Prepublished copy

Nature Environment & Pollution Technology

This is a peer-reviewed prepublished version of the paper

Prediction on the Level of Toxicity in Fruits and Vegetables based on PAHs using Machine Learning

Staphney Texina [1†], Sathees Kumar Nataraj [2], Alagammai Renganathan[1] and Kavitha Vasantha[2]

[1]Department of Math and Science, University of Technology Bahrain, Salmabad, Bahrain.

[2] Department of Mechatronics Engineering, University of Technology Bahrain, Salmabad, Bahrain.

[†]Corresponding author(s). E-mail(s): texinastaphney@gmail.com

Abstract

This study focuses on assessing the toxicity levels in fruits and vegetables based on the presence of polycyclic aromatic hydrocarbons (PAHs), particularly in regions affected by industrial and vehicular pollution where the particulate matter deposits on the plant surfaces. Traditional methods, including Gas Chromatography/Mass Spectrometry (GC/MS) and High-Performance Liquid Chromatography (HPLC), are used to measure PAH levels in fruits and vegetables which are found to be valuable but expensive and time-consuming. Although, the detection of toxicity relies on either expert knowledge or experimental analysis when compared with the limitations set by EFSA (European Food Safety Authority). Therefore, in this study, artificial intelligence techniques have been employed to evaluate the toxicity levels based on 16 PAHs. The PAHs concentrations in fruits and vegetables were collected from different articles corresponding to safe and unsafe dataset, then validated through statistical analysis. The validated dataset is classified using different machine learning algorithms. Based on the output from neural network, the level of toxicity is also scaled and compared with the targeted outputs. The promising results of the classification of toxicity using artificial intelligence methods are substantiated by an experimental study and validated through statistical methods. From the results, it can be observed that the machine learning algorithm has given classification accuracy more than 90% along with their degree of harmfulness. This research holds implications for food safety and public health, offering a novel approach to the interdisciplinary understanding of climate change by addressing the impact of environmental contaminants on the edibility of fruits and vegetables.

Keywords: Polycyclic aromatic hydrocarbons (PAHs), Environmental contaminants, Fruits & Vegetables, Statistical measures, Machine learning algorithm

Key Words	Polycyclic aromatic hydrocarbons (PAHs), Environmental contaminants, Fruits & Vegetables, Statistical measures, Machine learning algorithm
DOI	https://doi.org/10.46488/NEPT.2025.v24i02.D1690 (DOI will be active only after the final publication of the paper)
Citation of the Paper	Staphney Texina, Sathees Kumar Nataraj, Alagammai Renganathan and Kavitha Vasantha, 2025 Prediction on the Level of Toxicity in Fruits and Vegetables based on PAHs using Machine Learning. *Nature Environment and Pollution Technology*, 24(2), D1690. https://doi.org/10.46488/NEPT.2025.v24i02.D1690

1 Introduction

In recent years, there has been a growing concern about the impact of polycyclic aromatic hydrocarbons (PAH) on both environmental and public health (Abdel-Shafy & Mansour, 2016). These contaminants, which are generated from various anthropogenic and natural sources, have been found to have adverse

effects on ecosystems and human well-being. Multiple studies have shown that exposure to polycyclic aromatic hydrocarbons can harm human health, especially for vulnerable populations such as children, older adults, and individuals with existing health problems (Mallah et al., 2022; Organization, 2021; Singh & Agarwal, 2018). From the literatures, it is observed that fruits and vegetables are consumed in different forms for their nutritional values but, the growth of these fruits and vegetables are contaminated through pollution in different forms which results in the adsorption of Polycyclic Aromatic Hydrocarbons (Camargo & Toledo, 2003). Recent studies have shown that PAH contamination has an impact on public health, and mostly observed in urban areas due to the emission of PAHs from automobiles and cooking oil fumes. In Brazil, a case study on the impact of PAH contamination was examined in street food vendors that resulted in potential health risks such as diabetes, oxidative stress, cardiovascular and pulmonary disease, respiratory diseases, skin allergies and cancer among individuals(Deligannu & Muniandy, 2024). A study on the potential health risk due to PAH exposure in industrial areas was conducted in India. Soil samples were collected and assessed from two cities. PAH sources were identified as traffic emissions, industrial emissions and coal combustion for domestic livelihood. The health risk assessment resulted in a high potential risk of cancer due to the consumption of contaminated vegetables from these areas(Sankar et al., 2023). In China, a study conducted on a farmland indicated the presence of 16 PAHs in soil and crops with varying concentrations. It showed that leafy vegetable crops had higher PAH concentration in leaves compared to the roots and fruits whereas the fruit and vegetable crops showed higher PAH concentration in fruits than in roots or leaves(Cui et al., 2022). During the health risk assessment, it posed high carcinogenic risk in adult males and females based on the dietary intake.

From these studies, it is evident that PAH's analysis on consumables is necessary to be studied and detection of toxicity should be considered as an important measure to protect the environment. Contrarily, the toxicity of food products or consumables may be a potential threat to mankind putting individual lives at high health risk. This resulted us to initiate exploring and evaluating PAHs, as well as developing an intelligent system for detecting toxicity. PAHs are formed during the incomplete combustion of garbage, organic waste, sewage sludge, wood, gas etc. PAHs are composed of carbon and hydrogen atoms and contain two or more aromatic rings (Khalili et al., 2021a). The contamination of PAH is widespread in the environment both in terrestrial and aquatic organisms due to which the presence of PAH in food supply is considerably high(Paris et al., 2018).The contamination of PAH in agricultural and animal food products can occur during growth, transportation (exhaust from combustion engines) , storage and also when the food is smoked, grilled, roasted, fried and cooked (Paris et al., 2018). While there are over 100 recognized PAHs, the United States Environmental Protection Agency (UPESA) has identified only 16 as the primary concern(Abou-Arab et al., 2014) because these PAHs are unsafe and can enter variety of life on earth through inhalation, ingestion, and even through skin contact (Omoyeni et al., n.d.). PAH contamination in raw food such as fruits and vegetables are through soil, water and air(Paris et al., 2018). In addition to this, the amount of PAH concentration depends on environmental PAH (urban areas have high amount of PAH), soil characteristics (weak soil needs to be strengthened using chemical fertilizers) and physiological properties e.g., the longer the growth period of the plant, higher the absorption of PAH contaminants(Khalili et al., 2021a). Fruits and vegetables can get contaminated with PAHs when air particulate matter settle on their surfaces. Plants near industries or roads tend to have more PAH deposits, including Benzo[a]pyrene, dibenz[a,h]anthracene, and chrysene, compared to plants in rural areas(Ashraf & Salam, 2012). In fruits and vegetables, low molecular weight (LMW) PAHs and high molecular weight (HMW) PAHs are adsorbed by the waxy surface, particularly on outer leaves and fruit peels(Camargo & Toledo, 2003). The concentration of PAHs tends to be higher on these exposed surfaces. Studies reveal variations in PAH concentrations among different parts of plants, with root vegetables potentially having higher levels than stem vegetables(Zhong & Wang, 2002). Research in China identified factors affecting PAH levels in vegetables, including anthropogenic emissions, vegetable species, and wind direction(Jia et al., 2018). Common PAHs found in fruits and vegetables include fluorene, fluoranthene, pyrene, anthracene, phenanthrene, benzo(a)anthracene, and benzo(a)pyrene, with leafy and stem vegetables having higher concentrations(Zhong & Wang, 2002). Similarly, C. Choochuay (Choochuay et al., 2023) have analyzed the toxicity and health risk assessment based on the PAH concentration in Thai and Myanmar rice. From this study, it is identified that the level of PAHs with its toxicity and health risk assessment. The findings can be summarized as follows: a) The level

of PAHs in Thailand varied from 0.09 - 37.15 ng.g^{-1} with an arithmetic mean of 18.22 ± 11.76 ng.g^{-1}, whereas that in Myanmar varied from 0.07 – 150.73 ng.g^{-1} with an arithmetic mean of 34.70 ± 40.57 ng.g^{-1}.Due to increased food security concerns, numerous studies explore threats associated with consuming contaminated food(Abou-Arab et al., 2014).

In a 2021 study conducted by Khalil F et al. in Iran, the analysis of PAHs in fruits and vegetables revealed high concentrations of acenaphthene (135.1 ± 7.1 µg/kg) and naphthalene (114.1 ± 5.0 µg/kg), while benzo(k)fluoranthene, benzo(a)pyrene, benzo(g,h,i)fluoranthene, Indeno(1,2,3-cd)pyrene, and benzo(g,h,i)perylene were not detected(Khalili et al., 2021a). Another study by Alice Paris et al. in 2018 reported relatively low PAH levels ranging from 0.01 to 0.5 µg/kg in wet weight for fruits and vegetables(Paris et al., 2018). However, plants near roadways and urban areas can exceed the concentration of 5 µg/kg(Paris et al., 2018).

The experiments conducted in Pakistan and Saudi Arabia in 2013 by Mohammad W. Ashraf, root vegetables like carrot and potato exhibited high PAH concentrations of 13 µg/kg and 11 µg/kg, respectively, while turnip had concentrations of 10.9 µg/kg and 9.26 µg/kg. The study also observed higher contamination in the peels than the cores of fruits and vegetables, with cabbage having the highest concentration among leafy vegetables (Ashraf et al., 2013), (Ashraf, n.d.; Ashraf & Salam, 2012). In India, a study by Bishnoi N et al. in 2006 identified Anthracene, Naphthalene, Fluorene, Pyrene, Phenanthrene, and Fluoranthene as predominant PAHs in vegetable and soil samples. The use of an Isocratic High-Performance Liquid Chromatography (HPLC) system with UV detection revealed carcinogenic compounds such as BAP and dibenz(a,h)anthracene, with LMW-PAHs more abundant than HMW-PAHs(Narsi.R.Bishnoi et al., 2006).

A study in Jordan by Farh Al-Nasir in 2022 evaluated four vegetables, finding tomatoes with the highest concentration of 21.774 µg/kg and zucchini with the lowest concentration of 10.649 µg/kg(Al-Nasir et al., 2022a). In summary, the literature also emphasizes the use of various detection methods, including High-Performance Liquid Chromatography with fluorescence detection (HPLC-FLD) an excellent quantification and separation tool, Solid phase microextraction (SPME) a sensitive solvent-free sample preparation technology, Gas Chromatography with Mass Spectrometry (GC-MS) a method where two analytical tools combined to identify and measure the concentration of chemicals in food and environment, and Gas Chromatography with flame ionization detector (GC-FID) an analytical technique that is used to separate and analyse mixtures consisting of volatile compounds (Abou-Arab et al., 2014).

Similar to these technologies, chemical analysis have also been done using various techniques such as saponification/ ultrasonication, clean-up using a silica solid phase extraction cartridge and GC-MS, liquid-liquid extraction with solvents like n-hexane to determine the elements of eight PAHs (BaA, BkF, BbF, DahA, BaP, BghiP, IP, Chry) in fruits and vegetables. The results show that the PAH concentration in fruits is 0.67µg/kg and in vegetables it is 0.82µg/kg (Lee et al., 2019). In Egypt(Abou-Arab et al., 2014), a study made on the level of PAH in vegetables and fruits such as potatoes, spinach, apple and guava using GC-MS, showed high level concentration in spinach (8.977µg/kg), potatoes (6.196µg/kg), apple (2.867µg/kg) and guava (2.334µg/kg). The researcher concluded with preventive measures such as thorough washing, boiling, and peeling of skin of fruits and vegetables is effective in reducing the amount of PAH consumption(Abou-Arab et al., 2014). Okaba, Fidelis A,2020, reported that vegetables grown in Nigerian traffic routes were tested for PAH concentration and evaluated using GC-MS and AAS (Atomic Absorption Spectrophotometer) which determined the presence of high PAH in vegetables(Okaba et al., 2020). Although vegetables were boiled, it did not show notable difference ($p>0.05$) in the PAH concentration between fresh and boiled vegetables. Boiling the vegetables only reduced the mean concentration of PAHs(Okaba et al., 2020). Another study determined the concentration of PAH by growing plants in contaminated and uncontaminated soil, the results showed elevated levels of PAH in vegetables and fruits grown in contaminated soil(Samsøe-Petersen et al., 2002a). In samples of tomatoes and okra, the Σ16 PAH concentration was in the range of 2.12±1.5 and. 99.88±29.18 respectively. Also, naphthalene exhibited high concentration of 60% in vegetables (Omoyeni et al., n.d.) , (Tesi et al., 2021). The concentration of Σ16 PAH in vegetables was in the range of 532 to 2261 in leafy vegetables of southern Nigeria (Tesi et al., 2021). Ce-Hui Mo, 2009 reported that the determination of PAH and PAE (Phthalic Acid Esters) in vegetables in South China, indicated that PAE was present in higher amount than total PAH. However, due to the seasonal changes of PAHs in vapor and particulate matter in the region, more study is to be done to test variations of PAHs in various classes of vegetables (Mo et al., 2009). Therefore, from the review it can be observed that the analysis of PAH, and their effects on the health and environment are alarming. There are many studies on the identification of hazards and toxicity using machine learning algorithms; however,

there are no research on the statistical measures of the PAHs of the samples or measuring the PAH levels (Al-Nasir et al., 2022b; Khalili et al., 2021b; Samsøe-Petersen et al., 2002b). There are some studies such as, Vinay Kumar Pandey et.al, (2023) have used machine learning algorithms in the applications of food processing industry to identify the hazards associated with preservation of fruits and vegetables (Pandey et al., 2023a). PAHs are present in various fruits and vegetables due to factors like location, agricultural practices, and storage. This indicates the widespread presence of PAHs, necessitating monitoring. It is also observed from WHO, under the natural toxins in food (WHO/V. Gupta-Smith, 2023), stated that research experts review all the available study and suggested with an outcome based on level of health concern, which includes measures to prevent and control contamination. The authors have provided a detailed discussion on the future of machine learning algorithms in the food industry, the factors that affect the quality of food being preserved and assist in determining the optimal parameter combinations for deciding the maximum produce preservation.

Rajesh Megalingam et al., employed different machine learning like k- cluster, computer vision and artificial intelligence techniques along with colour classification to determine rotten food (Megalingam et al., 2019). Therefore, it can be observed from the above literature that PAHs adhere to the surfaces of fruits and vegetables, particularly on outer leaves and peels. It highlights variations in concentration of PAH among different plant parts and factors influencing PAH levels. While there are multiple methods to analyse the PAH values, artificial intelligence techniques have a potential to outreach in the field of toxicity detection. Analysing the PAH values using machine learning algorithm is one of the initial works to determine the toxicity of PAH in fruits and vegetables. It can also be perceived that Machine learning algorithms and various AI (Artificial Intelligent) techniques are used to examine the perishing nature of food, but there is no research conducted to measure the toxicity level nor predict the degree of harmfulness of PAH in fruits and vegetables. Hence, in this study an intelligent toxicity detection system has been developed to explore the impact of PAH toxicity in fruits and vegetables. Machine learning algorithms have been used to analyze PAH contamination (Vasantha et al., 2023) in fruits and vegetables, providing an efficient and accurate monitoring method. Machine learning algorithms can handle complex data and detect even trace levels of PAHs in which traditional methods have some limitations. By training models on historical data, the intelligent model can be used to predict contamination trends and circumstances for proactive measures ahead. Incorporating recent data and case studies highlights the critical issue of PAH contamination. Thus, the proposed system on toxicity detection can be helpful to society, ensuring food safety and protecting public health.

The proposed system depicted in Figure 1, utilizes machine learning algorithms to analyse the collected empirical data and provide results on toxicity. A detailed explanation on the implementation of the proposed system have been discussed in the subsequent sections of this article, where section 2.1 describes data collection, section 2.2 elaborates on the statistical analysis of the collected data followed by section 2.3which provides the design and evaluation of the machine learning algorithms used in this research and the results of the proposed system are also discussed in section 2.4.

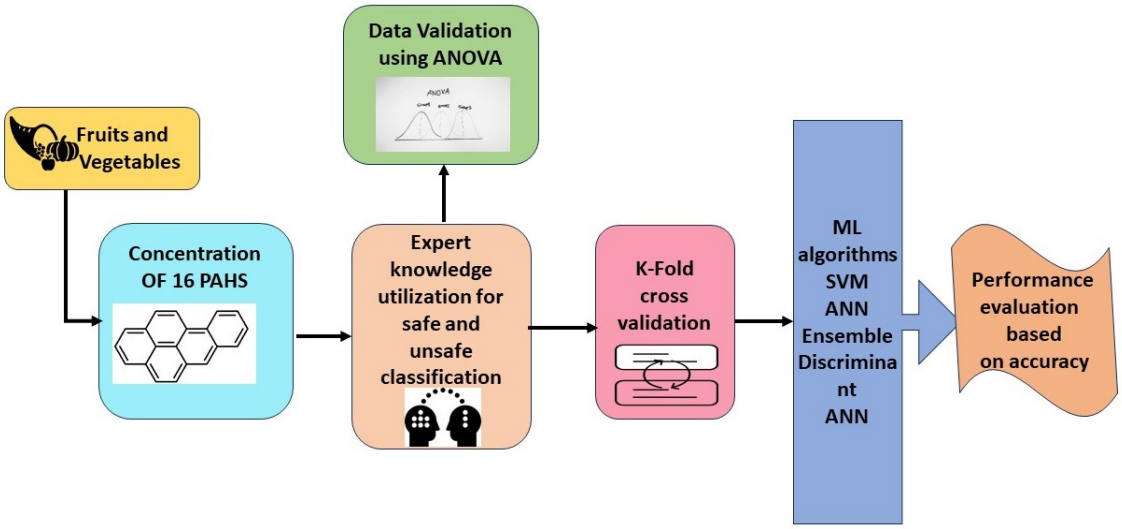

Fig. 1 The Proposed PAH based Toxicity Detection Model

2 Methodology

The main aim of this study is to employ artificial intelligence techniques for assessing the toxicity levels of fruits and vegetables based on 16 PAHs. A classification system is proposed, utilizing the machine learning algorithms such as Support Vector Machine (SVM), Ensemble, Regression, Discriminant, Tree, k-Nearest Neighbour(k-NN), Naïve Bayes, Artificial Neural Network (ANN) to classify the toxicity in fruits and vegetables. Additionally, the research aims to compare the outcomes of these models.

The classification model takes the concentration of 16 PAHs in fruits and vegetables as input and classifies the level of toxicity based on the machine learning algorithm. The subsequent section details the data collection process, it's validation, design and evaluation using machine learning algorithms.

2.1 PAH Data Collection,

The data collected for this study is based on the experimental analysis from different research in the field of environmental pollution, environmental contamination and toxicology, polycyclic aromatic compounds, toxic chemical hazards in food and feed (Paris et al., 2018).From the literature, it can be summarized that PAH deposit is found more on the surface of the fruits, leaves and vegetables than the inner tissues. As stated in WHO, "natural toxins need to be kept as low as possible to protect people". Therefore, in this study fruits and vegetables were considered for the toxicity detection using AI techniques. The PAHs corresponding to the proposed objectives were collected from 24 articles published in various platforms such as IEEE, Nature Environment and Pollution Technology, Elsevier, International Journal of Nutrition and Food sciences, MDPI, Journal of Environmental Science and Health and others. The total number of samples are 519. These samples represent different fruits and vegetables, and these are experimented from various parts of the world. Therefore, the data relating to concentration of PAH on leafy vegetables like spinach, jute and pumpkin leaves(Camargo & Toledo, 2003), (Khalili et al., 2021a), (Omoyeni et al., n.d.), (Tesi et al., 2021), (Mo et al., 2009), (Janska J et al., 2006), as well as small, medium and large sized vegetables(Khalili et al., 2021a), (Zhong & Wang, 2002), (Ashraf et al., 2013)(Ashraf, n.d.)(Al-Nasir et al., 2022a), (Lee et al., 2019), (Tuteja et al., 2011) and fruits (Camargo & Toledo, 2003), (Khalili et al., 2021a), (Narsi.R.Bishnoi et al., 2006)(Janska J et al., 2006), other leafy vegetables (romaine lettuce, Chinese cabbage and Shanghai green cabbage), stem vegetables (lettuce), seed and pod vegetables (broad bean), rhizome vegetables (daikon) were considered in this research. PAHs concentration of samples are collected from different experimental results reported in research articles are summarized in Table 1.

Table 1 Different Fruits and Vegetables for PAH Analysis

Categories of fruits & vegetables	No. of samples	References
Leafy vegetables	122	(A. Ramezan et al., 2019), (Samsøe-Petersen et al., 2002a)(Al-Nasir et al., 2022c), (Lee et al., 2019), (Mo et al., 2009), (Paris et al., 2018)
Root vegetables	108	(A. Ramezan et al., 2019), (Ashraf & Salam, 2012), (Ashraf et al., 2013)
Stem vegetables	157	(Al-Nasir et al., 2022c), (Ashraf, n.d.), (Ashraf & Salam, 2012), (Ashraf et al., 2013), (Janska J et al., 2006), (Jia et al., 2018)
Fruits	132	(A. Ramezan et al., 2019), (Samsøe-Petersen et al., 2002a), (Camargo & Toledo, 2003), (Paris et al., 2018)
Total	519	

The PAHs of fruits and vegetables collected in this analysis include Acenaphthene (Ace), Acenaphthylene(Aceph),Anthracene(An), Benzo[b] fluoranthene (BbF),

Benzo[g,h,i]perylene(BgP),Benzo[k]fluoranthene(BkF),Chrysene(Chr),Dibenz[a,h]anthracene(DBA),Fluoranthene(Flu),Fluorene(Fl),Indeno[1,2,3-c,d] pyrene(Inp), Phenanthrene(Ph), Pyrene(Pyr) and Naphthalene (Nfl). Among these PAHs, Scientific Committee on Food, European Food Safety Authority (EFSA) which is an agency that provides scientific advice to risk managers and communicates the risk associated with food chain, considers BaP, DBahA, BaA, BbF, BjF, BkF, CHR, BghiP, and IP as potentially carcinogenic and genotoxic compounds (Paris et al., 2018).

According to US Environmental protection agency (USEPA), fruits and vegetables have lesser concertation of PAH when compared to processed and unprocessed meat and meat products. The minimum and maximum recommended limit is 0.01 and 0.5 µg/kg (Paris et al., 2018). Based on the Maximum Contamination limit and expert knowledge, the 519 data have been sorted as safe and unsafe for consumption. The total number of data corresponding to safe and unsafe are 231 and 288 respectively. The segregated dataset is validated using ANOVA and the outcomes are reported in the subsequent section.

2.2 Statistical Analysis on PAHs

The examination of the 16 PAHs across the 519 samples reveals a non-linear pattern, making it challenging to determine any possible significance through visual inspection for categorizing the data as safe or unsafe. Consequently, adhering to established standards and limitations for specific PAHs, the data was categorized into safe and unsafe. This research endeavours to validate whether there is any significant difference between these segmented datasets(Frossard & Renaud, 2021), (Yang et al., 2020). In this study, the data collection techniques or experiments were not used to measure the PAHs. Instead, the toxicity of PAHs was assessed in fruits and vegetables using machine learning methods. The data were acquired from a variety of articles that use experimental measures. As a result, there is no possibility of missing data. In addition, the dataset has also been processed using Analysis of variance (ANNOVA) to access the level of significance. The preprocessing findings and variance analysis are reported in the subsequent sections.

In the initial analysis, the 16 PAHs of both safe and unsafe datasets were subjected to ANOVA statistical analysis, but the results failed to meet the required hypothesis. The hypothesis for the validation process is as follows:

1. Null hypothesis is that there is no significant difference among the samples.
2. Whereas the alternate hypothesis is that at least a sample should differ significantly from other samples.

The level of significance considered is 0.05. The null hypothesis will fail to accept if the probability value that is, p value is less than 0.05. The results indicated a significant difference between the safe and unsafe datasets, contradicting the latter hypothesis. Further investigation revealed that the variance stemmed from the missing PAHs in some samples, as researchers focused on measuring the main PAHs to determine toxicity. To address this, various statistical measures were employed to ascertain if a regression line adequately fits the data. The validation in this analysis involved calculating the sum of square values of the PAHs. The results of each analysis are detailed below.

From the ANOVA analysis, to determine the significance between safe and unsafe, the F value for the 519 samples resulted in 1.3067 and the p-value is 0.001 which evidently confirmed that there is a significant difference between the safe and unsafe data. Hence, the determination of toxicity in fruits and vegetables by expert knowledge and the standard limit has been validated by analysis of variance. The results of variance analysis done on the data set are consolidated in Table 2.

Table 1 ANOVA Measures Corresponding to Safe and Unsafe Data

Source of Variation	SS	df	MS	F	P-value	F crit
Between Samples	1E+09	518	2E+06	1.3067	0.0012	1.1556
Within Samples	9E+08	519	2E+06			

From the ANOVA analysis, to determine the significance between the samples of safe data, the F-value for 231 samples resulted in 0.6794 and the p-value is 0.9982. As the level of significance is more than the threshold limit, samples of the safe category fail to differ significantly from each other. The test results are

presented in Table 3.

Table 2 ANOVA Measures between Samples of Safe Data

Source of Variation	SS	df	MS	F	P-value	F crit
Between Samples	123.87	230	0.5386	0.6794	0.9983	1.2425
Within Samples	183.12	231	0.7927			

Similarly, from the ANOVA analysis, to determine the significance between the samples of unsafe data, the F-value for 288 samples resulted in 1.2918 and the p-value is 0.0151. As the p value is less than 0.05, samples of unsafe category differ significantly from each other. Therefore, it can be concluded that there is a significant difference between the concentration of PAHs of the samples corresponding to unsafe categories as the samples are collected from fruits and vegetables grown in different regions across the globe which are subjected to different environmental conditions like temperature variations, air pollution, water and soil quality.
. The test results are presented in Table 4.

Table 4 ANOVA Measures between Samples of Unsafe Data

Source of Variation	SS	df	MS	F	P-value	F crit
Between Samples	1E+09	287	4E+06	1.2918	0.0151	1.2145
Within Samples	9E+08	288	3E+06			

From the above statistical analysis, it has been observed that there is no significant difference between the samples for the safe data set whereas for the unsafe data set, there is a significant difference between the samples. This is due to a large variation between the PAH of fruits and vegetables, with scattered concentration of PAHs across the dataset. The minimum and maximum values of such PAHs are tabulated in Table 5. Therefore, the unsafe data has been analysed further to understand the distribution of PAHs in each sample of fruits and vegetables.

Table 5 Concentration Range of PAHs

PAHs	Min Concentration	Max Concentration
Nap	0	115.50
Pyr	0	1896.00
Phe	-0.03	209.00
Chr	0	2361.00
BaP	0	338.00
BbF	-0.05	2361.00
BaA	-0.25	176.00

But the above scenario is not encountered in the samples related to safe data. Hence, the data has been analysed using sum of square method(Nataraj et al., 2022). The following equation (1) has been used for the computation of sum of square of each element in the dataset.

$$\text{Sum of the squares} = \sum_{i=1}^{519}\sum_{j}^{16} x(i,j)^2 \qquad (1)$$

Where, i represents the row index ranging from 1 to 512 dataset.

j represents the column index ranging from 1 to 16 PAHs
$x(i,j)$ represents the value at the i^{th} row and ^{j}th column of the 512 x 16 dataset
Then the sum of square values is analysed using analysis of variance and the F value has been calculated as 0.9897 and the p-value as 0.05346. As the p value is greater than 0.05, samples fail to differ significantly from each other. The analysis of variance results is shown in Table 6.

Table 6 ANOVA Measures between Samples corresponding to Sum of Square values of Unsafe.

Source of Variation	SS	df	MS	F	P-value	F crit
Between samples	3E+13	287	9E+10	0.9898	0.5347	1.2145
Within Samples	3E+13	288	9E+10			

In this study, a stratified sampling technique has been employed, which ensures the samples were collected from different articles (representing various regions, and at different seasons, and multiple sources e.g., local markets, farms, urban, and rural areas). Therefore, this approach provides a significant representation of the dataset with high dimension, reducing the likelihood of sampling bias. In this analysis, cross-validation techniques have also been employed and used to train the machine learning models. The classification models use the feature normalization (bipolar normalization) and randomization techniques to prevent overfitting. Additionally, multiple methods have been used for the evaluation and presented the best performance metrics, such as sensitivity of each class, accuracy and misclassification rate. This comprehensive approach strengthens the validity of the findings and enhances the credibility of this research. The following section presents a detailed explanation on the modelling of the machine learning algorithms and the results.

2.3 Design and evaluation of machine learning algorithms for the toxicity detection system

Machine learning, a subset of Artificial Intelligence, leverages algorithms and data to emulate human brain functionality (Nataraj et al., 2021). Widely employed for pattern recognition, clustering, and signal processing, machine learning algorithms play a crucial role in prediction and clustering based on labelled or unlabelled datasets (Pandey et al., 2023b). In our analysis for toxicity detection, we have adopted simple and well-established learning algorithms to evaluate toxicity level of the fruits and vegetables. Based on expert knowledge and standard limits, the samples were categorized as safe and unsafe and validated using ANOVA analysis.

In this research, k-fold cross-validation is utilized for data segmentation, linear and non-linear classifiers are employed for data evaluation. The classification algorithms are chosen due to their wide acceptance and convenience in modelling and evaluating PAH datasets.
The designed models feature 16 inputs and 1 output, evaluated using different machine learning algorithms to classify toxicity in fruits and vegetables. The neural network model incorporates one hidden layer with 15 hidden neurons, utilizing the Levenberg-Marquardt backpropagation algorithm for weight updating. Error calculation is performed using the Mean Square Error method. The k-NN and SVM models are trained with standard parameters. The value of K varies from 0 to 10 for k-NN models, while a standard size is chosen as the sigma value for SVM models.

This comprehensive approach aims to utilize diverse machine learning techniques for robust toxicity detection, providing a foundation for effective analysis and decision-making in environmental assessments.

2.4 k-fold Cross-Validation

To assess the classification model and evaluate system performance, the k-fold cross-validation method is employed in this research. This sampling process serves to generalize the model and aids in selecting the most appropriate one for the task at hand. Well, there are many segmentation techniques that are used in the classification such as Hold Out, leave-one-out cross-validation, and k-fold cross validation.

It is well known that the usual methods such as Hold Out methods were used for large data set and the leave-one-out-cross-validation is very similar to k-fold cross validation and the random splitting such as 60-40, 70-30, 80-20 methods may lead to overfitting. In this analysis k-fold cross validation has been chosen, since these methods are powerful and have an ability to generalise the machine learning model, even if the data set/ feature set have limited samples. Also, 'k' provides equally sized validation for multiple epochs. Specifically, a 5-fold cross-validation is utilized, denoted by k=5, where the dataset is divided into 5 non-overlapping folds of equal size (Megalingam et al., 2019).

This method is instrumental in training and testing the model across all dataset subsets, effectively reducing variance (Sonwani et al., 2022). The PAH dataset, with its raw 519 x 16 features, undergoes segmentation into training and testing datasets through the five-fold cross-validation. The training dataset comprises 80% of the total dataset (415 x 16 features), while the testing dataset holds the remaining 20% (104 x 16 features). This five-fold method results in the creation of five distinct training and testing sets. Figure 2 describes the entire process of 5-fold cross validation (A. Ramezan et al., 2019), (Wong & Yeh, 2019).

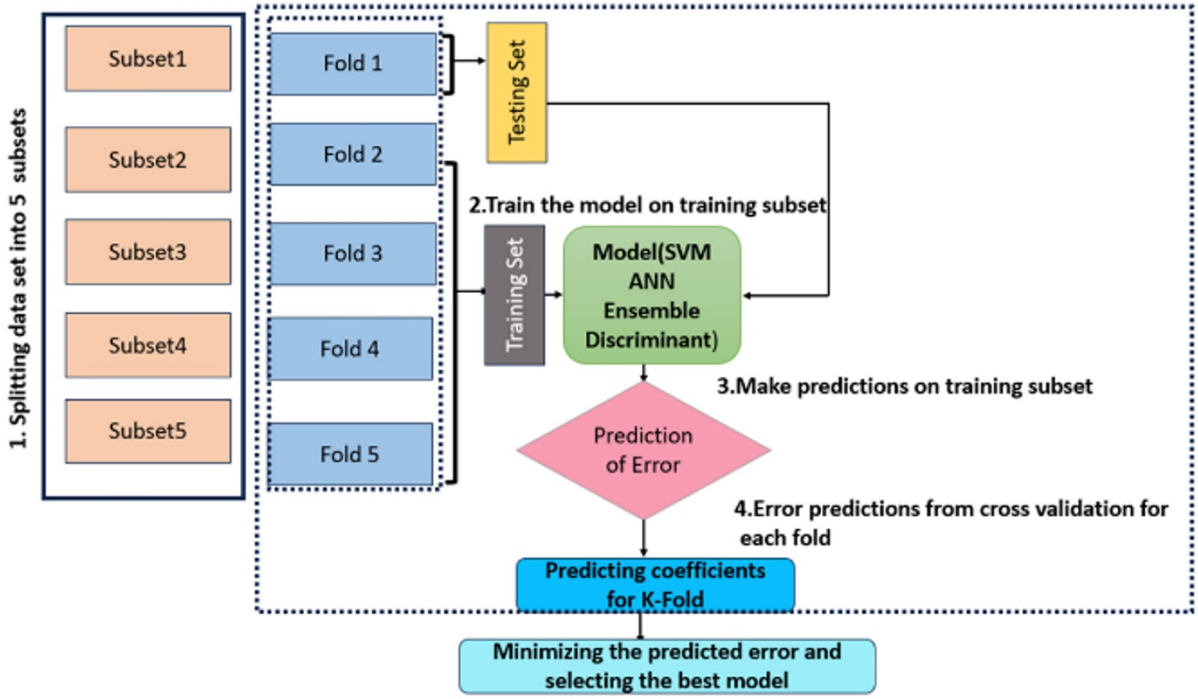

Fig. 2 Flowchart of K-Fold Cross Validation

The segmented training sets are concurrently trained using various algorithms, including feed-forward backpropagation-based Neural Network, Support Vector Machine (SVM), Ensemble, Regression, Discriminant, Tree, k Nearest Neighbour (k-NN), and Naive Bayes. Subsequently, the trained models undergo testing using the remaining five distinct 20% testing datasets.

The training and test results derived from the algorithms are discussed in detail in the subsequent section of this article, providing a comprehensive understanding of the model's performance across diverse evaluation scenarios.

3 Results and Discussion

The main aim of this research is to understand the levels of toxicity in fruits and vegetables using machine learning algorithm and analyse the levels of Polycyclic Aromatic Hydrocarbons (PAHs). The main objective revolves around the careful investigation of PAHs collected from various sources, validated by a comprehensive statistical analysis to correlate the data set, discerning between safe and unsafe data.

Analysis of the collected PAHs led to the development of a robust two-class pattern recognition model. The dataset of 16 input features is used as input to various machine learning algorithms, each of which is characterized by different technical specifications. Given the nature of this binary classification problem, only one output neuron is considered. Based on the outcomes, the results highlight that the machine

learning method achieves over 85% accuracy in classifying two-class problems. The following section provides details on the results of various machine learning approaches and how they contribute to the detection of toxicity in fruits and vegetables.

3.1 Classification of safe and unsafe data using MATLAB Classification Learner app

In this study, the development of machine learning algorithms for toxicity analysis on fruits and vegetables was undertaken using MATLAB's Classification Learner application (Wang et al., 2022). While some algorithms were directly implemented through the Classification Learner, others required custom parameterization. In this research, various classification algorithms have been examined in an attempt to establish a reliable and efficient model for detecting toxicity in fruits and vegetables. Various classification algorithms have been used in this evaluation process to determine a reliable and efficient model for the detection of toxicity in fruits and vegetables. The classification models were designed with 16 inputs and 2 outputs by applying respective approaches. The Discriminant Analysis classifier has been employed with linear (LDA- Linear discriminant analysis) and quadratic (QDA- quadratic discriminant analysis) discriminant approaches, which is ideal for the analysis of high dimensional data with limited significance in feature interactions (Le et al., 2020). This model is specifically suitable for toxicity classification based on its ability to handle various covariance structures, providing robust performance among various correlations of PAH. Ensemble models were designed with bagged trees and RUSBoosted trees that enhance the model's robustness and handles class imbalance(Ampomah et al., 2020). Ensemble model especially suited for the classification tasks involving toxicity detection and offering a balance between accuracy and computational efficiency. Whereas Kernel models were designed using SVM and logistic regression kernels. These models can capture complex patterns through non-linear transformations(Colkesen et al., 2016). An advantage of kernel model is its flexibility in handling non-linear relationships which makes its suitable for complex toxicity classification.

Other models such as KNN models were designed using fine, medium and coarse KNN configurations(Ali et al., 2020). This model provides an effective and simple approach for the detection of toxicity. Naive Bayes models are effective for classification of toxicity due to their ability to handle continuous data. This model was designed using Gaussian and kernel distributions(Pérez et al., 2009). Artificial Neural networks models provide high accuracy and speed and are also powerful for complex classification tasks (Tritsaris et al., 2021). Neural networks were designed using wide, bilayered and trilayered configurations. SVM models were designed using various kernel functions such as Linear SVM, Quadratic SVM, Cubic SVM, Fine Gaussian SVM, Medium Gaussian SVM and Coarse Gaussian SVM(Nanda et al., 2018). These models are suitable for toxicity detection due to its robustness and efficiency in binary classification tasks. Decision trees are a popular choice as they provide interpretable models for toxicity classification. These models were designed using fine, medium and coarse configurations(Vargaftik et al., 2021). Regression models were designed using binary GLM logistic regression, efficient logistic regression and efficient linear SVM (Wu & Yang, n.d.)which provides accurate, efficient predictions and are suitable for binary classifications. The design specification of the classification models are given in Table 7.

Table 7 Design specification of the classification models used in the toxicity detection system

Classification models	Classification algorithm	Total Cost (Validation)	Prediction Speed (obs/ sec)	Training time (sec)	Number of learners	Maximum Splits	Learning Rate	Parameters	Neighbors	Distance Metric	Distribution	Layers	Kernel
Discriminant Analysis Classifier	LDA	71	13,000	13.803									
	QDA	67	7,700	13.192									

Prepublished copy

Nature Environment & Pollution Technology

This is a peer-reviewed prepublished version of the paper

Ensemble Machine Learning Algorithms	Bagged Trees	25	3,100	4.7758	30	518							
	RUSBoosted Trees	31	3,500	5.5973	30	20	0.1						
Kernel Models	SVM Kernel	26	16,000	8.5592				Auto(Expansion)					
	Logistic Regression Kernel	34	12,000	7.3023				Auto(Expansion)					
K-Nearest Neighbors (KNN)	Fine KNN	20	9,000	2.6819					1	Euclidean			
	Medium KNN	24	8,400	2.3804					10	Euclidean			
	Coarse KNN	86	9,000	1.2029					100	Euclidean			
Naive Bayes	Gausian	56	7,200	8.8527							Gaussian		
	Kernel	54	3,500	8.0303							Kernel (Gaussian)		
Artificial Neural Network	Wide NN	15	15,000	6.8221								1(100 neurons)	
	Bilayered NN	12	21,000	6.5796								2(10 neurons each)	
	Trilayered NN	11	14,000	6.8536								3(10 neurons each)	
Support Vector Machine (SVM)	Linear SVM	44	5,000	2.3861									Linear
	Quadratic SVM	32	7,100	1.6382									Quadratic
	Cubic SVM	34	9,400	1.4974									Cubic
	Fine Gaussian SVM	22	7,000	1.4187									Gaussian
	Medium Gaussian SVM	32	5,900	1.4945									Gaussian
	Coarse Gaussian SVM	64	5,000	1.4597									Gaussian
Decision Trees	Fine Tree	17	3,400	13.21		100							
	Medium Tree	17	6,000	2.2963		20							
	Coarse Tree	24	7,900	1.515		4							
Regression Models	Binary GLM LR	14	3,800	3.2137									

	Efficient LR	14	7,200	2.21									
	Efficient Linear SVM	19	8,100	1.6321									

The design performance of the classification models are represented in Table 8 and the results are compared. From the results, it can be observed that based on the design and performance metrics, Coarse KNN model have high total cost (validation) of 86, prediction speed of 9,000 and Training time is 1.2029 sec compared to the other models. Whereas Trilayered Neural Network appears to be the best model for toxicity classification in fruits and vegetables with a lowest total cost (validation) of 11, with a high prediction speed of 14,000 obs/sec and a reasonable training time of 6.8536 sec. It is now evident that neural network models are proven to be efficient in terms of design, learning nonlinear patterns and prediction.

Table 8 The design performance of the classification models:

Model	Total Cost (Validation)	Prediction Speed (obs/sec)	Training Time (sec)
Linear Discriminant	71	13,000	13.803
Quadratic Discriminant	67	7,700	13.192
Bagged Trees	25	3,100	4.7758
RUSBoosted Trees	31	3,500	5.5973
SVM Kernel	26	16,000	8.5592
Logistic Regression Kernel	34	12,000	7.3023
Fine KNN	20	9,000	2.6819
Medium KNN	24	8,400	2.3804
Coarse KNN	**86**	**9,000**	**1.2029**
Gaussian Naive Bayes	56	7,200	8.8527
Kernel Naive Bayes	54	3,500	8.0303
Wide Neural Network	15	15,000	6.8221
Bilayered Neural Network	12	21,000	6.5796
Trilayered Neural Network	**11**	**14,000**	**6.8536**
Linear SVM	44	5,000	2.3861
Quadratic SVM	32	7,100	1.6382
Cubic SVM	34	9,400	1.4974
Fine Gaussian SVM	22	7,000	1.4187
Medium Gaussian SVM	32	5,900	1.4945
Coarse Gaussian SVM	64	5,000	1.4597
Fine Tree	17	3,400	13.21
Medium Tree	17	6,000	2.2963
Coarse Tree	24	7,900	1.515
Binary GLM Logistic Regression	14	3,800	3.2137
Efficient Logistic Regression	14	7,200	2.21
Efficient Linear SVM	19	8,100	1.6321
Min	11	3100	1.2029
Max	86	21000	13.803

As discussed, the proposed methodology allowed for a comprehensive evaluation of each classifier's performance in toxicity detection, with the chosen parameters tailored to the characteristics of the dataset and the objectives of the study. The results obtained from model 1 classification learner are shown in Table 9, presenting the minimum, average, and maximum classification accuracy from 10 trials.

Table 9 The Classification Accuracy of Toxicity detection using model 1 without applying Principal Component Analysis (PCA).

Classification models	Classification Accuracy (10 trials)		
Model Type	MIN (%)	MAX (%)	Average (%)
Discriminant	86.3198	87.0906	86.7052
Ensemble	81.1175	98.2659	92.3314
Kernel	93.4489	94.9904	94.2197
KNN	83.4297	97.1098	93.7058
Naive Bayes	89.2100	89.5954	89.4027
Neural Network	97.1098	97.8805	97.4952
SVM	70.3276	97.1098	89.8844
Tree	89.7881	91.1368	90.7996
Regression	92.6782	97.8805	95.2794

From the results, the Ensemble model has the highest accuracy of 98.2659% compared to other models. Ensemble models combine predictions from multiple machine learning models to enhance overall performance. On the other hand, the SVM model performed the least, achieving an accuracy of 70.3276%. This lower performance can be attributed to the challenge of finding the best hyperplane to separate different classes, especially in our dataset with non-linear characteristics and some PAH values being zero. The SVM model's sensitivity to parameters like the choice of kernel and regularization parameter (C) contributed to its specific challenges. Despite variations, all models maintained an average performance rating exceeding 85%. This summary provides a straightforward overview of the results without using any feature optimization techniques, emphasizing the highest classification accuracy of the Ensemble model, and the challenges faced by the SVM model in the context of the dataset complexity.

Hence, a classification model has also been developed by using PCA (Principal Component Analysis). The classification accuracy of the classification models is presented in Table 8. By applying PCA, the order of features is sorted according to the dominants of the feature, and no PAH values were neglected. Following this optimization, the ensemble model displayed performance ranging from a minimum of 55.4913% to a maximum of 97.88% over 10 different trials. overall, it has been observed that a performance rating of 86% has been achieved across all the trials.

Table 10 The Classification Accuracy of Toxicity Detection using model 1 with applying PCA.

Classification models	Classification Accuracy (10 trials)		
Model Type	MIN (%)	MAX (%)	Average (%)
Discriminant	86.3198	87.0906	86.7052
Ensemble	55.4913	97.8805	87.3732
Kernel	89.5954	95.3757	93.1920
KNN	72.4470	95.3757	91.2653
Naive Bayes	89.2100	94.4123	92.4855
Neural Network	94.7977	98.2659	96.1946
SVM	70.3276	97.1098	92.4213
Tree	89.7881	96.7245	94.0848
Regression	92.6782	97.8805	95.0867

From Table 10, it can also be observed that the traditional ANN model achieved a minimum classification accuracy of 92.42% and a maximum of 98.26%, this shows the feed forward back propagation using LM algorithm have proven the robustness in non-linear pattern recognition method. Similarly, the minimum accuracy of 72.447% and 95.37% of maximum accuracy are attained from the k-NN model. The SVM classifiers attained a minimum accuracy of 70.327%, and a maximum accuracy of 97.109%. The results indicate that the neural network models, trained using the Levenberg backpropagation algorithm, achieved a maximum accuracy of 98.26%.

Additionally, the confusion matrices corresponding to the classifier models are illustrated in figures 3 to 4 and elaborated in the following sections. Due to the extensive nature of discussing all observations from multiple classifiers, this article focuses on presenting the confusion matrix corresponding to the maximum classification accuracy achieved by the ANN model with PCA versus the minimum classification accuracy obtained by the SVM model without PCA. Figure 3 and Figure 4 presents the confusion matrix corresponding to the ANN model and SVM model respectively.

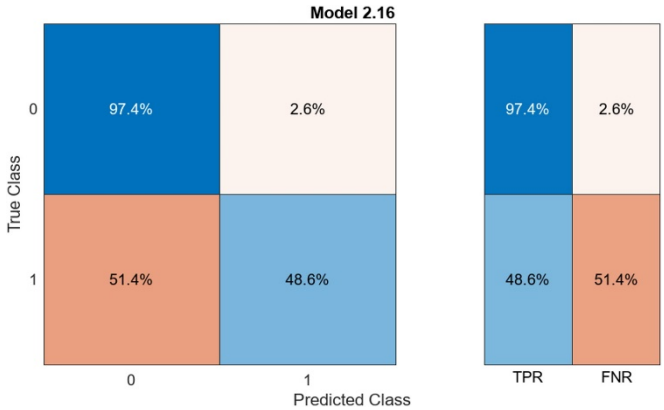

Fig. 3 Confusion Matrix of SVM Model with Minimum Accuracy of Classifier 1

Referring Figure 3 it can be observed that SVM Model (Without PCA) has the following sensitivity levels.

1. True Positive Rate for Pattern 1: 97.4%
2. True Positive Rate for Pattern 2: 48.6%
3. False Positive Rate for Pattern 1: 2.6%
4. False Positive Rate for Pattern 2: 51.4%

The two-class SVM model can also be used to determine the effectiveness of classification model for each class based on its sensitivity. The sensitivity of the "Safe data" category is 97.4%, indicating that the related samples were correctly classified with few occurrences of misclassification. Similarly, the sensitivity of the "Unsafe data" category is 48.6%, indicating that the classification model has a lower sensitivity and a misclassification accuracy of 51.4%. As a result of this, the SVM model developed for the two-class problem is not appropriate for the generalization of the toxicity detection system.

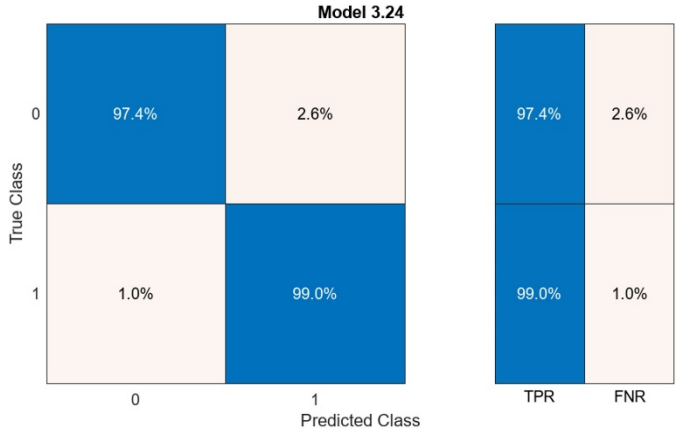

Fig. 4 Confusion Matrix of ANN Model with Maximum Accuracy of Classifier 2

Referring Figure 4 it can be observe that the ANN Model (With PCA):

- True Positive Rate for Pattern 1: 97.4%
- True Positive Rate for Pattern 2: 99.0%
- False Positive Rate for Pattern 1: 2.6%
- False Positive Rate for Pattern 2: 1.0%

The two-class ANN model can also be used to determine the effectiveness of classification model for each class based on its sensitivity. The sensitivity of the "Safe data" category is 97.4%, indicating that the related samples were correctly classified with few occurrences of misclassification. Similarly, the sensitivity of the "Unsafe data" category is 99.0%, indicating that with more accuracy, the related samples were correctly classified. As a result of this, the ANN model developed for the two-class problem is suitable for the generalization of the toxicity detection system. However, we have also considered the four-class problem for the generalization of the toxicity detection system. The classification model's effectiveness for the multiclass based on its sensitivity have been discussed in the following section.

Additionally, our observations reveal that the outputs from ANN models, using a binary activation function, ranged from -0.22699 to 2.8721 without rounding the net values or applying absolute values. Based on these findings, we further divided the dataset into four distinct sets corresponding to No Harm, Low Harm, Moderate Harm, and Severe Harm. These divided datasets were employed to model a multi-layer neural network for classifying different levels of toxicity. The split datasets were also validated using ANOVA for the F-value and P-value, both of which demonstrated significant differences between the four toxicity levels.

3.2 Level of toxicity using multi-layer neural network

The ANN model for the two-class problem achieved a maximum accuracy of 98.26%, with a misclassification of 6 samples. Regression analysis for the two-class problem was performed at various stages of training, validation, and testing, as illustrated in Figure 5(Faraw, 2015).

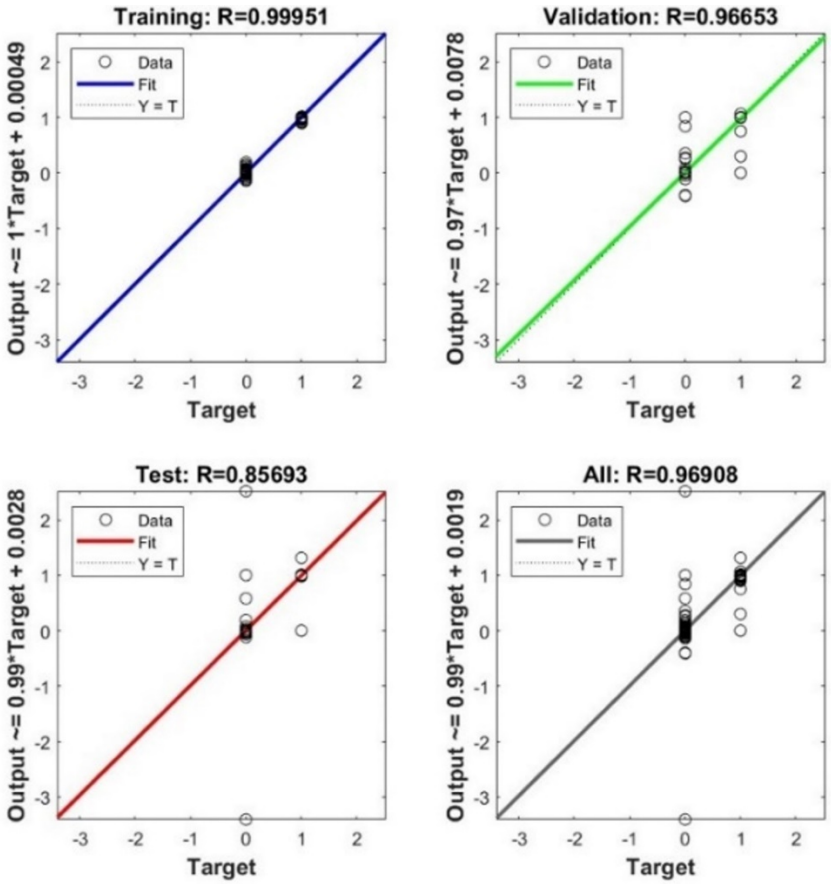

Fig. 5 Regression Analysis of two class Pattern Recognition using FFNN

The R values during these stages were found to be 0.99951, 0.96653, and 0.85693, respectively. These R values, being close to 1, confirm the robustness of the results(Faraw, 2015)(Judd et al., 2017). Additionally, Figure 6 displays a scattered plot revealing that the output neural network effectively discriminates between 0 and 1.

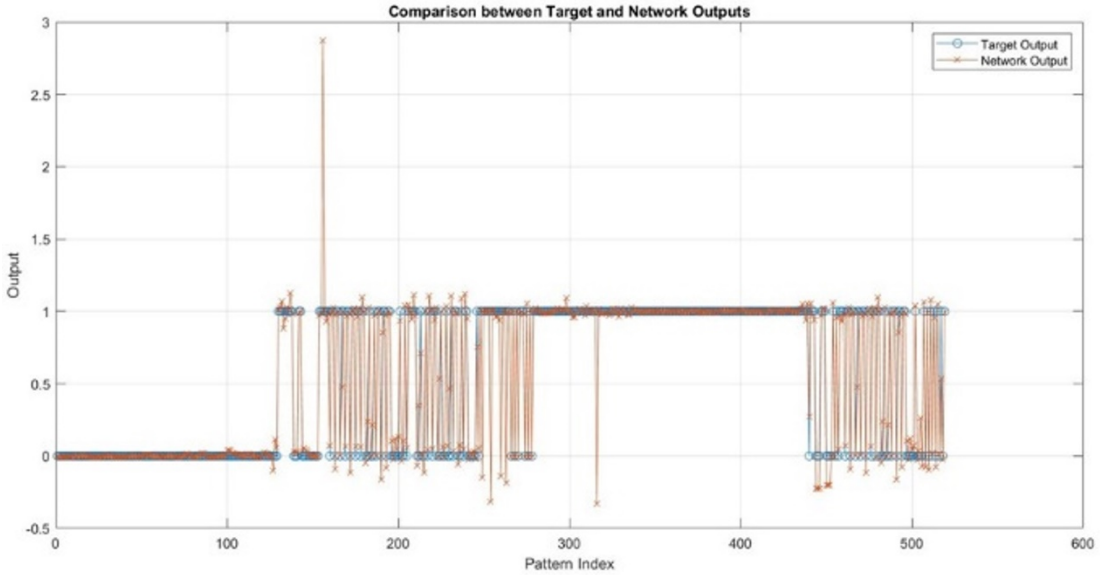

Fig. 6 Scattered plot of two class patterns comparing the actual and target outputs.

Leveraging this discrimination, we further divided the data into four sets and modelled a multi-layer NN for the four-class problem. The development of the multi-layer neural network model involved configuring 16 input neurons, one hidden layer with 15 hidden neurons, and 2 output neurons to address the four different toxicity patterns. With 1000 epochs and a goal parameter set to 1e-10, the MLNN model demonstrated with a performance of 97% accuracy with mean square error (MSE) of 9.99e-11 and 3% misclassification rate.

Analysing the confusion matrix in Table 9 revealed accurate classification for all 222 "no harm" samples, 9 out of 10 "low harm" samples, and 6 out of 7 "moderate harm" samples. Notably, one "low harm" sample was incorrectly categorized as "moderate harm," resulting in a 17% false negative rate. For the "severe harm" category, 272 out of 275 samples were correctly classified, with a 1% false positive rate indicating two samples falling into "no harm" and one into "low harm." This analysis justifies the MLNN model's effectiveness in determining the toxicity level of fruits and vegetables, even with a dataset of 513 samples (excluding 6 misclassified samples in the two-class problem).

Table 11 Confusion Matrix for MLNN for the Four Class Problem

		Actual output				**TP**	**FP**
		No Harm	**Low harm**	**Moderate harm**	**Severe harm**		
Target Output	**No Harm**	222	0	0	0	100%	0%
	Low harm	0	9	1	0	90%	10%
	Moderate harm	0	0	6	0	100%	0%
	Severe harm	2	1	0	272	99%	1%
False Negative		1%	10%	17%	0%	**97%**	**3%**

From the confusion matrix represented in Table 11, it is observed that there are false negative samples and false positive samples. Therefore, the sensitivity and specificity of the multi-classification system has been analyzed using equation 2 and equation 3.

$$sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (2)$$

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives} \quad (3)$$

The Sensitivity and specificity results are tabulated in Table 12 and discussed below.

Table 12 The Sensitivity and specificity results

Class	Sensitivity $= \frac{TP}{TP+FN}$	Specificity $= \frac{TN}{TN+FP}$
No Harm	$\frac{222}{222+0} = 1.00\ (100\%)$	$\frac{9+1+0+1+0+6+0+272}{9+1+0+1+0+6+0+272} = \frac{289}{289} = 1.00\ (100\%)$
Low Harm	$\frac{9}{9+1} = 0.90\ (90\%)$	$\frac{222+0+0+1+0+6+0+272}{222+0+0+1+0+6+0+272} = \frac{500}{500} = 1.00\ (100\%)$
Moderate Harm	$\frac{6}{6+0} = 1.00\ (100\%)$	$\frac{222+0+0+0+0+1+0+1+272}{222+0+0+0+0+1+0+1+272} = \frac{498}{498} = 1.00\ (100\%)$
Severe Harm	$\frac{272}{272+2} = 0.99\ (99.26\%)$	$\frac{222+0+0+0+9+1+0+6}{222+0+0+0+9+1+0+6} = \frac{238}{238} = 1.00\ (100\%)$

An understanding of the effectiveness of the classification model for every class can be established from the sensitivity and specificity results derived from the confusion matrix. The results are discussed as follows:
Level of Sensitivity

The ability of a model to distinguish examples of a specific class from all instances that genuinely belong to that class is measured by its sensitivity. Therefore, the classification model based on the multi-

layer neural network model with maximum classification accuracy has been chosen for the sensitivity and specificity analysis. From this analysis, considering the "No Harm" and "Moderate harm" classes: Sensitivity stands at 100%, indicating that the model accurately detects every "No Harm" and "Moderate harm" occurrences. Considering the "Low Harm" category, the model's sensitivity is 90%, indicating that 90% of "Low Harm" occurrences are accurately identified, with minimal instances of misclassification. Similarly, the sensitivity of the "Severe harm" class have 99.26%, indicated a high accuracy in identifying instances of "Severe harm", with a few misclassifications.

Level of specificity relates to how effectively the model can identify and reject samples that do not fall within a specific class. In this classification problem, every class has a specificity value of 100%, which strongly suggests that samples that do not belong to each class were successfully rejected by the model. From this analysis, with high sensitivity and specificity values across all classes, the classification model appears to be effective at identifying samples of each class overall, according to the results of the sensitivity and specificity analysis. There is certainly potential for development, nevertheless, particularly in accurately recognizing cases of the "Low harm" class, where the sensitivity is marginally lower than that of other classes. Additional examination and enhancement of the model could potentially enhance its efficacy in terms of increased classification precision.

Moreover, regression and scatter plots were used to assess the strength of the relationship between targeted outputs and the actual output. The regression plot and scatter plots are depicted in Figure 7 and Figure 8 for the four-class problem during training, validation, and testing stages, respectively (Faraw, 2015). The R values are 0.9919, 0.989, and 0.91 and these results further validate the robustness of the classification model. For future studies, increasing the sample size could enhance system stability and facilitate the generalization of the model for global applicability.

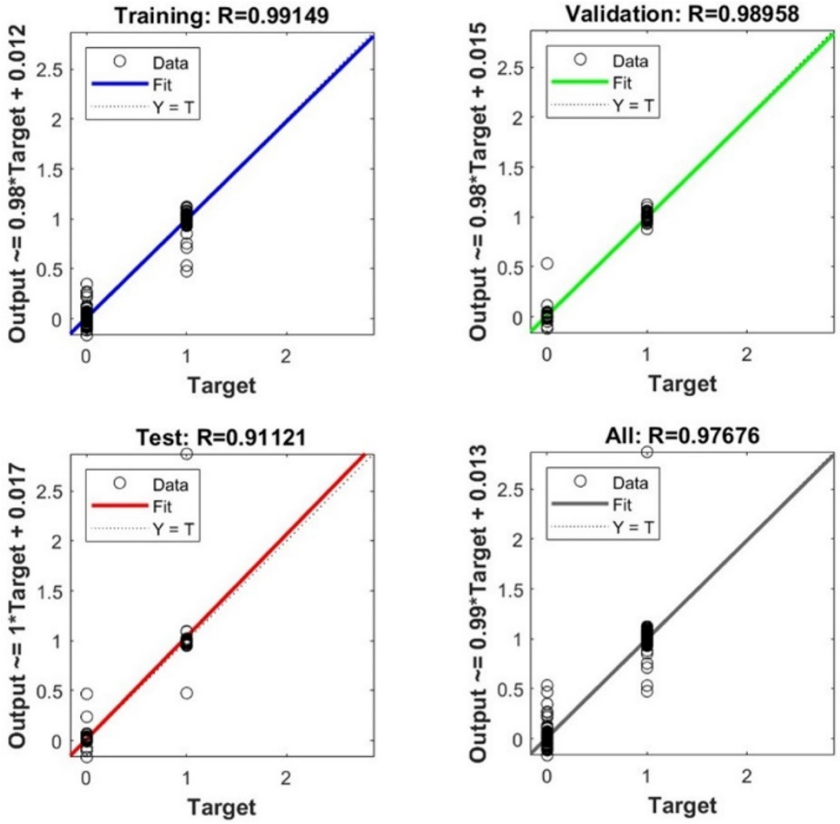

Fig. 7 Regression analysis of Multi Class Pattern Recognition using MLNN

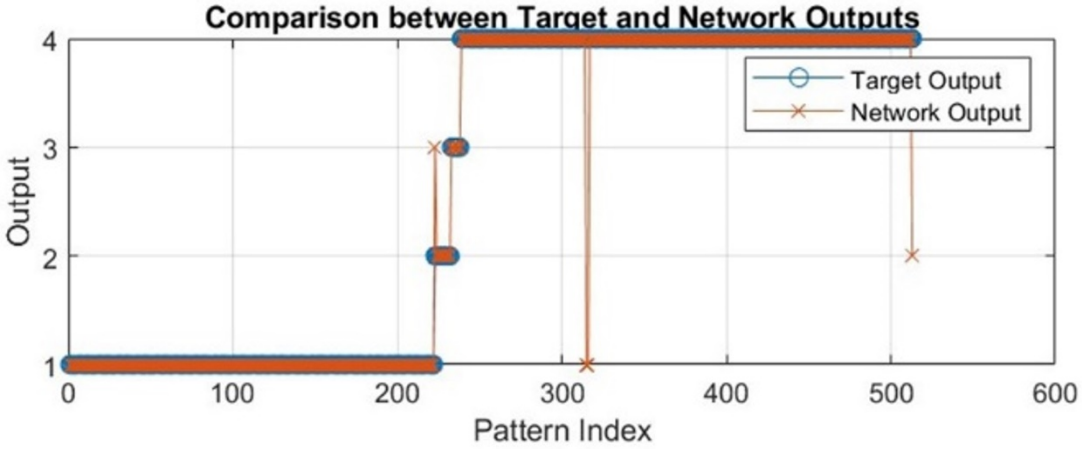

Fig. 8 Scattered plot of four class patterns comparing the Actual and Target outputs

4 Conclusion

This study thoroughly investigated the presence of Polycyclic Aromatic Hydrocarbons (PAHs) in fruits and vegetables, employing robust statistical measures for a comprehensive understanding of the dataset. The detection of toxicity in these consumables were successfully achieved through the implementation of machine learning algorithms, including Artificial Neural Network (ANN), k-Nearest Neighbors (K-NN), and Support Vector Machine (SVM). Remarkably, the medium k-NN and Cubic-k-NN models demonstrated 100% accuracy, while Quadratic SVM, Cubic SVM, and cosine k-NN models exhibited an accuracy of 92.3%. Despite the promising results from all three models, ANN classifiers emerged as the most accurate in predictions, especially given the binary class nature of the problem and the minimal number of samples considered for the toxicity detection system.

Furthermore, the outputs from the ANN models were investigated to determine the toxicity level of the samples, revealing highly promising results. To enhance the generalization of this toxicity classification system, future work will involve developing a real-time dataset with diverse feature extraction and optimization methods. The models trained using various machine learning algorithms showcased efficiency and provided substantial results, laying the groundwork for the development of a generalized prototype model. Standardizing the level of toxicity will enable a more precise representation of the severity of fruits and vegetables. The results of this study establish the viability of applying machine learning algorithms to predict toxicity in various products, paving the way for broader application in the future. Also, the performance of the models trained using different machine learning algorithms provides a solid foundation for the development of a standardized toxicity classification system. This standardization facilitates to provide precise decision of toxicity severity in products, thereby enabling informed decision-making and regulatory intervention.

Finally, the study has several limitations, such as those associated with the collected dataset. While the dataset used in this research presents samples from variety of regions, seasons, and sources, it may not accurately reflect the global diversity of fruits and vegetables. This constraint may affect the generalizability of our classification models. Expanding the methodology to include samples from different countries considering the climates, and farming practices might strengthen the analysis. In collaboration with research institutions and industries may allow a practical implementation of the research and it would also be beneficial to investigate longitudinal studies examining PAH contamination over multiple periods and seasons. This would provide a more complete understanding of temporal changes and their impact on contamination levels. Furthermore, establishing the study's limitations is crucial for interpreting the findings and directing future research. By addressing potential overfitting, the need for more diverse datasets, and other constraints, future work on this study will focus on expanding datasets, incorporating longitudinal data, leveraging advanced detection technologies and evaluation methods. This may enhance the reliability, stability of the classification models and applicability of PAH analysis in fruits and vegetables.

In summary, our research not only highlights the current state of PAH contamination in fruits and vegetable but also opens direction for future research and technological applications that can significantly enhance food safety and public health. By addressing these challenges and suggesting concrete solutions,

contributing to safer and healthier food safety.

References

A. Ramezan, C., A. Warner, T., & E. Maxwell, A. (2019). Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification. *Remote Sensing*, *11*(2), 185.

Abdel-Shafy, H. I., & Mansour, M. S. M. (2016). A review on polycyclic aromatic hydrocarbons: source, environmental impact, effect on human health and remediation. *Egyptian Journal of Petroleum*, *25*(1), 107–123.

Abou-Arab, A. A. K., Abou-Donia, M. A. M., El-Dars, F., Ali, O., & Hossam, A. (2014). Levels of polycyclic aromatic hydrocarbons (PAHS) in some Egyptian vegetables and fruits and their influences by some treatments. *Int J Curr Microbiol App Sci*, *3*(7), 277–293.

Ali, A., Alrubei, M., Hassan, L. F. M., Al-Ja'afari, M., & Abdulwahed, S. (2020). Diabetes classification based on KNN. *IIUM Engineering Journal*, *21*(1), 175–181. https://doi.org/10.31436/iiumej.v21i1.1206

Al-Nasir, F., Hijazin, T. J., Al-Alawi, M. M., Jiries, A., Al-Madanat, O. Y., Mayyas, A., A. Al-Dalain, S., Al-Dmour, R., Alahmad, A., & Batarseh, M. I. (2022a). Accumulation, Source Identification, and Cancer Risk Assessment of Polycyclic Aromatic Hydrocarbons (PAHs) in Different Jordanian Vegetables. *Toxics*, *10*(11), 643.

Al-Nasir, F., Hijazin, T. J., Al-Alawi, M. M., Jiries, A., Al-Madanat, O. Y., Mayyas, A., A. Al-Dalain, S., Al-Dmour, R., Alahmad, A., & Batarseh, M. I. (2022b). Accumulation, Source Identification, and Cancer Risk Assessment of Polycyclic Aromatic Hydrocarbons (PAHs) in Different Jordanian Vegetables. *Toxics*, *10*(11), 643.

Al-Nasir, F., Hijazin, T. J., Al-Alawi, M. M., Jiries, A., Al-Madanat, O. Y., Mayyas, A., A. Al-Dalain, S., Al-Dmour, R., Alahmad, A., & Batarseh, M. I. (2022c). Accumulation, Source Identification, and Cancer Risk Assessment of Polycyclic Aromatic Hydrocarbons (PAHs) in Different Jordanian Vegetables. *Toxics*, *10*(11), 643.

Ampomah, E. K., Qin, Z., & Nyame, G. (2020). Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement. *Information (Switzerland)*, *11*(6). https://doi.org/10.3390/info11060332

Ashraf, M. W. (n.d.). *Polycyclic Aromatic Hydrocarbons in Vegetables and Fruits produced in Saudi Arabia*.

Ashraf, M. W., & Salam, A. (2012). Polycyclic aromatic hydrocarbons (PAHs) in vegetables and fruits produced in Saudi Arabia. *Bulletin of Environmental Contamination and Toxicology*, *88*, 543–547.

Ashraf, M. W., Taqvi, S. I. H., Solangi, A. R., & Qureshi, U. A. (2013). Distribution and risk assessment of polycyclic aromatic hydrocarbons in vegetables grown in Pakistan. *Journal of Chemistry*, *2013*.

Camargo, M. C. R., & Toledo, M. C. F. (2003). Polycyclic aromatic hydrocarbons in Brazilian vegetables and fruits. *Food Control*, *14*(1), 49–53.

Choochuay, C., Deelaman, W., & Pongpiachan, S. (2023). Polycyclic Aromatic Hydrocarbons in Thai and Myanmar Rice: Concentrations, Distribution and Health Concerns. *Nature Environment and Pollution Technology*, *22*(3), 1097–1110. https://doi.org/10.46488/NEPT.2023.v22i03.002

Colkesen, I., Sahin, E. K., & Kavzoglu, T. (2016). Susceptibility mapping of shallow landslides using kernel-based Gaussian process, support vector machines and logistic regression. *Journal of African Earth Sciences*, *118*, 53–64. https://doi.org/10.1016/j.jafrearsci.2016.02.019

Cui, X., Ailijiang, N., Mamitimin, Y., Zhong, N., Cheng, W., Li, N., Zhang, Q., & Pu, M. (2022). *Pollution Levels, Sources and Risk Assessment of Polycyclic Aromatic Hydrocarbons in Farmland Soil and Crops Near*. https://doi.org/10.21203/rs.3.rs-1273637/v1

Deligannu, P., & Muniandy, T. (2024). Review on the Health Risk of Polycyclic Aromatic Hydrocarbon (PAH) Exposure Among Street Food Vendors. *European Journal of Theoretical and Applied Sciences*, *2*(1), 532–539. https://doi.org/10.59324/ejtas.2024.2(1).46

Faraw, J. J. (2015). *Practical Regression and ANOVA using R*.

Frossard, J., & Renaud, O. (2021). Permutation tests for regression, ANOVA, and comparison of signals: the permuco package. *Journal of Statistical Software*, *99*, 1–32.

Janska J, M. H., Tomaniova V, M. K., & Vavrova, M. (2006). Polycyclic aromatic hydrocarbons in fruits and vegetables grown in the Czech Republic. *Bulletin of Environmental Contamination and Toxicology*, *77*(4), 492–499.

Jia, J., Bi, C., Zhang, J., Jin, X., & Chen, Z. (2018). Characterization of polycyclic aromatic hydrocarbons (PAHs) in vegetables near industrial areas of Shanghai, China: Sources, exposure, and cancer risk. *Environmental Pollution*, *241*, 750–758.

Judd, C. M., McClelland, G. H., & Ryan, C. S. (2017). *Data analysis: A model comparison approach to regression, ANOVA, and beyond*. Routledge.

Khalili, F., Shariatifar, N., Dehghani, M. H., Yaghmaeian, K., Nodehi, R. N., & Yaseri, M. (2021a). The analysis and probabilistic health risk assessment of PAHs in vegetables and fruits samples marketed Tehran Chemometric. *Glob Nest J*, *23*, 497–508.

Khalili, F., Shariatifar, N., Dehghani, M. H., Yaghmaeian, K., Nodehi, R. N., & Yaseri, M. (2021b). The analysis and probabilistic health risk assessment of PAHs in vegetables and fruits samples marketed Tehran Chemometric. *Glob Nest J*, *23*, 497–508.

Le, K. T., Chaux, C., Richard, J. P., & Guedj, E. (2020). *An adapted linear discriminant analysis with variable selection for the classification in high-dimension, and an application to medical data*. https://www.sciencedirect.com/science/article/pii/S0167947320301225

Lee, Y.-N., Lee, S., Kim, J.-S., Patra, J. K., & Shin, H.-S. (2019). Chemical analysis techniques and investigation of polycyclic aromatic hydrocarbons in fruit, vegetables and meats and their products. *Food Chemistry*, *277*, 156–161.

Mallah, M. A., Changxing, L., Mallah, M. A., Noreen, S., Liu, Y., Saeed, M., Xi, H., Ahmed, B., Feng, F., & Mirjat, A. A. (2022). Polycyclic aromatic hydrocarbon and its effects on human health: An overeview. *Chemosphere*, *296*, 133948.

Megalingam, R. K., Sree, G. S., Reddy, G. M., Krishna, I. R. S., & Suriya, L. U. (2019). Food spoilage detection using convolutional neural networks and K means clustering. *2019 3rd International Conference on Recent Developments in Control, Automation & Power Engineering (RDCAPE)*, 488–493.

Mo, C.-H., Cai, Q.-Y., Tang, S.-R., Zeng, Q.-Y., & Wu, Q.-T. (2009). Polycyclic aromatic hydrocarbons and phthalic acid esters in vegetables from nine farms of the Pearl River Delta, South China. *Archives of Environmental Contamination and Toxicology*, *56*, 181–189.

Nanda, M. A., Seminar, K. B., Nandika, D., & Maddu, A. (2018). A comparison study of kernel functions in the support vector machine and its application for termite detection. *Information (Switzerland)*, *9*(1). https://doi.org/10.3390/info9010005

Narsi.R.Bishnoi, Mehta, U., & Pandit, G. G. (2006). Quantification of Polycyclic aromatic hydrocarbons in fruits and vegetables using high performance liquid chromatography. *Indian Journal of Chemical Technology*, *13*, 30–35.

Nataraj, S. K., Al-Turjman, F., Adom, A. H. B., R, S., M, R., & R, K. (2022). Intelligent Robotic Chair With Thought Control and Communication Aid Using Higher Order Spectra Band

Features. *IEEE Sensors Journal*, *22*(18), 17362–17369. https://doi.org/10.1109/JSEN.2020.3020971

Nataraj, S. K., M P, P., Yaacob, S. Bin, & Adom, A. H. Bin. (2021). Classification of thought evoked potentials for navigation and communication using multilayer neural network. *Journal of the Chinese Institute of Engineers*, *44*(1), 53–63. https://doi.org/10.1080/02533839.2020.1838950

Okaba, F. A., Daka, E. R., & Tulonimi, J. K. (2020). Evaluation of Polycyclic Aromatic Hydrocarbons and Toxic Elements in Some Vegetables Cultivated Along Roadsides in Port Harcourt and Environs. *Journal of Environmental Science, Toxicology and Food Technology*, *14*, 14–31.

Omoyeni, A. O., Maryrose, O. I., & Benedict, O. C. (n.d.). *Concentrations, Sources and Risk Assessment of Polycyclic Aromatic Hydrocarbons in Vegetables Cultivated in the Environs of Rivers Niger-Benue Lokoja, Nigeria*.

Organization, W. H. (2021). *Human health effects of polycyclic aromatic hydrocarbons as ambient air pollutants: report of the Working Group on Polycyclic Aromatic Hydrocarbons of the Joint Task Force on the Health Aspects of Air Pollution*. World Health Organization. Regional Office for Europe.

Pandey, V. K., Srivastava, S., Dash, K. K., Singh, R., Mukarram, S. A., Kovács, B., & Harsányi, E. (2023a). Machine Learning Algorithms and Fundamentals as Emerging Safety Tools in Preservation of Fruits and Vegetables: A Review. *Processes*, *11*(6), 1720.

Pandey, V. K., Srivastava, S., Dash, K. K., Singh, R., Mukarram, S. A., Kovács, B., & Harsányi, E. (2023b). Machine Learning Algorithms and Fundamentals as Emerging Safety Tools in Preservation of Fruits and Vegetables: A Review. *Processes*, *11*(6), 1720.

Paris, A., Ledauphin, J., Poinot, P., & Gaillard, J.-L. (2018). Polycyclic aromatic hydrocarbons in fruits and vegetables: Origin, analysis, and occurrence. *Environmental Pollution*, *234*, 96–106.

Pérez, A., Larrañaga, P., & Inza, I. (2009). Bayesian classifiers based on kernel density estimation: Flexible classifiers. *International Journal of Approximate Reasoning*, *50*(2), 341–362. https://doi.org/10.1016/j.ijar.2008.08.008

Samsøe-Petersen, L., Larsen, E. H., Larsen, P. B., & Bruun, P. (2002a). Uptake of trace elements and PAHs by fruit and vegetables from contaminated soils. *Environmental Science & Technology*, *36*(14), 3057–3063.

Samsøe-Petersen, L., Larsen, E. H., Larsen, P. B., & Bruun, P. (2002b). Uptake of trace elements and PAHs by fruit and vegetables from contaminated soils. *Environmental Science & Technology*, *36*(14), 3057–3063.

Sankar, T. K., Kumar, A., Mahto, D. K., Das, K. C., Narayan, P., Fukate, M., Awachat, P., Padghan, D., Mohammad, F., Al-Lohedan, H. A., Soleiman, A. A., & Ambade, B. (2023). The Health Risk and Source Assessment of Polycyclic Aromatic Hydrocarbons (PAHs) in the Soil of Industrial Cities in India. *Toxics*, *11*(6). https://doi.org/10.3390/toxics11060515

Singh, L., & Agarwal, T. (2018). Polycyclic aromatic hydrocarbons in diet: Concern for public health. *Trends in Food Science & Technology*, *79*, 160–170.

Sonwani, E., Bansal, U., Alroobaea, R., Baqasah, A. M., & Hedabou, M. (2022). An Artificial Intelligence Approach Toward Food Spoilage Detection and Analysis. *Frontiers in Public Health*, *9*, 816226.

Tesi, G. O., Iniaghe, P. O., Lari, B., Obi-Iyeke, G., & Ossai, J. C. (2021). Polycyclic aromatic hydrocarbons (PAHs) in leafy vegetables consumed in southern Nigeria: concentration, risk assessment and source apportionment. *Environmental Monitoring and Assessment*, *193*(7), 443.

Tritsaris, G. A., Carr, S., & Schleder, G. R. (2021). Computational design of moiré assemblies aided by artificial intelligence. *Applied Physics Reviews*, *8*(3).

https://doi.org/10.1063/5.0044511

Tuteja, G., Rout, C., & Bishnoi, N. R. (2011). Quantification of polycyclic aromatic hydrocarbons in leafy and underground vegetables: a case study around Panipat city, Haryana, India. *Journal of Environmental Science and Technology*, *4*(6), 611–620.

Vargaftik, S., Keslassy, I., Orda, A., & Ben-Itzhak, Y. (2021). RADE: resource-efficient supervised anomaly detection using decision tree-based ensemble methods. *Machine Learning*, *110*(10), 2835–2866. https://doi.org/10.1007/s10994-021-06047-x

Vasantha, K., Texina, S., Renganathan, A., & Natraj, S. K. (2023). Toxicity Detection in Water Based on Polycyclic Aromatic Hydrocarbons Using Machine Learning Algorithms. *2023 IEEE 8th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, 1–6. https://doi.org/10.1109/ICETAS59148.2023.10346267

Wang, Z., Chu, X., Li, D., Yang, H., & Qu, W. (2022). Cost-sensitive matrixized classification learning with information entropy. *Applied Soft Computing*, *116*, 108266.

WHO/V. Gupta-Smith. (2023, March 10). *Natural toxins in food.* WHO. https://www.who.int/news-room/fact-sheets/detail/natural-toxins-in-food

Wong, T.-T., & Yeh, P.-Y. (2019). Reliable accuracy estimates from k-fold cross validation. *IEEE Transactions on Knowledge and Data Engineering*, *32*(8), 1586–1594.

Wu, J., & Yang, H. (n.d.). *ACCEPTED BY IEEE TNNLS 1 Linear Regression Based Efficient SVM Learning for Large Scale Classification*.

Yang, Q., Williamson, A.-M., Hasted, A., & Hort, J. (2020). Exploring the relationships between taste phenotypes, genotypes, ethnicity, gender and taste perception using Chi-square and regression tree analysis. *Food Quality and Preference*, *83*, 103928.

Zhong, W., & Wang, M. (2002). Some polycyclic aromatic hydrocarbons in vegetables from northern China. *Journal of Environmental Science and Health, Part A*, *37*(2), 287–296.