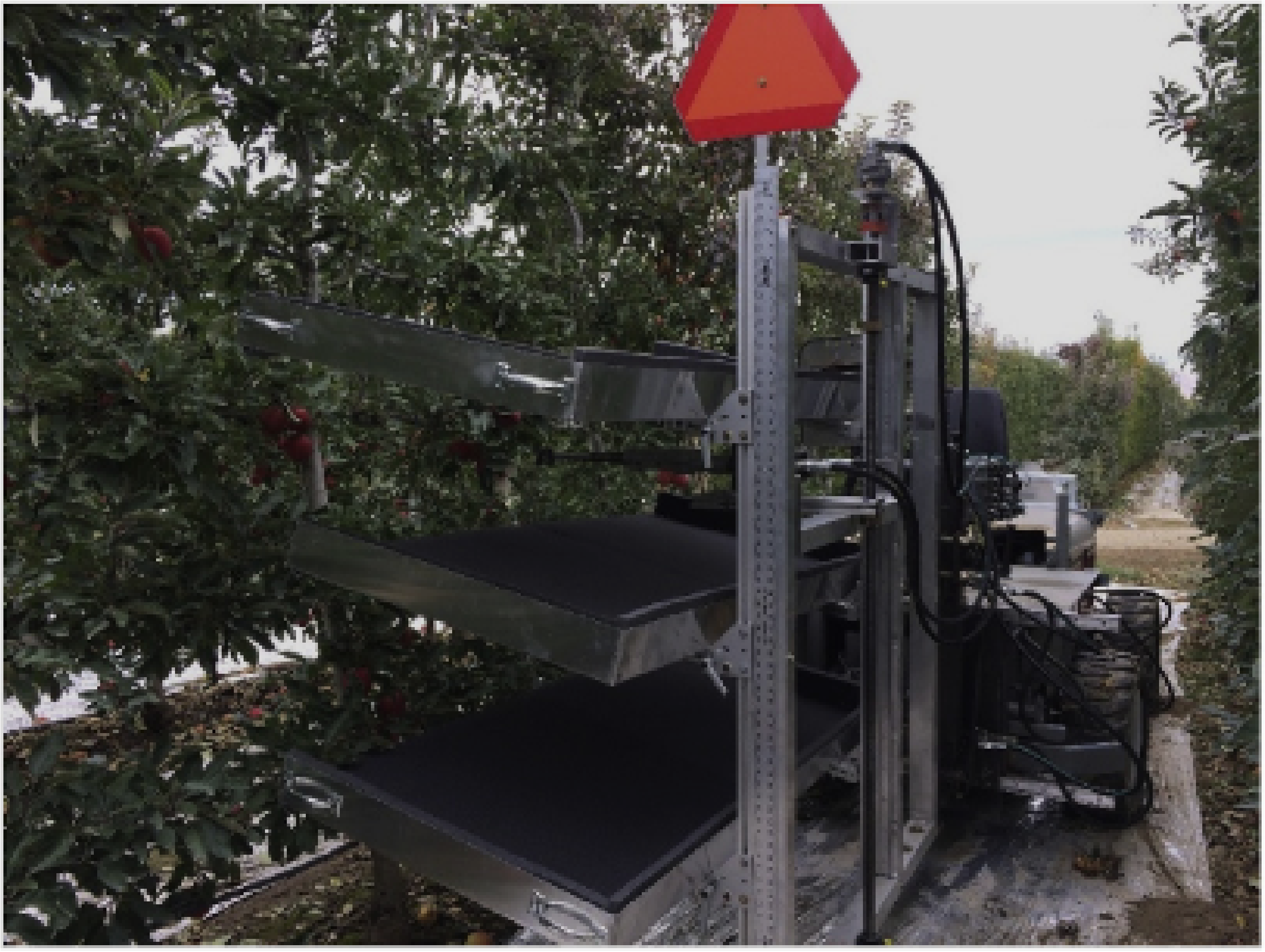


Fig. 5: Prototype of a Vibratory Harvester [16]

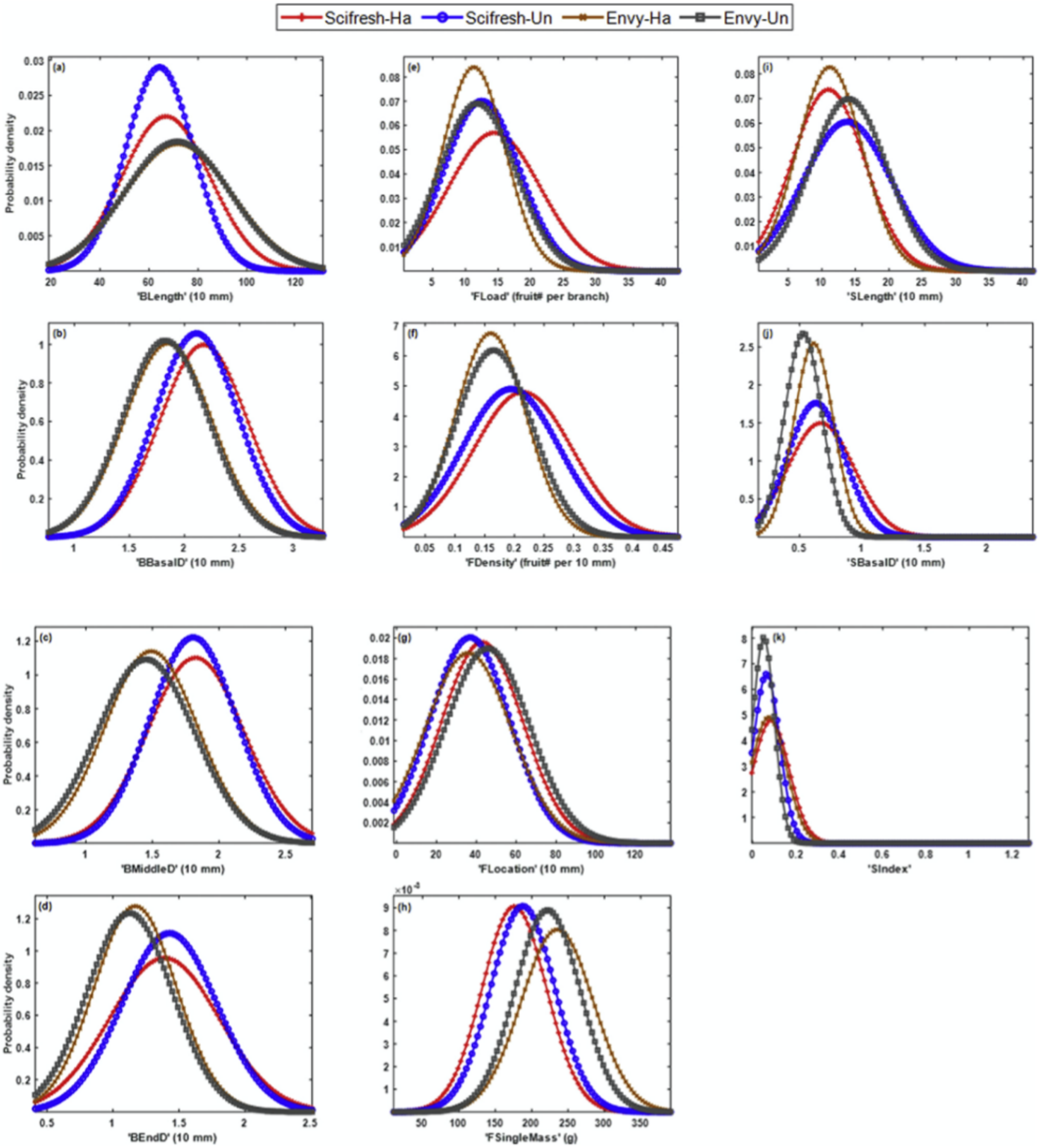


Fig. 6: Probability density of 11 canopy parameters [9]

III. Supervised Learning

A. Weighted KNN

The goal is to perform a binary classification on how the parameters may influence the output harvesting system.

Previous papers [8] mentioned that the authors were doing a binary classification on apples, but there is a lack of clarity on which classes the algorithm will classify and the overall flow of various steps that are used in developing a Supervised Machine Learning is given in the Figure 7. The algorithm will divide apples into two categories: ripe and unripe. All the previous journals [8,9] stated that they intend to classify mechanically harvested and mechanically unharvested apples, but it is pointless to classify an apple if it is manually plucked or done using automated machines because this does not provide us with any insights to provide valuable feedback to the growers.

Following a review of previous journals [8,9], it is clear that the apples will be classified as ripe or unripe. This problem's inputs and outputs were already known. After considering the various algorithms, the KNN algorithm was finalised due to its outstanding performance records in classifying the objects based on the nearest neighbours in different datasets and its predictions that the nearest objects have similar characteristics. Upon further consideration, a weighted KNN algorithm was finalised to classify the apples based on their mechanical harvesting. Weighted KNN is an enhanced version of the KNN algorithm.

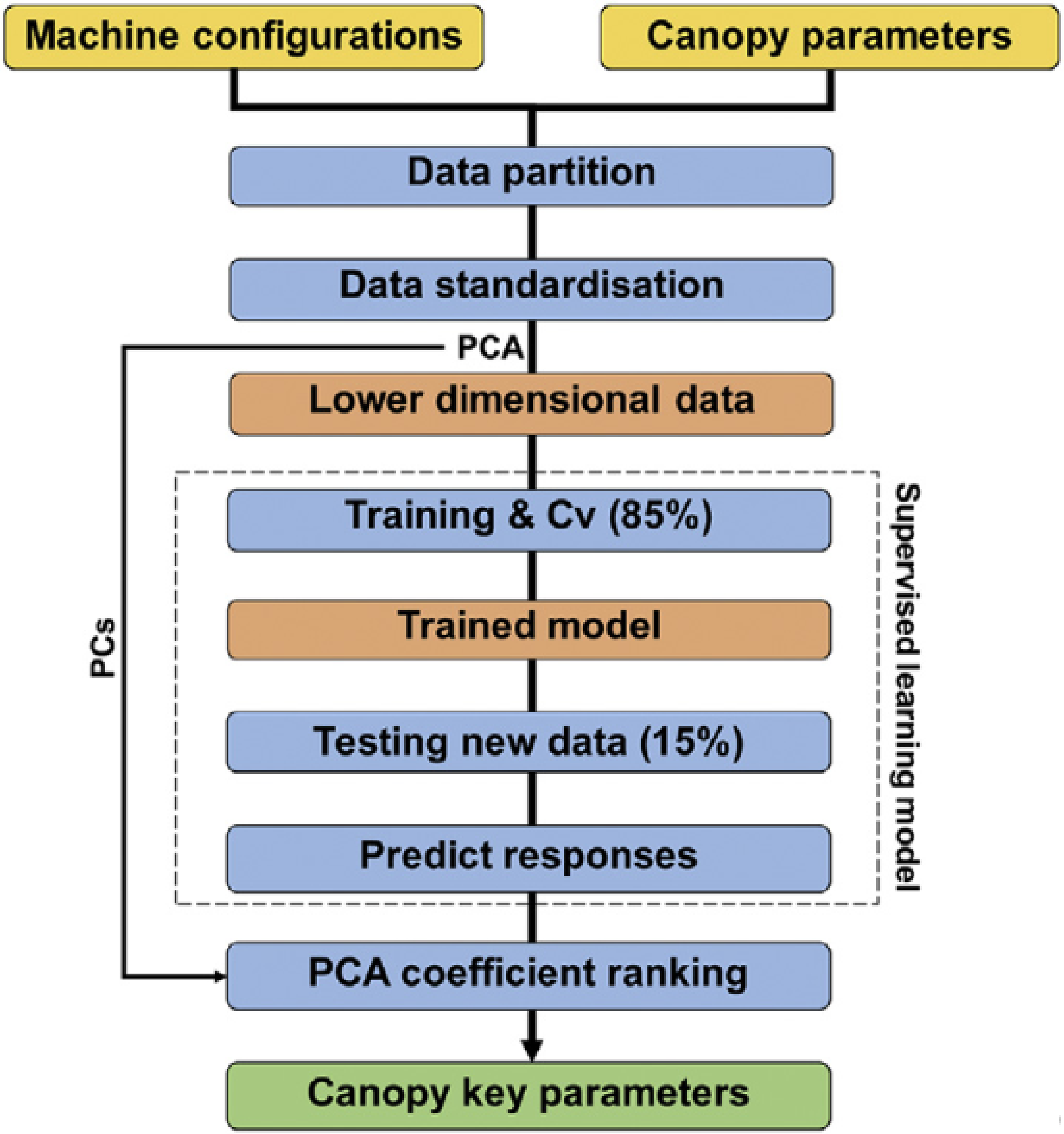


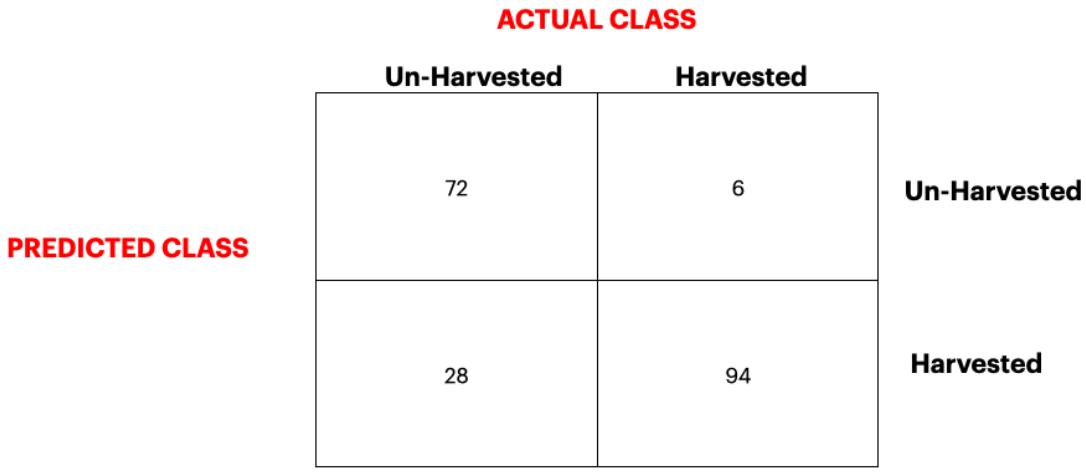
Fig. 7: Flowchart of a Supervised Machine Learning algorithm

The problem with KNN is with choosing the K value, if the K value is small, the algorithm is sensitive to the outliers, and if it is large it may include many points from other classes and the formulae for the weighted KNN algorithm is as follows.

$$y' = \underset{v}{\operatorname{argmax}} \sum_{(\mathbf{x}_i, y_i) \in D_z} w_i \times I(v = y_i)$$

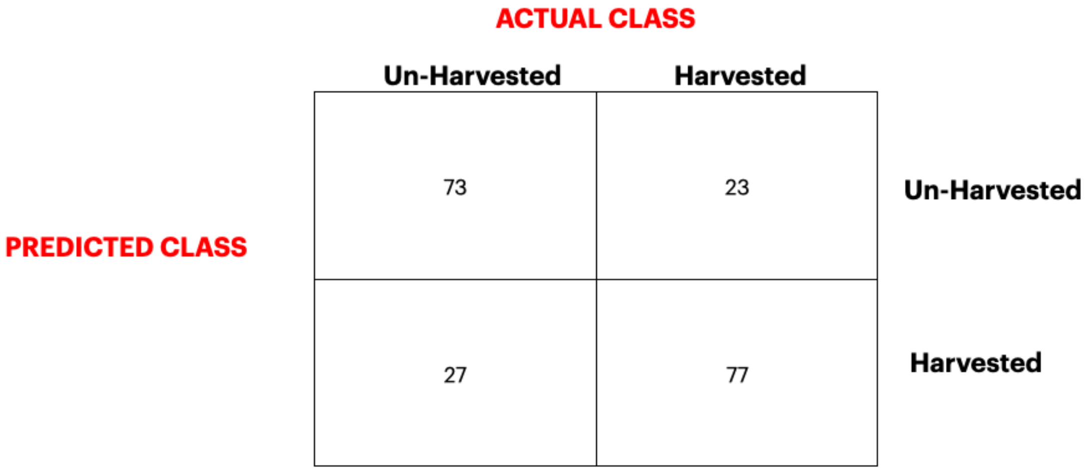
As a result, supervised learning [11] is implemented, and the weighted KNN algorithm is used to classify the inputs. This algorithm was chosen because of its outstanding performance and accuracy in classifying the input data points using k nearest neighbors [7].

The weighted KNN is used because it assigns weights to the nearest data points and classifies the new data point based on distance weighted voting. To accomplish this, the distance between the new data point and all other data points is calculated, and then which data points weigh more and are more closely related are used for weighted voting. Initially, all of the input parameters are normalized because each canopy parameter has a different metric measure, which may result in biased classification, so it is critical to normalize all of the parameters before proceeding [11]. Then 85% of the input data is used for training the algorithm out of which a part of data is used for validating the algorithm that is to tune the hyperparameters that are present in the algorithm. The validating data set is never used for training the algorithm and the remaining 15% of the input data is used for testing the performance of the model. The confusion matrix for the Sci-Fresh and Envy is given in the Figure 8 and 9.



		ACTUAL CLASS	
		Un-Harvested	Harvested
PREDICTED CLASS	Un-Harvested	72	6
	Harvested	28	94

Fig. 8: Normalised Confusion matrix(%) of Sci-fresh



		ACTUAL CLASS	
		Un-Harvested	Harvested
PREDICTED CLASS	Un-Harvested	73	23
	Harvested	27	77

Fig. 9: Normalised Confusion matrix(%) of Envy

The PCA algorithm is then applied to the input variables in order to decrease the number of predictor variables. The first principal component explains 80% of the variance [12] in our input data, and it is followed by the other principal components. As a result, the principal components that explain the majority of our input data will be considered for evaluation.

[13, Liu] investigated on Bayesian optimisation algorithm to estimate the optimal hyper parameter values like k value, weighted distance, distance over thirty distance metrics and finally the results show better accuracy by using “cityblock” as distance metric with “k=1” and the weighted distance can be calculated using the “squared-inverse” method.

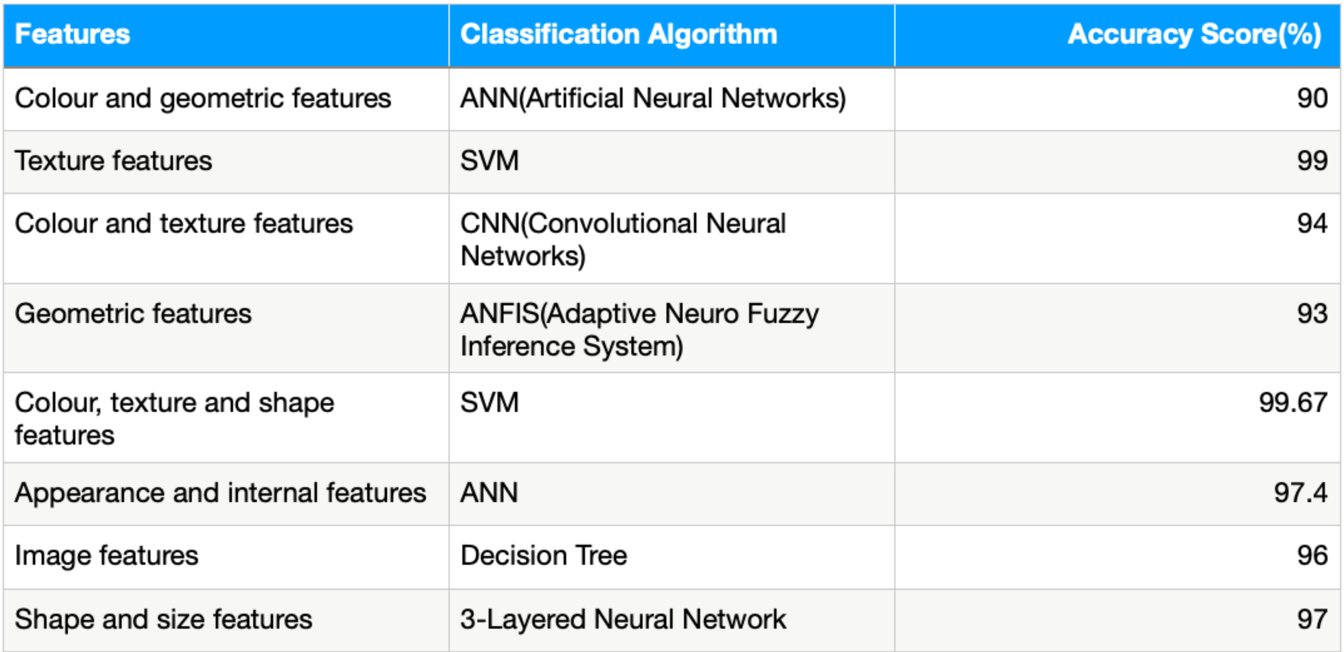
B. Model Testing and Evaluation

The model is initially trained using 5 fold cross-validation, which divides the input data into 5 equal subsets. In order to train the algorithm with different sets of training sets and determine the best hyperparameter values. The model is then evaluated using various performance measures such as accuracy score, ROC curve, and confusion matrix. The accuracy score formula is (TP + TN) / (TP + TN + FP +FN), where TP stands for True Positive, TN stands for True Negative, FP stands for False Positive, and FN stands for False Negative. Where TP represents “mechanically harvested” and TN represents “mechanically unharvested”. The Area under Curve in ROC tells the goodness of the model. The AUC value lies between 0 and 1 as the value of AUC increases the goodness of the model also increases.

After completing the training and cross-validation for weighted KNN model, the remaining 15% dataset is used to validate the performance of the model, which predicts responses to the inputs. The accuracy for the sci-fresh is between the range 81-90.7%, however, a larger data set could help in improving the prediction accuracies. The data size can be increased by combining the data from different harvesting conditions. However, the selected canopy parameters are dependent on the harvesting configuration.

Additionally a larger dataset could be used to improve the accuracy of both training-cv(cross-validation) and testing dataset. This can be done by combining the dataset that is available from different harvest treatments. However this may also decrease the accuracy score of the model because there are canopy parameters which doesnot have much impact on the output that is for classifying an apple whether it is a ripe or unripe. Hence, it is very important to select the canopy parameters that actually impact the harvest of an apple to get high accuracy score. Finally it is clear that weighted KNN algorithm can produce acceptable accuracy score for the given input values with or without using the Principal Component Analysis.

[14, Rath] conducted a similar study on classifying fruits using machine learning and image analysis by implementing different algorithms such as SVM, Decision Tree, ANN, and so on, and discovered that SVM and ANN produce high accuracy scores of around 99 percent and 97 percent, respectively. The accuracy scores of different algorithms is depicted in the Figure 10.



Features	Classification Algorithm	Accuracy Score(%)
Colour and geometric features	ANN(Artificial Neural Networks)	90
Texture features	SVM	99
Colour and texture features	CNN(Convolutional Neural Networks)	94
Geometric features	ANFIS(Adaptive Neuro Fuzzy Inference System)	93
Colour, texture and shape features	SVM	99.67
Appearance and internal features	ANN	97.4
Image features	Decision Tree	96
Shape and size features	3-Layered Neural Network	97

Fig. 10: Comparative analysis of different techniques applied for classification of Apples

As a result, rather than using the 11 canopy parameters as inputs to the algorithm, the images of the apples [15] can be used to determine whether the apple is ripe or unripe. This can be accomplished by building a neural network based on color, texture, shape, and features. Figure 11 depicts images for various harvest levels of an apple.

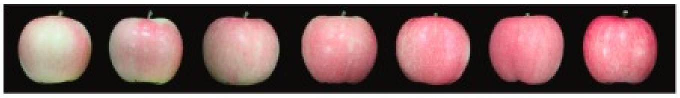
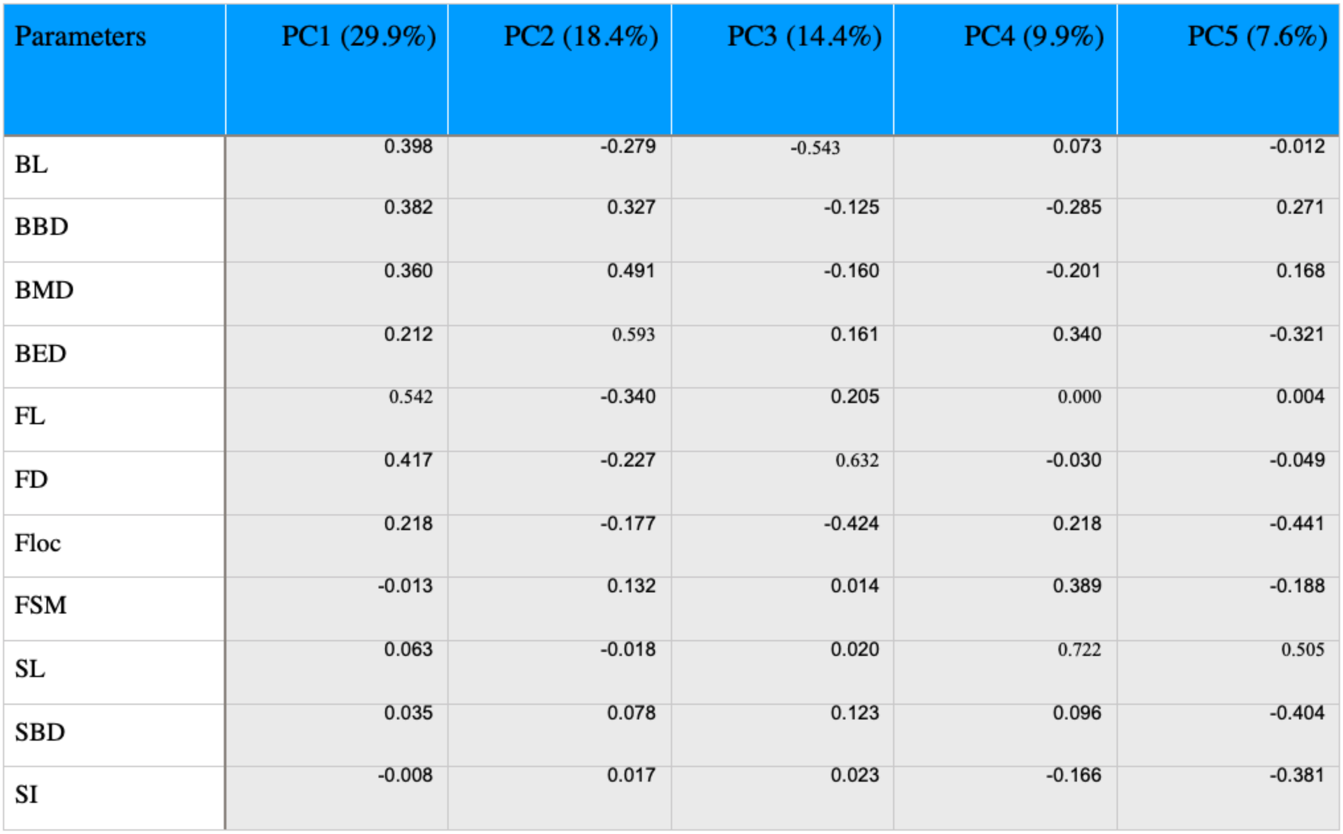


Fig. 11: Apples at different Harvest stages [15]

[8, Zhang] used a testing set and a training set to evaluate the algorithm. Overall, the model achieved an accuracy score of (85.9±0.2%) for "sci-fresh" and a score of (68.5±0.5%) for "envy" on the testing set. The AUC curve yielded similar results, with values ranging between 0.75-0.82 for "sci-fresh" and 0.66-0.83 for "envy." The principal component analysis is also used to reduce the dimensionality of the input features, but the results are nearly identical, with only a 1% difference in the accuracy score. As a result, the PCA may not have a large effect on the accuracy score and can be ignored to save the algorithm's computation time. [8, Zhang] examined the probability distribution function for all 11 canopy parameters, and their findings shows that the Fruit Load, Branch Base Diameter, and Shoot Length all have a significant impact on fruit removal (mechanically harvested or mechanically un-harvested).

[17, Tonguc] conducted a similar study on classifying fruits. The much-needed exercise is to differentiate the variety of fruits is to classify them. In most cases, fruits that come from the same family differ in the sense of size and colour only. In most cases, fruits that come from the same family differ in the sense of size and colour only. If the customer requires a particular variety of mango for a cause, there is no existing tool that helps them. In such cases, automated variety recognition helps in identifying the variety of the fruits. Automatic grading can be done by using image processing techniques and digital cameras. This experiment involves two parts, namely a protective box and a conveyor belt. By using the C# programming techniques, the software is created to determine the size and colour of the apples. This particular software takes around 1.3 seconds to inspect the apple image and the accuracy was found to be around 80%. For the Classification process, support vector machines (SVM) is used and the accuracy of detection was 100%. In recent years, image processing technique is expanded into the fruit industry due to their consistency and objective inspection technique. This technique in the fruit industry has been increasing during recent years.



Parameters	PC1 (29.9%)	PC2 (18.4%)	PC3 (14.4%)	PC4 (9.9%)	PC5 (7.6%)
BL	0.398	-0.279	-0.543	0.073	-0.012
BBD	0.382	0.327	-0.125	-0.285	0.271
BMD	0.360	0.491	-0.160	-0.201	0.168
BED	0.212	0.593	0.161	0.340	-0.321
FL	0.542	-0.340	0.205	0.000	0.004
FD	0.417	-0.227	0.632	-0.030	-0.049
Floc	0.218	-0.177	-0.424	0.218	-0.441
FSM	-0.013	0.132	0.014	0.389	-0.188
SL	0.063	-0.018	0.020	0.722	0.505
SBD	0.035	0.078	0.123	0.096	-0.404
SI	-0.008	0.017	0.023	-0.166	-0.381

Fig. 12: Coefficients of the first 5 PC's (PC1-PC5) for "Sci-fresh" with 11 canopy parameters

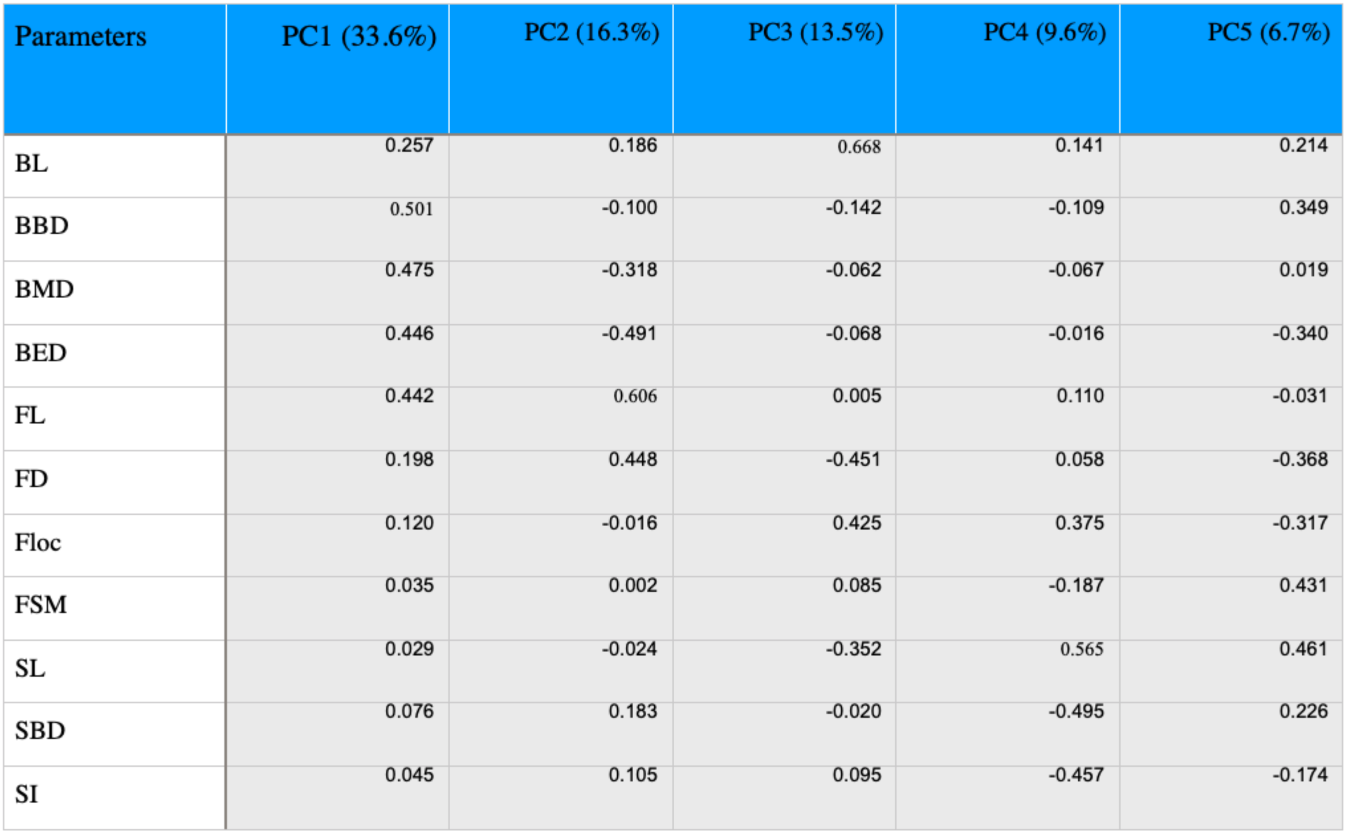
The grading of fruit involves its category, severity of the disease and contamination of the fruits. If the grading is done manually, it is a huge time taking process. It is important to adopt an automated grading system in this regard. An automatic image processing system is a reliable method for grading and sorting fruits.

IV. Unsupervised Learning

A. Principal Component Analysis

The main advantage of principal components is to reduce the dimensionality in the data set and to examine the co-efficient of PC(Principal Component) for efficient learning. The co-efficient values of the PC1 to PC5 are presented in the table4. The parameter represents the large co-efficient value in each of the columns which were correlated by the PC. The co-efficient value above 0.5 is considered as most relevant which is presented in bold type for each column. For example, in the PC1 column for Sci-fresh type, the Fruit load value is 0.542, so it is considered as a highly relevant canopy parameter among all other parameters in mechanical harvesting. [8] also indicates that the branches with higher fruit load were easier to be mechanically harvested and they are more suitable for vibratory mechanical harvesting. For PC2 and PC3 “BEndD”, “Blength” and “FDensity” are highly suitable canopy parameters because of their co-efficient value greater than 0.5. Previous results also suggest that shoot length influences the fruit removal results in vibratory mechanical harvesting. The key canopy parameters for Envy includes “BBasalD”, “FLoad”, “BLength” and “SLength” and the coefficient of first five Principal components for ”Sci-fresh” and ”Envy” are given in the Figure 12 and 13 respectively.

By aggregating all the results, the key canopy parameters include “FDensity” and “FLoad” in the fruit category, “BEndD” and “BBasalD” in the branch category and “SBasalD” and “SLength” in the shoot category. After performing the one-way Analysis of Variance, it shows that the key canopy parameters of Sci-fresh apples are showing significant differences between the mechanical harvested and un-harvested apples. Some External Parameters like the orchard trellising system and harvesting year could also possibly influence the results which were not discussed in this paper.



Parameters	PC1 (33.6%)	PC2 (16.3%)	PC3 (13.5%)	PC4 (9.6%)	PC5 (6.7%)
BL	0.257	0.186	0.668	0.141	0.214
BBD	0.501	-0.100	-0.142	-0.109	0.349
BMD	0.475	-0.318	-0.062	-0.067	0.019
BED	0.446	-0.491	-0.068	-0.016	-0.340
FL	0.442	0.606	0.005	0.110	-0.031
FD	0.198	0.448	-0.451	0.058	-0.368
Floc	0.120	-0.016	0.425	0.375	-0.317
FSM	0.035	0.002	0.085	-0.187	0.431
SL	0.029	-0.024	-0.352	0.565	0.461
SBD	0.076	0.183	-0.020	-0.495	0.226
SI	0.045	0.105	0.095	-0.457	-0.174

Fig. 13: Coefficients of the first 5 PC’s (PC1-PC5) for “Envy” with 11 canopy parameters

V. Conclusion and Future Work

This article has presented different methods that are present in Supervised Learning for doing binary classification on harvested and unharvested fruits and discussed in detail the concepts that are available in classification. Moreover, this paper also explained the lack of clarity in the previously published papers about the classes that the algorithm will classify. In the near future, the goal should be only on the canopy parameters that are actually impacting the output variable like Fruit Load, Branch Base Diameter and Shoot length. The advanced algorithms like ANN, SVM and minimum redundancy maximum relevance has to be considered for feature selection instead of Principal component analysis for achieving better accuracy scores.

References

[1] USDA 2020. National agricultural statistics database. Washington, DC: USDA National Agricultural Statistics Service.*Retrieved from https://quickstats.nass.usda.gov.*

[2] Peterson, D. L., Whiting, M. D., & Wolford, S. D. (2003). Fresh market quality tree fruit harvester: Part I. Sweet cherry. Applied Engineering in Agriculture, 19(5), 539 - 543.

[3] Karkee, M., Silwal, A., & Davidson, J. R. (2018). Chapter 10: Mechanical harvest and infield handling of tree fruit crops. In Q. Zhang (Ed.), Automation in tree fruit production: Principles and practice (pp. 179-233). Wallingford, UK: CABI.

[4] Whiting, M. D. (2018). Chapter 6: Precision orchard systems. In Q. Zhang (Ed.), Automation in tree fruit production: Principles and practice (pp. 93-111). Wallingford, UK: CABI.

[5] Zhang, Q., & Pierce, F. (2013). Agricultural automation fundamentals and practices.

[6] Stephan, J., Sinoquet, H., Dones, N., Haddad, N., Talhouk, S., & Lauri, P.E. (2008). Light interception and partitioning between shoots in apple cultivars influenced by training. Tree Physiology, 28(3), 331-342.

[7] Kurtulmus, F., Lee, W. S., & Vardar, A. (2014). Immature peach detection in colour images acquired in natural illumination conditions using statistical classifiers and neural network. Precision Agriculture, 15(1), 57-79.

[8] Zhang, X., He, L., Zhang, J., Whiting, M., Karkee, M., & Zhang, Q. (2020). Determination of key canopy parameters for mass mechanical apple harvesting using supervised machine learning and principal component analysis (PCA). Biosystems Engineering, 193, 247–263. https://doi.org/10.1016/j.biosystemseng.2020.03.006.

[9] Zhang, J., Whiting, M., Karkee, M. New Machine Learning Findings Reported from Washington State University [Determination of Key Canopy Parameters for Mass Mechanical Apple Harvesting Using Supervised Machine Learning and Principal Component Analysis]. (2020). In Journal of Engineering(p. 2124–). NewsRX LLC.

[10] He, L., Fu, H., Karkee, M., & Zhang, Q. (2017). Effect of fruit location on apple detachment with mechanical shaking. Biosystems Engineering, 157, 63-71.

[11] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[12] Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. Chemometrics and Intelligent Laboratory Systems, 2(1-3), 37-52.

[13] Liu, W., & Chawla, S. (2011). Class confidence weighted knn algorithms for imbalanced data sets. In Pacific-asia conference on knowledge discovery and data mining (pp. 345-356). Berlin, Germany: Springer.

[14] Behera, S.K., Rath, A.K., Mahapatra, A. et al. Identification, classification & grading of fruits using machine learning & computer intelligence: a review. J Ambient Intell Human Computer(2020). https://doi.org/10.1007/s12652-020-01865-8.

[15] Mengsheng Zhang, Bo Zhang, Hao Li, Maosheng Shen, Shijie Tian, Haihui Zhang, Xiaolin Ren, Libo Xing, Juan Zhao, a machine learning algorithm, Infrared Physics & Technology,Volume 111, 2020. https://doi.org/10.1016/j.infrared.2020.103529.

[16] He, L., Zhang, X., Karkee, M., & Zhang, Q. (2018). Fruit accessibility for mechanical harvesting of fresh market apples. ASABE Paper No. 1801007. St. Joseph, MI: ASABE.

[17] Tonguc G, Yakut AK (2009) Fruit grading using digital image processing techniques. Tarım Makinaları Bilimi Dergisi 5(1):93–101 Unser M (1986) Sum and difference histograms for texture classification. IEEE Trans Pattern Anal Mach Intell 1:118–125 https://doi.org/10.1109/TPAMI.1986.4767760doi.