

Clustering performance using k-modes with modified entropy measure for breast cancer

Nurshazwani Muhamad Mahfuz^{1,2}, Heru Suhartanto³, Kusmardi Kusmardi^{4,5}, Marina Yusoff^{1,2}

¹College of Computing, Informatic and Media, Kompleks Al-Khawarizmi, Universiti Teknologi MARA (UiTM), Shah Alam, Malaysia

²Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Kompleks Al-Khawarizmi, Universiti Teknologi MARA (UiTM), Shah Alam, Malaysia

³Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia

⁴Department of Anatomical Pathology, Faculty of Medicine, Universitas Indonesia/Cipto Mangunkusumo Hospital, Jakarta, Indonesia

⁵Human Cancer Research Cluster, Indonesia Medical Education, and Research Institute, Universitas Indonesia, Jakarta, Indonesia

Article Info

Article history:

Received Jul 10, 2023

Revised Jul 14, 2023

Accepted Aug 10, 2023

Keywords:

Categorical data

Clustering

Distance metric

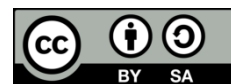
Entropy measure

Evaluation performance

ABSTRACT

Breast cancer is a serious disease that requires data analysis for diagnosis and treatment. Clustering is a data mining technique that is often used in breast cancer research to assess the level of malignancy at an early stage. However, clustering categorical data can be challenging because different levels in categorical variables can impact the clustering process. This research proposes a modified entropy measure (MEM) to enhance clustering performance. MEM aims to address the issue of distance-based measures in clustering categorical data. It is also a useful tool for assessing data loss in categorical clustering, which helps to understand the patterns and relationships by quantifying the information lost during clustering. An evaluation compares k-modes+MEM, k-means+MEM, DBSCAN+MEM, and affinity+MEM with conventional clustering algorithms. The assessment metrics of clustering accuracy, intra-cluster distance and fowlkes-mallow index (FMI) are employed to evaluate the algorithm performance. Experimental results show significant improvements. k-modes+MEM algorithm achieves a reduction in average intra-cluster distance and outperforms other algorithms in accuracy, intra-cluster distance, and FMI. The proposed algorithm can be extended to heterogeneous datasets in various domains such as healthcare, finance, and marketing.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Marina Yusoff

Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Kompleks Al-Khawarizmi

Universiti Teknologi MARA

Shah Alam, Selangor, Malaysia

Email: marina998@uitm.edu.my

1. INTRODUCTION

Cancer is a major contributor to global mortality, and breast cancer is a significant contributor, ranking as the second leading cause of cancer-related deaths among women [1]–[3]. Breast cancer is a prevalent and potentially life-threatening disease that requires accurate diagnosis and effective treatment strategies [4]. Data analysis techniques have advanced significantly in recent years which offers valuable insights into complex datasets [5]. Clustering is one of these techniques that has gained prominence in breast cancer research for its ability to uncover distinct patterns and relationships within the data, thereby aiding in improved decision-making and patient care.

Clustering relies on the concept of similarity or dissimilarity between data points. Traditionally, distance-based measures have been used to quantify the similarity between numerical variables. However, clustering categorical data presents unique challenges due to the absence of direct numerical values for distance computations. Categorical attributes represent non-quantifiable characteristics, so the notion of distance or similarity is not as straightforward. Appropriate measures are carefully selected to ensure accurate and meaningful results when clustering categorical data, such as breast cancer data [6], [7]. The nature of categorical data in breast cancer research makes it difficult to cluster using traditional methods. This is because categorical data cannot be easily converted into numerical representations, and similarity measures designed for numerical data may not be appropriate for categorical data. To overcome these challenges, specialized techniques and approaches have been developed that specifically address the challenges of clustering categorical data in breast cancer research.

Conventional clustering algorithms, such as k -modes, are designed for numerical data. However, clustering categorical data from breast cancer datasets presents unique challenges. Unlike numerical data, which can be used for distance-based computations, categorical data lacks inherent numerical values that can be employed in clustering algorithms. k -modes is an extension of k -means that solves this problem by assigning each data point to the cluster with the mode (most frequent value) of the same categorical attribute. It allows k -modes to be used for clustering categorical data while maintaining the same efficiency as k -means [8]. There are some challenges in clustering categorical data, including the number of categories in each variable that can vary, the categories may not be evenly distributed, and the categories may have different meanings or interpretations [9]. Despite these challenges, several clustering algorithms using different similarities such as the Jaccard index, Hamming distance, and Cosine similarity have been developed for categorical data. However, clustering categorical data can be a valuable tool for identifying patterns and relationships in breast cancer data that may not be apparent from other methods [10].

In this regard, modified entropy measures (MEM) are necessary to overcome these challenges and ensure reliable clustering outcomes. It influences the concept of entropy to quantify the dissimilarity between categorical variables. These measures capture the degree of uncertainty or randomness within the variables and provide a basis for evaluating the quality of clustering results [11]. It incorporates inherent limitations and constraints when employing measures based on entropy in the clustering of categorical data. MEM enables a more accurate assessment of the data distribution and helps identify the categories that significantly impact clustering results by considering each uncertainty of category and information content. It considers the unique characteristics of categorical variables and identifies the categories that have a substantial impact on the clustering substantially the accurate representation of the underlying patterns and relationships within the breast cancer data. In this paper, we propose the utilization of the MEM in conjunction with conventional clustering to enhance clustering performance as well as the reliability of the clustering outcome.

The remainder of the paper is organized as follows. Section 2 briefly reviews related work in k -modes clustering using entropy and breast cancer. Section 3 discusses the methodology involved, and section 4 proposes a method for performance criteria and comparison of computation involving the results. Section 5 is the discussion, which. Finally, section 6 concludes the paper with future improvements to future work.

2. RELATED WORK

Breast cancer is the most prevalent form of cancer among women worldwide [12], and extensive research has been conducted to uncover patterns, relationships, and subtypes within breast cancer datasets [13]. Various machine learning techniques have been explored for breast cancer classification, with promising results. For example, machine-learning approaches offer alternative prognostic tools, particularly in the Asian region, for breast cancer survival studies [14]. This means that they can be used to predict the likelihood of a patient surviving breast cancer. Additionally, the k -nearest neighbor (KNN) algorithm has shown superior accuracy in predicting breast cancer recurrence [15]. This means that it can be used to predict the likelihood of a patient's breast cancer recurring.

The Hamming distance is a metric often used to measure the similarity between two categorical vectors. It is calculated by counting the positions where the two vectors differ. A Hamming distance of 0 means that the two vectors are identical, while a Hamming distance of 1 means that the two vectors differ in exactly one position. The Hamming distance has several limitations. First, it does not distinguish between different types of differences [16] such as Hamming distance of 1 is the same whether the two vectors differ in the first or last positions. Second, the Hamming distance does not consider the order of the categories [17]. For example, a Hamming distance of 1 is the same whether the two vectors differ in the first and second positions or the second and first positions. Despite these limitations, the Hamming distance is a simple and effective metric for measuring the similarity between categorical vectors. It is often used in clustering algorithms, where it is used to group vectors that are similar to each other.

One specific clustering algorithm designed for categorical data is k -modes clustering. It identifies the most frequent values (modes) within each cluster and can handle incomplete categorical matrix datasets by assigning missing values to the most frequent value in the cluster. Studies have shown that k -modes clustering performs well in terms of accuracy, recall, adjusted rand index (ARI), and normalized mutual information (NMI) [18]. Furthermore, the utilization of refined initial points in the k -modes clustering algorithm has shown improved precision compared to the random selection method without refinement. This refinement process enhances the reliability and applicability of the algorithm in various data mining applications involving categorical data [19].

The effectiveness and stability of the proposed distance-entropy method have been validated through numerical and visual results [20]. Furthermore, graph clustering privacy preservation in the social internet of things (IoT) has been enhanced by incorporating structure entropy-based clustering techniques [21]. The research on machine learning techniques for clustering, classification, and prediction in breast cancer and other categorical data domains shows promise. However, further studies are necessary to develop more effective algorithms and validate previous research findings.

3. METHOD

The paper presents a new algorithm called k -Modes+MEM specifically designed for clustering breast cancer data. This algorithm combines two key techniques such as the k -modes clustering algorithm and the MEM. Figure 1 visually represents the overall process flow for handling categorical data in the proposed algorithm.

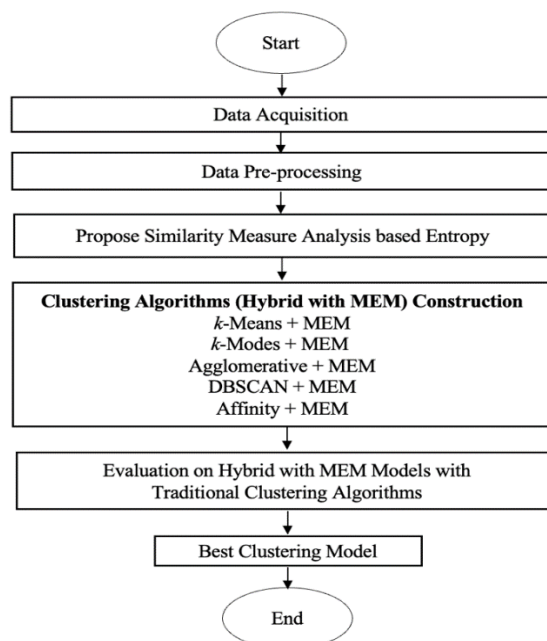


Figure 1. Process flow of clustering process

3.1. Data acquisition

The unclassified controlled information (UCI) dataset for breast cancer is accessed from the UCI machine learning repository website. It consists of 201 instances belonging to one class and 85 instances belonging to another. The dataset comprises categorical features with distinct scales: ordinal, nominal, and binary. The features consist of age, menopause, tumor size, inv-nodes, node caps, deg-malign, breast, breast quad, and irradiation.

3.2. Modified entropy measure algorithm

The MEM algorithm assists as a technique for clustering heterogeneous categorical data, which can be seen in Algorithm 1. When clustering categorical data, there is a risk of losing valuable information, which can be difficult in decision-making based on the data. To tackle this issue, the algorithm begins by evaluating the reliability of each feature within the dataset and determining how dependable it is.

The entropy distance measure is estimated to enhance the ability of the algorithm to handle heterogeneous categorical information and its associated information loss. The significance of a feature, essential for effective data clustering, is assessed. Steps 1 through 3 detail the process of determining feature reliability.

Once the feature reliability is established, the algorithm assigns weights to these features, containing different feature types from the questionnaire. Steps 4 to 6 handle the identification of weights for each feature. These steps consider the reliability of all features, including binary, nominal, Likert data, and ordinal data, in order to allocate weights accurately. Steps 7 to 33 elaborate on how the algorithm calculates the dissimilarity between two individuals using ordinal, nominal, Likert scale, and binary.

The subsequent phase involves computing the distances between various feature categories using their assigned weights and entropy, where entropy is constructed through a dissimilarity matrix based on respondent choices. Eventually, the distance matrix is generated from these entropy values. The calculation of the distance between two individuals is influenced by the differing feature weights that distinguish them.

Algorithm 1. Modified entropy distance measure

```

Input : Data:  $N \times D$  matrix
Output: Distance  $(x_i, x_j)$  for  $i, j \in \{1, 2, \dots, n\}$ 
1. For  $r = 1$  to  $d$  do
2. Reliability  $r = \frac{E_{O_r(s)}}{s}$ 
3. End for
4. For  $r = 1$  to  $d$  do
5. Weights,  $w = \frac{R}{\sum R}$ 
6. End for
7. For  $r_{ordinal} = 1$  to  $d_{ord}$  do
8. If  $i_r \neq j_r$  then
9.  $Dist(O_r(i_r), O_r(j_r))^2 = w \cdot \sum_{s=\min(i_r, j_r)}^{\max(i_r, j_r)} E_{O_r(s)}$ 
10. Else
11.  $Dist(O_r(i_r), O_r(j_r))^2 = 0$ 
12. End if
13. End for
14. For  $r_{Likert\ scale} = 1$  to  $d_{Likert\ Scale}$  do
15. if  $i_r \neq j_r$  then
16.  $Dist(O_r(i_r), O_r(j_r))^2 = w \cdot \sum_{s=\min(i_r, j_r)}^{\max(i_r, j_r)} E_{O_r(s)}$ 
17. Else
18.  $Dist(O_r(i_r), O_r(j_r))^2 = 0$ 
19. End if
20. End for
21. For  $r_{nominal} = d_{ord, Likert\ Scale} + 1$  to  $d$  do
22. If  $i_r \neq j_r$  then
23.  $\varphi(O_r(i_r), O_r(j_r))^2 = w \cdot \sum_{s=i_r, j_r} E_{O_r(s)}$ 
24. Else
25.  $Dist(O_r(i_r), O_r(j_r))^2 = 0$ 
26. End if
27. End for
28. For  $r_{binary} = d_{ord, Likert\ Scale} + 1$  to  $d$  do
29. If  $i_r \neq j_r$  then
30.  $\varphi(O_r(i_r), O_r(j_r))^2 = w \cdot \sum_{s=i_r, j_r} E_{O_r(s)}$ 
31. Else
32.  $Dist(O_r(i_r), O_r(j_r))^2 = 0$ 
33. End for
    
```

3.3. Proposed hybrid clustering method with modified entropy measure

The proposed method aims to enhance the clustering performance of breast cancer data by introducing a hybrid clustering method with the MEM algorithm as shown in Algorithm 2. Traditional clustering methods include k-means, k-modes, Agglomerative, DBSCAN, and affinity propagation are embedded with MEM. By default, k-modes is using Hamming distance as a similar measure while k-means, Agglomerative, DBSCAN, and affinity propagation are using euclidean distance.

MEM steps are added in each of the traditional clustering methods for a new means of calculation of distance measures. The hybrid of k –Modes with the MEM algorithm procedure is described in Algorithm 2. Steps 1-2 start with initializing the number of modes, k . The application of MEM in Step 3. Step 4-5 is allocating data objects to the closest cluster using a simple measure of dissimilarity and updating each cluster mode for every allocation.

Step 6 compares the dissimilarity value of each object to the mode. A data object should be relocated to the appropriate cluster and the mode of the second cluster modified if it is the closest mode to another cluster in Step 7. If any data objects experience a change in clustering, repeat Step 6 until nothing occurs.

Algorithm 2. Hybrid k-modes+MEM

Input: Data: $N \times D$ matrix, where N is the number of data points and D is the number of features, k : Number of clusters, $Max_iterations$: Maximum number of iterations

Output: Cluster assignments for each data point

1. **Start**
2. Choose an initial mode number k
3. **Apply modified entropy distance measure**
4. Allocate data objects to the closest cluster using a simple dissimilarity measure.
5. After every allocation, update each mode of the cluster.
6. After allocating all data objects to a cluster, compare the dissimilarity value of each object to the mode.
7. If a data object is the closest mode to another cluster, it should be moved to the appropriate cluster and the mode of the second cluster should be updated.
8. Repeat step 6 until no data objects change clustering.
9. **End**

3.4. Evaluation criteria

Performance evaluation plays a pivotal role in clustering analysis. It facilitates the assessment of clustering algorithm quality and effectiveness. The evaluation encompasses various metrics, such as clustering accuracy (CA), intra-cluster distance and FMI. These metrics contribute to an understanding of how well a clustering algorithm is performing and its ability to appropriately group and differentiate data points.

3.4.1. Clustering accuracy

Clustering accuracy is a metric that measures how accurately a clustering algorithm has classified data and ability to correctly group similar data points [22], [23]. A high CA value indicates a high level of agreement between the predicted and true clusters which suggests a more accurate clustering outcome. The following describes the formula for clustering accuracy.

$$CA = \frac{(TP+TN)}{TP+TN+FP+FN} \times 100\% \quad (1)$$

Where, true positive (TP) refers to instances where the model correctly predicted the positive class. False positive (FP) refers to situations in which the model predicted a positive class, but the actual class was negative. True negative (TN) indicates that the model correctly predicted the negative class. False negative (FN) represents cases where the model predicted a negative class, but the actual class was positive.

3.4.2. Intra cluster distance

The intra-cluster distance is a significant indicator of clustering quality. A lower intra-cluster distance implies that the data points within a cluster are highly similar which indicates a higher degree of cohesion and cluster compactness. The formula is as follows.

$$Intra - Cluster Distance = \frac{1}{N} \sum_{i=1}^n distance (point_i, centroid) \quad (2)$$

Where N represents the total number of data points in the evaluated clusters and i refers to the equation of a particular feature, indicating that it contributes to the intra-cluster distance calculation.

3.4.3. Fowlkes-mallows index

The fowlkes-mallows index is a metric that assesses the similarity between two clusters by comparing the clustering result to a known ground truth partition [24]. FMI produces a similarity score ranging from 0 to 1, with a higher score indicating a higher similarity between the two clusters. The formula for calculating the FMI is as shown in:

$$FMI = \frac{TP}{\sqrt{(TP+FP)(TP+FN)}} \quad (3)$$

where, true positive (TP) refers to instances where the model correctly predicted the positive class. False positive (FP) refers to situations in which the model predicted a positive class, but the actual class was negative. True negative (TN) indicates that the model correctly predicted the negative class. False negative (FN) represents cases where the model predicted a negative class, but the actual class was positive.

4. COMPUTATIONAL RESULTS

In this section, the research results are presented along with an explanation of the effectiveness of the proposed techniques compared to conventional clustering methods. A novel approach that utilizes a MEM as a distance metric is introduced to minimize the distance within clusters and enhance the overall quality of clustering. The main objective of this study is to improve clustering quality by incorporating entropy measures into conventional clustering techniques. The proposed approach outperforms conventional clustering methods in terms of cluster compactness and homogeneity.

4.1. Clustering accuracy

Figure 2 illustrates the comparison of algorithms for CA using different clustering algorithms. The *k*-Modes+MEM algorithm exhibits the highest accuracy among the compared methods, achieving an accuracy rate of 47% and 15.18% increase in accuracy compared to the conventional *k*-modes algorithm. Furthermore, *k*-means+MEM, DBSCAN+MEM, and Affinity+MEM also show an improvement in performance accuracy from the conventional clustering algorithm, except for Agglomerative+MEM.

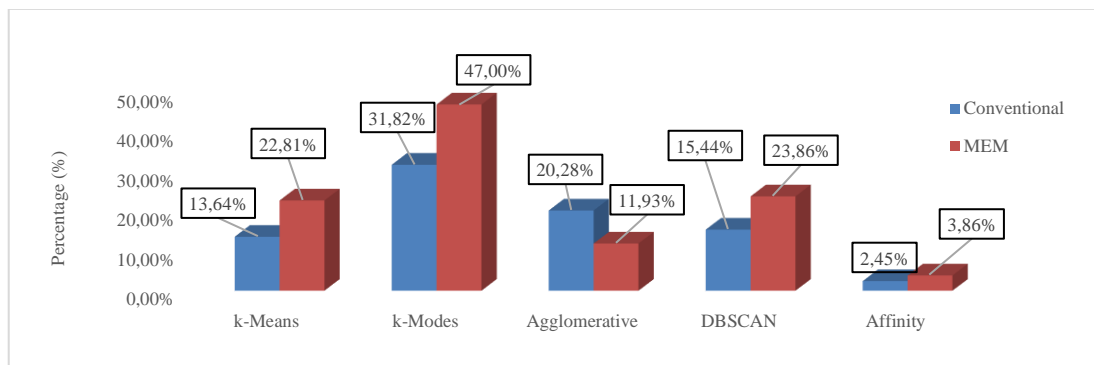


Figure 2. Comparison of distance metric for CA

4.2. Intra-cluster distance

Table 1 demonstrates the comparison of algorithms for average intra-cluster distance. The improved clustering performance of the *k*-Modes+MEM algorithm is evident from its lower intra-cluster distance value of 0.1614. A smaller average intra-cluster distance suggests that the data points within clusters are closer to each other. This implies tighter clustering and improved separation between clusters. This value indicates that the algorithm has successfully formed more compact and homogeneous clusters. In addition, other conventional clustering algorithms also show improvement when using the MEM method.

Table 1. Comparison of average intra-cluster distance

Clustering technique	Average of intra-cluster distance	Improvement (%)
k-means	2.4370	4.25%
k-means+MEM	2.2385	
k-modes	2.5034	87.89%
k-modes+MEM	0.1614	
Agglomerative	3.2622	9.85%
Agglomerative+MEM	2.6771	
DBSCAN	1.2912	15.91%
DBSCAN+MEM	0.9368	
Affinity	9.4456	80.98%
Affinity+MEM	0.9929	

4.3. Fowlkes-mallows index

Figure 3 indicates the comparison algorithms of distance metrics for FMI. FMI scores reveal that the *k*-modes+MEM clustering algorithm achieves the highest FMI score (FMI=0.76) compared to other algorithms. A high FMI value indicates a strong agreement between the clustering results and the true class labels. This demonstrates the ability of the *k*-modes+MEM algorithm to produce more accurate and reliable clustering results in comparison to the other methods evaluated. Furthermore, *k*-modes+MEM and Affinity+MEM shows improvement when using the MEM method.

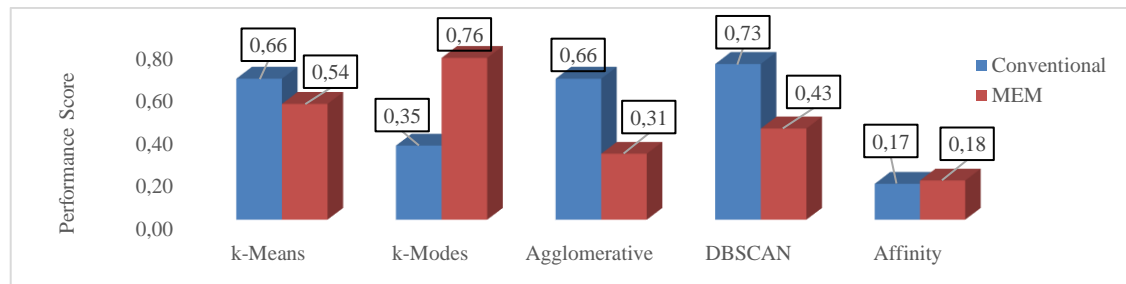


Figure 3. Comparison of distance metric for FMI

5. DISCUSSION

This research explores the impact of various clustering methods on clustering performance metrics. A novel clustering algorithm called the k -Modes+MEM algorithm is proposed to consider information for each data feature, which incorporates the MEM technique. The results demonstrate that the k -Modes+MEM algorithm generally outperforms other algorithms regarding accuracy, average intra-cluster distance, and FMI. This indicates that the algorithm creates more compact and homogeneous clusters, leading to better separation and more accurate clustering results [25], [26]. It also indicates superior convergence capabilities. The comparison with existing clustering methods confirms the superiority of the algorithm and highlights its potential as a reliable and effective approach for various clustering tasks. The proposed algorithm significantly advances the baseline method, reinforcing its effectiveness and value in achieving accurate and reliable clustering results. Despite the promising results, it is essential to acknowledge the limitations of this research which evaluated the performance of the framework on a single dataset which raises concerns about its generalizability to other datasets.

Hence, it is important to compare the performance of the algorithm with other benchmark datasets to ascertain its effectiveness. Nevertheless, the proposed k -Modes+MEM framework presents an efficient and effective solution for clustering categorical data, making a valuable contribution to categorical clustering. The utilization of the MEM technique further enhances the performance of the algorithms. The proposed algorithm outperforms the baseline method in various aspects. Consequently, it is well-suited for clustering scenarios based on heuristics.

6. CONCLUSION

In conclusion, a comparison was conducted between conventional clustering algorithms and clustering algorithms embedded with MEM. The objective of this hybrid approach was to leverage the strengths of both conventional clustering algorithms and clustering embedded with MEM. Metrics for accuracy, intra-cluster distance, and FMI were calculated to assess the effectiveness of the algorithms. The results consistently demonstrated that the hybridization algorithm outperformed state-of-the-art methods in most cases, demonstrating the potential for enhancing evaluation performance. k -Modes+MEM is expected to be tested on different heterogeneous data domains such as healthcare, finance, marketing, and more. This research contributes to the advancement of the field of clustering by providing evidence of the superiority of hybridized algorithms over traditional methods.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Strategic Research 100-RMC 5/3/SRP (053/2021), Institute for Big data analytics and artificial intelligence (IBDAAI), and Universiti Teknologi MARA for the financial support provided to this research project.




REFERENCES

- [1] A. Desiani, A. A. Lestari, M. Al-Ariq, A. Amran, and Y. Andriani, "Comparison of support vector machine and k-nearest neighbors in breast cancer classification," *Pattimura International Journal of Mathematics (PIJMath)*, vol. 1, no. 1, pp. 33–42, May 2022, doi: 10.30598/pijmathvol1iss1pp33-42.
- [2] M. Mangukiya, "Breast cancer detection with machine learning," *International Journal for Research in Applied Science and Engineering Technology*, vol. 10, no. 2, pp. 141–145, Feb. 2022, doi: 10.22214/ijraset.2022.40204.
- [3] S. F. Khorshid, A. M. Abdulazeez, and A. B. Sallow, "A comparative analysis and predicting for breast cancer detection based on data mining models," *Asian Journal of Research in Computer Science*, pp. 45–59, May 2021, doi: 10.9734/ajrcos/2021/v8i430209.




- [4] M. Yusoff, T. Haryanto, H. Suhartanto, W. A. Mustafa, J. M. Zain, and K. Kusmardi, "Accuracy analysis of deep learning methods in breast cancer classification: a structured review," *Diagnostics*, vol. 13, no. 4, p. 683, Feb. 2023, doi: 10.3390/diagnostics13040683.
- [5] R. S. Shankar, R. S. Chigurupati, P. Voosala, and N. Pilli, "An extensible framework for recurrent breast cancer prognosis using deep learning techniques," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 29, no. 2, pp. 931–941, Feb. 2023, doi: 10.11591/ijeecs.v29.i2.pp931-941.
- [6] Y. Tao *et al.*, "A comparative analysis of trajectory similarity measures," *GIScience and Remote Sensing*, vol. 58, no. 5, pp. 643–669, Jul. 2021, doi: 10.1080/15481603.2021.1908927.
- [7] Y. Donyatalab, F. K. Gündoğdu, F. Farid, S. A. Seyfi-Shishavan, E. Farrokhzadeh, and C. Kahraman, "Novel spherical fuzzy distance and similarity measures and their applications to medical diagnosis," *Expert Systems with Applications*, vol. 191, p. 116330, Apr. 2022, doi: 10.1016/j.eswa.2021.116330.
- [8] A. Sapegin and C. Meinel, "K-metamodes: Frequency-and ensemble-based distributed k-modes clustering for security analytics," in *Proceedings - 19th IEEE International Conference on Machine Learning and Applications, ICMLA 2020*, Dec. 2020, pp. 344–351, doi: 10.1109/ICMLA51294.2020.00062.
- [9] J. P. Sarkar, I. Saha, S. Chakraborty, and U. Maulik, "Machine learning integrated credibilistic semi supervised clustering for categorical data," *Applied Soft Computing Journal*, vol. 86, p. 105871, Jan. 2020, doi: 10.1016/j.asoc.2019.105871.
- [10] C. Zhu, L. Cao, and J. Yin, "Unsupervised Heterogeneous coupling learning for categorical representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 533–549, Jan. 2022, doi: 10.1109/TPAMI.2020.3010953.
- [11] M. K. Dhar, S. M. N. Hasan, T. R. Otushi, and M. Khan, "Entropy-based feature selection for data clustering using k-means and k-medoids algorithms," in *Proceedings - 2020 5th International Conference on Research in Computational Intelligence and Communication Networks, ICRCICN 2020*, Nov. 2020, pp. 36–40, doi: 10.1109/ICRCICN50933.2020.9296186.
- [12] D. H. Dr. Happy *et al.*, "Study on clinical presentation of breast carcinoma of 80 cases," *East African Scholars Journal of Medical Sciences*, vol. 5, no. 7, pp. 210–215, Jul. 2022, doi: 10.36349/easms.2022.v05i07.004.
- [13] S. S. Reddy, N. Pilli, P. Voosala, and S. R. Chigurupati, "A comparative study to predict breast cancer using machine learning techniques," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 27, no. 1, pp. 171–180, Jul. 2022, doi: 10.11591/ijeecs.v27.i1.pp171-180.
- [14] X. Xu, M. Jaber-Douraki, and N. A. Wallace, "Predicting the prognostic value of POLI expression in different cancers via a machine learning approach," *International Journal of Molecular Sciences*, vol. 23, no. 15, p. 8571, Aug. 2022, doi: 10.3390/ijms23158571.
- [15] I. K. A. Enriko, M. Melinda, A. C. Suliyani, and I. G. B. Astawa, "Breast cancer recurrence prediction system using k-nearest neighbor, naïve-bayes, and support vector machine algorithm," *Jurnal Infotel*, vol. 13, no. 4, pp. 185–188, Dec. 2021, doi: 10.20895/infotel.v13i4.692.
- [16] B. D. Verma, R. Pratap, and D. Bera, "Efficient binary embedding of categorical data using BinSketch," *Data Mining and Knowledge Discovery*, vol. 36, no. 2, pp. 537–565, Mar. 2022, doi: 10.1007/s10618-021-00815-y.
- [17] K. S. Dorman and R. Maitra, "An efficient k-modes algorithm for clustering categorical datasets," *Statistical Analysis and Data Mining*, vol. 15, no. 1, pp. 83–97, Feb. 2022, doi: 10.1002/sam.11546.
- [18] C. Zhang *et al.*, "MD-SPKM: A set pair k-modes clustering algorithm for incomplete categorical matrix data," *Intelligent Data Analysis*, vol. 25, no. 6, pp. 1507–1524, Oct. 2021, doi: 10.3233/ida-205340.
- [19] Y. Sun, Q. Zhu, and Z. Chen, "An iterative initial-points refinement algorithm for categorical data clustering," *Pattern Recognition Letters*, vol. 23, no. 7, pp. 875–884, May 2002, doi: 10.1016/S0167-8655(01)00163-5.
- [20] Y. Li and X. Chao, "Distance-Entropy: An effective indicator for selecting informative data," *Frontiers in Plant Science*, vol. 12, Jan. 2022, doi: 10.3389/fpls.2021.818895.
- [21] Y. Tian, Z. Zhang, J. Xiong, L. Chen, J. Ma, and C. Peng, "Achieving graph clustering privacy preservation based on structure entropy in social IoT," *IEEE Internet of Things Journal*, vol. 9, no. 4, pp. 2761–2777, Feb. 2022, doi: 10.1109/JIOT.2021.3092185.
- [22] C. Bepery, S. Bhadra, M. M. Rahman, M. K. Sarkar, and M. J. Hossain, "Improved mean shift algorithm for maximizing clustering accuracy," *Journal of Engineering Advancements*, vol. 2, no. 01, pp. 01–06, Jan. 2021, doi: 10.38032/jea.2021.01.001.
- [23] Q. Li, S. Wang, C. Zhao, B. Zhao, X. Yue, and J. Geng, "HIBOG: Improving the clustering accuracy by ameliorating dataset with gravitation," *Information Sciences*, vol. 550, pp. 41–56, Mar. 2021, doi: 10.1016/j.ins.2020.10.046.
- [24] C. Guyeux, S. Chrétien, G. B. Tayeh, J. Demerjian, and J. Bahi, "Introducing and comparing recent clustering methods for massive data management in the internet of things," *Journal of Sensor and Actuator Networks*, vol. 8, no. 4, p. 56, Dec. 2019, doi: 10.3390/jsan8040056.
- [25] M. Sharma and J. K. Chhabra, "Sustainable automatic data clustering using hybrid PSO algorithm with mutation," *Sustainable Computing: Informatics and Systems*, vol. 23, pp. 144–157, Sep. 2019, doi: 10.1016/j.suscom.2019.07.009.
- [26] M. Alswaitti, M. K. Ishak, and N. A. M. Isa, "Optimized gravitational-based data clustering algorithm," *Engineering Applications of Artificial Intelligence*, vol. 73, pp. 126–148, Aug. 2018, doi: 10.1016/j.engappai.2018.05.004.

BIOGRAPHIES OF AUTHORS






Nurshazwani Muhamad Mahfuz    is currently a Ph.D. student in Information Technology at Universiti Teknologi MARA (UiTM) Shah Alam. Her educational background consists of a Bachelor of Science (Statistics) (Hons) at Universiti Teknologi MARA (UiTM), Malaysia. She continued her education in a Master of Applied Statistics at Universiti Teknologi MARA (UiTM), Malaysia. Her current research interests include data mining, clustering, statistics, multi-view learning, artificial intelligence, optimization, and their application in timber. She can be contacted at email: 2017373123@student.uitm.edu.my.






Heru Suhartanto    graduated from Mathematics Department, Universitas Indonesia in 1986. He joined Computer Science Centre Universitas Indonesia in 1986; He got Ph.D. degree from Department of Mathematics, The University of Queensland (UQ, Australia) in 1998 with thesis on Parallel Iterated Multistep Runge-Kutta methods for solving large ordinary differential Equations. He got position as a Post Doctoral Fellow UQ until 2000. He was also an honorary Professor and adjunct Professor of School of ITEE UQ in 2014-3026, then 2017-2019. He got awards such as the 2010 University of Queensland Indonesia Alumni Award, UI Best Researcher, Dies Natalis ke-58 UI tahun 2008. Among his research interests are high performance computing, cloud computing, machine learning, and distance learning. He can be contacted at email: heru@cs.ui.ac.id.



Kusmardi M.Sc., Ph.D.    is a Lecturer and Researcher at Department of Anatomic Pathology Faculty of Medicine, Universitas Indonesia/Cipto Mangunkusumo Hospital, Drug Development Research Cluster of Indonesian Medical Education and Research Institute (IMERI), Human Cancer Research Cluster of IMERI, Universitas Indonesia. Research interest in immunology, pathology, cancer biology, herbal medicine, biomedical sciences, oncology, animal model for cancer research, biostatistic, and research methodology. He can be contacted at email: kusmardi.ms@ui.ac.id.



Marina Yusoff    is currently a senior fellow researcher at the Institute for Big Data Analytics and Artificial Intelligence (IBDAAI) and Associate Professor of Computer and Mathematical Sciences, at Universiti Teknologi MARA Shah Alam, Malaysia. She has a Ph.D. in Information Technology and Quantitative Sciences (Intelligent Systems). She previously worked as a Senior Executive of Information Technology at SIRIM Berhad, Malaysia. She is most interested in multidisciplinary research, artificial intelligence, nature-inspired computing optimization, and data analytics. She applied and modified AI methods in many research and projects, including deep learning, neural network, particle swarm optimization, genetic algorithm, ant colony, and cuckoo search for real-world problems and industrial projects. Her recent projects are data analytic optimizer, audio, and image pattern recognition. She has many impact journal publications and contributes as an examiner and reviewer to many conferences, journals, and universities' academic activities. She can be contacted at email: marina998@uitm.edu.my.