

# Clustering algorithms for analysing electronic medical record: A mapping study

Siti Nur Shahidah Zaman Shah<sup>1</sup>, Marshima Mohd Rosli<sup>1,2</sup>

<sup>1</sup>College of Computing, Informatics & Media, Universiti Teknologi MARA, Selangor, Malaysia

<sup>2</sup>Institute for Pathology, Laboratory and Forensic Medicine (I-PPerForM), University Teknologi MARA, Selangor, Malaysia

---

## Article Info

### Article history:

Received Oct 6, 2022

Revised Jan 18, 2023

Accepted Jan 30, 2023

---

### Keywords:

Clustering algorithm

Electronic medical record

Mapping study

Medical data

---

## ABSTRACT

Electronic Medical Records (EMRs) contain patients' history related to their medication, vaccine, test results and insurance information. EMRs need to be stored to facilitate the application of clinical treatment and prevention protocols. Clustering algorithms automate the process of information extraction and support health data management. Hence, in this mapping study, we systematically examine the literature on clustering algorithms used for analysing EMRs. We focus on studies published in 2016-2021 to present an overview of clustering techniques used in these studies to analyse medical data. We found 27 studies on clustering techniques, clustering technique problems and the evaluation parameters for analysing EMRs. However, although several studies have focused on this topic, only a few have taken the significant step of examining the clustering techniques used for analysing medical data particularly electronic medical record. Our results highlight that three clustering techniques have been used to analyse medical data, namely, the partitioning, the hierarchical and the density-based algorithms. We identified several clustering technique problems and 10 different evaluation parameters. The results suggest that researchers should focus on analysing medical data that will drive data-driven decision-making by management and promote a data-driven culture to ensure health care quality.

*This is an open access article under the [CC BY-SA](#) license.*



---

## Corresponding Author:

Marshima Mohd Rosli

College of Computing, Informatics & Media, Universiti Teknologi MARA

Selangor, Malaysia

Email: marshima@tmsk.uitm.edu.my

---

## 1. INTRODUCTION

Several public hospitals use an electronic medical record (EMR) system to manage patient data, such as data on patient care, diagnostic procedures, medical history and allergic conditions [1]–[3]. Owing to the increasing number of EMRs globally, medical information is too complex and voluminous to be processed and analysed using traditional methods. In recent years, the clinical information in EMRs is being fully used through data mining to identify patterns using a combination of machine learning and statistics [4]–[6].

Data mining has been conducted ever since its emergence in the late 1980s. It is a cross-disciplinary process that focuses on exploring very large datasets to uncover patterns [7]. Similarly, machine learning resides within artificial intelligence and is closely linked with data mining, its predecessor. Machine learning primarily uses learning algorithms to identify patterns in sample data and then uses these patterns to make decisions without being specifically programmed for this purpose [8].

Machine learning has been widely applied in medical research by mining EMRs using data mining methods to find patterns that can aid medical practitioners mainly in decision-making [9], [10]. Some machine learning techniques are regression, classification, clustering, dimensional reduction and reinforcement

learning. These techniques are commonly used as a tool for the solution of classification, forecasting and prediction problems [11].

Clustering, a common machine learning technique, involves grouping objects according to similarities in their data and classifying each into a specific group. The algorithm analyses the clustered data, identifies a group of data that share the same traits and develops a model from these data [12]. For example, clustering is used to group patients by gender, age and blood type based on their demographics.

In this study, we investigate the literature on clustering algorithms to identify challenges associated with analysing medical data in recent research. Prior reviews on clustering techniques have paid less attention to medical research. Therefore, we present the outcomes of a systematic mapping study to categorise recent research on clustering algorithms used to analyse medical data. Through this study, we aim to address this research gap by seeking answers to three main research questions: (RQ1) What are the studies that have investigated clustering techniques for analysing EMRs? (RQ2) What are the studies that have discussed problems associated with clustering techniques for analysing EMRs? (RQ3) What are the studies that have used evaluation parameters on clustering techniques?

Through this systematic mapping study, we provide researchers and practitioners with an overview of the existing clustering algorithms for analysing medical data. We also provide practical insights by exploring these research questions. Overall, this systematic mapping study contributes to the knowledge base on clustering algorithms for analysing medical data.

## 2. RELATED WORK

EMRs are the innovation that eliminate the need for traditional paper charts containing the patient's medical history [1]. The use of EMRs enables researchers to gain new insight into the depth of large datasets [13]. It is another way to learn new medical knowledge, apart from through conducting clinical and biological experiments. An EMR can store an extensive range of patient information, such as the history of medication, vaccination, test results and insurance information.

EMRs are used to store all past clinical information of each patient in detail during their treatment at medical institutions [14]. The incidence and clinical characteristics of diseases can be identified through mining all previous clinical information recorded in the EMR system [15]–[17]. Data mining has become an important technique to help researchers extract knowledge from large and complex data, such as medical records for patients [18]–[20].

Data mining is also termed 'knowledge discovery and data mining'. The data mining process mainly aims to extract essential and useful information from a database [7]. It involves data pre-processing techniques, such as data cleaning, integration, transformation and reduction; pattern discovery and knowledge evaluation [21]. Various data mining techniques, such as association, classification and clustering, can be used to extract interesting patterns from huge amounts of data.

The technique first used for data mining is association. It is a method used to detect interesting relationships between variables in large databases [22]. It can help to identify the rules relations in data collection between the medical authorities, such as medication, symptoms, disease and health condition [23], [24]. Next, classification is a method to extract a model with significant classes described. Some examples of classification techniques are the decision tree algorithm and the Naïve Bayes algorithm.

Clustering is another data mining technique. It involves grouping data with identical features into classes or clusters [22]. It can be applied in different fields, such as artificial intelligence, pattern recognition and neural networks [25]. In addition, clustering algorithms have been widely used for analysing medical data. There are several clustering techniques, such as the partitioning algorithm, the hierarchical algorithm and the density-based algorithm [26].

K-means is a known partitioning clustering algorithm commonly used to analyse EMRs. For example, it is used to cluster EMRs to identify the relevant features of patients with kidney failure and heart disease [27]. K-means is the most popular clustering algorithm because of its success in yielding efficient clusters, its ease of implementation, its improvement of prediction accuracy of and its flexibility [28].

The hierarchical algorithm is another clustering technique. It is quite popular owing to its visualisation capability [23]. It focuses on building a cluster hierarchy [29]. Hierarchical clustering groups data by type in a chain of importance, and it is used for probability analysis in healthcare [27]. This technique is commonly used for measuring physiological features across all patient medical records [30]. The last clustering algorithm technique is density-based clustering algorithm. Density-based spatial clustering of applications with noise (DBSCAN) is one of the most famous density-based clustering techniques. DBSCAN is more efficient in finding clusters, has the attribute of noise cancellation and is robust to outliers [31], [32]. The density-based algorithm is also broadly used on healthcare and medical datasets such as biomedical images. For example, DBSCAN is used on skin lesion images to detect homogeneous colour regions. In addition, DBSCAN is used to identify populations with dengue fever [29].

### 3. METHOD

A systematic mapping study is a type of well-defined methodology used to primarily identify and review all possible available research evidence that is relevant. Consequently, it is used to provide a clear explanation for specific research questions in a study. In addition, these questions should have been considered in the literature. In this mapping study, we adopted the guidelines suggested by researchers in software engineering [33], [34].

We categorised and structured empirical evidence based on studies that discussed the clustering algorithms used to analyse medical data. Thus, this study will help researchers and practitioners to gather, classify and aggregate results in related studies to identify research gaps and challenges for improvement in the body of knowledge. The mapping process consists of five activities: defining research questions, implementing the search strategy, selecting studies, keywording of abstracts and extracting data [35], [36].

#### 3.1. Research questions

Framing a clear, concise research question is an obligation for any systematic mapping study. Without a specified question, writing a well-written, well-focused review can be very onerous. For that reason, this study's research questions are developed using the PICOC (Population, Intervention, Comparison, Outcome, Context) model of Kitchenham *et al.* [34]. We constructed three research questions as follows:

RQ1: What are the studies that have investigated clustering techniques for analysing EMR?

RQ2: What are the studies that have discussed problems associated with clustering techniques for analysing EMR?

RQ3: What are the studies that have used evaluation parameters on clustering techniques?

#### 3.2. Search strategy

A thorough search is required for the process of identifying relevant literature. For this reason, a search strategy is necessary, which involves identifying the right search string with the relevant terms and keywords. This approach will ensure wide coverage and increase the chance of identifying the right publications. Therefore, in this mapping study, we constructed the search string as follows:

- Identify keywords in relevant papers.
- Search synonyms for identified keywords.
- Formulate search strings using Boolean OR to include alternative spellings and synonyms, and Boolean AND to combine major terms.

The search string: ("clustering technique" OR "clustering algorithm" OR "machine learning" OR "data mining") AND ("analyse") AND ("disease") AND ("electronic medical record" OR "EMR" OR "medical data") AND (LIMIT-TO (SUBJAREA, "COMP")) AND (LIMIT-TO (LANGUAGE, "English"))

In the search process, we included papers from conferences and journals on the research topics, such as the System & Software Journal and the Biomedical Informatics Journal. We limited our research to the computer science domain. We considered papers published in the five years from 2016 to 2021. We used the search string to conduct a primary search on SCOPUS, ScienceDirect, SpringerLink and ACM (Association for Computing Machinery) Digital. The search results are presented in Table 1.

Table 1. Primary search results

Online database	Search results	Duplicate papers	Relevant papers
SCOPUS	92	25	4
Science Direct	329	15	13
SpringerLink	96	9	4
ACM Digital	182	1	6
Total	699	50	27

#### 3.3. Selection of papers

The only inclusion criteria for the overall selection process were that the selected studies should be empirical studies related to clustering algorithms in the domain of medical data. Therefore, the literature search covered studies published within the 5-year period of 2016–2021. The specific inclusion criteria were as follows: (a) Peer-reviewed papers that provide evidence on analysing and evaluating EMR using clustering algorithms; (b) state the evaluation criteria for clustering techniques; and (c) use EMRs. The exclusion criteria were as follows: (a) Papers that provide evidence on the clustering techniques/algorithms but do not use EMRs or medical data and (b) are not in English.

On performing the automatic search, we obtained a set of 699 papers. In the screening process, first, we screened the title and abstract of each paper in order to exclude unrelated and identical documents, which reduced the number of papers to 121. Next, we scanned the introduction and conclusion of each paper, which reduced the first set to only 27 papers.

### 3.4. Classification scheme

First, we observed the keywords and concepts in the abstract of each paper. Next, we combined the keywords and concepts to produce an outline of the issues and challenges of the background research. We then selected suitable keywords to structure the schemes or categories. Last, we determined three different categories, which are clustering techniques for analysing EMRs (RQ1), problems in clustering techniques (RQ2) and evaluation criteria used to evaluate clustering techniques (RQ3).

### 3.5. Data extraction and mapping of studies

In this process, we used Excel spreadsheet for data logging and output production. We mapped the identified studies into three categories and obtained relevant information from these to answer the research questions. We tabulated the results into tables and calculated the publication frequencies in each category.

## 4. RESULTS

### 4.1. Clustering techniques for analysing EMR (RQ1)

Clustering algorithms have been widely used to analyse medical data [31], [37]–[40]. Three clustering techniques are commonly used for analysing EMRs: partitioning, hierarchical and density-based algorithms. K-means is a famous partitioning clustering algorithm that is commonly used for analysing EMRs [37]–[39], [41]. It is used to group a collection of patient records to identify the relevant features of patients with heart disease [42], [43]. In addition, some research efforts have been devoted to exploiting clustering techniques on data related to kidney failure [27]. Moreover, k-means clustering has also been used in a study to group patients with Parkinson's disease [44].

Next, the hierarchical algorithm is another clustering method discussed in recent studies owing to its visualisation capability, which is requested by many physicians [26]. Hierarchical clustering aims to construct a cluster hierarchy and can be regarded as a partitioning clustering sequence. Its groups data included in a type of chain of importance that are utilised for expectation in healthcare. Next, hierarchical clustering is commonly used for physiological measurements across all patients [30].

Density-based clustering algorithms are widely used to find clusters of nonlinear and arbitrary shapes based on the density of the connected points. DBSCAN is one of the most famous density-based clustering techniques [45]–[47]. The density-based algorithm is broadly used on healthcare and medical records such as biomedical images [48], [49]. DBSCAN is used to detect homogeneous colour regions in skin lesion images. It is also used to discover the hidden cluster and pattern for heart disease [50]. Further, it is used to determine the population with dengue fever [29].

We found 27 relevant studies, of which only 23 fulfilled the three inclusion criteria. We noticed that some had included more than one clustering technique. For example, k-means and DBSCAN were both used in one study [50]. Figure 1 presents the distribution of studies on clustering techniques for analysing EMR.

As Figure 1 shows, the number of studies on clustering techniques for analysing EMRs increased from 2016 to 2021. In 2019, the highest number of studies was published on the partitioning clustering technique, namely, 7 studies. Meanwhile, six studies were on hierarchical clustering and three on density-based clustering.

### 4.2. Clustering technique problems for analysing EMR (RQ2)

There are many problems related to the clustering technique. However, we only show the related and mentioned issues present in the papers without any clarification for categorization. In addition, papers that present more than one problem were counted in every category, and the total sums up to more than 27. Table 2 shows the frequency of problems in the clustering techniques.

As previously mentioned, we identified 27 studies that met the inclusion criteria. As illustrated in Table 2, the partitioning clustering algorithm is not recommended for datasets with noise and many outliers. This is because it is highly susceptible to noise and outliers with abnormal values [27]. The k-means partitioning clustering technique is biased to spherical clusters and is sensitive to the initial selection of centroids. Moreover, k-means is less efficient in partitioning the initial collection of data into subsets of different data distributions, including patients with different examination history [43].

Owing to space and time constraints, the hierarchical algorithm is sometimes not suitable for large data. If no data or limited data are available, hierarchical clustering may be a suitable option because the number of clusters is not predetermined. However, hierarchical clustering is expensive and not scalable. Therefore, it is not suitable for big data [29]. In addition, it is hard to detect the gaps between the objects.

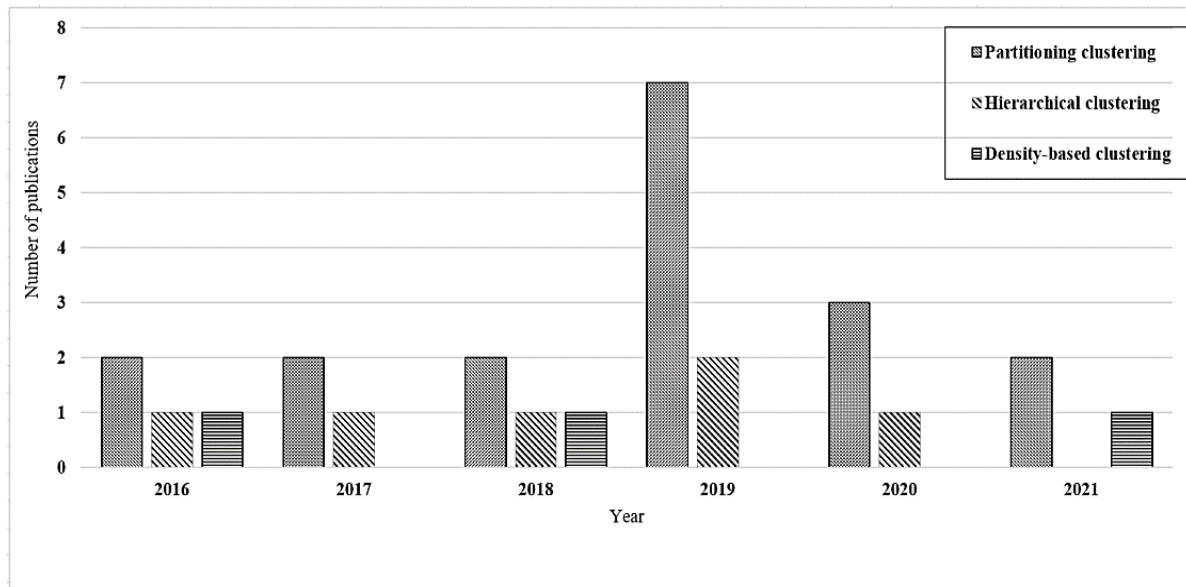


Figure 1. Distribution of studies on clustering techniques for analysing EMR

The process of density-based clustering is slow for large datasets. DBSCAN is able to find clusters with different sizes and forms, but it is weak in identifying clusters of variant density [42]. This limitation can be overcome by decomposing the clustering process in subsequent steps with the multiple-level DBSCAN algorithm. Another issue is that the DBSCAN algorithm fails to track cluster centres and cannot be simply trained and utilised with new data [32].

Table 2. Distribution of papers on clustering technique problems

Clustering techniques	Problems	No. of papers
Partitioning clustering	– P1: Highly susceptible to noise and abnormal values outliers.	10
	– P2: Biased to spherical clusters.	4
	– P3: Sensitive to the initial selection of centroids.	2
	– P4: Less efficient in partitioning initial collection of data into subsets.	2
Hierarchical clustering	– H1: Not suitable for big data owing to space and time limitations.	3
	– H2: Expensive and not scalable.	3
	– H3: Hard to detect the gaps between the objects.	1
Density-based clustering	– D1: Slow for large datasets.	3
	– D2: Weak in identifying clusters of variant density.	1
	– D3: Unable to track cluster centres.	2
	– D4: Cannot be simply trained and utilised with new data.	1
Total		31

#### 4.3. Evaluation parameters used in clustering techniques (RQ3)

All these algorithms are explained and analysed based on certain evaluation parameters, such as the number of clusters, cluster distance, epsilon value and threshold value. Table 3 illustrates the distribution of relevant papers for the common evaluation parameters used in 21 studies.

As Table 3 shows, several evaluation parameters are used in clustering techniques. The most used evaluation parameter is the number of clusters-9 papers applied this parameter to evaluate clustering performance. We found that two papers each used the sum of squared error, silhouette score, overall similarity, cluster density and threshold value as an evaluation parameter of the clustering technique. In addition, some other parameters can be used to evaluate the clustering technique, such as cluster distance, the Davies–Bouldin Index, epsilon value and neighbourhood size.

#### 4.5. Mapping

To provide an overview of clustering algorithms research analysing medical data in the literature, we present a map in Figure 2 that depicts the distribution of papers on clustering algorithms analysing medical

data according to their techniques, problems and common evaluation parameters used. This mapping study presents each paper in a bubble form, where the size and number of each bubble represent the frequency of papers classified for that category. Because a paper may contribute more than one technique, problem or evaluation parameters, each of these related papers is divided into more than one factor in each category. For that reason, the total number of paper counts in the bubble plot is not equal to the total of 27 relevant papers.

Table 3 Distribution of papers on evaluation parameters

No.	Evaluation parameter	No of papers
1.	Sum of Squared Error	2
2.	Silhouette Score	2
3.	Overall Similarity	2
4.	Cluster Density	2
5.	Cluster Distance	1
6.	Davies–Bouldin Index	1
7.	Epsilon Value	1
8.	Neighbourhood Size	1
9.	Number of Clusters	9
10.	Threshold Value	2
Total		21

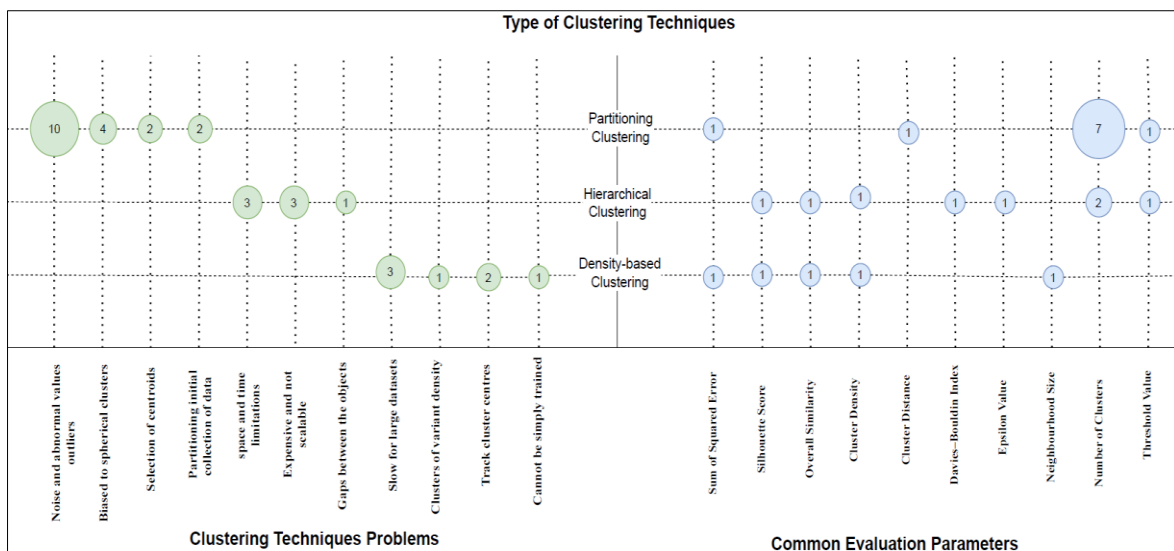


Figure 2. Distribution of clustering algorithms research by techniques, problems and evaluation parameters

5. DISCUSSIONS

In this study, we aimed to determine the status of recent research on clustering algorithms for analysing medical data. We performed a mapping study to identify the clustering techniques used, associated problems and the evaluation parameters used in these techniques and found 27 such studies. Almost all the reviewed papers were published between 2016 and 2021.

Thus, this study highlights the clustering techniques used to analyse EMRs. We found that 23 studies have applied clustering techniques for analysing EMRs, namely partitioning, hierarchical and density-based algorithms. Among all the techniques, the partitioning technique is the most popular. Some studies applied more than one clustering technique for data analysis. Further, 18 studies used partitioning clustering to cluster and analyse medical data, followed by five studies that used hierarchical clustering and three that used density-based clustering.

Next, this study emphasises the problems involved in these clustering techniques. Partitioning clustering is not recommended for noisy datasets because it is highly susceptible to noise. Next, hierarchical clustering is sometimes not suitable for big data because it is expensive and not scalable. Meanwhile, density-based clustering is weak in identifying clusters of variant density. Another issue is the failure to track cluster centres; further, it cannot be simply trained and utilised with new data.

Last, this study also highlights the evaluation parameters used in clustering techniques. Clustering techniques are analysed based on certain evaluation parameters. We identified 10 evaluation parameters: sum

of squared error, silhouette score, overall similarity, cluster density, cluster distance, the Davies–Bouldin Index, epsilon value, neighbourhood size, number of clusters and threshold value. We found that the number of clusters is the most popular evaluation parameter, with nine studies using it.

## 6. CONCLUSION

This systematic mapping study has provided an in-depth review of clustering algorithms used for analysing medical data. Concerning the limitations of the review, as we stated, according to the selection criteria, we included only studies that clearly examined clustering algorithms for analysing medical data. We excluded studies that described clustering algorithms but did not use EMRs or medical data. The results show that three main types of clustering techniques for analysing EMR which are partitioning, hierarchical and density-based clustering. Meanwhile, based on the mapping results, the most addressed problems of clustering techniques in the literature are issues that related with highly susceptible to noise and outliers with abnormal values. However, the results also reveal a lack of significant research in addressing evaluation parameter related with cluster distance, the Davies–Bouldin Index, epsilon value and neighbourhood size. We encourage researchers to consider clustering algorithms used to analyse medical data for improving the management of medical data.

## ACKNOWLEDGEMENTS

The authors would like to thank the Ministry of Education Malaysia and Universiti Teknologi MARA for their financial support to this project under Lestari Grant No. 600-RMC/MYRA 5/3/LESTARI (078/2020). We would also like to thank the College of Computing, Informatics & Media, Universiti Teknologi MARA, Selangor, Malaysia for all the supports.

## REFERENCES

- [1] G. Canino, P. H. Guzzi, G. Tradigo, A. Zhang, and P. Veltri, "On the analysis of diseases and their related geographical data," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 1, pp. 228–237, Jan. 2017, doi: 10.1109/JBHI.2015.2496424.
- [2] L. Waithera, J. Muhia, and R. Songole, "Impact of electronic medical records on healthcare delivery in Kisii teaching and referral hospital," *Medical & Clinical Reviews*, vol. 03, no. 04, 2017, doi: 10.21767/2471-299x.1000062.
- [3] R. Eden, A. Burton-Jones, A. Staib, and C. Sullivan, "Surveying perceptions of the early impacts of an integrated electronic medical record across a hospital and healthcare service," *Australian Health Review*, vol. 44, no. 5, pp. 690–698, 2020, doi: 10.1071/AH19157.
- [4] T. Giaeadi, "The impact of electronic medical records on improvement of health care delivery," *Libyan Journal of Medicine*, vol. 3, no. 1, p. 4, 2008, doi: 10.4176/071118.
- [5] G. Ferrante, A. Licari, S. Fasola, G. L. Marseglia, and S. La Grutta, "Artificial intelligence in the diagnosis of pediatric allergic diseases," *Pediatric Allergy and Immunology*, vol. 32, no. 3, pp. 405–413, 2021, doi: 10.1111/pai.13419.
- [6] D. Schlauch and E. Al., "Development of a real-time risk model (RTRM) for predicting in-hospital COVID-19 mortality," *Nucl. Phys.*, vol. 13, no. 1, pp. 104–116, 1959, doi: 10.1101/2021.04.26.21256138.
- [7] D. Verma and N. Mishra, "Analysis and prediction of breast cancer and diabetes disease datasets using data mining classification techniques," *Proceedings of the International Conference on Intelligent Sustainable Systems, ICISS 2017*, pp. 533–538, 2018, doi: 10.1109/ISS1.2017.8389229.
- [8] J. R. Koza, F. H. Bennett, D. Andre, and M. A. Keane, "Automated design of both the topology and sizing of analog electrical circuits using genetic programming," *Artificial Intelligence in Design '96*, pp. 151–170, 1996, doi: 10.1007/978-94-009-0279-4\_9.
- [9] J. M. Banda *et al.*, "Finding missed cases of familial hypercholesterolemia in health systems using machine learning," *npj Digital Medicine*, vol. 2, no. 1, 2019, doi: 10.1038/s41746-019-0101-5.
- [10] A. Pina *et al.*, "Virtual genetic diagnosis for familial hypercholesterolemia powered by machine learning," *European Journal of Preventive Cardiology*, vol. 27, no. 15, pp. 1639–1646, 2020, doi: 10.1177/2047487319898951.
- [11] S. Masrom, R. A. Rahman, N. Baharun, and A. S. A. Rahman, "Automated machine learning with genetic programming on real dataset of tax avoidance classification problem," in *Proceedings of the 2020 9th International Conference on Educational and Information Technology*, Feb. 2020, pp. 139–143, doi: 10.1145/3383923.3383942.
- [12] B. Huidong JIN and K.-S. Leung, "Scalable model-based clustering algorithms for large databases and their applications," 2002.
- [13] R. Wang, J. Zhao, L. Peng, B. Yang, L. Wang, and B. Li, "Medical entity recognition of esophageal carcinoma based on word clustering," *2018 International Conference on Security, Pattern Analysis, and Cybernetics, SPAC 2018*, pp. 348–353, 2018, doi: 10.1109/SPAC46244.2018.8965515.
- [14] V. Ehrenstein, H. Kharrazi, H. Lehmann, and C. O. Taylor, "Obtaining data from electronic health records," *Tools and Technologies for Registry Interoperability, Registries for Evaluating Patient Outcomes: A User's Guide*, pp. 1–92, 2019.
- [15] B. Song, Y. Feng, X. Li, Z. Sun, and Y. Yang, "Un-apriori: A novel association rule mining algorithm for unstructured EMRs," *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services, Healthcom 2017*, vol. 2017-December, pp. 1–6, 2017, doi: 10.1109/HealthCom.2017.8210792.
- [16] M. Hosni *et al.*, "A systematic mapping study for ensemble classification methods in cardiovascular disease," *Artificial Intelligence Review*, vol. 54, no. 4, pp. 2827–2861, 2021, doi: 10.1007/s10462-020-09914-6.
- [17] F. Ahmad, N. A. Mat Isa, Z. Hussain, and M. K. Osman, "Intelligent medical disease diagnosis using improved hybrid genetic algorithm - multilayer perceptron network," *Journal of Medical Systems*, vol. 37, no. 2, 2013, doi: 10.1007/s10916-013-9934-7.




- [18] E. Saleh *et al.*, "Learning ensemble classifiers for diabetic retinopathy assessment," *Artificial Intelligence in Medicine*, vol. 85, pp. 50–63, 2018, doi: 10.1016/j.artmed.2017.09.006.
- [19] S. Bashir, U. Qamar, F. H. Khan, and L. Naseem, "HNV: A medical decision support framework using multi-layer classifiers for disease prediction," *Journal of Computational Science*, vol. 13, pp. 10–25, 2016, doi: 10.1016/j.jocs.2016.01.001.
- [20] Y. Li, E. Porter, A. Santorelli, M. Popović, and M. Coates, "Microwave breast cancer detection via cost-sensitive ensemble classifiers: Phantom and patient investigation," *Biomedical Signal Processing and Control*, vol. 31, pp. 366–376, 2017, doi: 10.1016/j.bspc.2016.09.003.
- [21] M. Umamaheswari and P. I. Devi, "Prediction of myocardial infarction using K-medoid clustering algorithm," *Proceedings of the 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing, INCOS 2017*, vol. 2018-February, pp. 1–6, 2018, doi: 10.1109/ITCOSP.2017.8303128.
- [22] S. Babu *et al.*, "Heart disease diagnosis using data mining technique," *Proceedings of the International Conference on Electronics, Communication and Aerospace Technology, ICECA 2017*, vol. 2017-January, pp. 750–753, 2017, doi: 10.1109/ICECA.2017.8203643.
- [23] W. Sun, Z. Cai, F. Liu, S. Fang, and G. Wang, "A survey of data mining technology on electronic medical records," *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services, Healthcom 2017*, vol. 2017-December, pp. 1–6, 2017, doi: 10.1109/HealthCom.2017.8210774.
- [24] M. Liao, Y. Li, F. Kianifard, E. Obi, and S. Arcona, "Cluster analysis and its application to healthcare claims data: A study of end-stage renal disease patients who initiated hemodialysis epidemiology and health outcomes," *BMC Nephrology*, vol. 17, no. 1, 2016, doi: 10.1186/s12882-016-0238-2.
- [25] M. Sampath Premkumar and S. Hari Ganesh, "A median based external initial centroid selection method for k-means clustering," *Proceedings - 2nd World Congress on Computing and Communication Technologies, WCCCT 2017*, pp. 143–146, 2017, doi: 10.1109/WCCCT.2016.42.
- [26] R. Fang, S. Pouyanfar, Y. Yang, S. C. Chen, and S. S. Iyengar, "Computational health informatics in the big data age: A survey," *ACM Computing Surveys*, vol. 49, no. 1, 2016, doi: 10.1145/2932707.
- [27] A. Abugabah, A. Al Smadi, and A. Abuqabbah, "Data mining in health care sector: Literature notes," *ACM International Conference Proceeding Series*, pp. 63–68, 2019, doi: 10.1145/3372422.3372451.
- [28] N. Arora, S. Jain, and S. K. Verma, "Range clustering: An algorithm for empirical evaluation of classical clustering algorithms," *2016 9th International Conference on Contemporary Computing, IC3 2016*, 2017, doi: 10.1109/IC3.2016.7880242.
- [29] M. I. Razzak, M. Imran, and G. Xu, "Big data analytics for preventive medicine," *Neural Computing and Applications*, vol. 32, no. 9, pp. 4417–4451, 2020, doi: 10.1007/s00521-019-04095-y.
- [30] C. Fiarni, E. M. Sipayung, and S. Maemunah, "Analysis and prediction of diabetes complication disease using data mining algorithm," *Procedia Computer Science*, vol. 161, pp. 449–457, 2019, doi: 10.1016/j.procs.2019.11.144.
- [31] R. Delshi Howsalya Devi and P. Deepika, "Performance comparison of various clustering techniques for diagnosis of breast cancer," *2015 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2015*, 2016, doi: 10.1109/ICCIC.2015.7435711.
- [32] K. Magoev, V. V Krzhizhanovskaya, and S. V Kovalchuk, "Application of clustering methods for detecting critical acute coronary syndrome patients," *Procedia Computer Science*, vol. 136, pp. 370–379, 2018, doi: 10.1016/j.procs.2018.08.277.
- [33] D. Budgen, M. Turner, P. Brereton, and B. Kitchenham, "Using mapping studies in software engineering," *Proceedings of PPIG 2008*, vol. 2, pp. 195–204, 2008.
- [34] B. Kitchenham, P. Brereton, and D. Budgen, "The educational value of mapping studies of software engineering literature," *Proceedings - International Conference on Software Engineering*, vol. 1, pp. 589–598, 2010, doi: 10.1145/1806799.1806887.
- [35] B. A. Kitchenham, D. Budgen, and O. Pearl Brereton, "Using mapping studies as the basis for further research - A participant-observer case study," *Information and Software Technology*, vol. 53, no. 6, pp. 638–651, 2011, doi: 10.1016/j.infsof.2010.12.011.
- [36] K. Petersen, S. Vakkalanka, and L. Kuzniarz, "Guidelines for conducting systematic mapping studies in software engineering: An update," *Information and Software Technology*, vol. 64, pp. 1–18, 2015, doi: 10.1016/j.infsof.2015.03.007.
- [37] S. E. V. Haryanto, M. Y. Mashor, A. S. A. Nasir, and Z. Mohamed, "Identification of Giemsa stained of malaria using k-means clustering segmentation technique," *2018 6th International Conference on Cyber and IT Service Management, CITSM 2018*, 2019, doi: 10.1109/CITSM.2018.8674254.
- [38] P. Manivannan and P. I. Devi, "Dengue fever prediction using K-means clustering algorithm," *Proceedings of the 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing, INCOS 2017*, vol. 2018-February, pp. 1–5, 2018, doi: 10.1109/ITCOSP.2017.8303126.
- [39] M. Guftar, S. H. Ali, A. A. Raja, and U. Qamar, "A novel framework for classification of syncope disease using K-means clustering algorithm," *IntelliSys 2015 - Proceedings of 2015 SAI Intelligent Systems Conference*, pp. 127–132, 2015, doi: 10.1109/IntelliSys.2015.7361135.
- [40] C. Prathibhamol, A. Suresh, and G. Suresh, "Prediction of cardiac arrhythmia type using clustering and regression approach (P-CA-CRA)," *2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2017*, vol. 2017-Janua, pp. 51–54, 2017, doi: 10.1109/ICACCI.2017.8125815.
- [41] M. Thangamani, R. Vijayalakshmi, M. Ganthimathi, M. Ranjitha, P. Malarkodi, and S. Nallusamy, "Efficient classification of heart disease using K-means clustering algorithm," *International Journal of Engineering Trends and Technology*, vol. 68, no. 12, pp. 48–53, 2020, doi: 10.14445/22315381/IJETT-V68I12P209.
- [42] T. Cerquitelli, S. Chiusano, and X. Xiao, "Exploiting clustering algorithms in a multiple-level fashion: A comparative study in the medical care scenario," *Expert Systems with Applications*, vol. 55, pp. 297–312, 2016, doi: 10.1016/j.eswa.2016.02.005.
- [43] T. Cerquitelli, A. Servetti, and E. Masala, "Discovering users with similar internet access performance through cluster analysis," *Expert Systems with Applications*, vol. 64, pp. 536–548, 2016, doi: 10.1016/j.eswa.2016.08.025.
- [44] K. Poczeta, L. Kubus, and A. Yastrebov, "Multidimensional medical data modeling based on fuzzy cognitive maps and k-means clustering," *Procedia Computer Science*, vol. 176, pp. 118–127, 2020, doi: 10.1016/j.procs.2020.08.013.
- [45] M. F. Ijaz, M. Attique, and Y. Son, "Data-driven cervical cancer prediction model with outlier detection and over-sampling methods," *Sensors (Switzerland)*, vol. 20, no. 10, 2020, doi: 10.3390/s20102809.
- [46] K. Chandel, V. Kunwar, A. S. Sabitha, A. Bansal, and T. Choudhury, "Analysing thyroid disease using density-based clustering technique," *International Journal of Business Intelligence and Data Mining*, vol. 17, no. 3, pp. 273–297, 2020, doi: 10.1504/IJBIDM.2020.109297.
- [47] K. Mumtaz, M. Studies, and T. Nadu, "An analysis on density based clustering of multi dimensional spatial data," *Indian Journal of Computer Science and Engineering*, vol. 1, no. 1, pp. 8–12, 2010.






- [48] A. Al-Shammari, R. Zhou, M. Naseriparsaa, and C. Liu, "An effective density-based clustering and dynamic maintenance framework for evolving medical data streams," *International Journal of Medical Informatics*, vol. 126, pp. 176–186, 2019, doi: 10.1016/j.ijmedinf.2019.03.016.
- [49] A. Mansoul and B. Atmani, "Representative case-based retrieval to support case-based reasoning for prediction," *International Journal of Strategic Decision Sciences*, vol. 12, no. 3, pp. 14–35, 2021, doi: 10.4018/ijds.2021070102.
- [50] R. M. Liaqat, B. Mehboob, N. A. Saqib, and M. A. Khan, "A framework for clustering cardiac patient's records using unsupervised learning techniques," *Procedia Computer Science*, vol. 58, pp. 368–373, 2016, doi: 10.1016/j.procs.2016.09.056.

## BIOGRAPHIES OF AUTHORS



**Siti Nur Shahidah Zaman Shah**    received Bachelor Degree in Computer Science from Universiti Teknologi MARA in 2019. She received the Master degree in Computer Science from Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA in 2021. She can be contacted at email [sitnurshahidaa@gmail.com](mailto:sitnurshahidaa@gmail.com).



**Marshima Mohd Rosli**    is a senior lecturer at the Department of Computer Science, Universiti Teknologi MARA, Malaysia, where she has been a faculty member since 2007. Marshima graduated with Bsc (Hons) Information Technology from Universiti Utara Malaysia in 2001 and an M.Sc. in Real Time Software Engineering from Universiti Teknologi Malaysia in 2006. She completed her Ph.D in Computer Science from the University of Auckland, New Zealand, in 2018. Her research interests are primarily in the area of software engineering, artificial intelligent and data analytics. She can be contacted at email: [marshima@fskm.uitm.edu.my](mailto:marshima@fskm.uitm.edu.my)