


**Please cite the Published Version**

Ren, Aifeng, Zahid, Adnan, Zoha, Ahmed, Shah, Syed Aziz , Imran, Muhammad Ali, Alomainy, Akram and Abbasi, Qammer H (2020) Machine Learning Driven Approach Towards the Quality Assessment of Fresh Fruits Using Non-invasive Sensing. IEEE Sensors Journal, 20 (4). pp. 2075-2083. ISSN 1530-437X

**DOI:** <https://doi.org/10.1109/jsen.2019.2949528>

**Publisher:** Institute of Electrical and Electronics Engineers (IEEE)

**Version:** Accepted Version

**Downloaded from:** <https://e-space.mmu.ac.uk/624442/>

**Usage rights:**  In Copyright

**Additional Information:** This is an Author Accepted Manuscript of an article published in IEEE Sensors Journal.

**Enquiries:**

If you have questions about this document, contact [openresearch@mmu.ac.uk](mailto:openresearch@mmu.ac.uk). Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from <https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines>)

# Machine Learning Driven Approach Towards the Quality Assessment of Fresh Fruits Using Non-invasive Sensing

Aifeng Ren, Adnan Zahid, Ahmed Zoha, Syed Aziz Shah, Muhammad Ali Imran, *Senior Member, IEEE*, Akram Alomainy, *Senior Member, IEEE*, and Qammer H. Abbasi, *Senior Member, IEEE*

**Abstract**—In agriculture science, accurate information of moisture content (MC) in fruits and vegetables in an automated fashion can be vital for astute quality and grading evaluation. This demands for a viable, feasible and cost-effective technique for the defect recognition using timely detection of MC in fruits and vegetables to maintain a healthy sensory characteristic of fruits. Here we propose a non-invasive machine learning (ML) driven technique to monitor variations of MC in fruits using the terahertz (THz) waves with Swisstol12 material characterization kit (MCK) in the frequency range of 0.75 THz to 1.1 THz. In this regard, multi-domain features are extracted from time-, frequency-, and time-frequency domains, and applied three ML algorithms such as support vector machine (SVM), k-nearest neighbour (KNN) and Decision Tree (D-Tree) for the precise assessment of MC in both apple and mango slices. The results illustrated that the performance of SVM exceeded other classifiers results using 10-fold validation and leave-one-observation-out-cross-validation techniques. Moreover, all three classifiers exhibited 100% accuracy for day 1 and 4 with 80% MC value (freshness) and 2% MC value (staleness) of both fruits' slices, respectively. Similarly, for day 2 and 3, an accuracy of 95% was achieved with intermediate MC values in both fruits' slices. This study will pave a new direction for the real-time quality evaluation of fruits in a non-invasive manner by incorporating ML with THz sensing at a cellular level. It also has a strong potential to optimize economic benefits by the timely detection of fruits quality in an automated fashion.

**Index Terms**—Classification, Fruits, Machine Learning, Moisture Content, Terahertz Sensing.

## I. INTRODUCTION

**F**RUITS and vegetables comprise an essential part of the human diet as they are the primary source of dietary nutrients [1]. In recent years, the rising demand for fruits quality evaluation and sensory characteristics have posed significant challenges in the agriculture sector. Although, manual sorting and grading can be done for the quality assessments and freshness of fruits detection, still this method is significantly fickle, inconsistent, and tedious [1]. In this situation, the identification of any microbially contaminated fruits is quite challenging and may cause severe threats to human health by causing

numerous diseases [1]. Therefore, this vital issue strongly demands an intervention of innovative, viable and feasible technological solution that can closely and accurately monitor health status and the MC of fruits to ensure its freshness and quality [2]. To address this uncertainty, many significant contributions and techniques on devising non-destructive have been suggested to accurately determine the MC in fruits [2][3]. These techniques include magnetic resonance imaging (MRI), near-infrared spectroscopy (NIRS), hyper-spectral imaging, have been investigated and widely deployed to perform the qualitative analysis and composites level of fruits [2].

Regrettably, these techniques have mainly focused on determining the soluble solid content, acidity and physical attributes of fruits. Researchers, scientists and biotechnologists are of a strong interpretation that these aforementioned factors cannot be considered as sole quality parameters for fruits freshness [4]. Also, there are other limitations to these methods, including low resolution and low sensitivity to detect any variations at a cellular level in fruits [5]. To overcome these challenges, terahertz time-domain spectroscopy (THz-TDS) was introduced to acquire detailed information at a cellular level due to its high absorption ability, sensitivity and resolution as depicted in Fig. 1. Due to these characteristics, it can identify the early inception of nutrients contamination in fruits [6][7]. This technique is deemed to have a promising result to detect diminutive variations of MC in fruit slices compared to previous methods. However, it is costly and not portable [8].

Hence, aforesaid prevailing challenges in monitoring the internal morphology and biological complex traits in fruits at

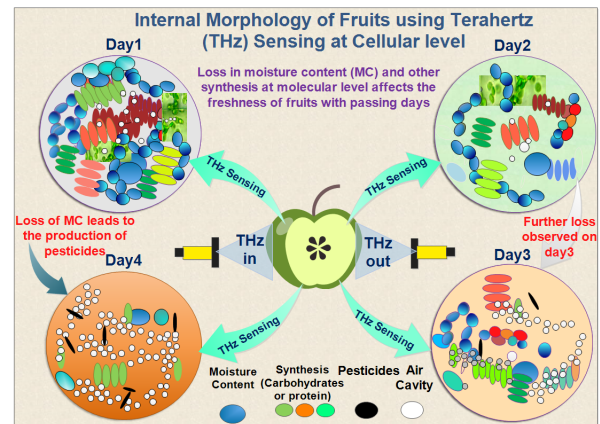


Fig. 1: Internal Structure of the Fruits for Four Days Using the THz Sensing at Cellular Level.

Aifeng Ren is with School of Electronic Engineering, Xidian University, Xi'an, Shaanxi, China (e-mail: afren@mail.xidian.edu.cn).

Adnan Zahid, Ahmed Zoha, Muhammad Imran, Qammer H. Abbasi are with School of Engineering, University of Glasgow, Glasgow, U.K. (e-mail: a.zahid.1@research.gla.ac.uk, Ahmed.Zoha, Muhammad.Imran & Qammer.Abbasi@glasgow.ac.uk).

Syed Aziz Shah is with School of Computing and Mathematics, Manchester Metropolitan University, U.K. (e-mail: azizshahics@yahoo.com)

Akram Alomainy is with School of Electronic Engineering and Computer Science, Queen Mary University of London, London, U.K. (e-mail: a.alomainy@qmul.ac.uk).

cellular level as shown in Fig. 1 enthralled researchers from diverse disciplines. Therefore, consciousness of quality control for fruits is seen as a potentially important application area for THz systems provided reliable machine learning (ML) techniques can be integrated with THz sensing equipment. Researchers from multidisciplinary studies emphasize that ML has been effectively applied in various disciplines and due to its high computing performance, it creates novel opportunities to fully comprehend the intensive data processes in numerous fields [6]. Some of the significant contributions achieved by applying ML in various disciplines are food security, meteorology, medicine, economic sciences etc. However, researchers are of a strong view that its potential in the discipline of quality assessment of fruits is still one of the least explored research areas until now [6].

In this paper, we report a novel, and non-invasive technique to closely monitor and foresee the future trends of MC in fresh fruits slices of apple and mango at molecular level by applying the THz waves in the frequency range of 0.75 THz to 1.1 THz [9]. For this purpose, we have performed in-lab experiments utilizing the fresh fruits slices of apple and mango and carefully observed the MC in fruits by using scattering parameters of THz waves. The observed data is further pre-processed and put into proposed ML algorithm. Thus, the integration of ML and THz have demonstrated the strong potential of evaluating the freshness of fruits in an automated fashion, which in turn, can help in reducing the health and purification expenses, and optimize economic benefits by maintaining the nutrients level and MC in fruits. The paper is organized as follows: Section II presents the data-collection and pre-processing methodology. Section III describes the feature extraction. This is followed in section IV by the feature selection technique. In Section V, performance of three classifiers are discussed. Finally, conclusion and future work are highlighted in Section VI.

## II. DATA COLLECTION AND PRE-PROCESSING METHODOLOGY

### A. Experimental Method

In this system, we employed a THz swissto12 MCK which was connected to Keysight Technologies Vector Network Analyzer (VNA) N5224A and with extender waveguide WM-250 (WR-1.0), enabling measurements to be performed in the terahertz frequency range of 0.75 THz to 1.1 THz as shown in Fig. 2. The MCK comprised of two circular waveguides with further two low-loss corrugated waveguide. The movable part of the MCK enabled the material under test (MUT) to accommodate the thickness of  $40\mu\text{m}$  to 4mm [10]. The scattering parameters (S-parameters) including reflection and transmission response ( $S_{11}$  and  $S_{21}$ ) was obtained after performing a fully two-port Short-Open-Load-Thru (SOLT) calibration, aiming to lessen any undesired noise while performing the measurements.

### B. Samples Preparation

In this study, two fresh fruits samples including mango and red apple were bought from Morrisons supermarket in

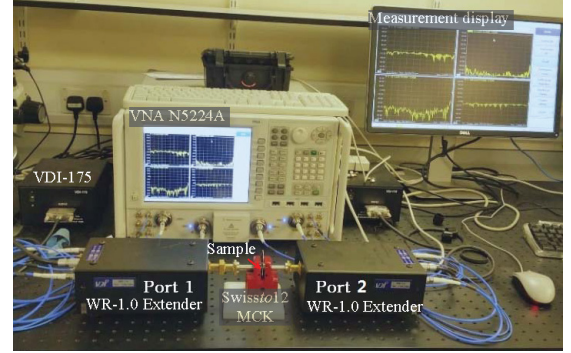


Fig. 2: Experimental set up of system for measuring the transmission response ( $S_{21}$ ) of apple and fruit slices.

Glasgow. Subsequently, two slices of each fruits were taken as samples with the average thickness between 2.3mm to 3.2mm. These slices were preserved carefully under the ambient temperature of  $18^\circ\text{C} \pm 0.1^\circ\text{C}$  and humidity of  $25\% \pm 2\%$ . The thickness of each slice was measured using a Vernier caliper at three different locations to ensure the evenness all over the whole slice and satisfy the threshold range of MCK.

### C. Data Collection

Before collecting observations, the weight of each slice was measured using a digital scale with an accuracy of 0.1mg to monitor the variation of MC in both apple and mango slices. The weight of each sample was converted into MC value with the passing time as follows [11].

$$MC_{(k)}^{(ij)} = \frac{M_{(k)}^{(ij)} - M_{(k)}^{(dry)}}{M_{(k)}^{(ij)}} \times 100\% \quad (1)$$

Where  $MC_{(k)}^{(ij)}$  and  $M_{(k)}^{(ij)}$  indicate the MC value and the mass of the  $k$ -th sample at the  $j$ -th measurement in the  $i$ -th day, respectively.  $M_{(k)}^{(dry)}$  denotes the mass of the  $k$ -th sample that dried out to the evaporation of MC over the course of four days. It was aimed to observe the various degree of MC in both samples at maximum location on slices. Hence, both samples were measured at four various locations with four different orientations. At each location, ten readings were recorded, and a total of 160 observations were collected for mango and apple slices on each day, respectively. This whole process was repeated for four consecutive days to monitor the variations of the freshness of fruit slices. At the end, there were 8 target data sets collected from VNA,  $S_{(k)}^{(n)}(f_i)$ , for two different slices ( $k=1,2$  numbered for apple slice and mango slice, respectively) in four consecutive days ( $n=1,2,3,4$ ) with frequency  $f_i$  ( $i=0:200$ ) in the range of 0.75 TH to 1.1 THz. Fig. 3 showed the mean transmission responses ( $S_{21}$ ) of the same day observations for apple and mango slices in four consecutive days,  $S_{(1)}^{(1)}(f)$ ,  $S_{(1)}^{(2)}(f)$ ,  $S_{(1)}^{(3)}(f)$ , and  $S_{(1)}^{(4)}(f)$ , respectively.

In Fig. 3(a) and (b), it can be observed that two different fruits slices including apple and mango showed distinctive transmission responses in the THz region from day 1 to 4. The difference in transmission response also revealed the presence of different degree of composites in the fruit, such as proteins,

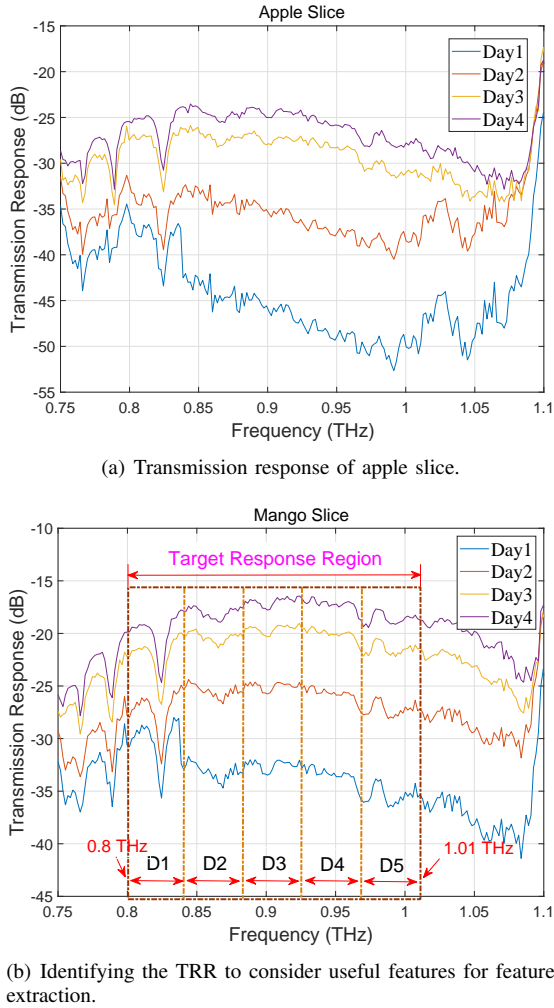


Fig. 3: Shows the process of obtaining TRR area to identify useful features for performing the classification process.

sugar and fats, causing the different absorption of the THz radiation. Upon a close analysis of Fig. 3, it was noticed that on day 1, transmission response was substantially low due to the presence of the high percentage of MC in fruits.

#### D. Data Pre-processing

Prior to any data analysis, it was essential to pre-process the observations to reduce the effects of the any unwanted background noise caused by the THz system hardware. This undesired noise would have produced false observations of MC in samples and hence affected the overall classification accuracy [12]. The simple method was to calculate the mean observations to reduce the random noise. However, the efficient technique was to consider wavelet-based method because decomposition in the wavelet domain could provide better de-noising performance for the ultrafast pulses of THz spectrum [13]. In our work, since observations obtained from VNA were in the frequency domain, so, these observations were converted to the time domain by applying Inverse Fast Fourier Transform (IFFT).

Subsequently, a three-level wavelet decomposition was adopted to pre-process the time-domain signals of all observations with the “heuristic” and soft threshold methods.

After de-noising pre-processing, the time-domain signals were arranged into a data matrix with the dimension of  $M \times N$  as (2) for time domain feature extraction. And the de-noising time-domain signals were performed Fast Fourier Transform (FFT) to obtain the frequency domain response and organized into  $M \times N$  matrix as (3) for frequency domain feature extraction.

$$s_{(k)}(n) = [s_{(k)}^{(1)}(n); s_{(k)}^{(2)}(n); s_{(k)}^{(3)}(n); s_{(k)}^{(4)}(n)] \quad (2)$$

$$S_{(k)}(f_i) = [S_{(k)}^{(1)}(f_i); S_{(k)}^{(2)}(f_i); S_{(k)}^{(3)}(f_i); S_{(k)}^{(4)}(f_i)] \quad (3)$$

Where  $k=1,2$  indicates apple slice and mango slice, respectively.  $M$  is equal to 640 and indicates the number of observations of the slice  $k$ .  $f_i$  denotes the  $i$ -th frequency point.  $N$  is equal to 201 and indicates the number of discrete time or frequency points.

In THz experiment system, the observations generated by MCK are raw data, and it can be distorted due to the system's imperfections and transmission losses at both ends of the THz region. So, these redundant and misdetection observations may affect the overall classification accuracy for different classifiers and produce forged information about freshness of slices. Hence, it is vital to identify target response region (TRR) from the overall region, as shown in Fig.3(b), to focus on meaningful observations and help to reduce computational loads for optimum classification results. Hence, two-sample t-test with statistical significance difference is performed on all observations of all different days [14]. It is noticed from cumulative distribution function (CDF) of the probability of t-test that the observations in the frequency range from 0.8 THz to 1.05 THz exhibited a significant difference with the value of probability  $p$  near to 0 between the different days based on MC of fruit slice. It further indicates that observations are more sensitive in this TRR area.

The block diagram of the proposed classification system for different days based on MC of fruit slices, as shown in Fig. 4. Since observations obtained from MCK were in the frequency domain, these observations were converted to the time domain and time-frequency domain features by applying IFFT and Short-Time Fourier transform (STFT), respectively, to obtain feature extraction. The description of three domain features is discussed in detail in next section individually.

### III. FEATURE EXTRACTION TECHNIQUE

#### A. Frequency Domain Features

The observations obtained from VNA were in frequency domain form, which can be used to extract the frequency domain features in TRR range directly. For this purpose, the variance of the Power Spectral Density (PSD) and the peak value of Cross Power Spectral Density (CPSD) were considered as given in (4) and (5), respectively [15].

$$\text{Var}\{S_{kk}(\omega)\} = \frac{1}{D} E \left\{ \left( S_{(k)}^{(n)}(\omega)^* \cdot S_{(k)}^{(n)}(\omega) \right) \right\} \quad (4)$$

$$\max\{S_{rk}(\omega)\} = \max \frac{1}{D} E \left\{ \left( R(\omega)^* \cdot S_{(k)}^{(n)}(\omega) \right) \right\} \quad (5)$$

In (4),  $S_{(k)}^{(n)}(\omega)$  is the transmission response matrix of the  $k$ -th fruit slice in  $n$ -th day.  $(\cdot)^*$  indicates conjugate transpose.



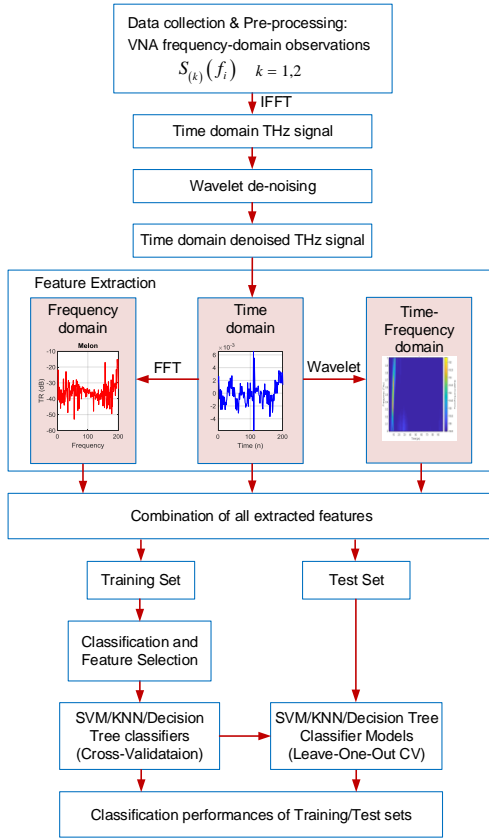


Fig. 4: The block diagram of proposed classification process.

$D$  is the width of the frequency windows.  $\omega$  is the angular frequency,  $\omega = 2\pi f$ . In (5),  $R(\omega)$  is the transmission response of the reference signal, which is measured by connecting the two halves of the corrugated waveguides of the MCK without sample between them together.

In the TRR range, five frequency windows,  $D_1, \dots, D_5$ , with equal width of each frequency window were taken into consideration to extract frequency-domain features for observations of fruits slices. The total area as shown in Fig. 3(b) would be from 0.8 THz to 1.01 THz. The selected region was investigated for all the observations from day 1 to day 4.

### B. Time Domain Features

The main purpose of extracting time domain features was to observe the transmission properties of time series of THz pulse referring to presence of MC in fruit slices from day 1 to 4. For waveforms in time domain, statistical features can be derived with an assumption that the signal is stationary [16][17]. In this study eleven time domain features were considered including mean, median, mean of absolute value (MAV), standard deviation (STD), mean of absolute deviation (MAD), skewness and kurtosis, Pearson correlation coefficient (PCC)[18], 25th percentile (Q1), 75th percentile (Q3), and Interquartile Range (IQR). In which, mean, median, MAV, STD and MAD are common properties used in statistics and probability theory. Skewness and kurtosis are two higher statistics terms, the former provides the symmetric information of data sets relative to the centre point, and the latter provides the flatness of data sets in a uniform distribution. PCC is used

to measure the linear relationship between the time-domain waveforms of the sample and reference signal [19][20]. Q1 and Q3 measure the value at the 25% and 75% location of the data set separately. IQR can be applied to measure the variability of a data set.

### C. Time-Frequency Domain Features

The time-frequency methods such as STFT and wavelet decomposition of the time-domain signal can provide detailed localization properties of THz pulses in time-frequency domain [20]. The wavelet-transform (WT) was more suitable for analyzing the short-duration pulse with fast and unpredictable changes to obtain interesting information [19]. After wavelet de-noising of the observations in time domain, the representation of the time-frequency domain was performed by three level wavelet decomposition using db8 wavelet, and a total of four sub-band signals are decomposed, in which one approximation coefficient,  $C_3(n)$ , and three detail wavelet coefficients,  $D_3(n)$ ,  $D_2(n)$ , and  $D_1(n)$ , can be applied to extract time-frequency domain features [21][22]. The features of the extracted wavelet coefficients can provide the energy distribution of the THz signal in time and frequency domain.

## IV. CLASSIFICATION RESULT AND FEATURE SELECTION

### A. Classification Accuracy for different feature datasets

In this study, three ML classifiers, namely support vector machine (SVM), K-nearest neighbour (KNN) and decision tree (D-tree) are considered to analyze the performances of various features datasets. These feature datasets are raw data collected from MCK, individual domain feature datasets of time, frequency, time-frequency, and hybrid combination dataset of three domain extracted features. In this regard, for each classifier, suitable parameters are selected to enhance the classification accuracy. So, for SVM classifier, Radial Basis Function (RBF) kernel is applied to set key parameters including the Gaussian kernel scale ( $\gamma$ ), and the optimum parameters of cost (C). To establish the appropriate values, series of values are assessed and eventually, 0.35 and 1 are chosen for ( $\gamma$ ) and (C), respectively, after the grid searching optimization of parameters  $\gamma(0.1 : 0.05 : 2)$  and  $C(0.5 : 0.5 : 2)$  [23]. For KNN classifier, we set the number of nearest neighbours  $k$  and the distance metric to 5 after examining the range of  $k(1:10)$  and Euclidean distance, respectively [24]. Lastly, for D-Tree, the maximum number of splits is set to 5. All the other parameters of three classifiers are retained as default values. Furthermore, all classifiers are trained by using 10-fold cross-validation to obtain the validation accuracy of classification of days for each fruit slice, and each type of dataset is partitioned into 80% and 20% training and testing data, respectively.

Table I depicts the comparison of all classifier's performance using various feature datasets. Upon close analysis, it is noticed that the performance of all three classifiers for raw data is unsatisfactory compared to other domain features performance. It also indicates the presence of less sensitive data or data littered with some random noise, which has severely affected the performance of classifiers and can be

TABLE I  
COMPARISON OF CLASSIFICATION PERFORMANCE OF DIFFERENT DAYS USING DIFFERENT DATASETS.

Slices	Classifier models	Classification Accuracy (%)					Execution time of Hybrid features (s)
		Raw data	Time-domain features	Frequency-domain features	Time-Frequency domain features	Hybrid features of three domains	
Apple	SVM	65.2	88.6	97.0	93.6	98.9	1.6506
	KNN	77.1	78.0	86.4	86.4	86.7	0.8274
	D-Tree	72.4	88.2	93.2	93.2	97.8	0.6776
Mango	SVM	70.1	92.0	93.4	93.4	93.4	1.9837
	KNN	79.7	87.7	86.4	86.4	88.6	0.6579
	D-Tree	76.1	91.6	92.5	92.5	93.6	0.8760

TABLE II  
CLASSIFICATION PERFORMANCE OF DIFFERENT DAYS USING SELECTED FEATURES

Slices	Feature selection methods	Classifier models	Number of selected features				Accuracy (%)	Execution time of selected features (s)
			Time-domain features	Frequency-domain features	Time-frequency domain features	Total no. of features		
Apple	SFS	SVM	9	6	4	19	99.8	1.2561
		KNN	9	5	4	18	88.2	0.5945
		D-Tree	11	8	0	19	99.5	0.2787
	Relief-F	SVM					99.1	1.4327
		KNN	4	10	3	17	98.9	0.5788
		D-Tree					99.8	0.5510
Mango	SFS	SVM	10	8	4	22	95.4	1.3456
		KNN	9	2	4	15	94.1	0.2678
		D-Tree	3	1	0	4	95.2	0.2702
	Relief-F	SVM					98.6	1.0870
		KNN	6	10	3	19	99.1	0.6229
		D-Tree					95.5	0.5861

substantially improved by observing a sensitive region. Moreover, Table I also shows an improvement of 17% for individual domain features compared to raw data. For instance, the classification performance of frequency- and time-frequency domains exceed time-domain. Likewise, frequency- and time-frequency domain features present the same level classification performance, which is due to that fact that the THz spectrum in the frequency domain can provide more comprehensive radiation information of samples including transmittance and absorption of THz waves. Furthermore, it is also perceived that the hybrid features of three domain exhibit the best classification performance i.e. 98.9% in SVM classifier for four consecutive days of the apple slice.

By close observations of SVM, KNN and D-Tree performance, it is evidently observed that results could be further improved by feature extraction in TRR range of the raw data. Moreover, using the state-of-the-art technique of feature selection method in [25], features could be combined optimally from multi-domains, and this also helped to reduce the computational time by eliminating less-informative features.

### B. Features Selection

The redundant or irrelevant features extracted from the previous processing can be removed through feature selection techniques to optimize the combinations of features for improving the classification performance and minimizing the computational cost for deployment. Feature selection techniques include filter methods based on the assessments of the relevance of features and wrapper methods based on the force search of different combinations of the feature space [25]. In this section, we applied two algorithms named sequential forward selection (SFS), and the Relief-based feature selection algorithms (Relief-F) to perform the feature selection [26]. SFS is the heuristic selection algorithm, which starts with an

empty set and combines one best feature in each step with high accuracy by using a classifier until the pre-defined number of features are added. Relief-F can offer a computationally efficient method to recognize the interactions of features and calculate the feature weights between 0 and 1 to rank, and select features without depending on the certain classifier [27]. Table II shows the classification results of different days for apple and mango slices with different feature selection algorithms and the number of selected features. The SFS and Relief-F algorithms yield over 2% improvement for classification of days for apple and mango slices with at least 12% features reduced out of the total of all extracted features. Interestingly, SFS with D-Tree classifier obtains the optimal feature combinations including only three time-domain and one frequency-domain features leading to about 2% improvement of accuracy for mango slice, compared to the results of extracted features in Table I. In addition, the execution time obtain in Table II indicates that selected features have not only enhanced the computational time but have also improved the classification accuracy considerably. The least improvement of the execution time on classification is Relief-F KNN of a mango slice, 5.3%, and the most one is SFS D-Tree of a mango slice which goes up to 69.2%.

## V. RESULTS AND DISCUSSIONS

Table III represents the confusion matrix and the classification accuracy for different days with different classifier models, based on features selection using SFS method for both training and test sets of apple and mango slices, respectively. For apple slice, the results illustrate an accuracy increment of 10% for both SVM and D-Tree models compare to KNN model in training set, whereas D-Tree model exhibits highest accuracy in the test set. Furthermore, for mango slice, both SVM and KNN displays better accuracy compare to D-Tree

TABLE III  
CONFUSION MATRIX AND ACCURACY OF CLASSIFIER MODELS FOR TRAINING AND TEST SETS OF FRUIT SLICES.

Slices	Classifier Models	Sample Sets	Days	Predicted				Accuracy of Each day (%)	Overall Accuracy (%)
				Day1	Day2	Day3	Day4		
Apple	SVM	Training Set	Day1	128	0	0	0	100	99.20
			Day2	0	62	1	1	96.88	
			Day3	0	0	64	0	100	
			Day4	0	0	2	94	97.92	
		Test Set	Day1	31	1	0	0	96.88	90.91
			Day2	3	12	1	0	75.00	
			Day3	0	0	15	1	93.75	
			Day4	0	0	2	22	91.67	
	KNN	Training Set	Day1	127	1	0	0	99.22	89.57
			Day2	16	42	5	1	65.63	
			Day3	0	6	51	7	79.69	
			Day4	0	0	7	89	92.71	
		Test Set	Day1	28	4	0	0	87.50	82.95
			Day2	3	11	0	0	68.75	
			Day3	0	0	13	3	81.25	
			Day4	0	1	2	21	87.50	
	Decision Tree	Training Set	Day1	128	0	0	0	100	99.15
			Day2	0	63	1	0	98.43	
			Day3	0	1	63	0	98.43	
			Day4	0	0	1	95	98.96	
		Test Set	Day1	32	0	0	0	100	95.45
			Day2	0	15	1	0	93.75	
			Day3	0	0	15	1	93.75	
			Day4	0	0	2	22	91.67	
Mango	SVM	Training Set	Day1	127	1	0	0	99.22	97.73
			Day2	0	61	3	0	95.31	
			Day3	0	4	92	0	95.83	
			Day4	0	0	0	64	100	
		Test Set	Day1	32	0	0	0	100	90.91
			Day2	0	12	4	0	75.00	
			Day3	0	3	21	0	87.50	
			Day4	0	1	0	15	93.75	
	KNN	Training Set	Day1	128	0	0	0	100	98.30
			Day2	0	61	3	0	95.31	
			Day3	0	2	94	0	97.92	
			Day4	0	0	1	63	98.44	
		Test Set	Day1	32	0	0	0	100	89.77
			Day2	0	10	6	0	62.5	
			Day3	0	3	21	0	87.5	
			Day4	0	0	0	16	100	
	Decision Tree	Training Set	Day1	128	0	0	0	100	92.61
			Day2	0	57	7	0	89.06	
			Day3	0	18	78	0	81.25	
			Day4	0	0	1	63	98.44	
		Test Set	Day1	32	0	0	0	100	94.32
			Day2	0	13	3	0	81.25	
			Day3	0	2	22	0	91.67	
			Day4	0	0	0	16	100	

model in training set, and D-Tree model achieves the best accuracy in the test dataset.

The performance of classifiers is assessed by applying two quality classification metrics including sensitivity (SENS) (also called true positive rate), and specificity (SPEC) [19][28]. The value of SENS represents the probability of the target class identified correctly in the total number of one target classes. Whereas, the value of SPEC expresses the probability of classifying the sample as non-target classes correctly in the non-target classes. The values of the two-quality metrics are found in the range from 0 to 1.

Table IV presents the performance of the proposed classifiers algorithms for different days using SFS method and indicates the presence of MC in both apple and mango slices. It can be observed that the D-Tree has shown substantial improvement in classification accuracy, indicating the freshness or staleness of both different slices. Furthermore, it is also depicted that the assessment parameter values of the SVM

classifier achieves better precision compare to KNN for apple and mango slices. From Table IV, it is worth observing that the classification of the days is equivalent to that of the percentage of the MC in the slices. It is due to this reason, different MC value leads to the different absorption strength of terahertz radiation in samples, which causes variations in the features for classification. Also, the results show that classification accuracy on days 2 and 3 is slightly more challenging than for day 1 and 4, especially for KNN classifier where the MC values falls in the range between 10% to 50%. It is because the MC value of day 1 is very high (freshness), whereas on day 4 (almost dried out) is very low, which leads to a clear separation of features in the feature space.

In a real-life application, it would be more convincing to obtain the classifier's performance by considering leave-one-observation-out-cross-validation technique for more accurate estimation of MC in fruits slices on different days. This propose method can be more efficient because in this process,

TABLE IV  
CLASSIFICATION ACCURACY WITH LEAVE-ONE-OBSERVATION-OUT TECHNIQUE BY MONITORING MC OF FRUIT SLICES.

Samples	Classes	Quality Metrics			MC (%)	Classifiers and test accuracy (%)		
		SVM	KNN	D-Tree		SVM	KNN	D-Tree
Apple slice	Day1	SENS	1	0.96	1	82.2	100	99.8
		SPEC	1	0.93	1			
	Day2	SENS	0.99	0.68	1	37.8	95.8	80.0
		SPEC	1	0.98	1			
	Day3	SENS	1	0.86	1	12.3	100	95.0
		SPEC	1	0.96	1			
	Day4	SENS	1	0.93	1	0.50	100	96.7
		SPEC	1	0.98	1			
Mango slice	Day1	SENS	1	1	1	77.9	100	100
		SPEC	1	1	1			
	Day2	SENS	0.80	0.71	0.86	48.7	89.5	81.3
		SPEC	0.98	0.96	0.97			
	Day3	SENS	0.90	0.84	0.92	15.0	85.7	80.8
		SPEC	0.94	0.93	0.97			
	Day4	SENS	0.96	0.95	1	0.24	100	99.3
		SPEC	0.99	0.98	1			

each observation is randomly selected from the dataset as the validation set, while the remaining observations are considered as the training set to generate the classifier model. This process is repeated until all the observations in the dataset are selected as validation set once. The cumulative classification results of days are calculated from leave-one-observation-out-cross-validation approach based on MC values of each day, as shown in Table IV. From table IV, it is found that the classification accuracy for day 1 and day 4 is nearly 100% which indicates the highest MC and the best freshness in day 1 and the lowest MC and dehydration in day 4. For day 2 and day 3, the classifiers can provide the quantitative reference based on MC value for the quality assessment of the fruit slices. Thus, the aim of applying the proposed technique is to develop the consistency of classifiers by observing all observations of both slices on different days.

## VI. CONCLUSION

This paper presented the novel and non-invasive sensing technique for the quality assessment of fresh of fruits by integrating machine learning ML with terahertz (THz) waves. For this purpose, scattering measurements of both apple and mango slices were obtained using Swissto12 MCK system. Multiple domain features such as time-, frequency-, and time-frequency domains were extracted and the classification was performed to determine the moisture content (MC) in fruits slices using SVM, KNN and D-tree classifiers. Results showed that by discarding unwanted features and applying comprehensive cross-validation technique, classification accuracy was substantially improved which eventually helped to reduce the computational time. Furthermore, it was perceived that, in most of the cases, the SVM classifier based on RBF kernel outperformed KNN and D-Tree classifier for both fruit slices. Thus, SVM demonstrated more precise quality assessments of fresh fruits by determining their MC more precisely for four consecutive days.

The proposed technique demonstrates the strong potential in the discipline of food and science technology by integrating ML with THz waves to assess real-time information of fruits on different days at cellular level. For future research, more fruits slices will be considered to extract additional features

by employing their electromagnetic parameters for the identification of different fruits in an automated and non-invasive manner.

## REFERENCES

- [1] O. W. Liew, P. Chong, B. Li, and A. K. Asundi, "Signature optical cues: emerging technologies for monitoring plant health," *Sensors*, vol. 8, no. 5, pp. 3205-3239, 2008.
- [2] H. Wang, J. Peng, C. Xie, Y. Bao, and Y. He, "Fruit Quality Evaluation Using Spectroscopy Technology: A Review," *Sensors*, vol. 15, no. 1, pp. 11889-11927, 2015.
- [3] M. Tonouchi, "Cutting-edge THz technology," *Nature Photonics*, vol. 1, pp. 97-105, 2007.
- [4] P. Butz, C. Hofmann, and B. Tauscher, "Recent Developments in Noninvasive Techniques for Fresh Fruit and Vegetable Internal Quality Analysis," *Journal of food science*, 79(9), R131-R141, 2005.
- [5] B., Ferguson and X. Zhang, "Materials for terahertz science and technology," *Nature Materials*, vol. 1, pp. 26-33, 2002.
- [6] K.G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine Learning in Agriculture: A Review," *Sensors*, vol. 18, no. 8, pp. 2674, 2018.
- [7] A. Ren, A. Zahid, D. Fan, X. Yang, M. Imran, A. Alomainy, and Q. Abbasi, "State-of-the-art terahertz sensing for food and water security – a comprehensive review," *Trends in Food Science and Technology*, vol. 85, pp. 241-251, 2019.
- [8] A. Zahid, H. T. Abbas, M. A. Imran, K. Qaraqe, A. Alomainy, D. R. S. Cumming, Q. H. Abbasi, "Characterization and Water Content Estimation Method of Living Plant Leaves Using Terahertz Waves," *Applied Sciences*, vol. 9, no. 14, 2781, 2019.
- [9] A. Zahid, H. Heidari, Chong Li, M. A. Imran, A. Alomainy, and Q. H. Abbasi, "Terahertz characterisation of living plant leaves for quality of life assessment applications," in *Baltic URSI Symposium (URSI)*, Poznan, 2018, pp. 117-120.
- [10] A. Khalid, D. Cumming, R. Clarke, C. Li, and N. Ridler, "Evaluation of a VNA-based material characterization kit at frequencies from 0.75 THz to 1.1 THz," in *Proceedings of IEEE 9th UK-Europe-China Workshop on Millimetre Waves and Terahertz Technologies*, 2016, pp. 31-34.
- [11] P. Nie, F. Qu, L. Lin, T. Dong, Y. He, Y. Shao, and Y. Zhang, "Detection of water content in rapeseed leaves using terahertz spectroscopy," *Sensors*, vol. 17, no. 12, pp. 2830, 2017.
- [12] X. Yin, B. Ng, and D. Abbott, "Terahertz Imaging for Biomedical Applications: pattern recognition and tomographic reconstruction," *Springer Science & Business Media*, 2012, pp. 65-90.
- [13] H. Li, D. Yuan, Y. Wang, D. Cui, and L. Cao, "Arrhythmia Classification Based on Multi-Domain Feature Extraction for an ECG Recognition System," *Sensors*, vol. 16, no. 10, pp. 1744, 20 Oct. 2016.
- [14] E. Berry, A. J. Fitzgerald, N. N. Zinov'ev, G. C. Walker, S. H.-Vanniasinkam, C. D. Sudworth, R. E. Miles, J. M. Chamberlain, and M. A. Smith, "Optical properties of tissue measured using terahertz-pulsed imaging," *Proc. SPIE 5030, Medical Imaging 2003: Physics of Medical Imaging*, (5 June 2003); doi: 10.1117/12.479993.
- [15] S. H. Von and F. W. Zwiers, "Statistical analysis in climate research," Cambridge University Press, 2001. ISBN 978-0-521-01230-0.



- [16] A. Phinyomark, P. Phukpattaranont, and C. Limsakul, "Feature reduction and selection for EMG signal classification," *Expert Systems with Applications*, vol. 39, pp. 7420-7431, 2012.
- [17] Javaid, H. Ali, N. Rashid, M. Tiwana, and M. W. Anwar, "Comparative Analysis of EMG Signal Features in Time-domain and Frequency-domain using MYO Gesture Control," in *Proceedings of the 2018 4th International Conference on Mechatronics and Robotics Engineering*, ACM, 2018, pp. 157-162.
- [18] H. Chen, X. Chen, S. Ma, X. Wu, W. Yang, W. Zhang, and X. Li, "Quantify Glucose Level in Freshly Diabetic's Blood by Terahertz Time-Domain Spectroscopy," *Journal of Infrared, Millimeter, and Terahertz Waves*, vol. 39, no. 4, pp. 399-408, 2018.
- [19] Siuly, X. Yin, S. Hadjiloucas, and Y. Zhang, "Classification of THz pulse signals using two-dimensional cross-correlation feature extraction and non-linear classifiers," *Computer Methods, and Programs in Biomedicine*, vol. 127, pp. 64-82, April 2016.
- [20] E. Berry, et al., "Time frequency analysis in terahertz pulsed imaging," In: Bhanu, B. and Pavlidis, I., (eds.) *Computer Vision: Beyond the Visible Spectrum*. Advances in Pattern Recognition. Springer Verlag, London, UK, 2005, pp. 290-329.
- [21] X. Yin, B. Ng, and D. Abbott, "Application of auto regressive models of wavelet sub-bands for classifying terahertz pulse measurements," *Journal of Biological Systems*, vol. 15, no. 4, pp. 551-571, 2007.
- [22] R. Haddadi, E. Abdelmounim, M. El Hanine, and A. Belaguid, "Discrete Wavelet Transform based algorithm for recognition of QRS complexes," in *International Conference on Multimedia Computing and Systems (ICMCS)*, Marrakech, 2014, pp. 375-379.
- [23] T. Noi, Phan, and M. Kappas, "Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery," *Sensors*, vol. 18, no. 1, pp. 18, 2018.
- [24] L. Hu, M. Huang, S. Ke, and C. Tsai, "The distance function effect on k-nearest neighbor classification for medical datasets," *Springerplus*, vol. 5, no. 1, pp. 1304, 2016.
- [25] S. Z. Gürbüz, B. Erol, B. Çağlıyan, and B. Tekeli, "Operational assessment and adaptive selection of micro-Doppler features," *IET Radar, Sonar & Navigation*, vol. 9, no. 9, pp. 1196-1204, 2015.
- [26] J. Pohjalainen, O. Räsänen, and S. Kadioglu, "Feature Selection Methods and Their Combinations in High-Dimensional Classification of Speaker Likability," *Intelligibility and Personality Traits, Computer Speech and Language*, vol. 29, no. 1, pp. 145-171, 2015.
- [27] R. J. Urbanowicz, et al., "Benchmarking Relief-Based Feature Selection Methods for Bioinformatics Data Mining," *Journa of Biomedical Informatics*, pp. 168-188, 2018.
- [28] Estefana Perez-Castano, et al., "Comparison of different analytical classification scenarios: application for the geographical origin of edible palm oil by sterolic (NP) HPLC fingerprinting," *Analytical Methods*, vol. 7, pp. 4192-4201, 2015.



**Aifeng Ren** received the Ph.D. degree in information and telecommunication engineering from Xi'an Jiaotong University, Xi'an, China, in 2007. He is currently an Associate Professor with School of Electronic and Engineering, Xidian University. He has authored/co-authored over 40 journal and conference publications. His research interests include embedded wireless sensor networks, terahertz sensing technology, complex brain network, and signal and image processing.



**Adnan Zahid** received his B.Sc. (Hons) in Electronics and Communications Engineering from Glasgow Caledonian University, and MSc degree in Electronics and Electrical Engineering from University of Strathclyde, in 2016. He is currently pursuing his Ph.D Research Degree in University of Glasgow. His research interests include Machine learning enabled Terahertz sensing for precision agriculture technology at cellular level.



**Ahmed Zoha** received the Ph.D. degree in electronic engineering from the University of Surrey, Surrey, U.K., in 2014, which was funded by the Engineering and Physical Sciences Research Council (EPSRC). Now he is an Assistant Professor with School of Engineering, University of Glasgow, Glasgow, U.K. His current research encompasses machine learning at the edge and representation learning using deep neural networks to drive intelligent reasoning.



**Syed Aziz Shah** is currently working as an Assistant Professor (lecturer) at Manchester Metropolitan University, UK. He completed his PhD from Xidian University China, in June 2018, and worked as a Post Doctorate Research Associate at University of Glasgow. He has (co)authored more than 25 technical articles in top-rank cross-disciplinary journals including IEEE and IET. His research interests include Machine Learning in Wireless Sensing, Radar Technology, Software Defined Radios, Antennas and Propagation, healthcare and agriculture technologies.



**Muhammad Ali Imran** (M'03-SM'12) received the M.Sc. (Hons.) and Ph.D. degrees from Imperial College London, U.K., in 2002 and 2007, respectively. He is the Voce Dean Glasgow College UESTC and a Professor of communication systems with School of Engineering, University of Glasgow. He is an Affiliate Professor at the University of Oklahoma, USA, and a Visiting Professor at 5G Innovation Centre, University academic and industry experience, working primarily in the research areas of cellular communication systems. He has been awarded 15 patents, has (co)authored over 400 journal and conference publications.



**Akram Alomainy** (S'04-M'07-SM'13) received Ph.D. degree in electrical and electronic engineering from the Queen Mary University of London (QMUL), U.K. 2007. He joined the School of Electronic and Computer Science, QMUL, in 2007, where he is a Reader with the Antennas and electromagnetics Research Group. His current research interests include small and compact antennas for wireless body area networks, radio propagation characterization and modelling, antenna interactions with human body, advanced antenna enhancement techniques for mobile and personal wireless communications, and advanced algorithm for smart and intelligent antenna.



**Qammer H. Abbasi** received his BSc and M. Sc. Degrees electronics and telecommunication engineering from the University of Engineering and Technology (UET), Lahore, Pakistan, and Ph.D. degree in electronic and electrical engineering from the Queen Mary University of London (QMUL), U.K., in 2012. In 2012, he was a Post-Doctoral Research Assistant with the Antenna and electromagnetics Group, QMUL. He is currently a Lecturer (Assistant Professor) with the School of Engineering, University of Glasgow, U.K. He has contributed to a over 250 leading international technical journal and peer reviewed conference papers, and 8 books. He received several recognitions for his research, which includes appearance in BBC, STV, dawnnews, local and international newspaper, cover of MDPI journal, most downloaded articles, UK exceptional talent endorsement by Royal academy of Engineering, National talent pool award by Pakistan, International Young Scientist Award by NSFC China, URSI Young Scientist award, National interest waiver by USA, 4 best paper awards and best representative image of an outcome by QNRF. He is an Associate editor for IEEE Journal of Electromagnetics, RF, and Microwaves in Medicine and Biology, IEEE Sensors, IEEE open access Antenna and Propagation, IEEE Access journal and acted as a guest editor for numerous special issues in top notch journals.