A New Approach to Signal Filtering Method Using K-Means Clustering and Distance-Based Kalman Filtering

Abstract: Human axillary odours taken by an electronic nose (e-nose) device that uses a Metal Oxide Semiconductor (MOS) sensor not only contains a gas signal from the pure source of the axillary odour but also has the potential to contain other substances such as perfume and deodorant. This situation requires noise reduction so that dirty data can be cleaned and produce better predictions without wasting a lot of data. The approach taken in this study is to detect data clusters and data centroids from each reference data. Dimensional reduction using Linear Discriminant Analysis (LDA) on the reference data is carried out, then look for the centroid of each data using K-Means Clustering and use it to be a good signal estimate and process using Kalman Filtering so that it can be used to process axillary odour data containing deodorant. The proposed method was tested by a stacked Deep Neural Network (DNN) approach and can increase accuracy by 18.95% and balanced accuracy by 11.87% compared to original invalid data before filtering. The proposed method is also tested by other classification methods and able to produce the highest accuracy with 79.29% in Support Vector Classifier (SVC) and Multi-Layer Perception (MLP), while other filtering methods only get the highest accuracy with 69.03% also in SVC and MLP. We also analysed the execution time of each tested methods.

Keywords: Electronic Nose, K-Means Clustering, Kalman Filtering, Noise Reduction, Signal Processing.

1 Introduction

In the human axilla, micro-organisms produce Volatile Organic Compounds (VOCs) mixed with sweat. VOCs are classified according to their molecular structure or functional groups including aliphatic hydrocarbons (among them chlorinated - halocarbons), aromatic hydrocarbons, alcohols, ethers, esters, and aldehydes [1]. Studies related to electronic nose (e-

nose) for the detection of VOCs in human body are still an interesting topic of discussion and are widely discussed by researchers. Accurate odour measurement is essential for many applications such as developing odour distribution models and estimating odour source locations based on odour data [2]. However, human axillary odours are still very susceptible to other scents that can interfere with the detection process, including the use of deodorants, alcohol, and so on. The output signal from the Metal Oxide Semiconductor (MOS) gas sensor not only contains a gas signal from the source of the odour but also contains noise data called noise. Noise causes inaccuracies in analysing data and estimating odour strength. The process of collecting axillary odour data must go through a noise cleaning process to remove odours other than odour before the detection process is carried out.

In a previous study, an e-nose consisting of multiple sensors and an intelligent analysis system was developed but the noise reduction of gas sensors was not well investigated [3], [4]. E-nose is also proposed for the detection of lung cancer using Principal Component Analysis (PCA) and Independent Component Analysis (ICA) methods [5]. Human body odour is used to track the improvement in the health status of athletes [6] but this study has not applied any signal filtering to process the odour signal data generated by the axilla. E-nose is also used for diabetes detection through VOCs produced by human breath [7]. This study used Discrete Wavelet Transform (DWT) as the pre-processing signal, but there is no comprehensive noise reduction study yet.

Detection of infectious respiratory diseases is proposed using e-nose by applying feature extraction and stacked DNN methods [8]. The accuracy results obtained are good, but here we use data that has been selected manually to remove outlier data so that the processed data is data that is clean from noise. Outlier detection in human axillary odour data is proposed by using an adaptive filter mechanism [9]. The method proposed here succeeds in separating valid data and invalid data, but has not been able to remove the noise contained in each invalid data. Invalid data that is detected will then be discarded and not used. In fact, the data is very large and has the potential for noise reduction processes to be carried out.

Framework for noise filtering using a fine-tuned Discrete Wavelet Transform (DWT) was proposed [10]. This study can reduce the noise generated by the e-nose by reducing the anomaly of the instantaneous signal increase and producing a fairly good accuracy. A combination of DWT and Long Short-Term Memory (LSTM) is proposed to reduce noise in meat data [11]. The proposed method obtains a higher accuracy than the conventional classification method by smoothing the e-nose signal and eliminating irregular signals.

However, the [10] and [11] were only optimal for data with noise generated by the data itself, not caused by a mixture of other substances.

Kalman Filtering, although it is a method that has been around for quite some time, is still widely discussed by researchers with some modifications. Kalman Filtering is used to perform noise filtering on electrical terminals by combining it with a Neural Network which can smooth the signal quite well but has not been tested on signals with too much noise and takes longer than other methods [12]. A Kalman Filter approach called Gaussian-Student's t mixture (GSTM) was also proposed which is claimed to be able to eliminate deep noise [13]. Kalman Filter is also combined with a probabilistic method for position tracking and finding existing anomalies, but this method is not optimal when applied to signal data and is not applied as noise reduction [14]. Kalman Filter has also been modified called Extended Kalman Filtering (EKF) for positioning using radar and lidar [15]. Kalman Filter is also integrated with the Support Vector Machine (SVM) to improve distance calculation [16]. An Adaptive Image Retrieval using Kalman Filter is also proposed for object image discovery by applying Red Green Blue (RGB) and gray level color analysis [17]. Kalman Filtering is also used to predict the spread of Covid-19 by applying ensemble learning and regression techniques to determine the distribution in the future [18].

Based on several previous studies, no one has discussed the removal of noise from the e-nose signal from data on human axillary odour contaminated with deodorants and perfumes. We propose this research to contribute to the development of e-nose implementation so that it can be a material for discussion and add references for further research. The approach taken in this study is to use Linear Discriminant Analysis (LDA) as a method for reducing dimensions, K-Means Clustering to detect clusters and centroids from each reference data which is deodorant-free axillary odour data and becomes the fundamental truth of clean signal data that is free of deodorant. Then the reference data is used in the Kalman Filtering process in order to obtain a signal that is close to the correct signal even though the data is actually data contaminated with deodorant. The purpose of this study was to determine the pure value of the tested axillary odour and to reduce odours other than the axillary odour so that the processed data is pure odour data and produces a higher level of accuracy.

2 Methodology

This section briefly explains the principle of the Electronic Nose (e-nose) system. Then, our proposed distance based Kalman filtering method details are also discussed.

2.1 Proposed Method

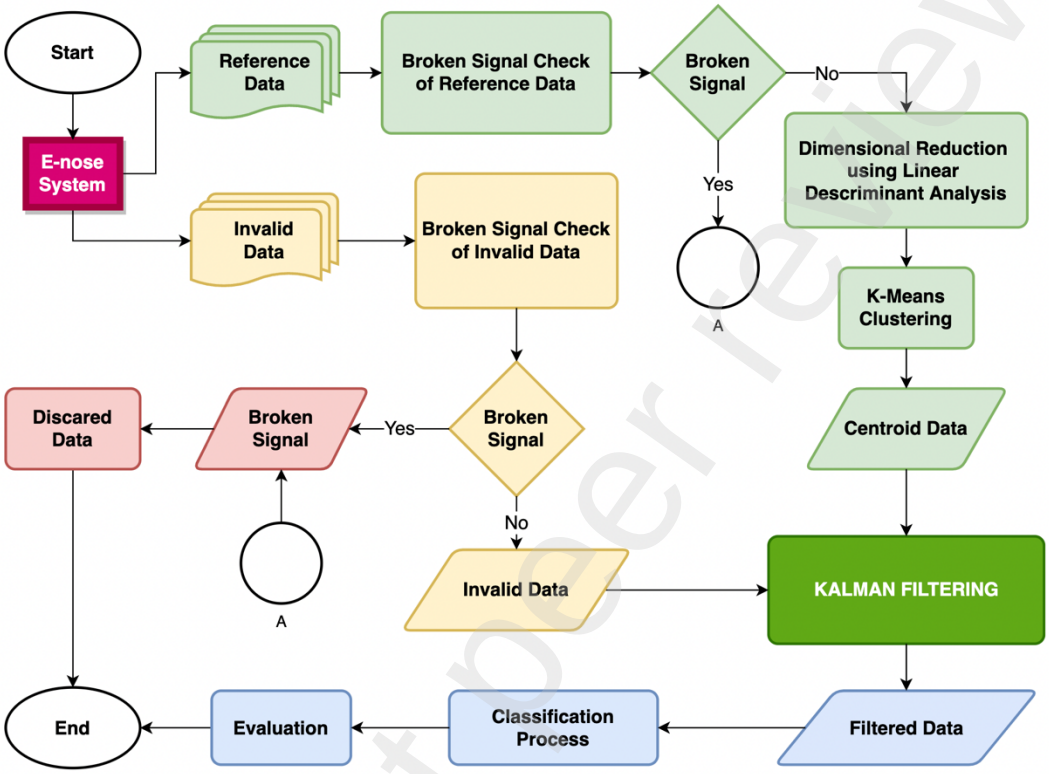

Figure 1 Proposed Method Mechanism

Figure 1 shows the approach used in this study is to detect clusters and centroids from each reference data which is deodorant-free axillary odour data to become the basis for deodorant-free clean signal data. Then the reference data is used to become the Original Estimate and the deodorant axillary odour data to become the Observation Estimate in the Kalman Filtering process so that a signal that is close to the correct signal is obtained even though the data is actually data contaminated with deodorant.

Human axillary odour data that is still clean or free from the use of deodorant is carried out by checking the broken signal using the threshold delta test method from any existing raw sensor data. Then the clustering process is carried out using K-Means Clustering. Then from the cluster, the centroid of each data line is determined in order to obtain the axillary centroid that is free from noise as a normal axillary signal image. Furthermore, the centroid of normal axillary odour data is entered as the Original Estimate parameter in the Kalman Filtering

process to calculate the actual sensor value discovery process in the axillary odour data containing deodorant.

Axillary odour data with deodorant is carried out by checking the sensor for broken signal as before. When the e-nose signal data is correct, then a feature extraction process is carried out for processing using Kalman Filtering. The data from Kalman filtering is entered into the classification process with the previous classification model in order to increase accuracy.

2.2 Electronic Nose (e-nose)

In this study, data was collected by taking samples of human axillary odour using an electronic nose (e-nose) which is an odour sensing device with gas sensors embedded in it. Gas sensors are electronic devices that detect various types of gases in the air and identify signals according to the required specifications [19]. This study uses Metal Oxide Semiconductor (MOS) sensor array in the e-nose system. The MOS Sensor consists of a sensing membrane and a transducer. Figure 2 (a) is a proposed e-nose design used in this study. The Odour from the human axilla is taken by using a hose to enter the e-nose system, in which there is a sensor that is controlled using a micro-computer. Then Figure 2 (b) is the resulting data stored in a computer for further analysis and classification processes using deep learning.

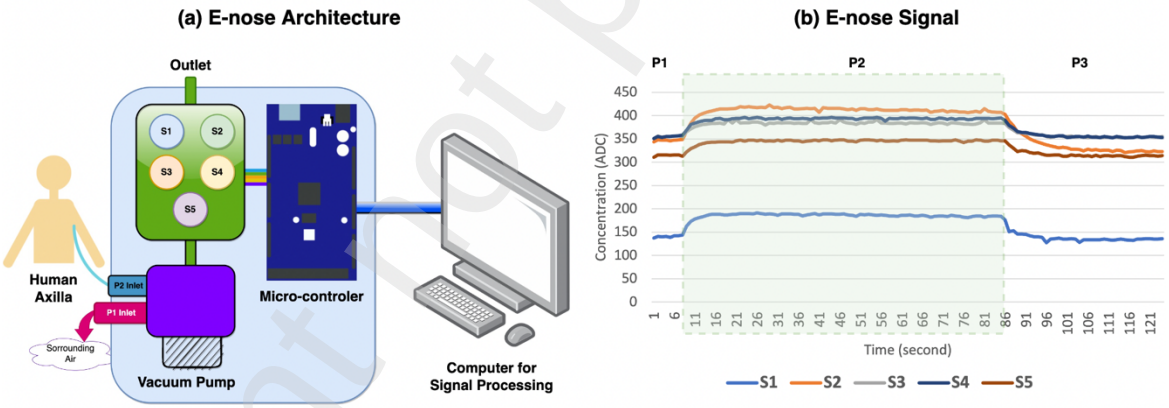

Figure 2 E-nose System (a) E-nose Data Collection, (b) Signal Result of e-nose

The resulting data is then saved into a Comma Separated Value (CSV) file for each subject. In the CSV file, the data is divided into 3 parts, namely P1, P2, and P3. P1 is a condition where the e-nose is identifying the surrounding air and has not taken data from human axillary sweat. P2 is data taken from odour and is the core data of the analysis and classification process in this study. P3 is the process of taking free air back by e-nose which aims to clean the sensor so that the next subject data collection process is not disturbed by the previous data.

E-nose records 8 seconds of P1 at the beginning of the data stored in the CSV file, then e-nose takes 78 seconds of P2, and then continues with 39 seconds of P3. However, what is used for the classification process is the 11th row from P2 to the end of P2, this is because at the beginning of the P2 data collection process, the sensor needs time intervals to get full axillary odour data from the previous one that took free air data. The movement of the sensor-generated by the e-nose from the human axilla is shown in Figure 2 (b).

2.3 Broken Signal Check Mechanism

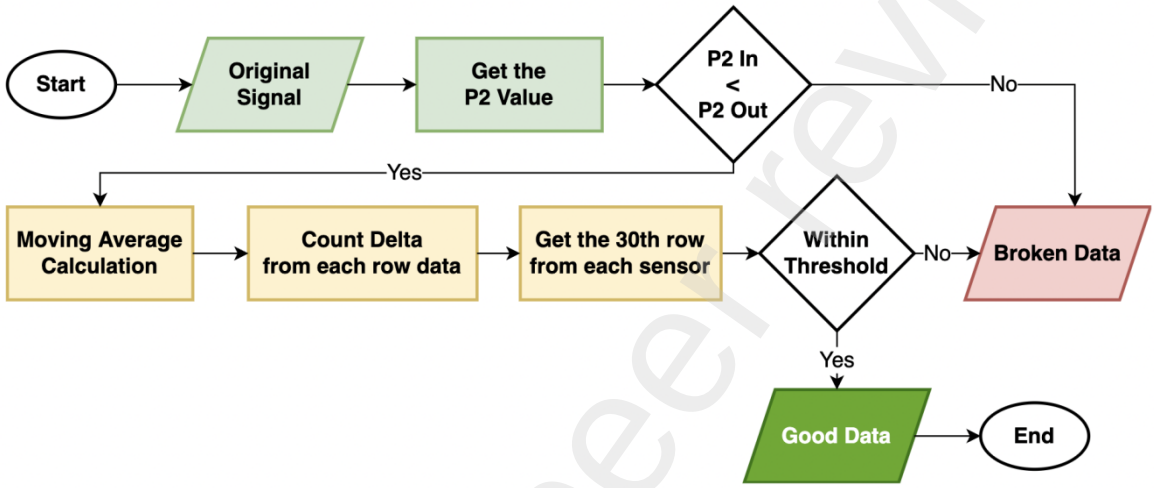

Figure 3 Broken Signal Check Mechanism (FilterMAV)

In addition to the use of deodorant which causes noise, retrieval of signals obtained from the human axilla taken using an e-nose is very susceptible to data collection errors. So, before doing the calculation process or signal analysis, it is necessary to go through a mechanism to check whether the signal is good or broken signal. Good means that the signals generated move according to the specifications of each sensor used, while broken signals mean that they do not match the specifications of the sensor. The fault signal checking mechanism is carried out by applying the mechanism in Figure 3.

The original signal data generated is taken from the P2 section, and then checked whether the beginning of P2 is lower than the end of P2. If the start of P2 is lower, then the signal is continued to the next check. Suppose the initial P2 is higher, then the signal is declared broken signal because the signal is indicated to have decreased not according to specifications. The mechanism for checking the next broken signal is to calculate the moving average value which can be done with the Equation (1), where $(\bar{a})$ is the value of the *moving average,* m is the number of data calculated in $(\bar{a})$, and n is the number of all data. Then the data with the lower

initial P2 signal is calculated as delta (Δ) in each sensor. The delta search process is carried out by applying Equation (2).

$$\overline{a} = \frac{1}{m}\sum_{i=0}^{n-1} x_{m-i} \quad (1)$$

$$\Delta_i = \overline{a}_{n-1} - \overline{a}_n \quad (2)$$

The data from P2 which has become ($\overline{a}$) is then taken from the 30th row to the end of P2. It is taken from line 30 because the beginning of P2 is a transition period from P1, so the situation is still not stable and causes the value of delta (Δ) to be too high and not in accordance with the sensor value which is already stable. The data is then checked whether it complies with the specified threshold. The determined threshold is the best result from various mechanisms and trials so as to find the maximum and minimum values for the appropriate sensor not too narrow so that a lot of data is wasted or too wide, which causes broken data to still enter. The threshold for each sensor s1, s2, s3, s4, and s5 sequentially determined is as follows:

- $m = 9$
- $\Delta_{max} = \{6, 6, 6, 6, 9\}$
- $\Delta_{min} = \{-5, -5, -5, -5, -6\}$

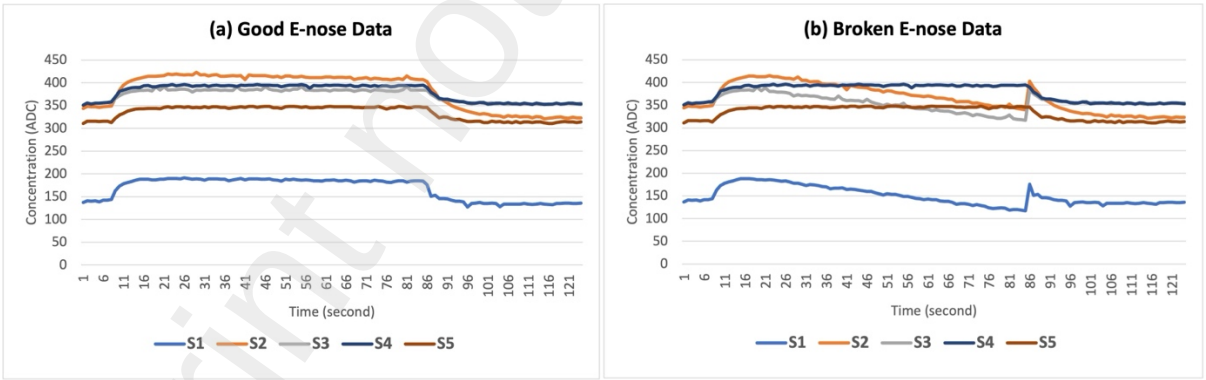

Figure 4 Difference of Signal: (a) Good Signal, (b) Broken Signal

From the results of the broken signal check mechanism, obtained signals that match the movement of the value of each sensor as good data, and those that do not fit are categorised as broken data. Figure 4 (a) shows an example of each good signal data and Figure 4 (b) broken signal data. Good data is continued to the next process, and broken data is not used because it interferes with the machine learning process carried out. The broken data is not data that

contains noise, but the broken data due to data collection errors or sensor errors and e-nose devices so that the data cannot be used.

2.4 Centroid Calculation Mechanism

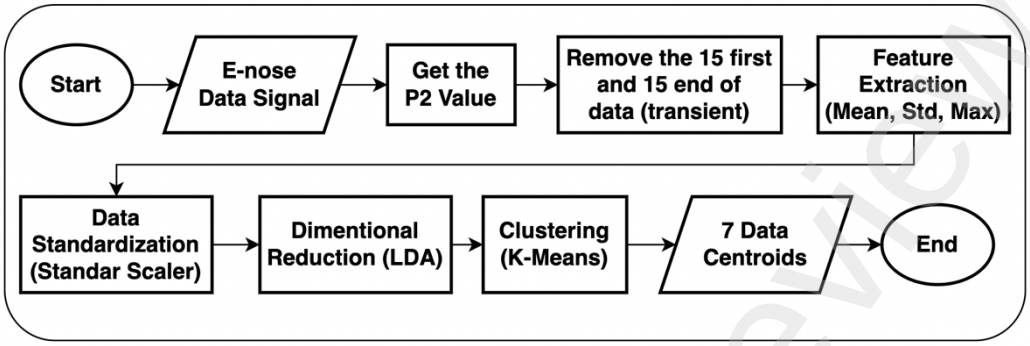

Figure 5 Centroid Calculation Mechanism

The centroid is sought which will be the measuring point for invalid data to be processed. To find the centroid of the 7 data classes is done using the mechanism in Figure 5. Each P2 sensor data is cut 15 beginning and 15 end to avoid transient data, namely data when moving from P1 to P2, and P2 to P3. Then the feature extraction process is carried out according to the previous reference data, using 3 statistical parameters, namely the Mean, Standard Deviation (Std), and Maximum value (Max). Then the data standardization process is carried out using the Standard Scaler. Then dimensional reduction was carried out using the same LDA as the data. Then, the clustering process is carried out to find the centroid of each class. The centroid calculation is also shown in Algorithm 1.

Algorithm 1 : Centroid Calculation

```
Input: Reference Data
Output: Centroid of reference data for each class, LDA_fit, KMeans_fit
```

```
RawDataReference ← load(referencedata.csv, "P2")[15:-15]
Sensor_columns ← ["S1","S2","S3","S4","S5"]

For sampling_id in RawDataReference.sampling_id do
      Data_sampling ← RawDataReference["sampling_id"] == sampling_id

      For sensor in Sensor_columns do
            Feature1 ← mean(Data_sampling["sensor"] == sensor)
            Feature2 ← standarDeviation(Data_sampling["sensor"] ==
            sensor)
            Feature3 ← maxValue(Data_sampling["sensor"] == sensor)
            Feature_sensor ← append(Feature1, Feature2, Feature3)
      end for

      Feature_sampling ← append(Feature_sensor)
```

```
end for

Feature_reference_data ← Feature_sampling

Standardized_fitur ← standarScaler(Feature_reference_data)
LDA_fit ← LDA.fit(Standardized_fitur, n_components = 2)
Reduced_fitur ← LDA_fit.transform(Standardized_fitur)
KMeans_clusters ← KMeans.fit_predict(Reduced_fitur, n_cluster = 7)
Centroid_reference_data ← KMeans_clusters. cluster_centers_

Return Centroid_reference_data, LDA_fit, KMeans_fit
```

2.5 Distance Between Invalid Data to Centroid of Reference Data

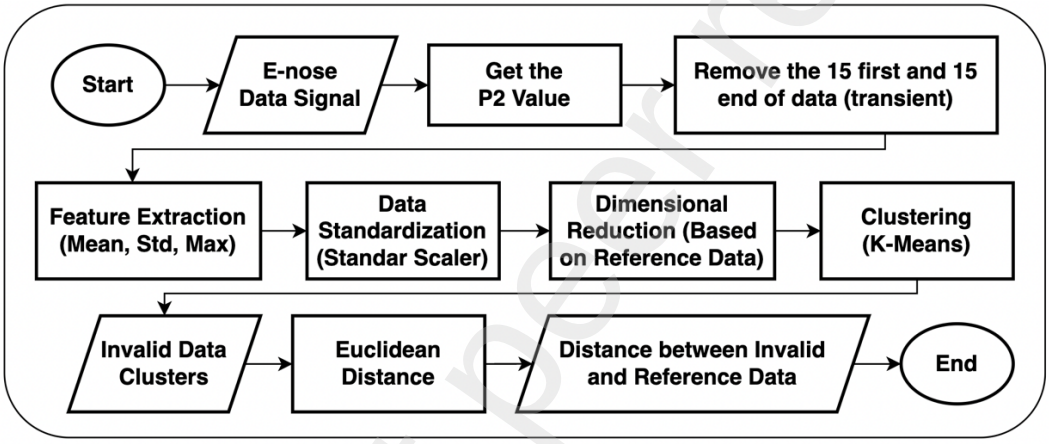

Figure 6 Distance Calculation Mechanism

After determining the invalid data to be used and cleaned or looking for valid values, the next step is to calculate the distance between the points of each invalid data and the centroid of each reference data class. The steps proposed in finding the distance between the invalid data point and the reference data centroid are shown in Figure 6.

After each reference data cluster is known, then the distance between the invalid data centroid points (contaminated with perfume and deodorant) to the normal data centroid (clean axilla) is calculated using the Euclidean Distance method which runs according to the Pythagorean theorem, where the distance of two data points is calculated by directly based on one straight line from data point 1 to other data. Euclidean Distance can be calculated by the Equation (3) and the distance mechanism is shown in Algorithm 2.

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad (3)$$

Algorithm 2 : Determining The Distance of Invalid Data to Reference Data Centroid

```
Input : Data Raw data invalid, Centroid_reference_data, LDA_fit,
KMeans_fit
Output : Cluster Data Invalid, Distance of invalid data to reference data
centroids.
```

```
CentroidDataReference ← load(Centroid_reference_data)
Sensor_columns ← ["S1","S2","S3",”S4","S5"]
Data_invalid ← load(datainvalid.csv, “P2”)[15:-15]

For sampling_id in Data_invalid.sampling_id do
      Data_sampling ← Data_invalid[“sampling_id”] == sampling_id
      For sensor in Sensor_columns do
            Feature1 ← mean(Data_sampling[“sensor”] == sensor)
            Feature2 ← standarDeviation(Data_sampling[“sensor”] ==
            sensor)
            Feature3 ← maxValue(Data_sampling[“sensor”] == sensor)
            Feature_sensor ← append(Feature1, Feature2, Feature3)
      end for
      Feature_sampling ← append(Feature_sensor)
end for
Feature_reference_data ← Feature_sampling

Standardized_fitur ← standarScaler(Feature_reference_data)
Reduced_fitur ← LDA_fit.transform(Standardized_fitur)
KMeans_invalid_clusters ← KMeans_fit.predict(Reduced_fitur)
Centroid_data_invalid ← KMeans_clusters.cluster_centers_

For cluster in Centroid_data_invalid do
      Distance_centroid[“cluster”] =
      euclideanDistance(Centroid_data_invalid[“cluster”],
      Centroid_reference_data)
end for

Return Invalid_clusters, Distance_centroid_invalid_to_reference_data
```

2.6 Distance Based Kalman Filtering

Finding the true value of the data containing noise is very necessary. This filtering process is one of the most widely used engineering tools [20]. Many modified filtering schemes have been developed to address the problem in various applications [21]. Each time the state of a system has to be estimated from noisy sensor information and some type of state estimator

is used to combine data from different sensors to produce an accurate estimate of the actual state of the system. In order to be able to produce signal data that is close to the actual value, it is necessary to analyse the axillary odour data that is clean and completely free from deodorant for the clustering process using K-Means Clustering and look for the centroid value of each data as the basis for the actual axillary odour value.

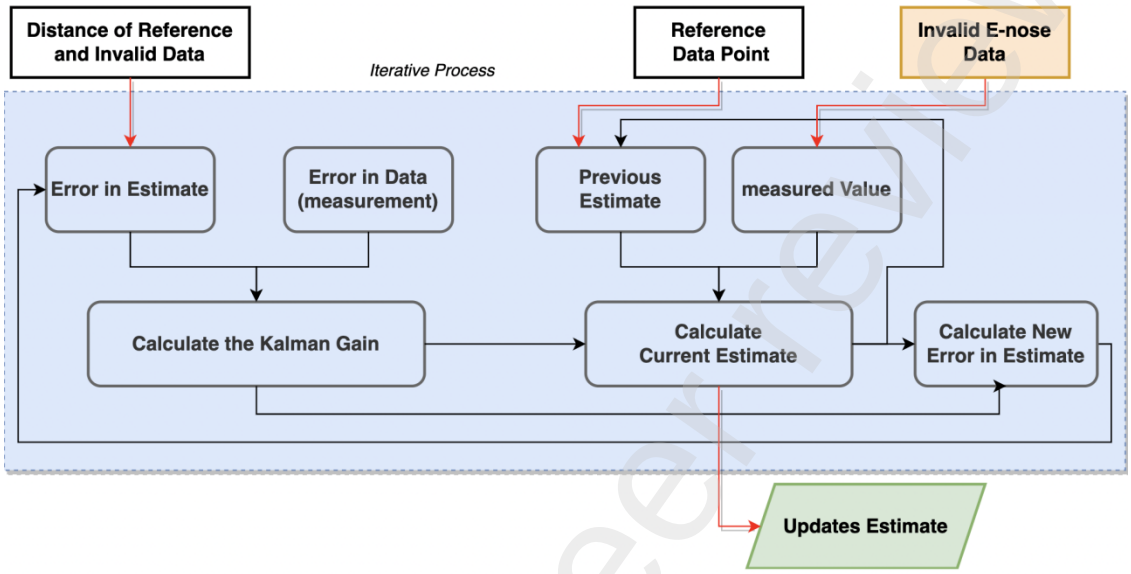

Figure 7 Distance Based Kalman Filtering

Figure 7 describes the algorithm of Kalman Filtering. There are three core processes that are carried out to find the estimated true value of a data, namely: (1) calculating Kalman Gain (KG), (2) calculating the current value estimate (EST_t), and (3) calculating a new error estimate (E_{EST_t}). Each of which can be calculated by Equation (4), Equation (5), and Equation (6) respectively. Kalman Gain reflects the relationship between measurement and estimation [21]. These measurements indicate which one is more reliable and should be "accepted" by the final estimate. The solution from Kalman Filtering is recursive where each updated state estimate is calculated from the previous estimate and the new input data, so only the previous estimate is saved for the next iteration process [22].

$$KG = \frac{E_{est}}{E_{est} + E_{mea}} \quad (4)$$

$$EST_t = EST_{t-1} + KG(MEA - EST_{t-1}) \quad (5)$$

$$E_{EST_t} = (1 - KG)(E_{EST_{t-1}}) \quad (6)$$

The axillary odour data containing deodorant is determined to be a measured value, which is then calculated as the actual axillary odour value according to the calculations mentioned above. Clean axillary odour that is free from deodorant and axillary odour contaminated with deodorant will then be carried out a process of finding the centroid which is the beginning of the error estimate value. This error estimate value is one of the initial determinants of the estimated data error in the axillary odour containing deodorant. The Kalman Filtering algorithm runs recursively to solve the problem of estimating the known state of the system based on certain mathematical models and noise measurement observations.

The data that has been filtered and is close to the original value from the pure axillary odour is then used for the detection of infectious respiratory diseases in order to increase accuracy and avoid detection errors. Prediction errors can be in the form of data that is actually negative but detected positive (false positive) or actually positive but detected negative (false negative).

3 Results and Discussion

3.1 Data Acquisition

The electronic nose (e-nose) data collection process in this study used subjects from various locations and conditions in the field and used several tools with the same sensor and anatomical system. Various kinds of situations and conditions of people and the environment in the field cause the data obtained to vary greatly even though they have used tools that have been determined with the same specifications and quality.

3.1.1 Reference Data

Reference data is data obtained under close supervision taken from predetermined subjects with various kinds of mechanisms. The first mechanism is normal axillary data taken from several people by ensuring that the human axillary odours are free from disturbing odours, including perfume odours, deodorant odours, or other odours that can interfere with the original smell of the axilla. This data is useful as a reference for normal, uncontaminated axillary data. Normal axillary data obtained amounted to 12 data.

The next data is data taken from the same subject, but deliberately mixed with perfume or deodorant with a predetermined concentration and time lag. So, the axilla odours are sprayed with a certain perfume or deodorant, then wait for a duration of 30 minutes and 60 minutes first and then the axillary odour data is taken using an e-nose. Subjects were asked to clean the body

and clothes first before being taken with a different concentration of odour, or it could be done on a different day provided that it met the predetermined criteria. These materials are then sprayed or applied (depending on how they are used) to the axilla area and then given a 30-minute and 60-minute pause before data collection is performed. The details of the reference data obtained are as written in Table 1 with a total of 30 data.

Table 1 Details of The Reference Data Obtained

Perfume or Deodorant	Time Interval		Total Data
	30 Minutes	60 Minutes	
DeoA	4	4	8
DeoB	6	6	12
Perfume	4	6	10
Total	14	16	30

Table 2 Distance of Each Centroid To The Normal Axillary Data Centroid

X	Y	Cluster	Distance to Normal Centroid
-2.4053	-3.1856	Normal	0.0000
1.1395	-1.0414	DeoA60m	4.1428
-4.3995	3.5564	DeoB30m	7.0307
-4.7269	-0.3669	DeoB60m	3.6517
6.1454	-8.0312	DeoA30m	9.8282
8.6686	3.6721	Perfume30m	13.0254
3.9661	4.5348	Perfume60m	10.0100

Data from the e-nose, which totals 5 MOS olfactory sensors. Then cut 15 initial data and 15 final data from each existing data sampling to avoid data change (transient) from P1 to P2. Then the feature extraction process is carried out using 3 statistical parameters, namely the mean, standard deviation, and Maximum value. Then the data normalisation process is carried out using the Standard Scaler. Because the data already has its own class, then dimensional reduction is used to find the best reduction value. Then, the clustering process is carried out to find the centroid of each class. The value of 7 cluster centroids and the distance of each centroid to the normal axillary data centroid generated are shown in Table 2.

3.1.2 Invalid Data

The data used in this study amounted to 5171 patient data, which consisted of 3030 valid data and 2141 invalid data. The valid and invalid labelling process was carried out manually

by 5 people using the majority voting method based on the results of observations. However, in this study, the invalid data used is the invalid result of the outlier detection model that has been determined. The model is the result of a gradual experiment that has been carried out using a variety of test scenarios.

Invalid data is indicated as invalid because it is detected as an anomaly from the existing dataset due to other odours such as using deodorant or perfume, then the sampling room contains other odours such as cigarette smoke, hand sanitiser and the sampling process is not in accordance with the procedure.

The outlier detection process uses a model that has been built using an adaptive filter method using Deep Neural Network (DNN) and self-feature extraction to overcome invalid data (outliers) which has a fairly good performance with a Balanced Accuracy of 90.4% (Purbawa et al., 2022). From the mechanism for determining outliers, it was found that from 5171 data, there were 1404 data that were detected outliers and were not included in the process of detecting infectious respiratory diseases. These data are automatically discarded because they are considered as outliers which will disturb the classification process. From this mechanism, outlier data of various types are still obtained, ranging from signals that go up and down drastically, signals that are very uneven, or signals that are reversed due to the user mistakenly entering the input and output intervals on the e-nose.

Because the research focuses on reducing noise caused by deodorants and perfumes, the results of the outlier detection need to be filtered again to remove irrelevant data. Then filtering is carried out using FilterMAV by utilising the smoothed movement of the sensor value, then a threshold is determined that can select irrelevant signals or the signal is already broken. From the previous data, which amounted to 1404 invalid data, after the filtering process with FilterMAV or a faulty signal check mechanism was carried out, then a selection process was carried out on data that had no class. Figure 8 (a) shows the data before FilterMAV and Figure 8 (b) after FilterMAV that 507 invalid data were obtained which will be processed.

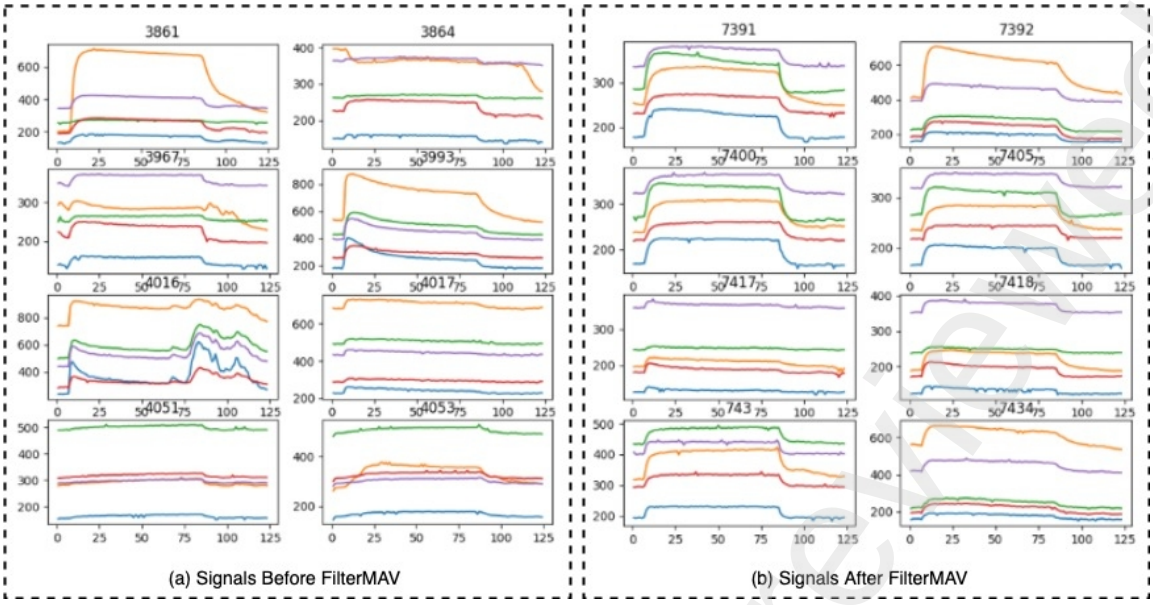

Figure 8 Data Visualization (a) Before FilterMAV, (b) After FilterMAV

3.2 Distance Between Invalid Data to Centroid of Reference Data

Table 3 Number of Invalid Data Per Class

Index Cluster	Cluster Invalid	Total
1	DeoB30m	46
2	DeoB60m	121
3	Perfume30m	127
4	Perfume60m	62
5	Normal	37
6	DeoA30m	90
7	DeoA60m	24
Total		507

In Table 3 we can see that the specified 507 invalid data can be divided into each cluster with different amounts. However, there were 86 invalid data that were still detected as normal data, which could be due to an insignificant increase in signal data and the distance between these data and the normal cluster centroid was not too far. Later, cases like this can be overcome by comparing the data centroid with the reference data centroid so that changes and adjustments to the new data do not stray too far and can be better. The visualisation results of invalid data are shown in Figure 9.

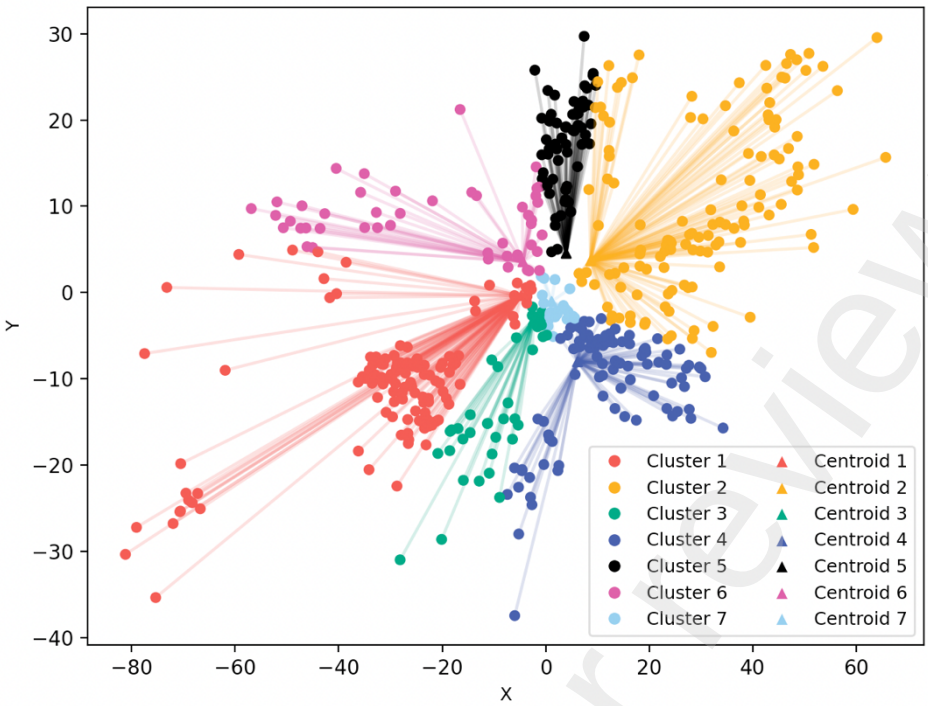

Figure 9 Invalid Data Visualisation in 7 Clusters

3.3 Distance Based Kalman Filtering Approach

As previously explained, the Kalman Filtering process carried out in this study utilises the original data from the invalid e-nose signal, the original data from the reference e-nose signal, the distance between the invalid data centroid points to the normal axillary data centroid, and the distance between the data cluster centroids. invalid to the normal axillary data cluster centroid. The proposed distance calculation in the calculation between centroids is to use Euclidean Distance.

Kalman Filtering works by utilising a set of measurement data which is then carried out an iterative process to get a smaller error value. Original data from normal axilla is used as additional reference data to carry out the iteration process and error values so that invalid data can be closer to the reference data or normal data so that it can reduce disturbing noise. Normal axillary data used as reference data amounted to 12 data. From the normal data, information is taken for each index and its sensors to serve as parameters in the Kalman Filtering process. The data is used as the value of the iteration process carried out in the Kalman Filtering process.

The process runs iteratively until a predetermined amount, which in this study proposes to use clean axillary data, as a parameter of how many times the iteration process is carried out. From the initial measurement using invalid axillary data, then iterating using normal axillary measurement data to find the last Estimate Value with the smallest possible Error Estimate

value. Each sampling_id produces 124 indexes, 5 sensor data and 12 normal axillary reference data, so the total number of calculations reaches 11,904 iterations. The last Estimate Value is then used for signal data that has been noise removal process using Kalman Filtering. An illustration of the calculation results of the proposed Kalman Filtering is shown in Figure 10.

```
KG        :  0.2959072152059306
EST_t     :  221.0
Eest_t    :  14.518452656222214
======================
KG        :  0.3226322812493825
EST_t     :  221.0
Eest_t    :  14.518452656222214
...
KG        :  0.6048951828821987
EST_t     :  190.0
Eest_t    :  8.46853256035078
======================
KG        :  0.6410800115570753
EST_t     :  188.0
Eest_t    :  7.692960138684903
======================
```

Figure 10 Calculation Results of The Proposed Kalman Filtering

Examples of comparison of the original invalid axillary signal data and the results of the filter process using the proposed method are shown in Figure 11 respectively. In Figure 11 (a) the graph looks very smooth and has a very significant increase when it enters P2 (axillary smell). Based on the experimental results and the determined majority voting results, these data are indeed characteristics of data contaminated with perfume or deodorant. Meanwhile, Figure 11 (b) is the estimation result of the proposed filter, which looks gentler but can reduce the intensity of other odours.

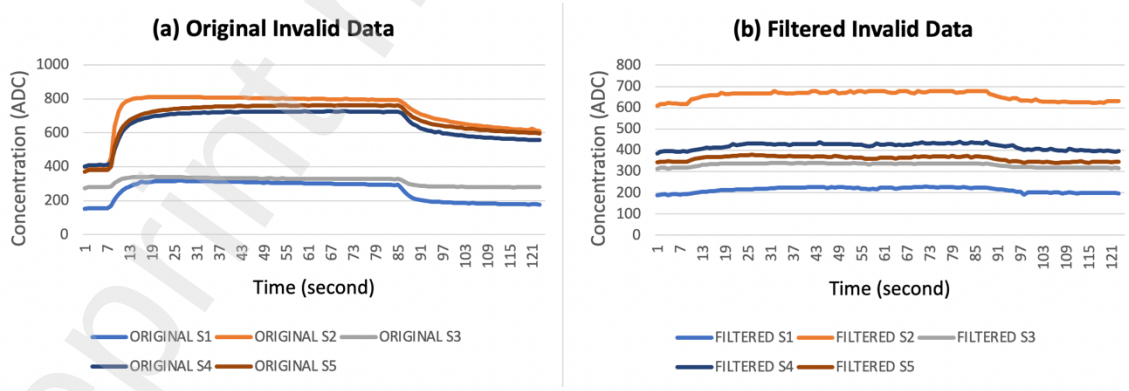

Figure 11 Signal Comparison of (a) Original Invalid Data Before Filter, (b) After Proposed Filter

Data that has not been filtered is called original data, while data that has been filtered is called filtered data. After the calculation process using Kalman Filtering is carried out, each of these data is used as a new dataset for the evaluation process.

3.4 Evaluation

Table 4 The Root Mean Square Errors (RMSE) Values

No.	RMSE_S1	RMSE_S2	RMSE_S3	…	RMSE_S8	MEAN
1	1.8628	1.5409	0.7018	…	1.1289	1.2512
2	0.3353	2.8094	0.5485	…	0.1699	1.1429
3	1.1679	3.2040	0.3031	…	0.8819	1.0701
…	…	…	…	…	…	…
505	0.3032	1.4538	0.1056	…	0.8894	0.8613
506	0.1970	2.8014	1.1004	…	1.3821	1.2243
507	0.1844	3.1504	1.0801	…	1.2859	1.0546
MEAN	0.5082	2.5341	0.7371	…	0.6802	1.0393

The evaluation process in Kalman Filtering can be done by consistently calculating filter-calculated covariances compared to the actual mean square errors [23]. Table 4 shows an example of the Root Mean Square Errors (RMSE) value obtained from each sensor on every invalid data that has gone through the filtering process with the Maximum Mean value of all sensors contained in the data is 1.8802, while the Minimum Mean value is 0.4666. The RMSE value obtained is quite good even though it still looks quite large, but keep in mind that mixing the axilla with the smell of perfume and deodorant really interferes with the olfactory process carried out by the e-nose. While the highest Mean RMSE Sensor value from all data is owned by S2 with a value of 2.5341, while the lowest Mean RMSE Sensor is owned by S4 with a value of 0.42908.

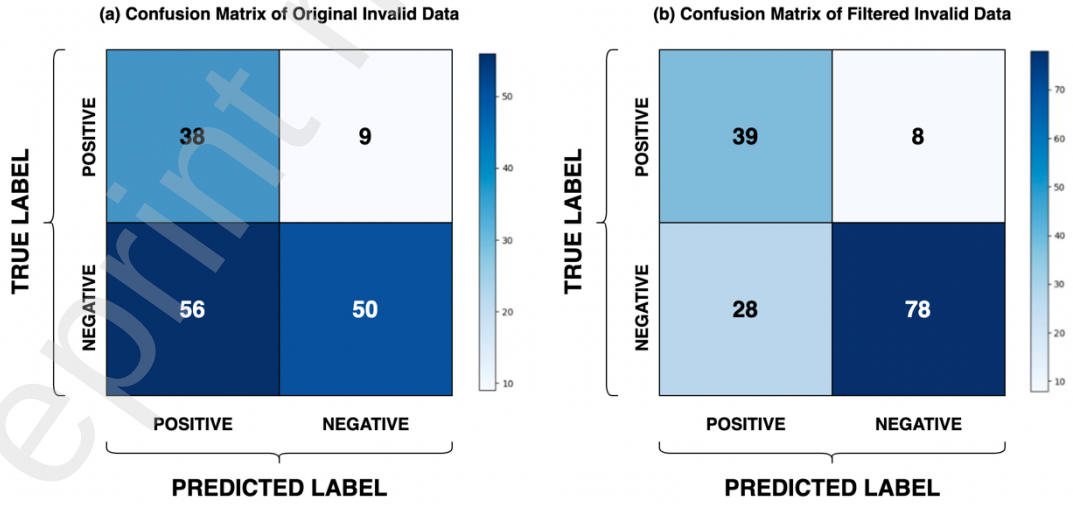

Figure 12 Confusion Matrix During Testing Phase of (a) Original Data, (b) Filtered Data

Furthermore, an evaluation process is carried out with a classification model of infectious respiratory diseases with proportion 70% of total data for training phase, and 30% for testing phase. The model used this time is a model that has been previously made using staked DNN [8]. In this model, five fine-tuned DNN models are used which are obtained from stacked hyperparameter tuning then, the output of each DNN model becomes the input of the meta-model in the form of a fully connected layer which results in a testing accuracy of 93.4%. However, it should be noted that here the model development uses data that has been selected manually and deletes invalid data, so that when the model is tried to predict truly invalid data directly (without any filtering process), the accuracy decreases. Figure 12 (a) shows the Confusion Matrix results from original invalid e-nose data (before filtering method). The testing accuracy obtained is only 57.52% with a Balance Accuracy (BA) of 62.59%. Then, the classification process is carried out using data that has been filtered. Figure 12 (b) shows the Confusion Matrix of the filtered invalid data (after filtering method) with a testing accuracy of 76.47% and Balanced Accuracy Testing of 74.46%. For detailed results from the Training and Testing data for each original data and filtered data, see Table 5.

Table 5 Detailed Results of The Training And Testing Phase

Evaluations	Original Data		Filtered Data		Improvement	
	Training	Testing	Training	Testing	Training	Testing
Sensitivity	0.3306	0.4043	0.6403	0.5821	30.97%	17.78%
Specificity	0.7358	0.8475	0.9023	0.9070	16.65%	5.95%
Precision	0.7455	0.8085	0.8091	0.8298	6.36%	2.13%
F1-Score	0.4581	0.5390	0.7149	0.6842	25.68%	14.52%
Accuracy	0.4520	0.5752	0.7994	0.7647	34.74%	18.95%
Balanced Accuracy	0.5332	0.6259	0.7713	0.7446	23.81%	11.87%

Next, the sampling_id analysis process was carried out which was false and correctly predicted on the original data and filtered data. From the testing evaluation results that have been carried out, it was found that there were 65 false data when using original data, but 47 data could be predicted to be true when using the proposed filtering method and could increase accuracy by 18.95%, and balanced accuracy by 11.865%.

Evaluation is also done by looking at the performance of other filtering methods, so that it can be seen whether the performance of the proposed method is better than the existing method. Other filtering methods used in the evaluation process include Moving Average

(MAV) with window = 7, Fast Fourier Transform (FFT), and Discrete Wavelet Transform (DWT) with several different thresholds and families. The full results are shown in Table 6.

Table 6 Classification Comparisons

Mechanism	Filtering Methods	Sensitivity	Specificity	Precision	F1 Score	Accuracy	Balanced Accuracy	BA Improvement
Data Training	Original Signal	0.331	0.736	0.746	0.458	0.452	0.533	0.00%
	Proposed Method	**0.640**	**0.902**	**0.809**	**0.715**	**0.799**	**0.771**	**23.81%**
	MVA	0.435	0.885	0.851	0.576	0.614	0.660	12.68%
	FFT	0.417	0.877	0.851	0.559	0.588	0.647	11.38%
	DWT (0.63,DB4)	0.411	0.862	0.830	0.549	0.582	0.636	10.31%
	DWT (0.63,bior1.3)	0.421	0.879	0.851	0.563	0.595	0.650	11.70%
	DWT (0.63,coif3)	0.412	0.875	0.851	0.556	0.582	0.644	11.05%
	DWT (0.63,dmey)	0.417	0.877	0.851	0.559	0.588	0.647	11.38%
	DWT (0.63,haar)	0.424	0.869	0.830	0.561	0.601	0.646	11.32%
	DWT (0.63,rbio1.3)	0.417	0.877	0.851	0.559	0.588	0.647	11.38%
	DWT (0.63,sym3)	0.423	0.893	0.872	0.569	0.595	0.658	12.46%
Data Testing	Original Signal	0.404	0.848	0.809	0.539	0.575	0.626	0.00%
	Proposed Method	**0.582**	**0.907**	**0.830**	**0.684**	**0.765**	**0.745**	**11.87%**
	MVA	0.435	0.885	0.851	0.576	0.614	0.660	3.41%
	FFT	0.417	0.877	0.851	0.559	0.588	0.647	2.11%
	DWT (0.63,DB4)	0.411	0.862	0.830	0.549	0.582	0.636	1.04%
	DWT (0.63,bior1.3)	0.421	0.879	0.851	0.563	0.595	0.650	2.43%
	DWT (0.63,coif3)	0.412	0.875	0.851	0.556	0.582	0.644	1.78%
	DWT (0.63,dmey)	0.417	0.877	0.851	0.559	0.588	0.647	2.11%
	DWT (0.63,haar)	0.424	0.869	0.830	0.561	0.601	0.646	2.05%
	DWT (0.63,rbio1.3)	0.417	0.877	0.851	0.559	0.588	0.647	2.11%
	DWT (0.63,sym3)	0.423	0.893	0.872	0.569	0.595	0.658	3.19%

It can be seen that the performance and increase in Balanced Accuracy (BA) resulting from the proposed method can be better than other filter methods. In Data Training, the increase in BA obtained from the proposed method reached 23%, followed by the Moving Average (MVA) method with an increase in BA of 12.68% from the original data. Meanwhile, during the testing process, the increase in BA performance from the proposed method reached 11.87%, followed by the MVA method with an increase in BA of 3.41% from the original data before

filtering. Details of the performance comparison during the testing process are shown in Figure 13.

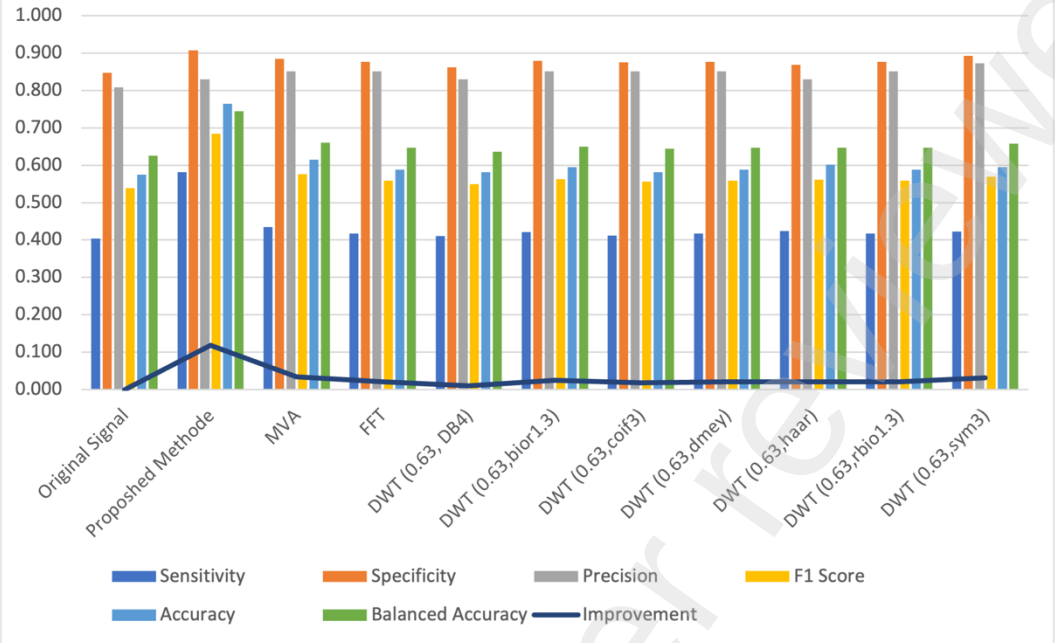

Figure 13 Accuracy Comparison of Other Methods

Table 7 Comparison of Invalid Data Accuracy With 6 Classification Methods

MECHANISM	CLASSIFIERS					
	LDA	RFC	LRC	SVC	DTC	MLP
Original Signal	0.6114	0.5957	0.6134	0.6903	0.5562	0.6824
Proposed Method	**0.7751**	**0.7751**	**0.7751**	**0.7929**	**0.7751**	**0.7929**
MVA	0.5937	0.5917	0.6114	0.6943	0.5306	0.6824
FFT	0.5878	0.6055	0.6174	0.6903	0.5128	0.6844
DWT (0.63, DB4)	0.5562	0.5838	0.6016	0.6903	0.5464	0.6903
DWT (0.63,bior1.3)	0.5680	0.5858	0.5957	0.6903	0.5325	0.6903
DWT (0.63,coif3)	0.5937	0.5996	0.6233	0.6923	0.5483	0.6864
DWT (0.63,dmey)	0.5976	0.5996	0.6213	0.6903	0.5503	0.6844
DWT (0.63,haar)	0.5976	0.5680	0.6154	0.6903	0.4734	0.6884
DWT (0.63,rbio1.3)	0.5917	0.6075	0.5897	0.6903	0.5385	0.6864
DWT (0.63,sym3)	0.5799	0.5937	0.6036	0.6903	0.5641	0.6884

In addition, the evaluation process is also carried out using other classification methods. Classification methods used for evaluation include Linear Discriminant Analysis (LDA), Random Forest Classifier (RFC), Logistic Regression Classifier (LRC), Support Vector Classifier (SVC), Decision Tree Classifier (DTC) and Multi-Layer Perception (MLP). Each of these classification methods is carried out by a hyper parameter tunning process using Gird Search in order to obtain optimal parameters. The results of these classification methods are

shown in Table 7. It can be seen that the proposed method is able to produce the highest accuracy and is quite stable with a minimum value of 77.51% accuracy in the LDA, RFC, LRC, and DTC classifications, and the highest results reach 79.29% in the SVC and MLP classification methods. While other signal filtering methods get the highest score of 69.03% and the lowest accuracy of 47.34%.

3.5 Weakness

As already explained, Kalman Filtering is an algorithm that runs using continuous iteration steps, so the more training data, the better the results but with a longer time. In this study, the proposed method pays attention to each row of e-nose signal data and has its own reference data centroid considered, so that the resulting time is longer. The consideration of this reference data is advantageous because the data that is processed is not only based on the data itself, but also looks at the centroid and reference data that has been given. The more lacuna data given, the more accurate the results will be, but with a longer duration because more iterations are needed according to the number of reference data provided. In contrast to other filtering methods, which only consider the value in the data itself to carry out the filtering process, so the time required is faster. The results of the calculation of the duration of each method are shown in Table 8.

Table 8 Comparison of Filtering Method Time Duration

Filtering Methods	Execution Time (s)
Proposed Method	**17.3939**
MVA	0.0331
FFT	0.0373
DWT (0.63,DB4)	0.0330
DWT (0.63,bior1.3)	0.0391
DWT (0.63,coif3)	0.0328
DWT (0.63,dmey)	0.0374
DWT (0.63,haar)	0.0460
DWT (0.63,rbio1.3)	0.0506
DWT (0.63,sym3)	0.0319

4 Conclusions

The proposed filtering method is able to produce a fairly good RMSE, with a maximum Mean of all sensors contained in the data is 1.8802, while the Minimum Mean value is 0.4666. While the highest Mean RMSE Sensor value from all data is owned by S2 with a value of 2.5341, while the lowest Mean RMSE Sensor is owned by S4 with a value of 0.4291. When

using the original data, the testing accuracy obtained is only 57.52% with a Balance Accuracy (BA) of 62.59%. Meanwhile, when using filtered data, the testing accuracy obtained is 76.47%, and BA is 74.46%. The proposed method can increase accuracy by 18.95%, and BA by 11.87%. The proposed method is also tested by other classification methods and able to produce the highest accuracy with 79.29% in SVC and MLP, while other filtering methods only get the highest accuracy with 69.03% also in SVC and MLP. However, the proposed method has a weakness where the calculation process is quite slow. This is because Kalman Filtering works in an iterative way, which means that the process will continue to run until all the data provided is processed. The more reference data, the more accurate it will be but the consequence will be the longer the process.

Ethics Approval and Consent to Participate

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Rumah Sakit Islam Jemursari Surabaya under Application No.001.EC.KEP.RSIAY.03.21 and Rumah Sakit Umum Daerah Dr. Soetomo under Application No. 0173/KEPK/IV/2021.

Acknowledgement

This research was funded by the Directorate of Reseach and Community Service, Directorate General of Higher Education, Research, and Technology, Indonesian Ministry of Education, Culture, Research and Technology under Penelitian Terapan Unggulan Perguruan Tinggi (PTUPT) Program, Institut Teknologi Sepuluh Nopember (ITS) under project scheme of the Publication Writing and IPR Incentive Program (PPHKI), and the Directorate General of Higher Education, Research, and Technology, and Indonesian Ministry of Education, Culture, Research and Technology under Matching Fund (MF) Kedaireka Program.

References

[1] Anand, S. S., Philip, B. K. and Mehendale, H. M., *Volatile Organic Compounds*, Third Edit., vol. 4. Elsevier, 2014.

[2] Wang, L., Su, S. W., Celler, B. G. and Savkin, A. V., "Modeling of a gas concentration measurement system," *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*, vol. 7 VOLS, no. February, pp. 6695–6698, 2005.

[3] Qi, H., Jun, L. and Hengwei, L., "A modified adaptive stochastic resonance for detecting faint signal in sensors," *Sensors*, vol. 7, no. 2, pp. 157–165, 2007.

[4] Tian, F., Yang, S. X., Xu, X. and Liu, T., "Circuit and Noise Analysis of Odorant Gas Sensors in an E-Nose," *Intelligent Systems for Machine Olfaction: Tools and Methodologies*, no. September 2004, pp. 102–124, 2011.

[5] Li, W., Jia, Z., Xie, D., Chen, K., Cui, J. and Liu, H., "Recognizing lung cancer using a homemade e-nose: A comprehensive study," *Computers in Biology and Medicine*, vol. 120, p. 103706, May 2020.

[6] Eamsa-Ard, T., Srikhirin, T. and Kerdcharoen, T., "Online Monitoring of Human Body Odor for the Improvement of Athlete's Health Status," in *2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, 2018, pp. 1–2.

[7] Sarno, R., Sabilla, S. I. and Wijaya, D. R., "Electronic Nose for Detecting Multilevel Diabetes using Optimized Deep Neural Network," *Engineering Letters*, vol. 20, pp. 31–42, 2020.

[8] Malikhah, Sarno, R., Inoue, S., Ardani, M. S. H., Purbawa, D. P., Sabilla, S. I., Sungkono, K. R., Fatichah, C., Sunaryono, D., Bakhtiar, A., Libriansyah, Prakoeswa, C. R. S., Tinduh, D. and Hernaningsih, Y., "Detection of Infectious Respiratory Disease Through Sweat From Axillary Using an E-Nose With Stacked Deep Neural Network," *IEEE Access*, vol. 10, pp. 51285–51298, 2022.

[9] Purbawa, D. P., Sarno, R., Malikhah, Ardani, M. S. H., Sabilla, S. I., Sungkono, K. R., Fatichah, C., Sunaryono, D., Parimba, I. S. and Bakhtiar, A., "Adaptive filter for detection outlier data on electronic nose signal," *Sensing and Bio-Sensing Research*, vol. 36, Jun. 2022.

[10] Wijaya, D. R., Sarno, R. and Zulaika, E., "Noise filtering framework for electronic nose signals: An application for beef quality monitoring," *Computers and Electronics in Agriculture*, vol. 157, pp. 305–321, Feb. 2019.

[11] Wijaya, D. R., Sarno, R. and Zulaika, E., "DWTLSTM for electronic nose signal processing in beef quality monitoring," *Sensors and Actuators B: Chemical*, vol. 326, p. 128931, Jan. 2021.

[12] Bai, Y., Wang, X., Jin, X., Zhao, Z. and Zhang, B., "A Neuron-Based Kalman Filter with Nonlinear Autoregressive Model," *Sensors*, vol. 20, no. 1, p. 299, Jan. 2020.

[13] Huang, Y., Zhang, Y., Zhao, Y. and Chambers, J. A., "A Novel Robust Gaussian–Student's *t* Mixture Distribution Based Kalman Filter," *IEEE Transactions on Signal Processing*, vol. 67, no. 13, pp. 3606–3620, Jul. 2019.

[14] Farahi, F. and Yazdi, H. S., "Probabilistic Kalman filter for moving object tracking," *Signal Processing: Image Communication*, vol. 82, p. 115751, Mar. 2020.

[15] Kim, T. and Park, T.-H., "Extended Kalman Filter (EKF) Design for Vehicle Position Tracking Using Reliability Function of Radar and Lidar," *Sensors*, vol. 20, no. 15, p. 4126, Jul. 2020.

[16] Lam, C. H., Ng, P. C. and She, J., "Improved Distance Estimation with BLE Beacon Using Kalman Filter and SVM," in *2018 IEEE International Conference on Communications (ICC)*, 2018, pp. 1–6.

[17] R, D., "Analysis of Adaptive Image Retrieval by Transition Kalman Filter Approach based on Intensity Parameter," *Journal of Innovative Image Processing*, vol. 3, no. 1, pp. 7–20, Mar. 2021.

[18] Singh, K. K., Kumar, S., Dixit, P. and Bajpai, M. K., "Kalman filter based short term prediction model for COVID-19 spread," *Applied Intelligence*, vol. 51, no. 5, pp. 2714–2726, May 2021.

[19] Chanonsirivorakul, R. and Nimsuk, N., "Analysis of Relationship between the Response of Ammonia Gas Sensor and Odor Perception in Human," *2020 8th International Electrical Engineering Congress, iEECON 2020*, vol. 1, 2020.

[20] Julier, S. J., Uhlmann, J. K. and Durrant-Whyte, H. F., "New approach for filtering nonlinear systems," *Proceedings of the American Control Conference*, vol. 3, no. June, pp. 1628–1632, 1995.

[21] Qu, J., Chai, Y. and Yang, S. X., "A real-time de-noising algorithm for e-noses in a wireless sensor network," *Sensors*, vol. 9, no. 2, pp. 895–908, 2009.

[22] Haykin, S., *Kalman Filtering and Neural Networks*, vol. 5. 2001.

[23] Jwo, D. J. and Cho, T. S., "A practical note on evaluating Kalman filter performance optimality and degradation," *Applied Mathematics and Computation*, vol. 193, no. 2, pp. 482–505, 2007.