Machine Learning Framework for the Detection of Anomalies in Aqueous Solutions Using Terahertz Waves

Adnan Zahid[1,2*], Kia Dashtipour[1], Ivonne E. Carranza[1], Hasan T. Abbas[1], Aifeng Ren[3], David R. S. Cumming[1], James P. Grant[1], Muhammad A. Imran[1], and Qammer H. Abbasi[1,*]

[1]James Watt School of Engineering, University of Glasgow, Glasgow, G12 8QQ, U.K
[2]Heriot Watt University, School of Engineering and Physical Science, Edinburgh. EH14 4AS, UK
[3]School of Electronic Engineering Xidian University, Xi'an, Shaanxi, China
[*]Qammer.Abbasi@glasgow.ac.uk
[+]these authors contributed equally to this work

ABSTRACT

Water is considered to be the most essential and vital resources to sustaining life. Ensuring its delivery to people with no intrusion of harmful impurities, safe, reliable, and in an affordable manner is one of huge challenge amid to the ongoing climate transformations. This demands to introduce a cost effective and notion of real-time monitoring system that can detect the microbiological contaminants in aqueous solutions in timely manner to protect the public and environment health. In this paper, the prospects of integrating non-invasive terahertz (THz) waves with machine learning (ML) enabled technique is studied. The research explores a method of using Fourier transform Infrared Spectroscopy (FTIR) system to observe the absorption spectra and characteristics of three solvents solution, including salt, sugar and glucose with various quantity in aqueous solutions in the frequency range of 1 THz to 20 THz. In this study, due to the different molecular configuration and vibration modes of substances, distinct absorption spectra peaks were achieved for different concentrations of solvent solutions at certain sensitive THz region. Moreover, using measurements observations data, meaningful features are extracted and incorporated four algorithms such as random forest (RF), support vector machine (SVM), decision tree (D-tree) and k-nearest neighbour (KNN). The results demonstrated that RF obtained a higher accuracy of 84.74% in identifying the substance in aqueous solutions. Moreover, it was also found that RF with 97.98%, outperformed other classifiers for estimation of salts concentration added in aqueous solutions. However, for sugar and glucose concentrations, SVM exhibited a higher accuracy of 93.11% and 96.88%, respectively, compared to other classifiers. Thus, proposed technique incorporating ML with THz waves, may be significant in providing an efficient, cost-effective and real-time monitoring for water quality detection system.

Introduction

In a rapidly developing and modern world, the importance and preservation of clean water without any harmful impurities for the overall global health, environmental protection, and economic development cannot be undervalued[1]. Providing sufficient and affordable water in a safe and reliable way with limited resources is a huge challenge of mounting sternness as the demand increases with a rising population[1,2]. Also, fresh and unpolluted water is worsened by climate transformations, more regular droughts in many parts of the word, and by water pollution, making it more demanding and costly to handle[1,2]. Mostly, the general consensus among the scientific community is that the emergence of infectious diseases such as tuberculosis, measles, and other lethal illness, often detected and caused by microbiological and micro-chemical contaminants in tap and drinking water sources, which cannot be detected by naked eyes, leading to jeopardize the public health and safety[2–4].

Hence, the aforementioned facts underline the imperative demand for the fast, reliable, secure, and sensitive technological solutions and vitalization of resources for the precise detection and regular monitoring of water contaminants in vulnerable population to discourage any short and long-term health consequence[2–4]. In recent times, many researchers have put tremendous efforts and suggested variety of directions[2] and technologies[4] to address the substantial issues as discussed earlier. In this regard, significant achievements have been obtained through the initiation of water quality sensors, microfluids sensors, model-based event detection, advanced vibrational spectroscopy and miniaturized biosensors, and widely improved the water contamination detection qualitatively[3,4]. However, these different approaches have some advantages and limitations. For example, the

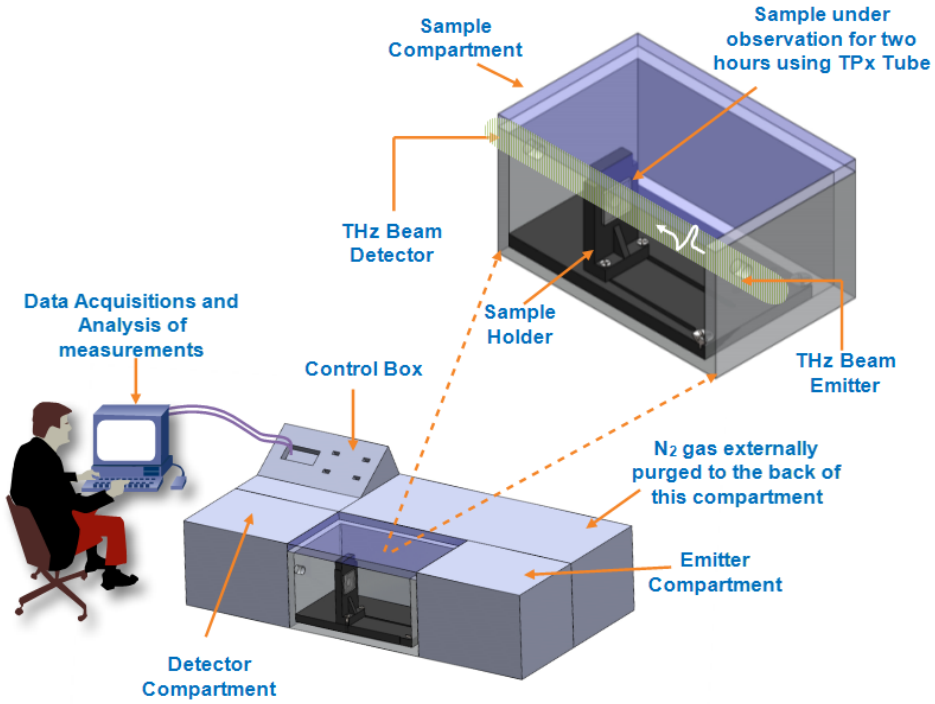


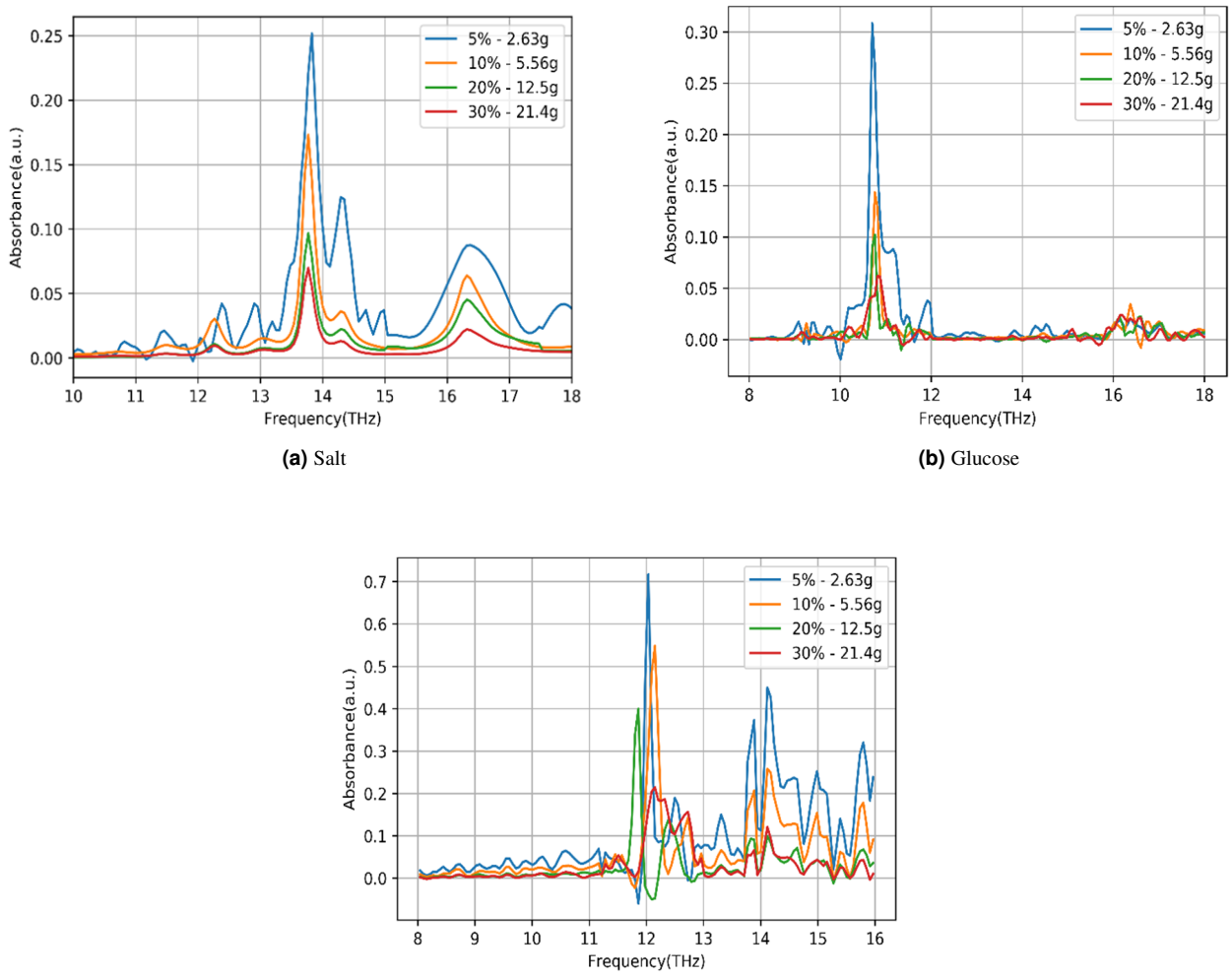
Figure 1. Isometric schematic of Fourier Transform Infrared Spectroscopy (FTIR) system, with the sample compartment above pointing the THz beam generated at the source, passing through the TPx tube to assess the constituents of salt, sugar and glucose in aqueous solutions. The observations for each sample ran up to 2 hours approximately.

deployment of numerous water quality sensors seems not very feasible due to high installation cost, time consuming, low response detection and less reliability[2–4].

In addition, results from model-based event detection have indicated certain error rate due to the low sensitivity, providing inadequate symptoms of contaminates in water[2]. Some researchers have also considered Infrared (IR) for the swift detection of impurities in solvents[5,6]. Though it has obtained considerable advancements and yield satisfactory results[6]. However, there are some limitations and have mainly focused on the theoretical calculations to observe the characteristics of impurities added in solvents and absorption features[6]. Thus, this technique is transpired as inappropriate and feasible for precise detection of contaminants in pure water at molecular level and have markedly minimize its suitability[6]. Despite in-depth theoretical attempts and substantial significant advancements over the past years, the microscopic frameworks leading to the numerous anomalies or contaminants of water, often considered as the compact substance or a primary biological solvent, remain from being fully comprehended by the researchers in physical and biological sciences[3–6]. Consequently, the concerning effects of poor contamination technique instantly require developing a more robust, qualitative, less operating costs, and high sensitivity quantification of contaminants in solvents in a non-invasive manner[4,6]

With this motivation and limitations found in previous techniques, this paper proposes a realistic method and application of Fourier transform Infrared Spectroscopy (FTIR) as depicted in Fig. 1 enabled by machine learning (ML)[7] that can provide the approximate prediction and detection of even the smallest of contaminants in distilled water due to high sensitivity and non-destructive nature and can also produce high optical throughput[6]. This technology includes terahertz (THz), which has achieved tremendous achievements in diverse field such as diagnostic applications of dental and skincare medical imaging, invisible hazard and vulnerable items, material characterizations, and telecommunications[8–11]. For this purpose, an integration of ML with THz can create a dynamic opportunity to uncover, measure, and thoroughly understand the data-intensive procedure in to minutely observe the absorbance spectra of different solutions in aqueous solutions[9]. The significant contributions of this work are as follows:

a) This paper suggests a novel technique by employing a FTIR setup that provides a THz frequency range of interest operating from 1 THz to 20 THz to precisely determine the various solvents constituents' characteristics in aqueous solutions



(a) Salt

(b) Glucose

(c) Sugar

Figure 2. The terahertz (THz) absorption spectra of salt, glucose and sugar constituents'characteristics in aqueous solutions in the frequency range of 0.8 THz to 18 THz

non-invasively.

b) The proposed methodology also suggests the ML driven approach to proactively determine the presence of any anomalies or impurities in aqueous solutions in real time to protect the environment, include early alerts to protect the public health, and reduce any superfluous costs.

c) In this study, by integrating THz with ML, we explore not only identifying the various constituents' in aqueous solutions, but also to determine the amount of impurities in each constituent solution by establishing ML algorithm technique.

d) Finally, this paper presents a notable and distinctive contributions of THz technology with ML in assessing the impurities in aqueous solutions at cellular level.

Results

In this work, the focus was mainly to observe the THz absorption spectra (AS) for three various distinct solutions as explained earlier. These measurements were performed in Terahertz Laboratory, at University of Glasgow with great care. The time taken for observing AS was 2 hours in order to obtain maximum point data to minutely observe in the THz region for any impurities added in distilled water. The number of data points calculated as 338 was collected for every sample. During the measurement process which lasted for 120 minutes for every sample, and 2102 scans were obtained for each sample.

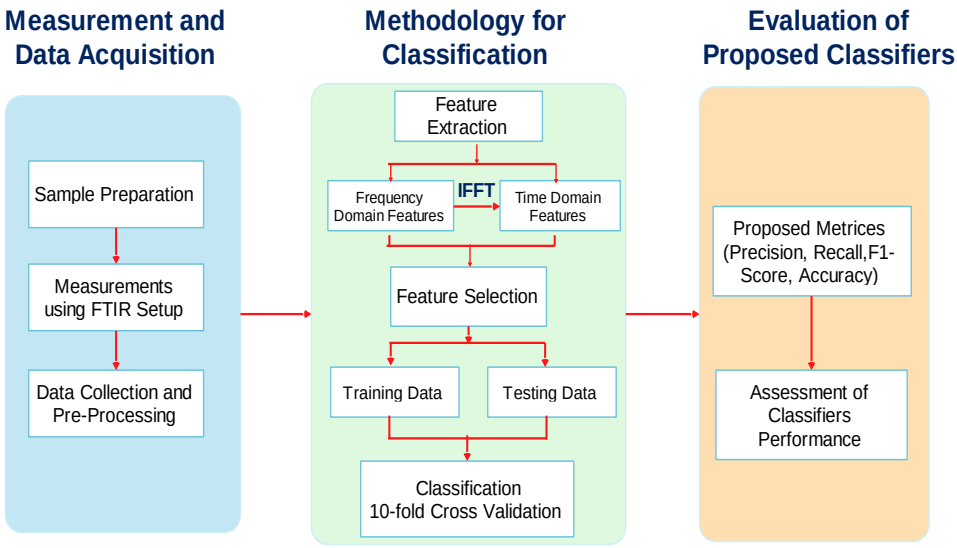


Figure 3. The methodological approach of proposed algorithm for the classification procedure.

The whole process was repeated for all three samples and data was pre-processed using Matlab 2019a, whereas python was used for ML classification in the form supervised learning. While performing the measurements, it was very important to monitor the (N2) gas on regular basis to ensure the continuous flowrate to the compartment, avoiding any irregular behaviour of constituents added in aqueous solutions. The absorption spectra of all three samples including, salt, sugar and glucose in Fig 2. (a), (b), and (c), respectively.

From Fig. 2, we have attempted to minutely observe the absorption spectra of molecular-scale dynamics of distinguished constituents' concentrations in aqueous solutions. In particular Fig. 2 (a), it describes the characteristics absorption peaks of different salts concentrations falling in the range of 13 THz to 15 THz range markedly indicating a more sensitive frequency region for precise detection of distinguished concentrations pattern reaching to 0.25, 0.17, 0.09, and 0.06 approximately for 5%, 10%, 20% and 30% concentrations, respectively. However, considering the range from 15 THz to 17 THz, it can be observed though it has distinct response for distinguished concentrations, but the peaks of absorption spectra for 5%, 10%, 20% and 30% are substantially lower than aforementioned range. This occurrence is attributed to the high sensitivity and strong penetration feature of THz that has depicted diminutive variations of salt concentrations in aqueous solutions at different region.

Upon a close analysis of Fig. 2 (b) and (c), it is also depicted that both glucose and sugar have exhibited a distinct response for various concentrations in different THz region. Notably, sugar concentrations display a more discernible response compare to glucose and this is clearly discovered by THz waves. Furthermore, this prominent and distinguished results showed a distinctive characteristics and functional properties of both sugar and glucose concentrations in aqueous solutions since sugar is mainly compound of various synthesis whereas, glucose is considered to be pure. These results reveal the significant influence of added ingredients in the aqueous solutions and interestingly, and also provide a promising method of the rapid and effective identification of elements for the various concentrations. However, the main objective of this study is also to establish a computationally competent and reliable method for estimating water quality variables using THz waves that reduces labour and the cost of accurately measuring these various parameters.

For this purpose, ML algorithm[7] has been developed to identify any unknown irregularities or anomalies in pure distilled water. In addition, it is also aimed to detect the exact amount of concentrations of mysterious impurities added to the distilled water. Thus, an effective, automated, and precise quantitative detection of harmful contaminants at molecular level in water is utmost of significance to provide early warnings to protect public health.

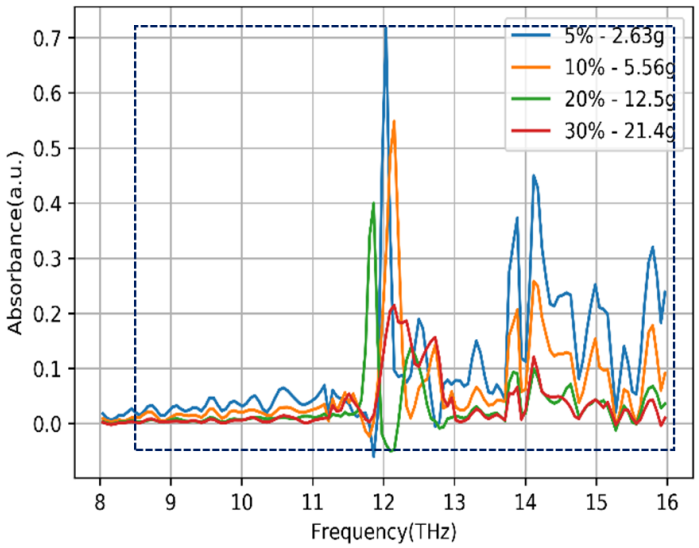


Figure 4. Identification of sensitive region to consider only relevant and meaningful features for the feature extraction process.

Feature Extraction Procedure

While taking the measurement, it was noticed that observations collected using FTIR setup were appeared to be little irregular and unwanted excessive variations. The occurrence of this noticeably undesired and counterfeit observations may have given the fictitious information about any impurities or concentrations level added into the aqueous solutions. Furthermore, useful observations would also have a fruitful impact on overall classification outcome. In addition, presence of any imprecise minerals or chemicals in water can be highly harmful for overall public health. In this regard, it was significant to discover the sensitive frequency region (SFR) in the THz region as shown in Fig. 4 with minimum intrusion of any external factors in the observations, contributing to obtain the maximum information about the smallest particles of constituents in water. For this purpose, specific region (SR) was established for each constituents' samples ranging from 8 THz to 17 THz out of whole region. Within this specific region, absorption spectra peaks for low and high concentrations can be easily discerned with little overlap. Researchers have suggested and applied many features extraction techniques to execute the classification accuracy[12].

Since the observations collected from setup was in frequency domain, so, it was converted to time-domain region using an Inverse Fast Fourier Transform (IFFT) to initiate the possibility of acquiring the statistical features of observations. Out of 338 features observations, 34 valuable features were extracted collectively by looking at both frequency and time domain features and are summarised in Table 1. It shows time-domain features such as mean, standard deviation (STD), skewness and kurtosis were useful for distribution of data, discovering any irregularities of examined area, and obtaining an evenness to a distribution of data, respectively[13],[14]. Q3 and Q1 showed how the observation data were dispersed in the two sides of the median[15,16]. The statistical domain features proven to be helpful for choosing most relevant and meaningful features, contributing to the accurate identification and concentrations quantities in aqueous solutions[14–16]. In this study, frequency domain features were also employed such as special entropy and spectral power. The block diagram of the proposed classification system for different days based on multi-domain features extraction approach is shown in Fig 3.

Classification Methodology

In this research, four classification techniques were considered used namely, support vector machine (SVM), K-nearest neighbour (KNN) and random forest (RF), and decision tree (D-tree)[17]. In this study, considering the measurements obtained using THz waves, two scenarios have been considered and developed a classifier model for them. The concept behind the formulation of these two scenarios is to adapt the real-life situations where the purity of water is extremely essential for the safety of human health. Keeping this significant aspect in mind, the performance of all four classifiers were analysed and tested for accurate identification of impurities and to trace the specific amount of constituents' concentrations in aqueous solutions for the given data. For this purpose, the measured dataset was randomly separated into training and testing with division of 70% and 30%, respectively.

For this purpose, Python SciKit library was used as it has been widely utilized in data-science discipline[18]. All the

Time-Domain Features		Frequency Domain Features	
Minimum	2	Special Entropy	3
Maximum	2	Spectral Power	9
Mean	2		
Standard Deviation	2		
Skewness	2		
Kurtosis	2		
25^{th} percentile (Q1)	2		
75^{th} percentile (Q 3)	2		
Range	2		
Root Mean Square	2		
Quartile	2		
Total features	22	Total features	12

Table 1. Significant Feature Extraction Techniques Using Time and Frequency Domain Features.

measurements data was converted into CSV format so that they can be easily processed by Scikit library. The dataset for different constituents with various concentrations are properly labelled to execute the supervised ML technique. In this regard, 10-fold cross validation technique was considered to critically analyse ML algorithms where each dataset are examined as test data and the remaining were taken as training data and this process was continued until all dataset are tested, resulting in the average results across all the repetitions.

In comparison to other ML techniques, the KNN algorithm is well-known for its simplicity and ease of operation[19]. This technique operates by evaluating the testing data to the training data. In this scenario, K sample are assigned to a feature of training data and subsequently, testing data is allocated to k sample that closely matches the new data. Thus, tuning this fundamental parameter of k-sample plays a significant role in achieving the ultimate performance of this classifier[19–21]. Furthermore, the SVM operates mainly on two classes and is formally defined by dividing hyperplane as a discriminatory classifier. The hyperplane acts as a decision borderline for classification of datasets between two classes. Equation 1 represents how the SVM operates[22].

$$\begin{aligned} \bar{w} \cdot u + \bar{b} > 0 \\ \bar{w} u + \bar{b} < 0 \end{aligned} \tag{1}$$

In above equation, 'w' indicates the weight vector, u displays the input vector and b denotes a constraint. Furthermore, random forest is a set of trees for making decisions. Every tree allows performance prediction by searching for features found during the training process[17]. The majority of prediction is the final prediction for the Random Forest[17].

Feature Selection

In applications such as performing the measurements and dealing with various instruments, possibility of some superfluous and extraneous features is increased which may result in lowering the classification performance. Therefore, it was essential to eliminate those features in order to enhance the classification performance of proposed classifiers as well as reducing the computational costs for deployment. To do so, three feature selection techniques namely, sequential forward selection (SFS), and Relief based selection algorithm (Relief-F) which are widely used are considered to accomplish the feature selection procedure[23]. In SFS method, at the start, empty features are being replaced by some noticeable features which helps to enhance the overall accuracy[23]. Compare to SFS, Relief-F can present a relatively effective approach by considering the function relationships for evaluating the weights of features for appropriate classification and selection instead of relying on different classifiers[24]. Just as precision and recall, individually, are incapable of covering all key aspects of accuracy, thus, F1-score employ the cumulative mean approach to show its performance. By this way, all aspects are considered and demonstrate the overall accuracy. The higher the score, the better the accuracy. Applying these feature selections has considerably yield an improvement of 5%, 4%, 7% and 3% in RF, SVM, D-Tree and KNN, respectively. Furthermore, the additional advantage of feature selection is the further reduction of overall number of features needed for the optimal set, hence computation weights is also optimized for optimal results.

Evaluation of Classifier Performance using Metrices

In this section, the performance of all proposed classifiers was evaluated by using four commonly metrices such as, accuracy, precision, recall (also known as true positive), and F1-score[25]. Table 2 presents the list of classifier performance metrices.

Classification Performance by Applying Tenfold Cross Validation					
Classifier Algorithm	Solvent	Quality Metrices			Accuracy %
		Precision	Recall	F1-Score	
Random Forest	Glucose	0.74	0.82	0.78	84.74%
	Salt	0.61	0.64	0.62	
	Sugar	0.67	0.71	0.69	
SVM	Glucose	0.61	0.64	0.62	74.57%
	Salt	0.67	0.64	0.65	
	Sugar	1.00	1.00	1.00	
D-Tree	Glucose	0.67	0.71	0.69	78.81%
	Salt	0.73	0.69	0.71	
	Sugar	1.00	1.00	1.00	
KNN	Glucose	0.68	0.74	0.71	80.08%
	Salt	0.75	0.69	0.72	
	Sugar	1.00	1.00	1.00	

Table 2. Classification Performance of all three Classifiers using Tenfold Cross Validation

Performance of Random Forest by Applying Tenfold Cross Validation					
Solvent	Conc.	Quality Metrices			Accuracy %
		Precision	Recall	F1-Score	
Salt	5%	1.00	0.98	0.99	81.53%
	10%	1.00	0.0.98	0.99	
	20%	0.98	0.98	0.98	
	30%	0.94	0.98	0.96	
Sugar	5%	0.92	1.00	0.95	96.61%
	10%	0.80	1.00	0.88	
	20%	0.81	0.73	0.76	
	30%	0.81	0.82	0.81	
Glucose	5%	0.90	0.96	0.93	87.02%
	10%	0.99	0.96	0.97	
	20%	0.95	0.92	0.94	
	30%	0.99	0.97	0.98	

Table 3. Classification Performance of RF by Applying Tenfold Cross Validation

Here, precision metric is employed to evaluate the precision of one of the classifications relative to all other classifications. In addition, recall or sensitivity values shows the possibility of occurring accurate classification of categorised classes from the remaining classes. Finally, F1-score is employed to obtain the average between the Precision and Recall metrices. In this study, the key objective of using these commonly agreed metrices was primarily to detect any potential misclassification, resulting in inaccurate details about the presence of impurities in aqueous solutions[25].

Discussion

This section presents the metrics evaluation of classifiers technique using various feature selection techniques. It is perceived that after selecting the relevant features, execution time taken by classifiers for performing ten-fold cross-validation was considerably reduced. The ten-fold cross-validation is also more suitable for the given phenomena because the dataset is not very large and is often the reality with water quality datasets. In cross-validation, the data is separated into k subsets and is repeated overall the available datasets, given that K-1 subsets as training set and 1 subset as testing set. Though, due to iterations, this method is considered as computationally intensive technically challenging, however, it is seemingly suitable for the given data. Table 2 depicted the quality metrices performance for all proposed classifiers ranging from 0 to 1, indicating the estimation of impurities solutions detection added to the aqueous solutions. By analysing the results in Table 3, 4, 5, and 6 it can be noticed that RF showed a higher accuracy 84.74% for the identification of unknown ingredients in aqueous solutions followed by KNN, D-Tree and SVM, D-Tree and KNN, with 80.08%, 78.81%, and 74.57%, respectively. Intriguingly, the

Performance of SVM by Applying Tenfold Cross Validation					
Solvent	Conc.	Quality Metrices			Accuracy %
		Precision	Recall	F1-Score	
Salt	5%	1.00	0.84	0.91	86.59%
	10%	1.00	0.81	0.89	
	20%	0.88	0.90	0.88	
	30%	0.87	0.93	0.89	
Sugar	5%	0.86	0.97	0.92	93.11%
	10%	0.95	0.89	0.92	
	20%	0.94	0.98	0.96	
	30%	1.00	0.87	0.93	
Glucose	5%	0.92	0.95	0.94	96.88%
	10%	1.00	0.96	0.98	
	20%	0.89	0.95	0.92	
	30%	1.00	0.97	0.99	

Table 4. Classification Performance of SVM by Applying Tenfold Cross Validation

Performance of D-Tree by Applying Tenfold Cross Validation					
Solvent	Conc.	Quality Metrices			Accuracy %
		Precision	Recall	F1-Score	
Salt	5%	1.00	0.84	0.91	82.10%
	10%	1.00	0.78	0.88	
	20%	0.81	0.81	0.81	
	30%	0.62	0.85	0.72	
Sugar	5%	1.00	0.97	0.98	93.01%
	10%	1.00	0.97	0.98	
	20%	0.93	0.96	0.95	
	30%	0.93	0.97	0.95	
Glucose	5%	0.95	0.97	0.96	95.51%
	10%	1.00	0.96	0.98	
	20%	0.98	0.95	0.93	
	30%	1.00	0.97	0.99	

Table 5. Classification Performance of D-Tree by Applying Tenfold Cross Validation

assessment of metrices for the sugar displayed an effective performance, showing 1 for all classifiers except RF, revealing that sugar is compound of other ingredients. The obtained results by KNN model also shows adequate performance considering the absorption spectra of glucose and sugar in different concentrations as both glucose and sugar molecules broadened in aqueous solutions. Furthermore, despite the distinctive complexities and chemical dynamics emanating from biomolecular vibrations and constituents of distinct solvents, the evaluation of classifiers can be deemed as relatively efficient and is certainly above the alarming stage. Considering a real-life scenario, the proposed classifier methodology can be substantial by using the amalgamation of highly sensitive and good penetration feature of THz with ML approach to detecting the contagious contaminants in pure water

The proposed study, in addition to discovering unknown contaminants in aqueous solutions, also quantifies and unravel the estimate prediction of quantity of contaminants added in aqueous solutions. For this purpose, classifiers model was developed, and their efficiency was assessed using the quality metrices. Upon a close inspection of results attained in Table 3, 4, 5, and 6, it was noticed that RF showed relatively enhanced performance compared to other algorithms, showing 97%, 95% and 85.24% for salt, glucose and sugar concentrations, respectively. Also, despite having virtually molecular configuration and morphological structure of glucose and sugar, all classifiers as observed, proved to have reasonable predictions of different concentrations levels for both glucose and sugar in aqueous solutions, ranging from 87.02% to 96.01%. For salt concentrations, RF outperformed other algorithms, showing a prediction accuracy of 97.98% for distinct concentrations in aqueous solutions. Hence, it was concluded that the Random Forest yields considerable reliability and promising accuracy results in both scenarios, recognizing the substances as well as its precise concentrations solutions in aqueous solutions compared to other classifiers.

Performance of KNN by Applying Tenfold Cross Validation					
Solvent	Conc.	Quality Metrices			Accuracy %
		Precision	Recall	F1-Score	
Salt	5%	1.00	0.84	0.91	82.10%
	10%	1.00	0.77	0.87	
	20%	0.80	0.80	0.80	
	30%	0.61	0.85	0.71	
Sugar	5%	1.00	0.97	0.98	93.01%
	10%	1.00	0.97	0.98	
	20%	0.93	0.96	0.95	
	30%	0.93	0.97	0.95	
Glucose	5%	1.00	0.86	0.93	95.51%
	10%	1.00	0.87	0.80	
	20%	0.71	0.91	0.80	
	30%	0.85	0.84	0.84	

Table 6. Classification Performance of KNN by Applying Tenfold Cross Validation

However, some limitations can adversely affect the machine learning algorithms, resulting in degrading the overall performance. This unintended situation appeared to have rarely occurred because of selecting inadequate variables due to its high intricacy. Nonetheless, machine learning based models are still a feasible substitute to the physically dependent modelling in predicting the realistic scenarios, where small error can be fatal to the public health and safety. Keeping this mind, in this study, the strong aim of applying cross-validation technique was to evaluate the consistency of proposed classifiers by minutely assessing the absorption spectra characteristics of different substances concentrations in aqueous solutions, providing real-time monitoring of unknown substance, and can detect early symptoms of contamination's in water. Furthermore, these preliminary results obtained from the amalgamation of ML with THz waves have the potential to curtail any microbiological contaminants in aqueous solutions and mitigate their harmful effects on human health.

Conclusion

In this research study, the use of non-invasive THz feature and ML enabled optimized technological solution was presented to detect various substances and their distinct concentrations in aqueous solutions. In this process, the FTIR system measured the absorption spectra and characteristics of salt, sugar and glucose solutions with varying levels for two hours and collected 338 data points for every specimen and regarded them as features. Since the observations were recorded at laboratory, there might be the possibility of some distortion in measurements. To prevent this, we performed features selection to discard any spurious that may yield forged observations of substance concentrations in aqueous solutions, given the public protection. The selection of meaningful and significant features drastically enhanced the classifier performance for detecting the substance solutions in aqueous solutions. Furthermore, the comprehensive cross-validation methodology exhibited in most cases, RF model showed reliability, and achieved highest classification accuracy in identifying the salt solutions and its quantity in aqueous solutions, compared to other classifiers. Moreover, KNN, D-Tree and SVM displayed substantial performance particularly for sugar and glucose concentrations in aqueous solutions.

These preliminary results showed a notable relationship of THz waves with machine learning (ML) techniques. It also fully reveals the significant influence of ML and its process reliability in terms of detecting the substance solutions as well as their concentrations in aqueous solutions. The outcome of this work has the potential to a play vital role by providing unprecedented and cost-effective opportunity in real-time monitoring to enhance a detection of impurities in water and potentially contribute to the protection of public health.

Methods

Setup

In this experimental setup, a Bruker 66 V/S series FTIR system was employed to accomplish the measurement of various constituent's in aqueous solution as shown in Fig. 3[26,27]. FTIR is a powerful analytical technique, providing a label-free, non-destructive method to show a rapid behaviour for any redundant impurities in solutions. The system was equipped with a DLaTGS/Polyethylene (PE) detector and a 6-micron Mylar beam splitter was employed to perform the measurements[26,27]. The spectral distribution of the beam-splitter and detector was 1–21 THz, and 0.3–21 THz, respectively[26]. To prevent any formation

Sample	Weight in Grams	Concentrations (%)
Table Salt (NaCl)	2.63g $\pm$ 0.1	5% $\pm$ 0.1
	5.56g $\pm$ 0.1	10% $\pm$ 0.1
	12.5g $\pm$ 0.1	20% $\pm$ 0.1
	21.4g $\pm$ 0.1	30% $\pm$ 0.1
White Granulated Sugar	2.63g $\pm$ 0.1	5% $\pm$ 0.1
	5.56g $\pm$ 0.1	10% $\pm$ 0.1
	12.5g $\pm$ 0.1	20% $\pm$ 0.1
	21.4g $\pm$ 0.1	30% $\pm$ 0.1
Glucose	2.63g $\pm$ 0.1	5% $\pm$ 0.1
	5.56g $\pm$ 0.1	10% $\pm$ 0.1
	12.5g $\pm$ 0.1	20% $\pm$ 0.1
	21.4g $\pm$ 0.1	30% $\pm$ 0.1

Table 7. Preparation of various specimens including salt, sugar and glucose concentrations to observe the THz response

of the vacuum in sample compartment, and absorption of THz power, Nitrogen (N_2) gas was externally purged into the sample compartment. However, some variations in different specimens was noticed and reasons is explained in further section. The flowrate of (N_2) was set to 600 L/hr as specified in FTIR system guidelines[26]. Owing to high transparency and zero losses in the THz spectrum, a distinct device Polymethylpentene (PMP) commonly known as TPX tube was employed for testing the various specimens. Considerable attention and precautionary measures were taken to ensure that the sampling device is positioned in the same location every time in sample compartment to reduce the distortion in measurements[27].

Sample Preparation

In this study, three various specimens were considered for measurements including table salt, pure sugar, and glucose as described in Table 7. Salt and sugar were bought from Holland Barrett glucose was ordered from Sigma-Aldrich, respectively. To prepare a solution of every solvent, a 50ml of distilled water was taken and mixed with different concentrations of salt, sugar and glucose such as 5%, 10%, 20% and 30% using (1)[28,29]:

$$Conc = \frac{\text{gram}}{\text{gram} + 50\text{ml}} \times 100\% \tag{2}$$

These solutions were prepared at room temperature set to 23 °C by adding 2.63g, 5.5g, 12.5g and 21.4g to 5%, 20%, 30% and 30%, respectively. The weights of all specimens were carefully calculated using an electronic scale with at least count of 0.1mg. Before placing mixture into the TPX tube, solutions were properly stirred for 3 to 5 minutes approximately to ensure they are being fully dissolved in distilled water. While filling the TPX tube filled with all solutions, great attention was given so that it should be filled up to 11.6ml just in lined to beam-splitter to obtain the maximum and accurate information. All the measurements were performed at an atmospheric temperature of 23 °C.

References

1. Alvarez, P., Chan, C., Elimelech, M., Halas, N. & Villagrán, D. Emerging opportunities for nanotechnology to enhance water security. *Nat. Nanotechnol.* **13**, 634–641, DOI: 10.1038/s41565-018-0203-2 (2018).
2. Zulkifli, S. N., Rahim, H. A. & Lau, W.-J. Detection of contaminants in water supply: A review on state-of-the-art monitoring technologies and their applications. *Sensors Actuators B: Chem.* **255**, 2657 – 2689, DOI: https://doi.org/10.1016/j.snb.2017.09.078 (2018).
3. Shalit, A., Ahmed, S., Savolainen, J. & Hamm, P. Terahertz echoes reveal the inhomogeneity of aqueous salt solutions. *Nat. Chem.* **9**, 273–278, DOI: 10.1038/nchem.2642 (2017).
4. Najah Ahmed, A. *et al.* Machine learning methods for better water quality prediction. *J. Hydrol.* **578**, DOI: 10.1016/j.jhydrol.2019.124084 (2019).
5. Son, J.-H. Terahertz electromagnetic interactions with biological matter and their applications. *J. Appl. Phys.* **105**, 102033, DOI: 10.1063/1.3116140 (2009). https://doi.org/10.1063/1.3116140.
6. Song, C. *et al.* Terahertz and infrared characteristic absorption spectra of aqueous glucose and fructose solutions. *Sci. Reports* **8**, 2–9, DOI: 10.1038/s41598-018-27310-7 (2018).

7. Liakos, K. G., Busato, P., Moshou, D., Pearson, S. & Bochtis, D. Machine learning in agriculture: A review. *Sensors* **18**, DOI: 10.3390/s18082674 (2018).

8. Zahid, A. *et al.* Characterization and water content estimation method of living plant leaves using terahertz waves. *Appl. Sci.* **9**, DOI: 10.3390/app9142781 (2019).

9. Zahid, A. *et al.* Machine learning driven non-invasive approach of water content estimation in living plant leaves using terahertz waves. *Plant Methods* **15**, 1–13, DOI: 10.1186/s13007-019-0522-9 (2019).

10. Federici, J. F. *et al.* THz imaging and sensing for security applications - Explosives, weapons and drugs. *Semicond. Sci. Technol.* **20**, DOI: 10.1088/0268-1242/20/7/018 (2005).

11. Woodward, R. M. *et al.* Terahertz pulse imaging in reflection geometry of human skin cancer and skin tissue. *Phys. Medicine Biol.* **47**, 3853–3863, DOI: 10.1088/0031-9155/47/21/325 (2002).

12. Yin, L., Chen, F., Zhang, Q. & Ma, X. Arrhythmia classification based on multi-domain feature extraction. *J. Physics: Conf. Ser.* **1237**, 022062, DOI: 10.1088/1742-6596/1237/2/022062 (2019).

13. Chen, H. *et al.* Quantify Glucose Level in Freshly Diabetic's Blood by Terahertz Time-Domain Spectroscopy. *J. Infrared, Millimeter, Terahertz Waves* **39**, 399–408, DOI: 10.1007/s10762-017-0462-2 (2018).

14. Mittleman, D. M., Jacobsen, R. H. & Nuss, M. C. T-ray imaging. *IEEE J. Sel. Top. Quantum Electron.* **2**, 679–692 (1996).

15. Siuly, Yin, X., Hadjiloucas, S. & Zhang, Y. Classification of thz pulse signals using two-dimensional cross-correlation feature extraction and non-linear classifiers. *Comput. Methods Programs Biomed.* **127**, 64 – 82, DOI: https://doi.org/10.1016/j.cmpb.2016.01.017 (2016).

16. Dutta, S., Chatterjee, A. & Munshi, S. Correlation technique and least square support vector machine combine for frequency domain based ecg beat classification. *Med. engineering & physics* **32**, 1161—1169, DOI: 10.1016/j.medengphy.2010.08.007 (2010).

17. Thanh Noi, P. & Kappas, M. Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using sentinel-2 imagery. *Sensors* **18**, DOI: 10.3390/s18010018 (2018).

18. Hao, J. & Ho, T. K. Machine learning made easy: A review of scikit-learn package in python programming language. *J. Educ. Behav. Stat.* **44**, 348–361, DOI: 10.3102/1076998619832248 (2019). https://doi.org/10.3102/1076998619832248.

19. Saçlı, B. *et al.* Microwave dielectric property based classification of renal calculi: Application of a knn algorithm. *Comput. Biol. Medicine* **112**, 103366, DOI: https://doi.org/10.1016/j.compbiomed.2019.103366 (2019).

20. Hu, L. Y., Huang, M. W., Ke, S. W. & Tsai, C. F. The distance function effect on k-nearest neighbor classification for medical datasets. *SpringerPlus* **5**, DOI: 10.1186/s40064-016-2941-7 (2016).

21. Li, K., Gu, Y., Zhang, P., An, W. & Li, W. Research on knn algorithm in malicious pdf files classification under adversarial environment. In *Proceedings of the 2019 4th International Conference on Big Data and Computing*, ICBDC 2019, 156–159, DOI: 10.1145/3335484.3335527 (Association for Computing Machinery, New York, NY, USA, 2019).

22. Shaikhina, T. *et al.* Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. *Biomed. Signal Process. Control.* **52**, 456 – 462, DOI: https://doi.org/10.1016/j.bspc.2017.01.012 (2019).

23. Pohjalainen, J., Räsänen, O. & Kadioglu, S. Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits. *Comput. Speech & Lang.* **29**, 145 – 171, DOI: https://doi.org/10.1016/j.csl.2013.11.004 (2015).

24. Urbanowicz, R. J., Olson, R. S., Schmitt, P., Meeker, M. & Moore, J. H. Benchmarking relief-based feature selection methods for bioinformatics data mining. *J. Biomed. Informatics* **85**, 168 – 188, DOI: https://doi.org/10.1016/j.jbi.2018.07.015 (2018).

25. Gibson, R. M., Amira, A., Ramzan, N., de-la Higuera, P. C. & Pervez, Z. Multiple comparator classifier framework for accelerometer-based fall detection and diagnostic. *Appl. Soft Comput.* **39**, 94 – 103, DOI: https://doi.org/10.1016/j.asoc.2015.10.062 (2016).

26. Saha, S. C. *et al.* Terahertz frequency-domain spectroscopy method for vector characterization of liquid using an artificial dielectric. *IEEE Transactions on Terahertz Sci. Technol.* **2**, 113–122, DOI: 10.1109/TTHZ.2011.2177172 (2012).

27. Saha, S. C. Application of terahertz spectroscopy to the characterization of biological samples using birefringence silicon grating. *J. Biomed. Opt.* **17**, 067006, DOI: 10.1117/1.jbo.17.6.067006 (2012).

28. Ren, A. *et al.* Machine learning driven approach towards the quality assessment of fresh fruits using non-invasive sensing. *IEEE Sensors J.* **20**, 2075–2083 (2020).

29. Ren, A. *et al.* Terahertz sensing for fruit spoilage monitoring. In *2019 Second International Workshop on Mobile Terahertz Systems (IWMTS)*, 1–4 (2019).

Acknowledgements

This research was funded under EPSRC DTA studentship which is awarded to A. Z. for his PhD.

Author contributions statement

Conceptualization, A. Z., K. D., and Q. H. A.; software, A. Z., K. D, and A. R.; resources, I. E. C., D. R. S. C, and J. G.; writing-original draft preparation, A. Z. and K. D.; writing-review and editing, A. Z., H.T.A, A. R., M. A. I., and Q. H. A.; supervision, Q. H. A. and M. A. I.; project administration, Q. H. A. All authors read and approved the final manuscript.

Additional information

Competing interests: The authors declare no competing interests.