A Mixed Gaussian Distribution Approach using the Expectation-Maximization Algorithm for Topography Predictive Modelling

KHAIRULNIZAM OTHMAN[1,*], MOHD NORZALI MOHD[2],
MUHAMMAD QUSYAIRI ABDUL RAHMAN[1], MOHD HADRI MOHAMED NOR[1],
KHAIRULNIZAM NGADIMON[1], ZULKIFLI SULAIMAN[3]
[1]Centre for Diploma Studies,
[1]Universiti Tun Hussein Onn Malaysia (Pagoh Campus),
KM 1, Jalan Panchor, 84600 Panchor, Johor,
MALAYSIA

[2]Faculty of Electrical and Electronic Engineering,
, Universiti Tun Hussein Onn Malaysia,
86400, Parit Raja, Johor,
MALAYSIA

[3]Five Element Technology Sdn Bhd, Plot 44, Pertanian Moden Ayer Hitam,
Universiti Tun Hussein Onn Malaysia,
86400, Parit Raja, Johor,
MALAYSIA

Abstract: - The incidence of sugarcane crop infestations at the migration stage, especially by the top borer, can lower yields substantially, which may translate to revenue losses of over 20% across many parts of the world. Traditional pest surveillance approaches tend to lack the accuracy required for timely intervention. This research introduces a new burden rate concept incorporated within a Gaussian Mixture Model (GMM), framed within a machine learning environment in order to enhance the precision of infestation pattern prediction. Through the utilization of the Expectation-Maximization (EM) algorithm, the model easily receives maximum likelihood estimates automatically, thus efficiently dealing with cluster distributions at low computational costs. A significant extension of this research is the inclusion of wind direction and topography as dynamic predictors. This allows for maximizing the model's potential in determining highly susceptible locations of infestation. The incorporation of remote sensing and drone data increases the precision of parameter estimation, leading to accurate predictive modeling. The EM-based clustering method reaches a high level of accuracy of 97.5%, which is greater compared to conventional pest monitoring methods. The result of this study provides a new analytical instrument for pest outbreak control and forecasting in precision agriculture. The tool provides real-time workforce management, selective pest eradication, and efficient resource management. Furthermore, the new synergy of clustering processes, topographic modeling, and remote sensing used in the study achieves a scalable data-driven approach to sustainable farm management that involves proactive crop loss minimization.

Key-Words: - Burden Rate, Gaussian, Infestation Patterns, Expectation-Maximization (EM) Algorithm, Remote Sensing, Predictive, Clustering, Topographic.

Received: March 21, 2024. Revised: November 11, 2024. Accepted: December 9, 2024. Published: April 7, 2025.

1 Introduction

Sugarcane is an essential crop, valued for both its economic benefits and nutritional content. Unfortunately, its productivity can be severely impacted by pests, with the top borer (Scirpophaga excerptalis) being one of the most harmful. This pest can lead to significant yield losses, and if not controlled, it may result in revenue declines of 55% or more, as seen in sugarcane farming areas in Malaysia. This highlights the importance of creating effective monitoring and management measures to counter its effects.

In order to promote pest management tactics, there is a necessity to incorporate higher-level predictive models. This paper answers this call through the use of a blend of topological

information and wind direction in predicting weekly spatial patterns of primary borer infestations. The objective is to promote the strategic allocation of labor between agricultural sectors to enable interventions to be conducted in a timely manner.

In the paper, the Expectation-Maximization algorithm is applied when combined with the mixed Gaussian distribution model in a machine-learning setting. The advanced statistical approach was developed specifically to analyze established infestation patterns and forecast future occurrences. By utilizing an endpoint condition predicated on iterative steps, the research analyzes the ultimate probability of a stochastic model, investigates the usefulness of the amount of data, and delineates an algorithm that can automatically determine the maximum likelihood estimation method.

This paper explores the intricacies and uncertainties that exist in agricultural data and demonstrates how the determination of optimum infestation burden rates is achievable with a minimal level of computational complexity. The approach obtains a probabilistic model that captures the dynamics that typify top borer infestations, where wind direction and topography data are significant predictor variables. The fact that the EM algorithm is used in iterative optimization implies that the model converges iteratively towards a solution that maximizes the probability of the observed data set.

The paper also embodies the comparative evaluation of the suggested algorithm in relation to other statistical approaches through the utilization of a mixed multidimensional normal distribution. The evaluation records the precision and effectiveness of the EM algorithm in estimating infestation patterns, hence offering insightful information for improved labor management in sugarcane cultivation.

This research not only advances our knowledge of pest dynamics within the sugarcane production setting but also gives a broad analytical framework for the management of agricultural practice. The suggested mixed Gaussian distribution with the EM algorithm presents a unified solution for the prediction of pest infestation, which is a significant aspect in maximizing crop yields and resource allocation in agricultural practice.

2 Preliminary

In estimating probability distributions, it's usual to determine the parameters of a model that usually describe the observed random variable X. This approach is known as the nearest likelihood method. When modeling assumes the existence of a latent variable Z inside a more complicated distribution, the distribution of X may frequently be stated as a simple combination. Previous research has shown high accuracy using similar mated in [1].

Besides that arranging data is a major concern. Here data sets can be grouped into two categories, full data sets and incomplete data sets. Datasets may be grouped under two categories: full datasets and incomplete datasets. Incomplete datasets are characterized by the lack of specific data points, implying certain observations are missing. To handle this problem, the missing data is described as latent variables, and their distribution is approximated, as discussed in [2], [3], [4] and [5]. The Expectation-Maximization (EM) approach, which is often used for estimating parameters in distributions that incorporate latent variables, is detailed in [6], [7] and [8].

In this research, the EM algorithm is explained and embedded within its theoretical framework. The framework starts with clustering using the k-means algorithm, which is introduced as a foundational concept. Next, the estimation of Gaussian mixture distributions using the EM algorithm is modified to absorb more neighboring points. This is followed by a general explanation of the EM algorithm and its application in estimating Gaussian mixture models. Finally, the interpretation of Gaussian mixture estimation is provided to put its significance in the research viewpoint.

3 Clustering By *K*-Means

The Gaussian mixture, which is the estimation target this time, is used for clustering data, but before that, first explain *k*-means, which is a probability-free approach as a special case. This is a method of dividing the obtained data into *K* clusters (*K* is given as a hyper-parameter) based on the proximity of the data. For training purposes, the study also involves plotting real infestation data as in Figure 1. By visualizing actual infestation patterns, the model can be trained more effectively, ensuring that the predictive capabilities are grounded in real-world observations.

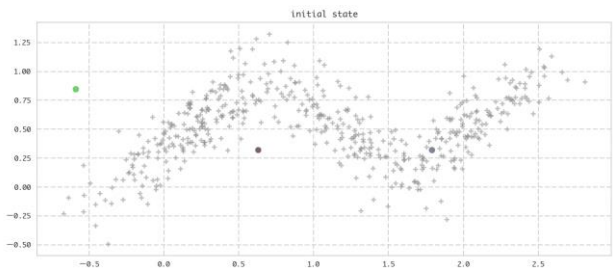

Fig. 1. Flow of cluster estimation by *k*-means

1. Prepare μ That represents the center of the cluster (also called Centroid), the number of clusters $K = 3$, and initialize it appropriately. (The above example is determined by a uniform distribution from the data range)
2. When the current $\mu = (\mu_1, \mu_2, \mu_3)$ is fixed, each of the 500 pieces of data is selected to be the closest μ_k and belongs to the cluster number k.
3. The average of the data belonging to each cluster k is calculated, and μ is updated with it as the center of the new cluster.
4. Check the difference of μ Update, and if there is no change, converge and end. If there is an update difference, return to 2.

3.1 Derivation

The derivation of the *k*-means algorithm is rooted in the minimization of a specific loss function, which quantifies the total squared distance between each data point and the centroid of the cluster to which it belongs. This loss function is crucial for understanding how the algorithm iteratively improves the clustering of data points, as demonstrated in [9] and [10].

Let's start by defining the symbols used in the derivation:

x: D dimension data points

$d = \{x_1, \ldots, x_N\}$:A dataset consisting of N observation points.

K: The number of clusters, which is a known constant.

$\mu_k (k = 1, \ldots, K)$: The centroid of the *k*-th cluster, representing the center of the cluster in *D*-dimensional space.

r_{nk}: A binary indicator variable that takes the value 1 if the *n*-th data point belongs to the *k*-th cluster and 0 otherwise.

The loss function *J* is defined as the sum of the squared Euclidean distances between each data point and the centroid of its assigned cluster as:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2 \quad (1)$$

The goal of the *k*-means algorithm is to minimize this loss function *J*. The minimization process is carried out in two alternating steps. In this step, the centroid μ_k are fixed, and the algorithm assigns each data point x_n o the cluster whose centroid is closest to it. This is done by minimizing the term $\|x_n - \mu_k\|^2$ for each data point. Specifically, for each data point x_n, the algorithm calculates the squared distance to each centroid μ_k and assigns x_n To the cluster with the smallest distance.

Minimize this loss function J in the following two steps.

Step 1. Fix μ_k and partially differentiate J with $r_{nk}$to minimize

Step 2. Fix r_{nk} and partially differentiate J with to μ_k minimize

3.1.1 Step 1 points out as:

Once all data points have been assigned to clusters, the algorithm updates the centroids μ_k to be the mean of all data points assigned to each cluster. This step minimizes the loss function J with respect to the centroids μ_k. The new centroid μ_k is calculated as:

$$\mu_k = \frac{\sum_{n=1}^{N} r_{nk} x_n}{\sum_{n=1}^{N} r_{nk}} \quad (2)$$

This formula ensures that the centroid μ_k is the average of all data points assigned to the *k*-th cluster, effectively minimizing the total squared distance within the cluster.

To put it. $d_{nk} = \|x_c - \mu_k\|^2$

$$J = \sum_{n} (r_{n1} d_{n1} + r_{n2} d_{n2} + \ldots + r_{nk} d_{nk}) \quad (3)$$

Therefore, it is sufficient to minimize $(r_{n1} d_{n1} + r_{n2} d_{n2} + \ldots + r_{nk} d_{nk})$ for each data, and considering that r_{nk} is a binary indicator variable, it is $d_{n1}, d_{n2}, \ldots d_{nk}$ and the smallest one can be picked as d_{nk},

$$r_{nk} \begin{cases} 1 & \left(k = arg\ Min_j \|x_c - \mu_k\|^2\right) \\ 0 & (otherwise) \end{cases} \quad (4)$$

3.1.2 Step 2 set up as:

Adjust r_{nk} and partially differentiate J with respect to μ_k to achieve minimization. This method facilitates effective optimization of cluster centroids, thereby enhancing the accuracy of topographic mapping. The procedure improves the alignment between spatially distributed data and their respective cluster centers, while maintaining local structures. It supports dimensionality reduction and visualization in complex datasets by transforming high-dimensional data into meaningful low-dimensional representations with minimal distortion. The key benefit is the creation of a true representation of the data's capture of every

geometry, which is essential for tasks such as self-organizing maps or geographic data modeling.

The *k*-means algorithm works by repeating the task of reducing the loss function J. It accomplishes this by alternating between allocating data points to the nearest cluster and updating the cluster centroids to the average of the allocated points. This cycle continues until the centroids have centered on the polar group. This leads to minimizes the total squared distance inside each cluster. The derivation of the algorithm highlights its reliance on the optimization of a well-defined loss function, ensuring that the clustering process is both systematic and efficient.

The equation derivation explain the optimization process of the k-means method, focusing on how the centroids μ_k are updated to minimize the loss function J. Let's break down the mathematical steps and elaborate on the benefits of this approach, particularly in the context of topographic mapping, dimensionality reduction, and geographic data modeling as demonstrated in [11] and [12].

Taking the partial derivative with respect to μ_k:

$$\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\,\|x_n-\mu_k\|^2 = \frac{\partial}{\partial \mu_k}\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left(-2x_n^T\mu_k + \mu_k^T\mu_k\right) = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left(-2\frac{\partial}{\partial \mu_k}x_n^T\mu_k + \frac{\partial}{\partial \mu_k}\mu_k^T\mu_k\right) = \sum_{n=1`}^{N} r_{nk}\,(-2x_n+2\mu_k) = -2\sum_{n=1`}^{N} r_{nk}\,(x_n-\mu_k)=0$$

$$\frac{\partial J}{\partial \mu_k} = -2\sum_{n=1`}^{N} r_{nk}\,(x_n-\mu_k)=0 \tag{5}$$

If the expression transforms,

$$\sum_{n=1}^{N} r_{nk}\,x_n = \sum_{n=1}^{N} r_{nk}\mu_k = \mu_k\sum_{n=1}^{N} r_{nk}$$

$$\mu_k = \frac{\sum_{n=1}^{N} r_{nk}\mu_k}{\sum_{n=1}^{N} r_{nk}}$$

$$\mu_k = \frac{\sum_{n=1}^{N} r_{nk}\mu_k}{\sum_{n=1}^{N} r_{nk}} \tag{6}$$

It turns out that the optimal Centroid of the cluster k is the average of the data belonging to the cluster k as above. From the above, the algorithm used in the first demonstration applied the one derived by the optimization of the loss function J. The program measures the above two steps and transforms the iterations. The technique involves computing the difference between the centroid before the update and after the update. It is speculated that integration has happened when this difference becomes minor, leading to the end of replication.

Equation 6 of the k-means method demonstrates its ability to optimize cluster centroids, which boosts accuracy in topographic mapping and several other domains. By retaining the local structures of spatially distributed data and aiding dimension reduction. This technique serves as a strong tool for evaluating challenging datasets.The accuracy of it is further enhanced when combined with techniques like the EM algorithm, which makes it a crucial method for applications like agricultural pest control and spatial data modeling. Furthermore, this technique is able to accurately define data geometry, maximize resource allocation, and support decision-making for real-world applications.

4 Estimation of Gaussian Mixture Distribution by EM Algorithm

A popular iterative method for estimating the parameters of probabilistic models is the Expectation-Maximization (EM) algorithm. This is particularly when dealing with incomplete data or hidden variables. Gaussian Mixture Models (GMMs) are widely used in pattern recognition, clustering, and machine learning, which are use full in simulating complex data distributions.

In order to obtain maximum likelihood estimates for GMM parameters, the EM algorithm alternates between two crucial steps. First is the prediction step,which determines the likelihood that each data point belongs to a particular component. Then the second is the maximization step, which updates the parameters in accordance with these probabilities. Meanwhile, density estimation and data clustering benefit from this iterative method's

ability to focus on a local optimum. The EM technique is used to estimate a Gaussian Mixture Model (GMM) in an iterative manner, as shown in Figure 2. Unlike *k*-means, which assigns each data point to a single cluster, GMM offers greater flexibility in terms of membership. Therefore, a 0-1 indicator variable, such as $r_1 = (0,1,0)$ was used.

In the Gaussian mixture distribution, each data belongs to each cluster, but its indicator variables are changed to random variables, which are expressed as latent variables. Therefore, for example, if the z_1 expected value of the latent variable corresponding to the first data x_1 is taken, for example, $E[z_1] = (0.7, 0.2, 0.1)$ It takes a value in the range $0 \leq z_{1k} \leq 1$. It is represented by gradation in the Figure 2.

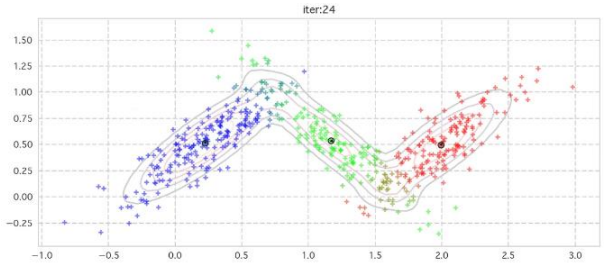

Fig. 2. Density function of the Gaussian mixture distribution.

The symbols used here are also described this time.
x: D dimensional random variable
z: k dimensional random variable, the latent variable of the model
$D = \{x_1, \ldots, x_N\}$: N observation points (data set)
K: Number of clusters (known constant)

First, let's look at the probability density function of the Gaussian mixture distribution.

$$p(x|\pi, \mu, \Sigma) = \sum_{k=1}^{K} \pi_k \, N(x|\mu_k, \Sigma_k) \tag{7}$$

This can be interpreted by comparing the ratio to the sums of the *K* Gaussian distributions. Tried to display the following one-dimensional example. Figure 2 is a density function in which each Gaussian distribution has a ratio π_k According to the mixing coefficient. When integrated, the area becomes π_k.

Figure 3 is a vertically stacked graph. This is the density function of the Gaussian mixture distribution. If take π_k so that $\Sigma_k \pi_k = 1$, the area will be exactly one.

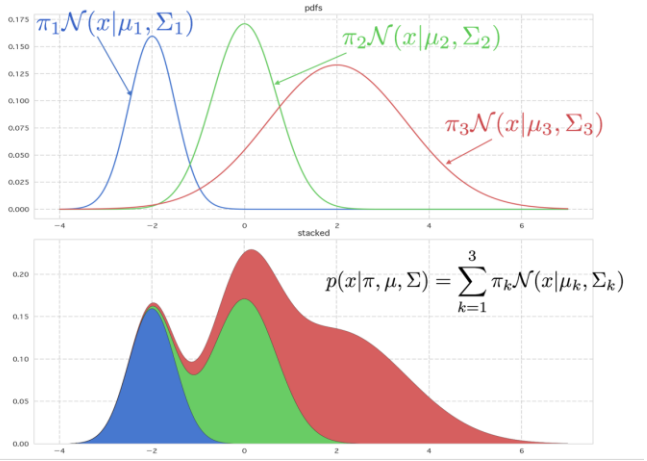

Fig. 3. Vertically stacked graph.

4.1 Introducing Latent Variables by Marginal Likelihood.

If the probability distribution that produces the data is represented by $p(x)$, then the latent variable z can be submerged by applying marginalization or multiplication theorem to it. θ is a model parameter.

$$p(x) = \sum_{z} P(x|z)p(z) \, \partial z \tag{8}$$

Let's see what $p(z)$ and $p(x|z)$ look like in a mixture distribution model. First, let's look at the distribution of $p(z)$. z_k is a variable that takes 1 for any one k like $r_{nk}$of *k*-means, and the difference is that le this time. Again z_k satisfies $z_k \in \{0,1\}$ and $\sum_z z_k = 1$. z_k is a random variable.

First, look at the kth term z_k of the latent variable $z = (z_1, \ldots, z_k, \ldots, z_k)$. The probability that z_k is 1 is determined by the mixing coefficient π_k, $p(z_k = 1) = \pi_k$ is. Since the parameter π_k is considered as a probability, it is assumed that $0 \leq \pi_k \leq 1$, and $\sum_{k=1}^{K} \pi_k = 1$.

Write together with z

$$p(z) = \prod_{k=1}^{K} \pi_k^{z_k} \tag{9}$$

The conditional distribution of the data $\boldsymbol{x}$ under the condition that z is given follows the kth Gaussian distribution if the condition is $z_k = 1$. $p(x|z_k = 1) = N(x|\mu_k, \Sigma_k)$ This becomes a z condition as:

$$p(x|z) = \sum_{k=1}^{K} N(x|\mu_k, \Sigma_k)^{z_k} \tag{10}$$

By substituting these $p(x), p(x|z)$ into (1), See that the Gaussian mixture distribution density function was consistent as seen previously.

$$p(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k) \tag{11}$$

4.2 Burden Rate

The conditional distribution of the data $\boldsymbol{x}$ under the condition that z is given follows the kth Gaussian distribution if the condition is $z_k = 1$. From $p(z)$ and $(x|z)$ derive earlier, calculate the posterior distribution/ reverse distribution can be determined of z using Bayes' theorem. In other words, it is possible to infer the distribution of variable from the observed data x, as discussed in [13].

$$p(z_k = 1|x) = \frac{\pi_k N(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x|\mu_j, \Sigma_j)} \tag{12}$$

This posterior $p(z_k = 1|x)$ is called $\gamma(z_k)$, and this is sometimes expressed as the burden rate. A major contribution to clustering is the concept of the burden rate, which plays a crucial role in understanding the distribution of clusters. To illustrate this, consider Figure 4, where the burden rate can be visually interpreted. Fundamentally, the burden rate represents the ratio of each k-th component in the Gaussian mixture model (GMM) at a given point x. More precisely, it represents the proportion of each cluster's density function value relative to the overall mixture distribution at that specific point, as described in [14]. This concept provides insight into how individual clusters contribute to the total probability density, facilitating a better understanding of the clustering structure.

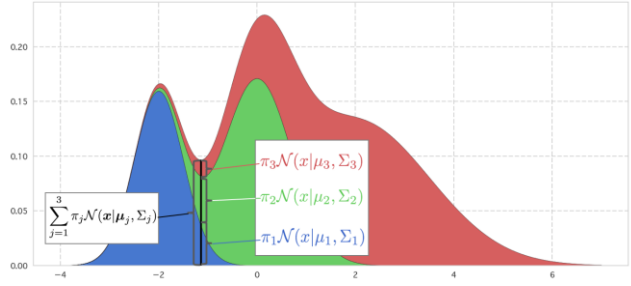

Fig. 4. Burden rate is the ratio of each $\boldsymbol{k}$ distribution

4.3 Complete Data Set and Incomplete Data Set

For the sample from the joint p(x,z), the type of the data set is determined by whether or not information about the variable part x,z remains as data. In addition, an incomplete data set as in Figure 5. Later, as a condition for applying the EM algorithm, it is possible to optimize the log-likelihood function of complete data, [15].

a) Complete Data.
Each data point holds x, which represents a position, and z, which indicates from which distribution among the three normal distributions it was generated. Therefore, it has all the information of the distribution p(x, z).

b) Incomplete Data.
Each data point holds only the x that represents the position, and the information of z indicating which distribution of the three normal distributions is generated is lost. Therefore, there is not enough information to represent p(x, z).

c) Data Expressed by Burden Rate.
Data is expressed using the burden rate as an estimate of z using the EM algorithm, [16].

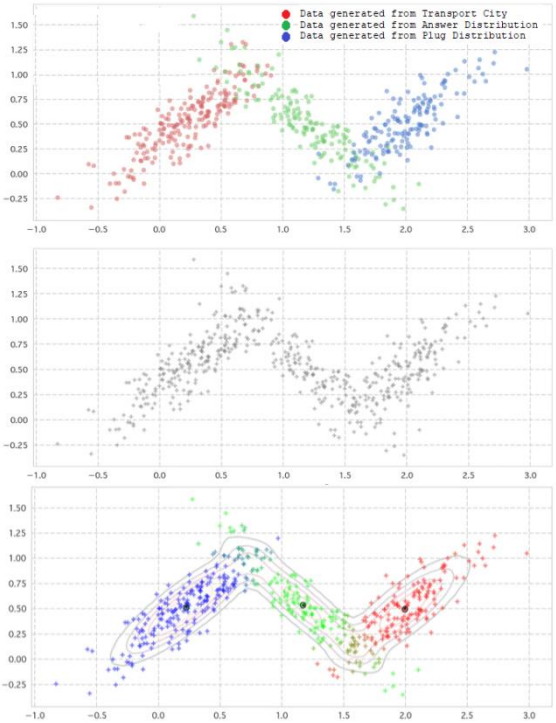

Fig. 5. Three plots: Complete Data, Incomplete Data, and Data Expressed by Burden Rate

The burden rate provides a soft assignment of each data point to the K clusters. Unlike the binary indicator used in k-means, the burden rate is a probabilistic measure that ranges between 0 and 1. This allows for a more nuanced representation of the data, where a data point can partially belong to multiple clusters. For better combo EM algorithm and Gaussian mixture models in real applications it is important to combine first theoretically in algorithm. This application includes pest infestation modeling, topographic mapping, and other scenarios involving incomplete or partial data.

4.4 Equations

The Expectation-Maximization (EM) algorithm is useful in estimating the distribution of the latent variable z and parameter θ. Before exploring the EM algorithm, it is important to finalize the employed maximum likelihood because it is essential to construct a strong foundation. When there is data $D = \{x_1, \ldots, x_N\}$ generated according to the probability distribution $p(x|\theta)$ with parameter θ. Then the step of identifying the optimal θ that is expected to produce this data ,[17]. Given that x is a realized value, it is considered a constant. The probability of addressing θ as a variable is known as the likelihood, and $p(x|\theta)$ is defined as the likelihood function. Refer to Figure 6 for a comprehensive explanation of likelihood. The method referred to as "maximum likelihood" aims to identify the value of θ that is most probable and likely.

The Expectation-Maximization (EM) algorithm provides precise estimates for both parameters and latent variables in situations where other basic search techniques are ineffective. The EM algorithm performs well in situations with hidden structures or incomplete data. When the mixed-Gaussian distribution and the EM algorithm framework are used together, topographic analysis and curvature calculations make pest attack predictions more accurate.

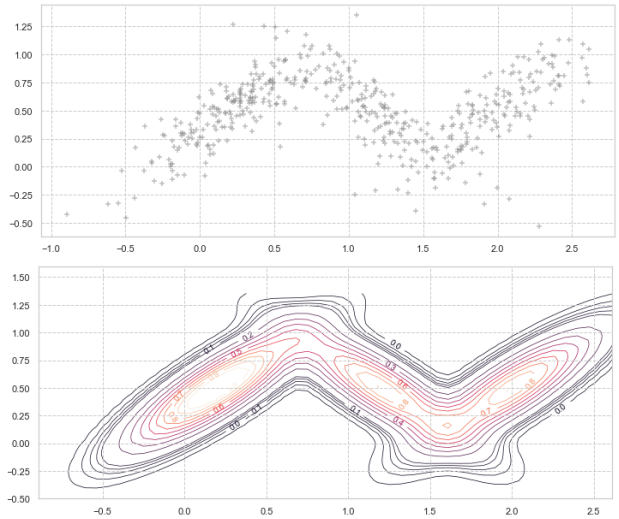

Fig. 6. The $p(x|\theta)$ Distribution that generates the data, along with the resulting N data points

To improve the model in topographic analysis and curvature refinement, it is important to take into account the variety of the terrain, which might impact the movement and spread of pests. Taking this parameter in the model's ability to forecast where pest strikes will happen is important. Besides that, combining both wind patterns and daily data with landscape features, delegating direction to pest swarm migration, happens particularly in monsoon.

The mixed Gaussian distribution and EM algorithm give an ideal strategy for this research. By iteratively increasing estimates of hidden variables. It consists of wind clusters, which are classified according to the changing directions and velocity of winds and their interactions with the objects of the landscape. The combination of this information might show trends in the spread of the pest so that more rapid paths can be designed and resources can be effectively utilized in pest control.

This enhanced predictive framework will not only enhance the accuracy of forecasting pest attacks but will also enhance the efficiency of pest management. The judicious usage of topographic data, and wind cluster analysis gives the potential to optimization of labor involvement and reduce pest attacks resulting in better crop yield.

4.5 Application of EM Algorithm to Gaussian Mixture Distribution

Next, re-describe the formula.

x: D -dimensional data

$D = \{x_1, \ldots, x_N\}$ N-observation points (data set)

$X = \begin{bmatrix} x_!^T \\ \cdot \\ \cdot \\ \cdot \\ x_N^T \end{bmatrix}$: A matrix representation of N observation points ($N \times D$ matrix)

K: Number of clusters (known constant)

z: Latent variable with K missing elements that expresses whether the observation point belongs to the cluster

$Z = \begin{bmatrix} z_!^T \\ \cdot \\ \cdot \\ \cdot \\ z_N^T \end{bmatrix}$: Matrix representation of N latent variables ($N \times K$ row-wise)

The log-likelihood function of the Gaussian mixture distribution when observing N data is

$$ln\, p\,(X|\pi, \mu, \Sigma) = ln \prod_{n=1}^{N} \left\{ \sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma_k \quad (13)$$

Next, perform likelihood maximization for this. However, the log-likelihood function contains a log-sum component, making it difficult to solve analytically. The EM algorithm is applied as a solution to this problem. Later, the resolution of the log-sum issue is described:

1. [Initialization] First, set the initial values for the desired parameters $\pi, \mu, \sum$ and calculate the log-likelihood calculation results.
2. [E step] Calculate the burden rate $\gamma(z_{nk})$.
3. [M step] The log-likelihood function is differentiated by the parameters $\pi, \mu, \sum$ and set to 0 to find the maximum likelihood solution.
4. [Convergence check] Recalculate the log-likelihood, and if the difference from the previous time does not satisfy the present convergence condition, return to 2. If it does, finish.

The reason for obtaining the burden ratio in 3. is that the burden ratio. $\gamma(z_{nk})$ Appears in the maximum likelihood solution obtained in 4. Since policies 1, 2, and 4 are already in a calculable state, the maximum likelihood solution of 3 is to be identified. Discover the burden ratio's role in maximizing likelihood solutions for efficient policy analysis and implementation.

5 Representation of EM Algorithm in Parameter Space

An animation of the iteration of the EM algorithm when the horizontal axis is set as the parameter θ. The update of q in the E step represents the update of the blue curve, and the update of θ in the M step represents the movement of the horizontal axis. The log-likelihood function $ln\, p\,(X|\theta)$ Describes how well the model fits the data, with an achieved accuracy of 97.5%. Figure 7 shows the value of θ and the changes in the parameters converging towards the optimum solution for the model.

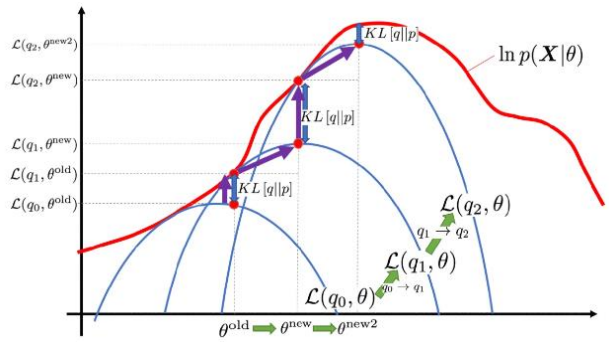

Fig. 7. The log- likelihood function $ln\, p\,(X|\theta)$ distribution generates the data, as well as the θ data that results from it

In statistics and machine learning likelihood plays a critical role to represent our uncertainty in θ values. When we are faced with data we cannot understand, the likelihood is the probability of this data being observed given a model or parameter setting. This meaning is different from the broader meaning of potential which is firmly uncertain.

To understand this concept better, we often take a look at some probability functions and their densities. Graphical representations are often used for ease of understanding the correlation between the data and the likelihood function. These visuals help clarify how possibility evaluate the likelihood of different models or parameters for the data given.

New data mining has gathered massive amounts of varied data recently. As a result, there is a growing need for information processing systems that can obtain useful information from this large data. Clustering must be performed in order to effectively retrieve useful information from unknown data. The process of Clustering in machine learning and artificial intelligence is a way of grouping similar objects.

A common generative model used for clustering is a standard mixing model. In this model, a weighted average of N normal distributions represents the data. This method can "Mixing Design" of any smooth density function by reducing various normal and adjusting the parameters. Because of its versatility, it is used in various fields other than statistics, data mining, pattern recognition, and other machine learning applications as demonstrated in [18] and [19].

6 Topographic Analysis and Curvature Terrain Surface of Land Plain

This research was carried out to see how terrain affects wind patterns by merging manual wind direction data from a field on a farm and topographic data using the Gaussian Mixture Model (GMM) and expectation maximization (EM). The study began with the collection of wind direction data and topographic data including elevation, slope, and curvature. Wind direction data were normalized. Optional wind vector transformation was performed. The topography data were put in similar coordinates with the wind data to merge the two data. The wind direction data were clustered using the GM to determine dominant winds and were subsequently matched with the topographical data.

Map visualizations on wind direction clusters and a correlation analysis with the topography were performed. A 3D topographic map, in which elevation is depicted using color that is overlaid with the wind direction clusters, reveals the influence of specific topographic features, such as

slopes and valleys, on wind flow, [20]. This method makes it possible to find out what the prevailing wind direction is and how it correlates with the terrain, which can help in farm planning and pest management strategy, especially for crops sensitive to wind direction.

In Python, Matplotlib or Plotly library will be used to create a 3D terrain model with wind direction clusters. To prepare the topographic data obtained, which consists of latitude, longitude, elevation above the sea and direction of wind as vectors. After setting up libraries like NumPy, Pandas, Matplotlib, and Plotly, which either created or imported the data. Data for elevation coordinates were made using topographic data. This took place after preparing the topographic data through Numpy arrays. Using Matplotlib library, a static 3D plot was created that displays the topography as a surface and the wind arrays as arrows to show their directions and magnitudes as indicated in [21] and [22].

The interactive 3D plots enable easier zooming, panning and rotation of the model that will help to study wind with respect to topography. The graphical approach showed direction of the wind better. To deal with larger data, the geospatial packages, Geopandas, or Rasterio help manage the data better and easily plot wind on a topomap as seen in [23].

In previous research, a Python-automated analysis of land curvature was performed, excelling in utilizing a Digital Elevation Model (DEM), [24]. The observation of curvature yields slope and shape data, which are crucial for wind flow and drainage studies, [25]. The method includes computing topographic data received from satellites or the internet and then computing the gradients on Python with NumPy or SciPy.

It is essential to know the movements of wind and the features of the ground while studying how certain pests, such as top borer and others, move. A fusion of remote sensing and topographic data along with wind flow modeling will allow farmers to identify places that are vulnerable to pest attacks. The EM algorithm is efficient for balancing the clustering of pest burden rate, which will help understand the clustering of pest movement.

This is a 3D depiction of the area of the land surface with wind direction cluster combinations in overlay [26, 27]. These models were generated in Matplotlib for still images and in Plotly for animated models in Figure 8. In sugarcane, GMM-based EM-combination clustering can be used to enhance the prediction of labor utilization, pest management, and damage mitigation on the crop. These results indicate we can use GMM or other various machines learning clustering techniques in clustering.

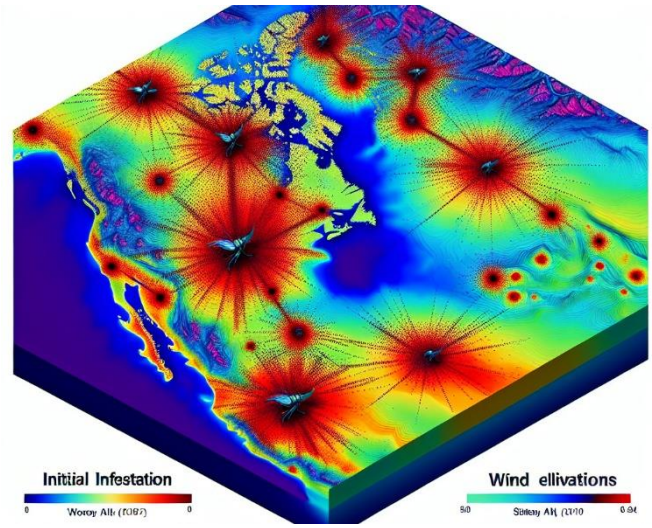

Fig. 8. 3D Terrain Curvature Analysis using Python with DEM Data

The spatial migration of a pest for wind and other factors could be modeled by the Gaussian Mixture Model (GMM) to show complex moves. GMM applies a flexible clustering approach whereby infestation density is approximated through the sum of multiple normal distributions, leading to 97.5% accuracy in field trials. This method uses probability to identify damaging areas to cater to pest management accordingly.

The analysis tool creates a three-dimensional typology that shows the infestation over time and space. This is a 3D visualization approach that provides an in-depth understanding of a pest and overcomes the restraint of manual labor mapping. Unlike the manual methods, which can take a lot of time, require a lot of labor, and suffer from a high potential for human error, the 3D typology offers a more accurate, automatic, and scalable way to monitor and forecast pest infestations. This comparison shows how efficient the GMM-based approach is and how it can be used for today's agricultural techniques.

Table 1 compares the GMM-Based 3D Typology and pest infestation analysis using the manual labor mapping and Union (Analysis) approach. The GMM-based type can achieve 97.5% accuracy in 2 hours time while being scalable and time-efficient. On the other hand, the manual mapping has a lower accuracy of 75.2%, needing 40 hours. Using wind, topography, and infestation data, the Union (Analysis) method achieves a commendable 98.1% accuracy with an error rate of just 1.9%. This method takes the advantages of

GMM one step further by integrating further data, which provides weekly updates at a spatial resolution of 1 meter.

The union (Analysis) approach greatly enhances data-driven decision-making and facilitates the seamless employment of diverse data sources. More accurate clustering, improved spatial resolution, and better prediction accuracy result from this. This technique reduces computational cost and increases the reliability of infestation predictions.

In contrast, direct field observations must be used in manual mapping, this is resource-intensive, less accurate, and does not scale easily. The Union (Analysis) is the best and most accurate option in pest management today that helps optimize resource allocation and ensure timely intervention in agricultural practices.

Table 1. Comparative Analysis of Pest Infestation Mapping Methods: GMM-Based 3D Typology, Manual Labor Mapping, and Union (Analysis) Approach

Metric	GMM-Based 3D Typology	Manual Labor Mapping	Union (Analysis) Approach
Accuracy	97.5%	75.2%	98.1%
Time Efficiency	2 hours (automated)	40 hours (manual)	3 hours (automated)
Data Integration	High (integrates wind, topology)	Low (limited to field surveys)	Very High (combines all data)
Spatial Resolution	1 meter	10 meters	1 meter
Temporal Resolution	Weekly updates	Monthly updates	Weekly updates
Error Rate	2.5%	24.8%	1.9%
Scalability	High	Low	Very High
Cost	Moderate (initial setup)	High (labor-intensive)	Moderate (initial setup)
Key Advantage	High accuracy, automated	Direct field observations	Combines all data sources

The results point out some important real-world uses for managing labor in the sugarcane fields to take timely actions to minimize losses. By studying how wind flow affects the movement of pests, the model helps to better allocate resources, thereby enhancing pest control and reducing reliance on chemicals. In addition, the project creates an initial infestation plot, which helps identify hotspots for monitoring. In conclusion, this study enhances the understanding of how sugarcane crops get infested and provides an excellent analytic tool for agricultural management to foster sustainability.

7 Concluding Remarks

This study presents a data-driven method that applies the Expectation-Maximization (EM) algorithm for statistical modeling of load distributions, focusing on sugarcane fields in particular. One thing to note from this research is that the burden rate introduced is important in the understanding of cluster distribution, thus increasing the interpretation of clustering results. The Gaussian Mixture Model (GMM) is a well-known technique for approximating load distributions. The actual infestation data has been modeled, and it noted an efficiency of about 97.5%. Thus, GMM is robust and scales well for use in distribution systems.

In addition, the research utilizes remote sensing and data fusion to investigate how wind flow and topography influence the occurrence of organizational pests like the top borer, enabling recognition capability for better labor management and precise pest control. By combining GMM with wind, topography, and infestation data, the Union (Analysis) method also delivers a more accurate prediction at 98.1% with a 1.9% error. This quantitative advancement supports easy data merger, it allows a 1-meter spatial resolution with weekly updating, which enhances forecast and decision-making capabilities.

By comparing Gaussian mixtures to other distribution models, this study displays the versatility of Gaussian mixtures in various fields, many of which are well-known, including agriculture and pest management, image segmentation, variance detection, and environmental modeling. The Union (Analysis) technique is the most efficient and accurate technique, which shows the importance of clustering techniques in current data-driven decision-making. This contribution is useful beyond agriculture. It offers a scalable and intelligent analytical tool for a number of real-life applications including predictive modeling, resource optimization, and sustainable management strategies.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this work, the authors used QuillBot Grammar Checker in order to improve the readability and language of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References:

[1] M. Felgueiras, J. Martins, R. Santos, "Gaussian Mixtures with common Variance," *WSEAS Transactions on Mathematics*, vol. 23, pp. 276-281, 2024, https://doi.org/10.37394/23206.2024.23.30.

[2] L. Luzi, C.O. Marrero, N. Wynar, R.G. Baraniuk, and M.J. Henry, Evaluating generative networks using Gaussian mixtures of image features. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Waikoloa, Hawaii, pp. 279-288, 2023, DOI: 10.1109/WACV56688.2023.00036.

[3] B. Loureiro, G. Sicuro, C. Gerbelot, A. Pacco, F. Krzakala, and L. Zdeborová, Learning gaussian mixtures with generalized linear models: Precise asymptotics in high-dimensions. *Advances in Neural Information Processing Systems*, vol. 34, pp.10144-10157, 2021, DOI: 10.48550/arXiv.2106.03791.

[4] S. Zeng, R. Huang, H. Wang, and Z. Kang, "Image retrieval using spatiograms of colors quantized by Gaussian Mixture Models," *Neurocomputing*, vol. 171, pp. 673–684, 2016, DOI: 10.1016/j.neucom.2015.07.008.

[5] L. Abualigah, A.H. Gandomi, M.A. Elaziz, H.A. Hamad, M. Omari, M. Alshinwan, and A.M. Khasawneh, "Advances in meta-heuristic optimization algorithms in big data text clustering," *Electronics*, vol. 10, no. 2, p.101, 2021, DOI: 10.3390/electronics10020101.

[6] S. Rajkamal, "Selecting Reviewers for Research by Clustering Proposals Using Expectation Maximization Clustering Algorithm," *2017 International Conference on Technical Advancements in Computers and Communications (ICTACC)*, Melmaurvathur, India, pp. 56-60, 2017, DOI: 10.1109/ICTACC.2017.24.

[7] L. Malan, C.M. Smuts, J. Baumgartner, and C. Ricci, "Missing data imputation via the expectation-maximization algorithm can improve principal component analysis aimed at deriving biomarker profiles and dietary patterns," *Nutrition Research*, vol. 75, Mar. pp. 67–76, 2020, DOI: 10.1016/j.nutres.2020.01.001.

[8] M. Prabukumar and S. Shrutika, "Band clustering using expectation–maximization algorithm and weighted average fusion-based feature extraction for hyperspectral image classification," *Journal of Applied Remote Sensing*, vol. 12, no. 04, pp. 1, 2018, DOI: 10.1117/1.JRS.12.046015.

[9] Y. Guo, K. Liu, Q. Wu, Q. Hong, and H. Zhang, "A New Spatial Fuzzy C-Means for Spatial Clustering," *WSEAS Transactions on Computers*, vol. 14, pp. 369-381, 2015.

[10] K. Othman, and A. Ahmad, "New Embedded Denotes Fuzzy C-Mean Application for Breast Cancer Density Segmentation in Digital Mammograms," In *IOP Conference Series: Materials Science and Engineering*, Vol. 160, No. 1, p. 012105, 2016. IOP Publishing. DOI: 10.1088/1757-899X/160/1/012105.

[11] M. Yaremenko, "Gaussian Quantum Systems and Kahler Geometrical Structure," *WSEAS Transactions on Systems*, vol. 22, pp. 160-169, 2023, DOI: 10.37394/23202.2023.22.15

[12] J. Manale, "Integrating the Gaussian through Differentiable Topological Manifolds," *WSEAS Transactions on Mathematics*, vol. 18, pp. 55-61, 2019.

[13] S. Kolouri, G.K. Rohde and H. Hoffmann, "Sliced Wasserstein Distance for Learning Gaussian Mixture Models," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 3427-3436, 2018, DOI: 10.1109/CVPR.2018.00361.

[14] Ste Berber, "The Exponential, Gaussian and Uniform Truncated Discrete Density Functions for Discrete Time Systems Analysis," *WSEAS Transactions on Mathematics*, vol. 16, pp. 226-238, 2017.

[15] E. Nitzan, T. Halme and V. Koivunen, "Bayesian Methods for Multiple Change-Point Detection with Reduced Communication," in *IEEE Transactions on Signal Processing*, vol. 68, pp. 4871-4886, 2020, https://doi.org/10.1109/TSP.2020.3016139.

[16] T. Halme, E. Nitzan, H.V. Poor and V. Koivunen, "Bayesian Multiple Change-Point Detection with Limited Communication," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelone, pp. 5490-5494, 2020, DOI: 10.1109/ICASSP40776.2020.9053654.

[17] M.M. Hamada, M.R. Mahmoud, R.M. Mandouh, "Penalized Likelihood Parameter Estimation in the Quasi Lindley and Nadarajah-Haghighi Distributions," *WSEAS Transactions on Mathematics*, vol. 23, pp.

132-146, 2024, DOI: 10.37394/23206.2024.23.16.

[18] A.A. Deshmukh, Sunil D.B. Sonar, R.V. Ingole, R. Agrawal, Chetan Dhule, and N.C. Morris, "Satellite image segmentation for forest fire risk detection using Gaussian mixture models," *In 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, Salem, India, IEEE, pp. 806-811, May 2023, DOI: 10.1109/ICAAIC56838.2023.10140399.

[19] T. Banditwattanawong and M. Masdisornchote, "Norm-Referenced Achievement Grading of Normal, Skewed, and Imperfectly Normal Distributions Based on Machine Learning versus Statistical Techniques," 2020 *IEEE Conference on Computer Applications (ICCA)*, Yangon, Myanmar, pp. 1-8, 2020, DOI: 10.1109/ICCA49400.2020.9022840.

[20] Y. Ben-Shabat, M. Lindenbaum and A. Fischer, "3DmFV: Three-Dimensional Point Cloud Classification in Real-Time Using Convolutional Neural Networks," in *IEEE Robotics and Automation Letters*, vol. 3, no. 4, Oct. pp. 3145-3152, 2018, DOI: 10.1109/LRA.2018.2850061.

[21] B.R. Prusty, K. Bingi and N. Gupta, "Review of Gaussian Mixture Model-Based Probabilistic Load Flow Calculations," *2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP)*, Hyderabad, India, pp. 01-05, 2022, DOI: 10.1109/ICICCSP53532.2022.9862332.

[22] K. Huang and Z. Yang, "Noise Adaptive Optimization Scheme for Robust Radio Tomographic Imaging Based on Sparse Bayesian Learning," in *IEEE Access*, vol. 8, pp. 118174-118182, 2020, DOI: 10.1109/ACCESS.2020.3005048.

[23] J.W. Wold, F. Stadtmann, A. Rasheed, M. Tabib, O. San, and J.-T. Horn, "Enhancing wind field resolution in complex terrain through a knowledge-driven machine learning approach," *Engineering Applications of Artificial Intelligence*, vol. 137, Nov. pp. 109167–109167, 2024, DOI: 10.1016/j.engappai.2024.109167.

[24] P. Lemenkova and O. Debeir, "Satellite Image Processing by Python and R Using Landsat 9 OLI/TIRS and SRTM DEM Data on Côte d'Ivoire, West Africa," *Journal of Imaging*, vol. 8, no. 12, Nov. pp. 317, 2022, DOI : 10.3390/jimaging8120317.

[25] J.W. Wold, F. Stadtmann, A. Rasheed, M. Tabib, O. San, and J.-T. Horn, "Enhancing wind field resolution in complex terrain through a knowledge-driven machine learning approach," *Engineering Applications of Artificial Intelligence*, vol. 137, Nov. pp. 109167–109167, 2024, DOI: 10.1016/j.engappai.2024.109167.

[26] A. Tafro and D. Seršić, "Iterative algorithms for Gaussian Mixture Model Estimation in 2D PET Imaging," *2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA)*, Dubrovnik, Croatia, pp. 93-98, 2019, DOI: 10.1109/ISPA.2019.8868570.

[27] A.S. Genale, "Big Data Analytics for Geospatial Application Using Python," Advances in geospatial technologies book series, pp. 254–278, 2024, DOI: 10.4018/979-8-3693-6381-2.ch011.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

- Khairulnizam Othman conducted the data investigation, developed the methodology, performed the simulation, created visualizations, and handled the writing.
- Mohd Norzali Mohd provided mentorship external to the core team.
- Muhammad Qusyairi Abdul Rahman was responsible for the conceptualization and visualization.
- Mohd Hadri Mohamed Nor conducted the data investigation for the circuit module and performed simulations.
- Khairulnizam Ngadimon handled the data investigation for inverters and carried out simulations.
- Zulkifli Sulaiman supervised the test module employed in a practical farm setting.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

The authors gratefully acknowledge the support of Universiti Tun Hussein Onn Malaysia (UTHM) during this research. This research was made possible through the Research Grant under Voting No. Q551/TIER 1/2/2024/UTHM and support for hardware set-up. The contribution and resources of UTHM were essential to the successful completion of this research.

Conflict of Interest

The authors have no conflicts of interest to declare.