Contents lists available at ScienceDirect

Sensing and Bio-Sensing Research

journal homepage: www.elsevier.com/locate/sbsr

Variational autoencoder analysis gas sensor array on the preservation process of contaminated mussel shells (Mytilus edulis)

Cendra Devayana Putra [a], Achmad Ilham Fanany Al Isyrofie [b,g], Suryani Dyah Astuti [b,c,g,*], Berliana Devianti Putri [f], Dyah Rohmatul Ummah [b], Miratul Khasanah [c], Perwira Annissa Dyah Permatasari [d], Ardiyansyah Syahrom [e]

[a] *Institute of Information Management, National Cheng Kung University, Tainan, Taiwan*
[b] *Magister of Biomedical Engineering, Department of Physics, Faculty of Science and Technology, Airlangga University, Surabaya 60115, Indonesia*
[c] *Department of Chemistry, Faculty of Science and Technology, Airlangga University, Surabaya 60115, Indonesia*
[d] *Department of Mathematic, Faculty of Science and Technology, Airlangga University, Surabaya 60115, Indonesia*
[e] *Medical Devices and Technology Centre Universiti Teknologi Malaysia, Johor 81310, Malaysia*
[f] *Department of Health, Faculty of Vocational Studies, Airlangga University, Surabaya 60282, Indonesia*
[g] *Department of Physics, Faculty of Science and Technology, Airlangga University, Surabaya 60115, Indonesia*

ARTICLE INFO

Keywords:
Gas sensor array
Mussel shells (*Mytilus edulis*)
Formalin
PCA
SVM
Electronic nose

ABSTRACT

Mussel shells is a macro zoobenthos that lives on soft substrates in the mud (infauna) and is classified as a bivalve. This research detects formalin in mussel shells utilizing an Electronic Nose comprised of gas sensor's array. The samples used were formalin mussel shells with several concentrations from 100 ppm to 500 ppm with the addition of 100 ppm. The research was conducted using six sensors with a sampling time of 120 s. The output voltage from each sensor is then clustered based on principal component analysis and classified using several techniques, which are support vector machine, decision tree and random forest. We demonstrate that all classifiers have an accuracy of 1. The phenomenon occurs because all feature representations can produce enough information to classify data. Principal component analysis achieves the best score in preserving the local structure. PCA can keep an average of 33% nearest data in the same neighbourhood. While variational autoencoder can keep 14% nearest data in the same neighbour, and autoencoder can keep 8% nearest data in the same area.

1. Introduction

Mussel shell (*Mytilus edulis*) is a marine animal with a tiny shape of about three to five millimetres and a pale brown body and belongs to the type of soft animal (mollusc). The development of mussel shells as a nutritious food ingredient for people has excellent prospects. Some coastal fishermen sell mussel shells in fresh, unprocessed form. New mussel shells can be obtained by buying directly from the merchant. For mussel shells to look fresh when purchased by consumers, traders or sellers often preserve their food with preservatives. Preservatives in food must be appropriate, both in the form of type and dosage.

Formaldehyde is commonly used in the food industry as a preservative for seafood, including mussels, to prevent spoilage and extend their shelf life. It is a chemical preservative that is not produced by living organisms and is not present in nature. Using a steady dose of formaldehyde to increase the shelf life of mussels is the cheapest and simplest method. The World Health Organization (WHO) believes the permissible formaldehyde consumption for a typical adult is between 1.5 and 14 mg per day (7.75 mg per day) [1]. Ingestion of formaldehyde in high doses may cause harm to the gastrointestinal system, kidneys, liver, and lungs, as well as cancer [2]. WHO also published a document which provides information on the safety of various food additives, which state that the formaldehyde is considered to be a genotoxic carcinogen. This substance can damage DNA and potentially lead to cancer [3]. Therefore, Indonesia prohibits the usage of formaldehyde inside the food. However, many businesspeople continue to employ formaldehyde as an extra chemical in their applications. To prevent the effect of formalin, many international and national policy measures on formaldehyde to reduce consumer and worker exposure levels [4]. Except from that, many different countries also set the regarding the

[*] Corresponding author at: Department of Physics, Faculty of Science and Technology, Airlangga University, Surabaya 60115, Indonesia.
E-mail address: suryanidyah@fst.unair.ac.id (S.D. Astuti).

https://doi.org/10.1016/j.sbsr.2023.100564
Received 11 February 2023; Received in revised form 3 April 2023; Accepted 24 April 2023
Available online 28 April 2023

maximum allowed levels of formalin in food products, including those meant for export. In the United States, for example, the Food and Drug Administration (FDA) has set a maximum limit of 10 ppm (parts per million) for formaldehyde in seafood [5], while the European Union has set a limit of 2.5 ppm for fishery products [6]. Formalin is delicate, reactive, colourless, exceedingly pure, and inexpensive, making it difficult for humans to detect [7].

Due to the inadequacies of human senses in detecting whether mussels contain formalin and salt, researchers are interested in inventing electronic nose-equipped detection equipment. The electronic nose (*E*-Nose) consist of numerous electronic gas sensors sensitive and selective to volatile chemicals contained in the main chamber of food sample [8]. *E*-Nose is a promising technology since it can expedite the identification of various food varieties at low manufacturing costs. The *E*-nose is a portable solution with benefits such as compact size and cheap ownership cost.

E-Nose technology is outfitted with sensor arrays, data gathering systems, signal processing units, data storage capabilities, and artificial intelligence to identify and analyze numerous vaporized compounds. Multiple kinds of gas sensors, including optical gas sensors, catalytic, electrochemical, polymer, metal oxide semiconductor (MOS), field effect transistor (FET), and piezoelectric sensors, are employed in e-nose technology [9]. *E*-Nose uses one most proper pattern recognition methods to classifying kind of odour [10]. Therefore, an analysis of the selection of classification techniques analysis is necessary.

Previous research has widely used e-noses and followed by classification techniques to ascertain the food products' quality, such as fruit [8], rice [11], meat [12], fish [13] and bacteria detection [14]. [15] success to adopt e-nose technology to determine the bread freshness level. The results show their analysis can achieve up to 98% accuracy. [11] proposed a method to determine the *aspergillus* sp. contamination level. They used least square regression (PLSR) and achieved 98% accuracy. [12] analyze the purity level of meat by using the e-nose instrument and ensemble method. Their results show ensemble method achieves 95.71% accuracy. [16] analyze seafood freshness level of *solea senegalensis*, *mullus barbatus*, and *sepia officinalis* by using an e-nose followed by a k-nearest neighbour classifier.

This research focuses on distinguishing fresh and contaminated mussel shells with e-nose, which is analyzed using several methods. First, we manually collect samples from fishermen in the Indonesian coastal city of Surabaya as the closest sea then we carry out the sample sensing process using an e-nose. Third, we used computational analysis for analysing the sensing data.

2. Material and methods

2.1. Formalin dilution preparation

The required concentration of the solution is according to the test. To get the appropriate concentration of the solution, it is necessary to dilute the solution. This can be done by determining the amount of solution to be made and then calculating the amount of initial solution to be diluted [17]. If a solution is diluted, the volume will increase, and the concentration will decrease, but the total amount of solute will be constant. This dilution can be obtained from the following formula:

$$V_1 \times K_1 = V_2 \times K_2$$

where V_1 represents the initial diluted volume solution (mL), K_1 represents the initial solution concentration (M), V_2 represents the prepared volume solution (mL), and K_2 represents the prepared solution concentration (M).

To make 40% formalin dilution, this research calculates the volume of the original solution (V_1) needed to make a formalin solution with several concentrations from 100 ppm to 500 ppm with the addition of 100 ppm. Then, take the volume (V_1) using a sterile pipette, then put it into a 100 mL beaker glass. Add distilled water into the beaker glass until it matches the volume (V_2) that will be used in the study, then stir using a spatula to mix homogeneously. Each sample has three replications. Then, pour the results of the dilution into a spray bottle that has been provided. Then, close the spray bottle using aluminium foil and a plastic warp. Finally, label the spray bottle according to the concentration to avoid confusion.

2.2. Sample preparation of mussel shells

The sample consists of mussel shells collected from fishermen in the Indonesian coastal city of Surabaya. Twenty samples were taken from each of two groups of mussel shells: the control group without the addition of formalin and the treatment group with the addition of formalin at various doses, from 100 ppm to 500 ppm with the addition of 100 ppm. The mussel shell sample weighs 10 g.

2.3. Gas sensor

The *E*-Nose technology used in this research has six TGS sensors, which are 2600, 2612, 2611, 2602, 2620, 826 types. The E-Nose technology is supported by an Arduino Mega 2560 and a Jupyter-enabled data-collecting system which will linked into a computer device. Each of the sensor able to detect a distinct gas [18]. The TGS-2600 can detect airborne contaminates (hydrogen, ethanol, etc.). It is capable of detecting methane, propane, and isobutane. The TGS-2611 can detect natural gas and methane. TGS-2602 can detect air contaminants (VOC, ammonia, H2S, etc.). The TGS-2620 can detect solvent and alcohol vapours. TGS-826 can detect ammonia. Coward. 1. Describe the experimental setup for *E*-Nose used in this work.

Before the detection series, preheat the E-nose for 30 min. The sequential acquisition procedure has three phases: baseline, acquisition, and refining. Tube 3 sucks in a clean air as control and flows the air into the chamber's input hose through a valve that shuts tubes 1 and 2, preventing clean air from mingling with sample smells. All sensors are in a steady condition for the duration of the 60-s baseline procedure. During capture, the valve shuts tube three and opens tube 1, enabling the target odour to enter the chamber. As the sample odour slowly fills the chamber, the sensor produces a specified output voltage. The sample duration is one hundred seconds. During the rinse operation, valves seal tubes 1 and 3 while leaving tube two open, enabling sample smells to return to the sample tube. Purifying the chamber's gas is the objective of the 120-s flushing procedure. When the gas of interest is present in the chamber, the detection process begins because each of the gas sensor can produce a voltage output. The gas flow's rate is 0.9 L/m.

2.4. Treatments

Each collected sample then put into a 10 mL glass beaker. Then the odour from samples will be detected by six TGS gas sensors, which are 2600, 2612, 2611, 2602, 2620, 826 types. Data from the sensor will be transferred through USB cable to the e-nose, and using Jupyter notebook software, a data acquisition system will be linked to a computer. The resulting data is entered into a file and can be viewed in Microsoft Excel. The output voltage from each sensor is then extracted based on feature extractions and classified using several methods.

3. Computational analysis

3.1. Feature extractions

3.1.1. Principal component analysis (PCA)

The mathematical approach which turns a huge value of correlated variables into a less value of un-correlated variables without sacrificing crucial information is known as PCA [19]. The objective of this method is to simplify by lowering the observable variable's dimensionality. This is accomplished by decorrelating the independent variables by changing the original independent variables into brand-new, uncorrelated variables. Termed constituents the first coordinate is the first principal component derived from the first biggest eigenvalue, and the second coordinate is the second principal component derived from the second largest eigenvalue. After obtaining components of the PCA findings that are independent of multicollinearity, these components become new independent variables that are regressed or evaluated using regression analysis to determine their influence on the dependent variable.

In this investigation, the *E*-Nose data range values were too large. Before performing PCA, we thus conducted data normalization to scale the data to the (0,1) range. In machine learning, data normalization is required for greater precision. A Min-Max-Scaler was used to normalize E-Nose replies to Python. The min-max scaler formula is:

$$x_{sclaed} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

After using a min-max scaler, the data were scaled between the minimum and maximum values (0,1). PCA was then used for the feature extraction procedure.

To perform PCA algorithm, the output of mussel shells sensors, a data matrix ($X = [x_1, x_2, \ldots, x_N]$), where N represents the total number of samples and x_i represents the i^{th} sample, are extracted. The mean of data matrix can be calculated by:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Subtract the mean of data matrix as

$$D = \{d_1, d_2, \ldots, d_N\} = \sum_{i=1}^{N} x_i - \mu$$

Calculate the covariance matrix which population mean is unknown as

$$var(x) = \frac{1}{N-1} D \times D^T$$

Calculate the eigenvectors V and eigenvalues λ of $var(x)$. Rank eigenvectors based on their respective eigenvalues. Then choose the eigenvectors with the largest value of eigenvalues $W = \{v_1, \ldots, v_k\}$. The selected eigenvectors (W) represents the projection space. Finally, in lower dimensional space (W) as $Y = W^T D$, all samples are projected.

3.1.2. Autoencoder and variational autoencoder

Autoencoders and variational autoencoders are neural network types that use an encoder-decoder strategy. Encoder turns the data of high dimensional into low dimensional, whereas decoder transforms vice versa. C_{AE} represents layer values that are compressed and need low dimensional data. To improve the network parameters, each unit's weights and biases will be modified and the network learns the $x = x_{out}$ out identity function. As the loss function, the autoencoder determines the differences between x and x out. One of the frequently used loss function in autoencoders is Mean Squared Error (MSE). MSE represents the mean position data value. The function of an auto-loss encoder is represented by Eq. 12.

As the loss function, the autoencoder determines the difference between x and x_{out}. The common loss function in auto-encoder is Mean Square Error (MSE). MSE represents the mean position data value. The expression below explains the auto-loss encoder's function.

$$f_{loss} = \left(W^T(W(x) + b) + b', x\right)$$

$W(\bullet)$ represents the encoder or decoder weight. b represents the encoder or decoder bias. Based on the information provided by the formula, the encoder output may be calculated by

$$C_{AE} = W_{AE} * I + b$$

It is crucial for the auto-encoder to locate a superior low dimensional data to initialized the weight of auto-encoder, which can be done by using random or RBM distribution. A resulting value are always significantly off when weights are initialized at random. RBM can generate weights and biases based on input's data of latent data structure, enabling backpropagation to avoid bad local minimums to some extent. An RBM-initialized automatic encoder can achieve the intended outcome more effectively.

3.1.3. Truncated singular value decomposition

The Truncated Singular Value Decomposition (SVD) of a matrix A can be obtained using the standard SVD algorithm, which involves computing the transpose of A, applying eigenvalue decomposition to $A * A^T$ and $B * B^T$, where B is the multiplication between matrix A and V, and truncating the resulting U, Σ and A matrices to retain only the k largest singular values and corresponding eigenvectors [20].

3.2. Classification techniques

3.2.1. Support vector machines (SVM)

Popular for constructing hyperplanes, also known as flat borders, are SVM classifiers. This hyperplane divides space into homogenous segments. Therefore, SVM is strong enough to construct intricate associations. Kernel modifications enable SVM classifiers to divide data into higher functional domains [21]. The kernel of the SVM may be represented as:

$$K\left(\overrightarrow{x_i}, \overrightarrow{x_j}\right) = \varphi(\overrightarrow{x_i}) \times (\overrightarrow{x_j}).$$

From the equation, ϕ(x) denotes a function which may shift the feature vectors xi and xj and combine them into a single feature. Numerous SVM core features have been created to categorize various data domains. The linear SVM classifier has little influence on data manipulations. A polynomial SVM kernel of order d augments the data transformation with a simple nonlinear. Radial-based kernels are another SVM kernel extremely close to ANNs and can categorize many types of data well [22,23].

This investigation seeks to differentiate between fresh and infected mussel shells. In machine learning, SVM is thus classed as a supervised learning algorithm that examines a given dataset and identifies data's patterns. Classification and regression analysis can be conducted using this technique [24,25]. Cortes and Vapnik [26] introduced SVM as a efficacious yet precise classification method based on statistical learning theory.

3.2.2. Random forest (RF)

RF [27] is a significant improvement of bagging, including the collecting and averaging of several decorated trees to increase prediction accuracy and control overfitting by fitting a group of decision tree classifiers to distinct subsamples of a dataset. RF performs similarly to boosting for many issues and is simpler to tweak and train. As a result, RF are prevalent and integrated in several software products.

We construct B decision trees using distinct samples, and for regression, we use majority votes for classification and arithmetic mean [28]. Our procedure is demonstrated in the following mathematical required steps. First, draw a bootstrap sample Z^* of size N from the training data. Grow a RF tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached. Select m variables at random from the p variables. Pick the best variable/split-point among the m. Split the node into two daughter nodes. Output the ensemble of trees $\{T_b\}_1^B$.

3.2.3. Decision tree

Decision tree algorithms are members of the supervised learning algorithm family. Decision trees construct classification or regression models as tree-like structures [29]. It splits the dataset into smaller groups while generating decision trees progressively [30]. Final output is a decision tree consisting of decision and leaf nodes. At least two branches emanate from a decision node (such as Outlook) (eg Sunny, Cloudy, Rainy). Leaf nodes (such as Play) represent classifications or options. The root node is the tree's highest decision node, which corresponds to the best predictor. Decision trees can process the qualitative and quantitative input.

The ID3 method of constructing decision trees by J.R. Quinlan employs a top-down greedy search across the space of viable branches without backtracking. ID3 constructs decision trees using information gain and entropy [31]. From the root node upwards, decision trees are constructed, and the data should be partitioned into subsets containing instances with comparable values (homogeneous). Using entropy, the ID3 method calculates sample homogeneity [32]. Considering the sample, if it is uniform, the entropy is zero, and when it is equally distributed, the entropy is one.

The calculation of two types of entropy is a must for constructing a decision tree. The initial entropy is calculated using an attribute. The, used formula to compute the initial entropy is

$$E(S) = \sum_{i=1}^{c} -p_i log_2 p_i$$

Then, the second entropy using the two attributes 's frequency table. The formula to calculate second entropy is

$$E(T,X) = \sum_{c \in X} P(c)E(c)$$

After the data set is separated, the entropy decrease contributes to the information increase. Depending on the characteristic. The purpose of building a decision tree is discovering the qualities which give most information.

3.3. Evaluation metrics

We put two focus evaluation in this experiment. The first evaluation matrix is focus on freshness prediction and the second evaluation is focus on feature extraction/dimension reduction. Evaluation for freshness prediction is accuracy.

Evaluation for feature extraction follows Huang's research in 2022 which is focus on dimension reduction evaluation. We used one evaluation method and marge two evaluation methods. We direct evaluation method is used because these methods are acceptable. However, we marge two other method because the assumption is same, computational efficiency and scalability [33].

4. Results and discussions

TGS sensors are thick-film's type of metal oxide semiconductors that offer a low cost, a long lifetime, and high sensitivity to target gas while employing a simple electrical circuit. These sensors are ideal for use in explosive and hazardous gas leak detectors. When a crystal of metal oxide, such as SnO2, is heated to a high enough temperature in air, negatively charged oxygen is adsorbed onto the crystal's surface. The donor electrons on the crystal surface travel to the adsorbed oxygen, leaving the space charge layer with a positive charge. This results in the formation of a surface potential that serves as a barrier to electron passage. Current passes across the junctions (grain boundaries) of the SnO2 microcrystals inside the sensor. At grain boundaries, adsorbate oxygen generates a potential barrier that prevents the free movement of charge carriers. The sensor's electrical resistance may be traced to this potential barrier. In the presence of a deoxidizing gas, the surface density of negatively charged oxygen drops, hence decreasing the height of the grain boundary barrier. The sensor resistance diminishes as the barrier height falls.

4.1. Gas sensor array (GSA) responses

The GSA Response Test is designed to determine the *E*-Nose sensor's response value. Each E-Nose GSA generates an output voltage as a characteristic pattern based on the sample's properties. We can witness the unique patterns formed by each sensor based on variations in the resultant voltage levels. Fig. 2. shows the output of each sensor for fresh and contaminated mussel samples.

Depending on the pattern features, each GSA in the *E*-nose generates output voltage with a unique pattern. We can witness the unique patterns formed by each sensor based on variations in the resultant voltage levels. Fig. 1. depicts the findings of the GSA reaction in samples of fresh chicken with and without *E. coli* contamination. 2.

E-Nose is a technology which detects and identifies items based on their odours. The generated signal response by the e-nose might appear as a connection between the mussel shell's concentration and odour. When the shell produces a greater concentration of gas, E-nose will generate the higher voltage. Alternatively, the variations in the voltage signal reveal scents with distinct patterns for each chicken variant. Coward. Fig. 2. demonstrates that the TGS 2611 and TGS 2602 are the most sensitive sensors. The TGS 2611 is sensitive to methane, and the resulting voltage difference is evident. Due to formalin's very strong odour and the sensor's sensitivity, the TGS 2611 sensor's output in the formalin mussel sample reached its maximum. Using the H2S-detecting sensor TGS 2602, we determined that formalin-free mussel shells generated the greatest quantity of H2S.

4.2. Feature extractions

Previous research used PCA to analyze correlation across features [8,16]. Previous research used linear discriminant analysis to fusion features [11]. Previous research proposed Statistical Feature Extraction [12].

4.2.1. PCA results

PCA may be used to decrease the number of variables in a collection. Typically, newly formed variables are difficult to comprehend. PCA has proved most effective in applications that emphasize data reduction rather than interpretation, such as picture compression. PCA modifies the dataset to produce a new collection of variables known as main and uncorrelated components. The PCA approach is carried out by exploring the covariance matrix to establish each variable's correlation. Then, determine the eigenvalues of each variable using the covariance matrix. The eigenvalues reflect the data at the newly generated coordinates (principal components). Table 1 displays the results of the eigenvalue computation, while Fig. 3. illustrates the link between eigenvalues and main components. Component 1 is notably distinct from component 2 as seen in Fig. 3.

The present research design revealed two conditions. The first was based on watching and identifying poultry, whereas the second was on *E. coli*-contaminated chicken. He was further separated into two subconditions depending on whether the chicken was healthy or unhealthy. Data were marked for categorizing good vs unhealthy chicken using supervised machine learning. After statistical feature extraction, MATLAB is used for data labelling (section 1.1). The sample data's statistical characteristics are shown in Table 1.

4.2.2. Autoencoder and Variational autoencoder

Autoencoder and variational autoencoder are feature extraction algorithms that are frequently employed for dimension reduction. Ten iterations of our successful autoencoder model for feature extraction required 21,394 milliseconds. Until the loss value approaches zero, both the loss data train and data validation graphs exhibit the same downward slope. Our figure also indicates that there is no chance of the model being overfit. The graph of loss function of each epoch is shown in Fig. 4.

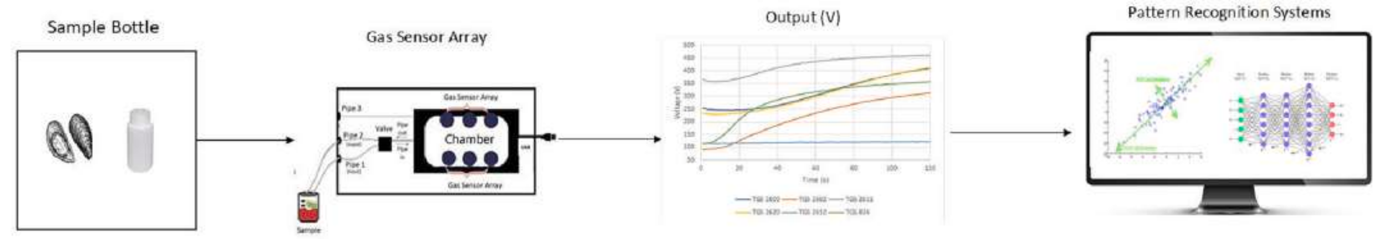

Fig. 1. Diagram of E-nose experimental setup.

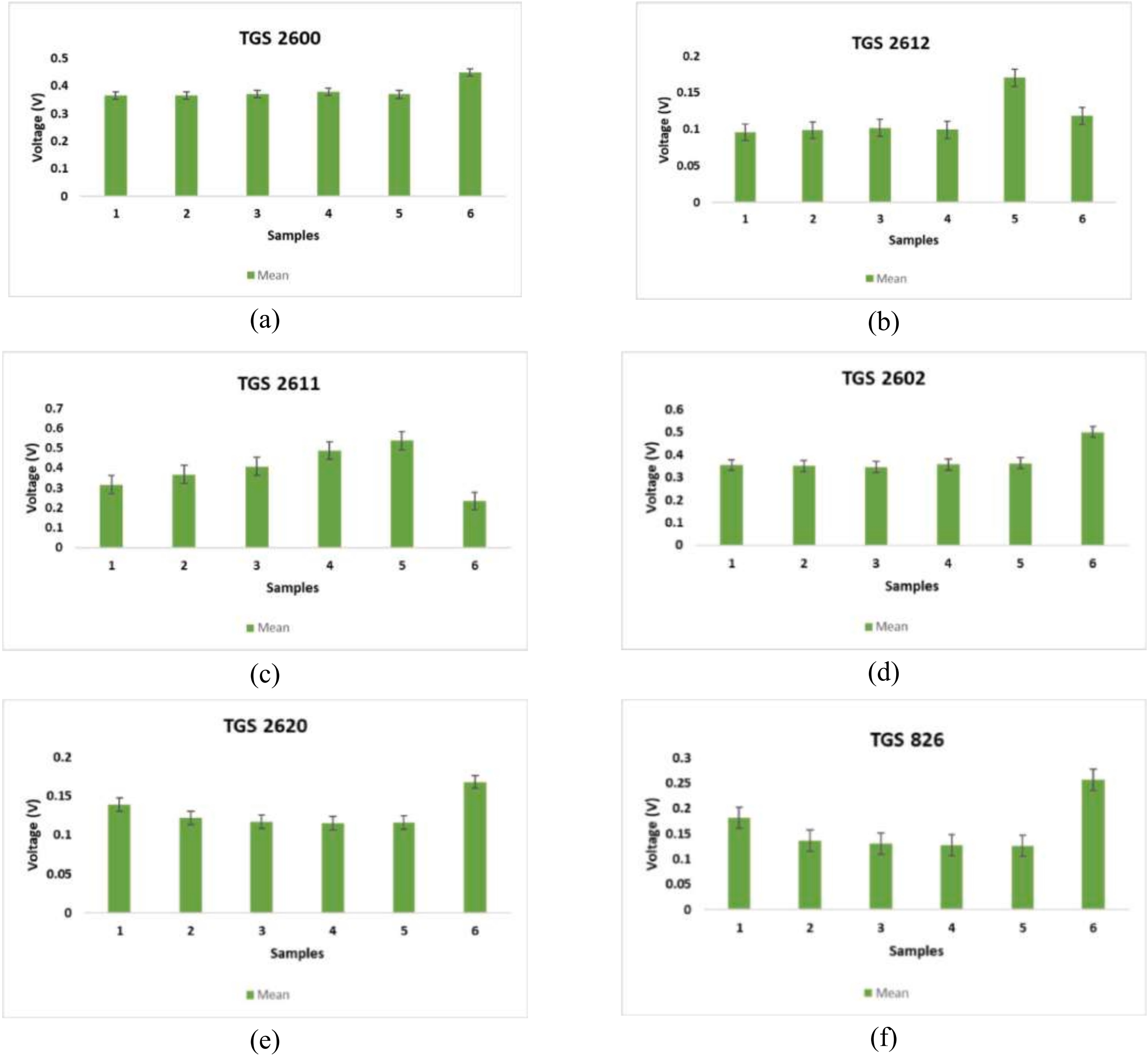

Fig. 2. (a) TGS 2600 result, (b) TGS 2612 result, (c) TGS 2611 result, (d) TGS 2602 result, (e) TGS 2620 result, and (f) TGS 826 result.

4.3. Freshness prediction

Freshness prediction is a part classification objective which aim to differentiate mussel condition (Fresh or contaminated). Feature collected from feature collected procedure then used as input for freshness prediction. We examine three kind of classification techniques: ensemble learning, machine and deep learning. Used machine learning technique in this research is SVM, RF, decision tree, and K-NN. Used ensemble learning in this research is using voting of combination between SVM, linear regression, and decision tree. Used deep learning method in this research is DNN and MLP.

Table 2 demonstrates that all classifiers have an accuracy of 1. This

Table 1
Result of eigenvalue computation.

Component	Eigenvalue	Proportion	Cumulative
1	4.6774	0.780	0.780
2	0.6146	0.102	0.882
3	0.4281	0.071	0.953
4	0.2109	0.035	0.989
5	0.0647	0. 011	0.999

phenomenon occurs in many cases of the chemical domain, especially using odour's receptor. an example, a combination of odour sensors and deep learning achieve 98% accuracy to classify chicken condition [34]. Beside, odour sensor followed by K-Nearest neighbour able to achieve perfect evaluation which is 100% accuracy [16]. [8] also shows 1.00 accuracy is possible to be achieved if a class is convergent grouped in a location. The phenomenon occurs because all feature representation can produce enough information for classify data.

Beside performance, this research record training time and prediction time to know the fastest classification technique for mussel. Fig. 5. show the training time of many classifications' technique. The fastest classifier in training and testing is decision tree. This occurs because the decision tree technique is very dependent on the depth of the tree, yet the data requires a shallow depth. The worst classifier is ensemble learning. It is fit to our mind, when others machine learning only used one classifier, ensemble learning need to train three classifiers. Training time. K-NN be the worst classifier for prediction. This condition occurs because K-NN is memory-based classifier which need to calculate when a test data is coming. Almost all classifiers need more time to train except decision tree and K-NN. K-NN faced this condition because K-NN does not do any calculation during training (memory-based).

Table 3 show the including deep learning method and a full-time processing comparison. All deep learning model need more than 11 s to fit the dataset. The fastest deep learning approach is deep neural network.

Tables 4 and 5 shows training and prediction time for transformed data (PCA and AE). The results indicate most classifier need more time to

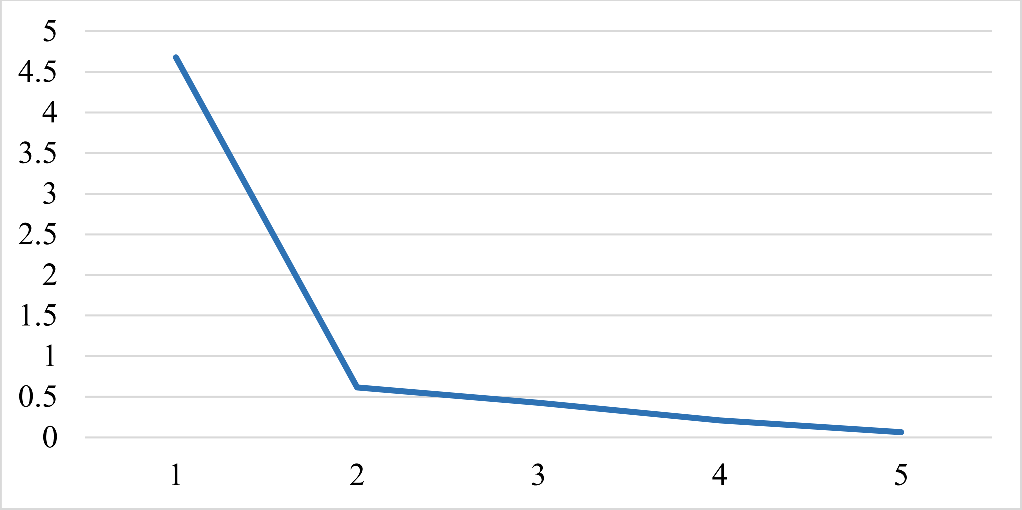

Fig. 3. Graph of relationship between eigenvalue and principal component.

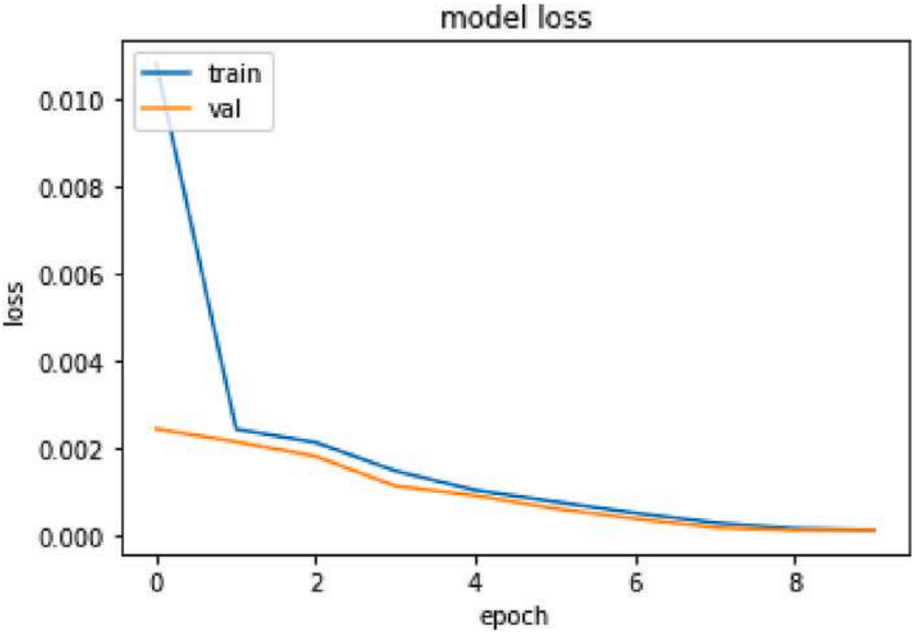

Fig. 4. MSE score of each epoch.

Table 2
Freshness prediction evaluation.

Prediction Method	Origin Data (no dimension reduction)	Principal Component Analysis	Autoencoder	Variational Autoencoder	Truncated Singular Value Decomposition
SVM	1.00	1.00	1.00	1.00	1.00
Decision Tree	1.00	1.00	1.00	1.00	1.00
RF	1.00	1.00	1.00	1.00	1.00
K-NN	1.00	1.00	1.00	1.00	1.00
Ensemble (SVM, LR, DT)	1.00	1.00	1.00	1.00	1.00
DNN	1.00	1.00	1.00	1.00	1.00
MLP	1.00	1.00	1.00	1.00	1.00

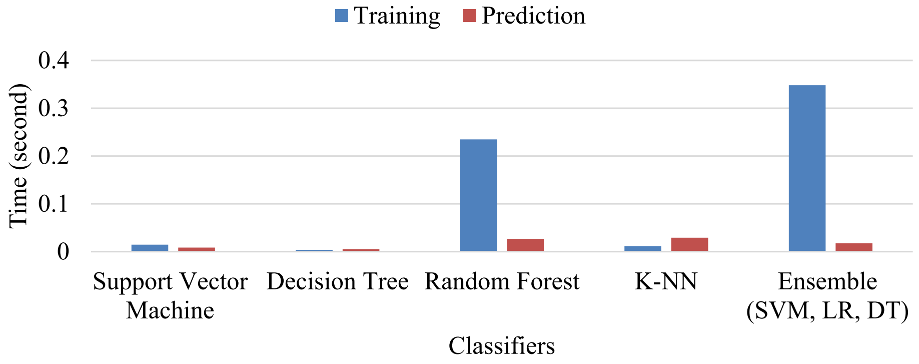

Fig. 5. Training time.

Table 3
Time processing comparison.

Method	Training Time (second)	Prediction Time (second)
SVM	0.0145	0.0083
Decision Tree	0.0036	0.0052
RF	0.2348	0.0266
K-NN	0.0117	0.0291
Ensemble (SVM, LR, DT)	0.3481	0.0174
DNN	11.0839	1.0646
MLP	11.8315	1.3147

Table 5
Time processing comparison with autoencoder.

Method	Training Time (second)	Prediction Time (second)
SVM	0.0344	0.0119
Decision Tree	0.0042	0.0064
RF	0.3066	0.0266
K-NN	0.0082	0.0248
Ensemble (SVM, LR, DT)	0.3440	0.0243
DNN	28.6630	22.3423
MLP	29.0374	24.8549

fit when dimension of data is reduced (to be 2 dimension). It is because lower dimension data is having a complexity problem. Even the dimension is lower, it does not mean the processing time is faster. Besides, we found SVM and K-NN need lower time when the input data is transformed data. Table 6 shows encoded vector evaluation.

Fig. 6. shows both deep learning model (deep neural network and multi-layer perception) has similar condition. They fit to dataset in the third epoch. However, a deep learning neural network model is more generalizable than multi-layer. In the first epoch, deep learning data.

4.4. *Detailed analysis of feature extractions*

4.4.1. *Analysis of data reconstruction*

A dimension reduction technique must be able to return a value to its original form (reconstruction). Reconstruction is required to determine whether information is lost when one dimension is merged with another. Fig. 7 shows RMSE score of each feature extraction. We compared the reconstruction values to displacement distance computations, such as RMSE. We discovered that the autoencoder lost the least amount of knowledge. The RMSE for the autoencoder was only 0.001, indicating that the average difference between the starting value and the reconstruction was only 0.001. The value obtained by the autoencoder is superior to the PCA technique utilized in numerous research.

Table 4
Time processing comparison with principal component analysis.

Method	Training Time (second)	Prediction Time (second)
SVM	0.0103	0.0061
Decision Tree	0.0040	0.0052
RF	0.2968	0.0225
K-NN	0.0064	0.0237
Ensemble (SVM, LR, DT)	0.3830	0.0185
DNN	16.6689	1.8271
MLP	14.9556	1.8624

Table 6
Encoded vector evaluation.

Method	Local Structure Preservation	Quantitative Global Structure Preservation
Principal Component Analysis	**0.331**	**0.99**
Autoencoder	0.085	0.89
Variational Autoencoder	0.144	0.97
T-SVD	0.308	0.99

Our research gives new insight for feature extraction in the chemical domain. Different findings have emerged from continuing studies in the field of computer science. We discovered that principal component analysis outperforms the autoencoder and T-SVD models [35]. We also

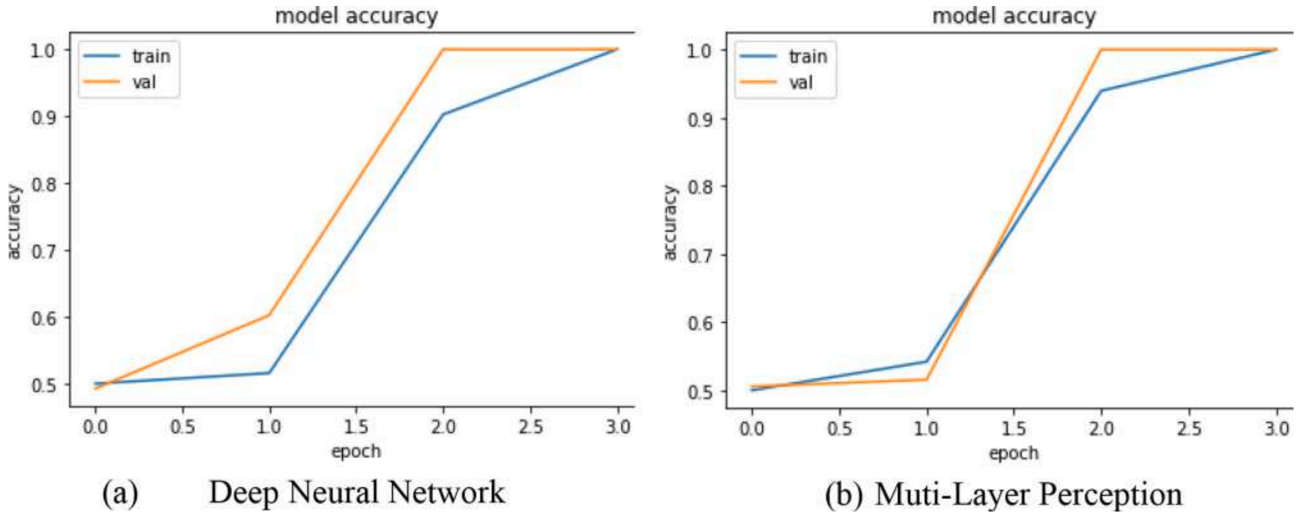

Fig. 6. Accuracy of each epoch.

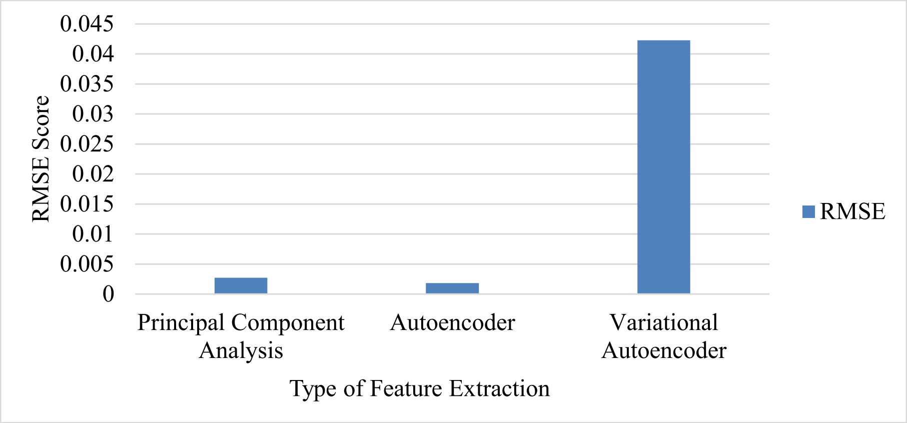

Fig. 7. RMSE score of each feature extraction.

find linear transformation is able implemented in the chemical domain using odour receptors, especially for mussels.

4.4.2. *Analysis of encoded vector*

Encoded vector analysis is a method for determining a data's potential level of analysis when it is in the encoded phase. Encoded vector analysis is a relatively new technique that is still being developed and evaluated in the field of natural language processing. As such, there may not be many papers that have used this specific technique. There are several research that used encoded analysis. Yang Liu et al. [36] used text encoding in their classification tasks for representing documents of different length, subject matter, and language. Andrew M Dai et al. [37] used paragraph vector that learns fixed-length vector representations of variable-length pieces of text, such as paragraphs and documents. Daniel Cer et al. [38] trained encoding sentences into embedding vectors that specifically target transfer learning. Fig. 8 shows the graph of the encoded vector analysis. The analysis is simplified by comparing the three approaches in two dimensions. We observed that the number of class groups for formalin-free shellfish was the same for all three ways, and that the number of class groups for formalin-free shellfish was nearly the same for all three methods. One formalin-formalin shellfish group and between six and seven non-formalinated shellfish groupings were identified. The principal component analysis, autoencoder, and T-SVD each had six non-formalinated shellfish groups; the variational auto encoder had seven. In addition to disparities in the number of groups, we discovered variances in the degree of divergence between the three techniques.

Principal component analysis provides the most comprehensive coverage, followed by autoencoder and variational autoencoder. A high level of convergence facilitates the achievement of a local optimal by a classifier. In contrast, the variational autoencoder produces a greater number of variational and clear encoded vectors. Clearer data facilitates the classifier's attainment of the global optimal.

Encoded vector was evaluated by following [33] evaluation technique. Local structure preservation requires that high dimensional space neighbours remain low dimensional space neighbours. It is conserved when the local neighbourhoods in the high dimensional space resemble local neighbourhoods in the low dimensional space. Local Structure Preservation is one of several measures that can be used to evaluate the quality of vector embeddings, along with measures such as semantic similarity, syntactic coherence, and downstream task performance. The preservation of global structure necessitates the maintenance of relative positions between clusters and larger-scale manifold structures. Quantitative Global Structure Preservation (QGSP) evaluates how well the text encoding method preserves the global structure of the original group. Specifically, QGSP measures how well the vector embeddings preserve the pairwise similarity relationships between all pairs of groups in a corpus. Both Local Structure Preservation and QGSP higher score generally indicates better performance.

Principal component analysis achieves the best score in preservation

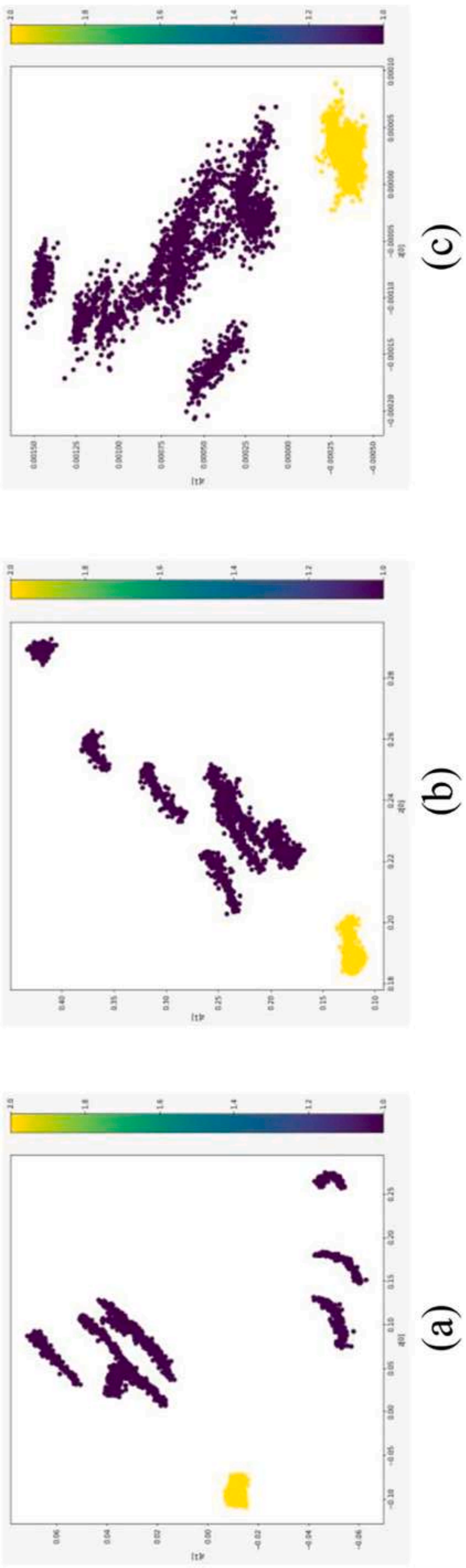

Fig. 8. 2D-Encoded vector of (a)Principal Component Analysis, (b) Autoencoder, and (c) Variational Autoencoder, (d) T-SVD.

of the local structure. PCA can keep an average of 33% nearest data in the same neighbourhood. While variational autoencoder can keep 14% nearest data in the same neighbourhood, and autoencoder can keep 8% nearest data in the same neighbourhood. While T-SVD can keep 30% nearest data in the same neighbourhood. Based on these proportion scores, PCA outperforms local structure preservation.

The principal component analysis also achieves the best score in the global quantitative structure of preservation. Spearman correlations show all feature extraction has positive correlation with original shape. It indicates PCA, AE, VAE, and T-SVD can handle feature extraction task. However, PCA achieve better representation than AE, VAE, and T-SVD.

5. Conclusions

This research has proven that GSA can identify gas's type inside the sample. It is showed by the presence of a variation in the patterns of output voltage of the sensor in each of the sample variation. We demonstrate that all classifiers have an accuracy of 1. The phenomenon occurs because all feature representation can produce enough information for classify data. Beside performance, this research record training time and prediction time to know the fastest classification technique for the mussel. The fastest classifier in training and testing is the decision tree. This is due to the fact that the decision tree technique is highly dependent on the depth of the tree, whereas the data demands a shallow depth. The worst classifier is ensemble learning. Principal component analysis achieves the best score in preserving the local structure. PCA can keep an average of 33% nearest data in the same neighbourhood. While variational autoencoder can keep 14% nearest data in the same neighbourhood, and autoencoder can keep 8% nearest data in the same neighbourhood. Based on these proportion scores, PCA outperforms local structure preservation. The principal component analysis also achieves the best score in the global quantitative structure of conservation. The limitation of this study is the voltage generation from the gas sensor is the only observation. There is no comparison to other analytical techniques such as Gas Chromatography Mass Spectrometry (GCMS) for determined gas compound's composition.

Future research

Prediction of freshness and dimension reduction were conducted. The freshness forecast reaches complete accuracy (1.00) because the infected mussel shells from Kenjeran Beach are clustered and near to class. We will continue to monitor the issue by searching for contaminated shells on more beaches, including Kepetingan Beach and Madura Strait. Other beaches projected to not having a complete accuracy of 1.0, which give us more observation. Dimensionality reduction is also effective. We wish to discover the best dimensionality reduction of the contaminated shell. It turns out that almost every dimension reduction may be enhanced.

CRediT authorship contribution statement

Cendra Devayana Putra: Conceptualization, Methodology, Validation. **Achmad Ilham Fanany Al Isyrofie:** Conceptualization, Methodology, Validation. **Suryani Dyah Astuti:** Methodology, Validation, Supervision. **Berliana Devianti Putri:** Methodology, Validation. **Dyah Rohmatul Ummah:** Conceptualization, Methodology. **Miratul Khasanah:** Methodology, Validation. **Perwira Annissa Dyah Permatasari:** Methodology, Validation. **Ardiyansyah Syahrom:** Conceptualization, Methodology, Validation, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could be perceived as having influenced the work described in this paper.

Data availability

The authors are unable or have chosen not to specify which data has been used.

Acknowledgements

We are gratified for the International Research Collaboration Funding of Airlangga University no. 355/UN3.15/PT/2023.

References

[1] F. Nowshad, M.N. Islam, M.S. Khan, Concentration and formation behavior of naturally occurring formaldehyde in foods, Agric. Food Secur. 7 (2018) 1–8, https://doi.org/10.1186/s40066-018-0166-4.

[2] S. Laly, E. Priya, S. Panda, A. Zynudheen, Formaldehyde in seafood: A review, Fish. Technol. 55 (2018) 87–93.

[3] JMPR, Pesticide residues in food — 2010 Toxicological evaluations sponsored jointly by FAO and WHO, World Health. 2 (2010) 595.

[4] O.F.T.H.E. Council, Regulation (EU) No 524/2013 of the European Parliament and of the council, Fundam. Texts Eur. Priv. Law. (2020) 1–123, https://doi.org/10.5040/9781782258674.0009.

[5] U.S. Food and Drug Administration, CFR - Code of Federal Regulations. Https://Www.Accessdata.Fda.Gov/Scripts/Cdrh/Cfdocs/Cfcfr/Cfrsearch.Cfm?Fr=182.20 721, 2019, pp. 1–10. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfCFR/CFRSearch.cfm?fr=170.3&SearchTerm=170.3.

[6] C.J. Stock, L. Carpenter, J. Ying, T. Greene, Gabapentin versus chlordiazepoxide for outpatient alcohol detoxification treatment, Ann. Pharmacother. 47 (2013) 961–969, https://doi.org/10.1345/aph.1R751.

[7] M.S. Hoque, L. Jacxsens, B. De Meulenaer, A.K.M.N. Alam, Quantitative risk assessment for formalin treatment in fish preservation: food safety concern in local market of Bangladesh, Procedia Food Sci. 6 (2016) 151–158, https://doi.org/10.1016/j.profoo.2016.02.037.

[8] Q. Liu, N. Zhao, D. Zhou, Y. Sun, K. Sun, L. Pan, K. Tu, Discrimination and growth tracking of fungi contamination in peaches using electronic nose, Food Chem. 262 (2018) 226–234, https://doi.org/10.1016/j.foodchem.2018.04.100.

[9] A.D. Wilson, M. Baietto, Advances in electronic-nose technologies developed for biomedical applications, Sensors. 11 (2011) 1105–1176, https://doi.org/10.3390/s110101105.

[10] A.I.F.A. Isyrofie, R. Afifudin, Y. Susilo, S. Kholimatussa'diyah Winarno, S.D. Astuti, Role of bacterial types and odor for early detection accuracy of bacteria with gas array, in: AIP Conference Proceedings vol. 2554, No. 1, AIP Publishing LLC, 2023, p. 060003.

[11] S. Gu, Z.H. Wang, W. Chen, J. Wang, Early identification of aspergillus spp. contamination in milled rice by E-nose combined with chemometrics, J. Sci. Food Agric. 101 (2021) 4220–4228, https://doi.org/10.1002/jsfa.11061.

[12] S. Wakhid, R. Sarno, S.I. Sabilla, The effect of gas concentration on detection and classification of beef and pork mixtures using E-nose, Comput. Electron. Agric. 195 (2022), 106838, https://doi.org/10.1016/j.compag.2022.106838.

[13] S.D. Astuti, S.D. Al Isyrofie, A.I.F. Nashichah, R. Kashif, M. Mujiwati, T. Susilo, Y. A. Syahrom, Gas Array sensors based on electronic nose for detection of tuna (Euthynnus Affinis) contaminated by pseudomonas aeruginosa, J. Medical Signals Sens. 12 (4) (2022) 306.

[14] A.A.S. Pradhana, S.D. Astuti, M. Khasanah, R.K.D. Ardianti, Detection of gas concentrations based on age on Staphylococcus aureus biofilms with gas array sensors, in: AIP Conference Proceedings 2314, 2020, 030012, https://doi.org/10.1063/5.0034112.

[15] B. Botre, D. Gharpure, Analysis of volatile bread aroma for evaluation of bread freshness using an electronic nose (E-nose), Mater. Manuf. Process. 21 (2006) 279–283, https://doi.org/10.1080/10426910500464677.

[16] S. Grassi, S. Benedetti, L. Magnani, A. Pianezzola, S. Buratti, Seafood freshness: e-nose data for classification purposes, Food Control 138 (2022), 108994.

[17] J.T. Nordeide, Accuracy of body mass estimates of formalin-preserved fish – a review, J. Fish Biol. 96 (2020) 288–296, https://doi.org/10.1111/jfb.14146.

[18] S.D. Astuti, Y. Mukhammad, S.A.J. Duli, A.P. Putra, E.M. Setiawatie, K. Triyana, Gas sensor array system properties for detecting bacterial biofilms, J. Med. Signals Sens. 9 (2019) 158–164, https://doi.org/10.4103/jmss.JMSS_60_18.

[19] S. Narasimhan, S.L. Shah, Model identification and error covariance matrix estimation from noisy data using PCA, IFAC Proc. 37 (2004) 511–516, https://doi.org/10.1016/s1474-6670(17)38783-9.

[20] N. Halko, P.G. Martinsson, J.A. Tropp, Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions, SIAM Rev. 53 (2011) 217–288, https://doi.org/10.1137/090771806.

[21] P. Borowik, L. Adamowicz, R. Tarakowski, P. Wacławik, T. Oszako, S. Ślusarski, M. Tkaczyk, Application of a low-cost electronic nose for differentiation between pathogenic oomycetes pythium intermedium and phytophthora plurivora, Sensors (Switzerland). 21 (2021) 1–16, https://doi.org/10.3390/s21041326.

[22] H. Feng, W. Wang, B. Chen, X. Zhang, Evaluation on frozen shellfish quality by Blockchain based multi-sensors monitoring and SVM algorithm during cold storage, IEEE Access. 8 (2020) 54361–54370, https://doi.org/10.1109/ACCESS.2020.2977723.

[23] S.I. Sabilla, R. Sarno, K. Triyana, K. Hayashi, Deep learning in a sensor array system based on the distribution of volatile compounds from meat cuts using GC–MS analysis, Sens. Bio-Sensing Res. 29 (2020), 100371, https://doi.org/10.1016/j.sbsr.2020.100371.

[24] X. Wei, Y. Zhang, D. Wu, Z. Wei, K. Chen, Rapid and non-destructive detection of decay in peach fruit at the cold environment using a self-developed handheld electronic-nose system, Food Anal. Methods 11 (2018) 2990–3004, https://doi.org/10.1007/s12161-018-1286-y.

[25] K. Brudzewski, S. Osowki, T. Markiewicz, Classification of milk by means of an electronic nose and SVM neural network, Sensors Actuators B Chem. 98 (2004) 291–298, https://doi.org/10.1016/j.snb.2003.10.028.

[26] C. Cortes, V. Vapnik, Suport vector network, Kluwer Acad. Publ. 20 (1995) 273–297.

[27] W. Deng, Z. Huang, J. Zhang, J. Xu, A data mining based system for transaction fraud detection, 2021 IEEE Int, Conf. Consum. Electron. Comput. Eng. ICCECE 2021 (2021) 542–545, https://doi.org/10.1109/ICCECE51280.2021.9342376.

[28] S.D. Astuti, M.H. Tamimi, A.A.S. Pradhana, K.A. Alamsyah, H. Purnobasuki, M. Khasanah, Y. Susilo, K. Triyana, M. Kashif, A. Syahrom, Gas sensor array to classify the chicken meat with E. coli contaminant by using random forest and support vector machine, Biosens. Bioelectron. X. 9 (2021), 100083, https://doi.org/10.1016/j.biosx.2021.100083.

[29] B. Charbuty, A. Abdulazeez, Classification based on decision tree algorithm for machine learning, J. Appl. Sci. Technol. Trends. 2 (2021) 20–28, https://doi.org/10.38094/jastt20165.

[30] Y. Huang, Y. Lan, S.J. Thomson, A. Fang, W.C. Hoffmann, R.E. Lacey, Development of soft computing and applications in agricultural and biological engineering, Comput. Electron. Agric. 71 (2010) 107–127, https://doi.org/10.1016/j.compag.2010.01.001.

[31] W. Xiaohu, W. Lele, L. Nianfeng, An application of decision tree based on ID3, Phys. Procedia 25 (2012) 1017–1021, https://doi.org/10.1016/j.phpro.2012.03.193.

[32] R. Tanone, H.B. Prasetya, Designing and implementing an organoleptic test application for food products using android based decision tree algorithm, Int. J. Interact. Mob. Technol. 13 (2019) 134–149, https://doi.org/10.3991/ijim.v13i10.9669.

[33] H. Huang, Y. Wang, C. Rudin, E.P. Browne, Towards a comprehensive evaluation of dimension reduction methods for transcriptomic data visualization, Commun. Biol. 5 (2022), https://doi.org/10.1038/s42003-022-03628-x.

[34] A.I.F. Al Isyrofie, M. Kashif, A.K. Aji, N. Aidatuzzahro, A. Rahmatillah, W. Winarno, Y. Susilo, A. Syahrom, S.D. Astuti, Odor clustering using a gas sensor Array system of chicken meat based on temperature variations and storage time, Sens. Bio-Sensing Res. 37 (2022), 100508, https://doi.org/10.2139/ssrn.4124077.

[35] X. Li, T. Zhang, X. Zhao, Z. Yi, Guided autoencoder for dimensionality reduction of pedestrian features, Appl. Intell. 50 (2020) 4557–4567, https://doi.org/10.1007/s10489-020-01813-1.

[36] Y. Liu, M. Lapata, Learning structured text representations, Trans. Assoc Comput. Linguist. 6 (2018) 63–75, https://doi.org/10.1162/tacl_a_00005.

[37] A.M. Dai, C. Olah, Q.V. Le, Document Embedding with Paragraph Vectors, 2015, pp. 1–8. http://arxiv.org/abs/1507.07998.

[38] S. Smetanin, M. Komarov, Deep transfer learning baselines for sentiment analysis in Russian, Inf. Process. Manag. 58 (2021), 102484, https://doi.org/10.1016/J.IPM.2020.102484.