Health and Technology (2024) 14:141–154
https://doi.org/10.1007/s12553-023-00805-8

ORIGINAL PAPER

Unsupervised machine learning for disease prediction: a comparative performance analysis using multiple datasets

Haohui Lu[1] · **Shahadat Uddin**[1]

Received: 18 September 2023 / Accepted: 21 November 2023 / Published online: 29 December 2023

Abstract

Purpose Disease risk prediction poses a significant and growing challenge in the medical field. While researchers have increasingly utilised machine learning (ML) algorithms to tackle this issue, supervised ML methods remain dominant. However, there is a rising interest in unsupervised techniques, especially in situations where data labels might be missing — as seen with undiagnosed or rare diseases. This study delves into comparing unsupervised ML models for disease prediction.

Methods This study evaluated the efficacy of seven unsupervised algorithms on 15 datasets, including those of heart failure, diabetes, and breast cancer. It used six performance metrics for this comparison. They are Adjusted Rand Index, Adjusted Mutual Information, Homogeneity, Completeness, V-measure and Silhouette Coefficient.

Results Among the seven unsupervised ML methods, the DBSCAN (Density-based spatial clustering of applications with noise) showed the best performance most times (31), followed by the Bayesian Gaussian Mixture (18) and Divisive clustering (15). No single model consistently outshined others across every dataset and metric. The study emphasises the crucial role of model and performance measure selections based on application-specific needs. For example, DBSCAN excels in Homogeneity, Completeness and V-measure metrics. Conversely, the Bayesian Gaussian Mixture is good in the Adjusted R and Index metric. The codes used in this study can be found at https://github.com/haohuilu/unsupervisedml/.

Conclusion This research contributes deeper insights into the unsupervised ML applications in healthcare and encourages further investigations into model selection. Subsequent studies could harness genuine disease records for a more nuanced comparison and evaluation of models.

Keywords Disease prediction · Performance comparison · Unsupervised machine learning · Healthcare dataset

1 Introduction

Machine learning (ML), a subfield of artificial intelligence, leverages computational methods to address challenges using historical data and information without requiring significant alterations to the fundamental process [1]. ML algorithms boast diverse applications, such as automated text classification [2], project analytics [3], spam email filtering [4], marketing analytics [5], and disease prediction [6]. They are primarily of two categories: supervised learning and unsupervised learning, with some researchers also acknowledging reinforcement learning algorithms that learn data patterns to respond to specific environments. Nevertheless, supervised learning and unsupervised learning are the most recognised types. The critical difference between these two categories lies in the existence of labels within the training data subset [1]. Supervised ML relies on labelled data. The dataset includes input features and corresponding output labels, allowing the algorithm to learn a mapping function to make predictions for test data or unseen data [7]. In contrast, unsupervised ML deals with unlabelled data, the dataset only consists of input features but no output labels. This method discovers patterns or clusters autonomously, without direct instructions [8].

The data science research community has recently shown an amplified interest in medical informatics, with disease prediction being a key area of focus [9]. Disease prediction plays a critical role in modern health. It allows for early

✉ Shahadat Uddin
shahadat.uddin@sydney.edu.au

Haohui Lu
haohui.lu@sydney.edu.au

[1] School of Project Management, Faculty of Engineering, The University of Sydney, Level 2, 21 Ross Street, Forest Lodge, NSW 2037, Australia

treatments and improves patient outcomes. ML is a robust tool for predicting disease risk within intricate health data. ML methods can learn from past data to predict future disease risks. Many studies are comparing the performance of supervised ML in the disease prediction domain [10–14].

Nonetheless, there are limited comparative studies on unsupervised ML in the disease prediction domain, as it has not gained as much popularity as supervised ML [9]. Data labels are not always available, particularly in cases where patients have undiagnosed or rare diseases. Vats et al. [15] compared the unsupervised ML techniques for liver disease prediction. They employed DBSCAN (Density-based spatial clustering of applications with noise), *k*-means, and Affinity Propagation to compare their prediction accuracy and computational complexity. Antony et al. [16] proposed a framework that compares different unsupervised ML methods for chronic kidney disease prediction. Alashwal et al. [17] investigated various unsupervised methods for Alzheimer's prediction, aiming to identify suitable techniques for patient grouping and their potential impact on treatment. Our research uncovered a gap in research, specifically a lack of thorough comparative studies of unsupervised learning algorithms across various types of disease prediction. As such, this research aims to evaluate the performance of different unsupervised ML algorithms in predicting diseases. It uses a variety of conditions, including heart failure, diabetes, and breast cancer, focusing on employing unsupervised ML techniques, such as *k*-means, DBSCAN and Agglomerative Clustering for disease prediction. The objective is to compare predictive performance by considering several performance measures, such as the Silhouette coefficient, Adjusted Mutual Information, Adjusted Rand Index, and V-measure. These measures are crucial in identifying the most effective approach for handling different datasets with numerous parameters. The key contributions of this research include:

- Comprehensive analysis and comparison of various unsupervised ML algorithms for disease risk prediction, using diverse benchmark datasets and performance measures.
- Identify the top-performing unsupervised ML method for healthcare researchers and stakeholders, which will eventually help select suitable techniques for enhanced disease risk prediction.

2 Methods

ML algorithms are primarily categorised into supervised and unsupervised learning based on the presence or absence of labels within the given data. Supervised learning uses labelled data, while unsupervised learning uses unlabelled data to discover patterns or clusters. This study focuses on different unsupervised learning methods in the disease prediction domain. They are partitioning clustering, model-based clustering, hierarchical clustering and density-based clustering.

2.1 Unsupervised machine learning algorithms

Unsupervised ML, also known as clustering, involves grouping data into clusters based on the similarity of their objectives within the same cluster while ensuring that they are dissimilar to objects in other clusters [8]. Clustering is a type of unsupervised classification since there are no predefined classes.

Figure 1 shows how unsupervised ML techniques classify three groups in a two-dimensional dataset. The dataset consists of 100 randomly generated data points divided into three groups based on their similarity. Different colours represent the clusters. On the scatter plot, the clusters are represented by different circles, and circular bounds have been placed around each cluster to visualise their boundaries better.

2.1.1 Partitioning clustering

Partitioning clustering requires the analyst to specify the number of clusters that should be generated. The *k*-means clustering is the most widely used method of partitioning clustering algorithms [18]. Figure 2 demonstrates the processes for the standard *k*-means clustering algorithm. The first step involves selecting *k* points as the initial centroids. After that, we need to classifying data points based on the distance to the centroids of these *k* clusters. Then, recomputing the centroid of each cluster based on classified points and repeating these steps until the centroids do not change. This study uses two popular *k*-means variants: classic *k*-means and Mini batch *k*-means [19].

2.1.2 Model-based clustering

Model-based clustering is another unsupervised ML method. It is a probabilistic approach to clustering that uses Gaussian Mixture Models (GMMs) to represent data as a mixture of Gaussian distributions [20]. GMM is a probabilistic model that attempts to fit a dataset to a combination of different Gaussian distributions. It evaluates the likelihood of each data point belonging to each cluster, as opposed to classic *k*-means clustering, which allocates each data point to a single cluster. This enables a more flexible and accurate representation of data distributions, mainly when dealing with overlapping or non-spherical clusters [20]. Figure 3

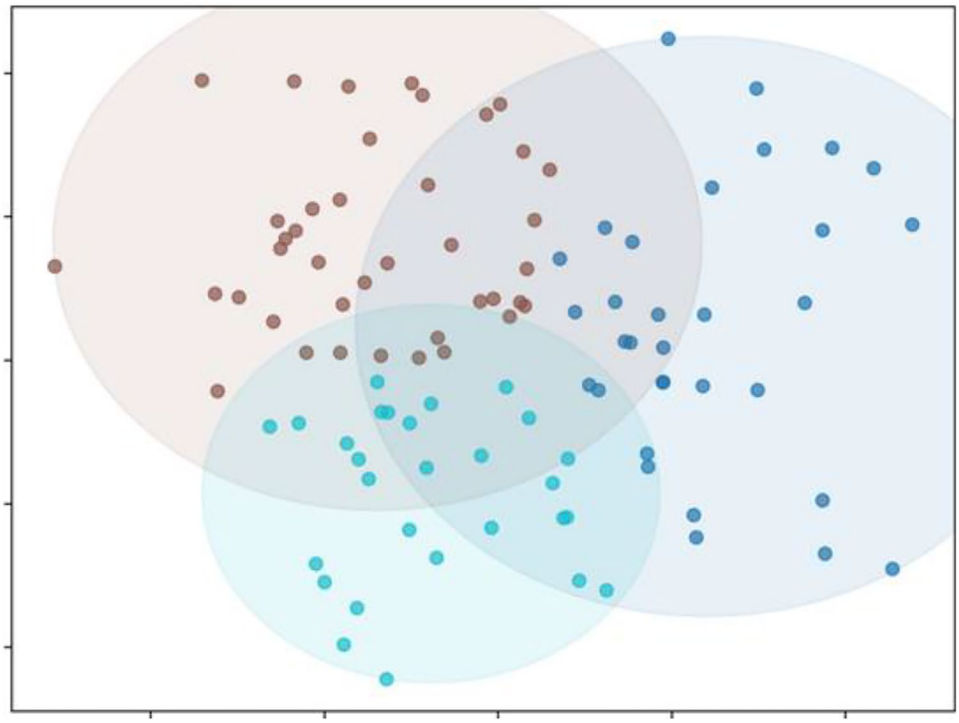

Fig. 1 An example of unsupervised learning

shows how a GMM model with four components is fitted to the data, and the resulting clusters are coloured. The GMM's Gaussian distributions are shown by ellipses, demonstrating each distribution's spread and direction and the probabilistic character of the clustering process. We also use Bayesian Gaussian Mixture (BGM) [21] for performance comparison.

2.1.3 Hierarchical clustering

Hierarchical clustering generates a group of nested clusters arranged in a hierarchical tree structure. This can be represented through a dendrogram, a tree-like diagram that documents the sequence of merges or splits [22]. Figure 4 shows

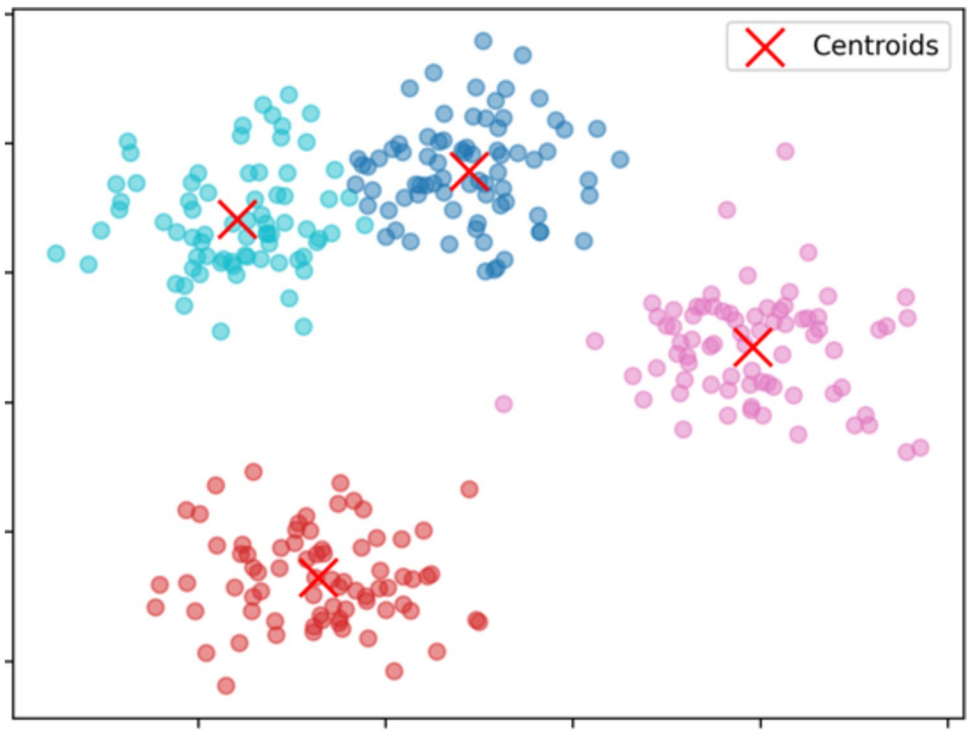

Fig. 2 Demonstration of the standard *k*-means clustering algorithm

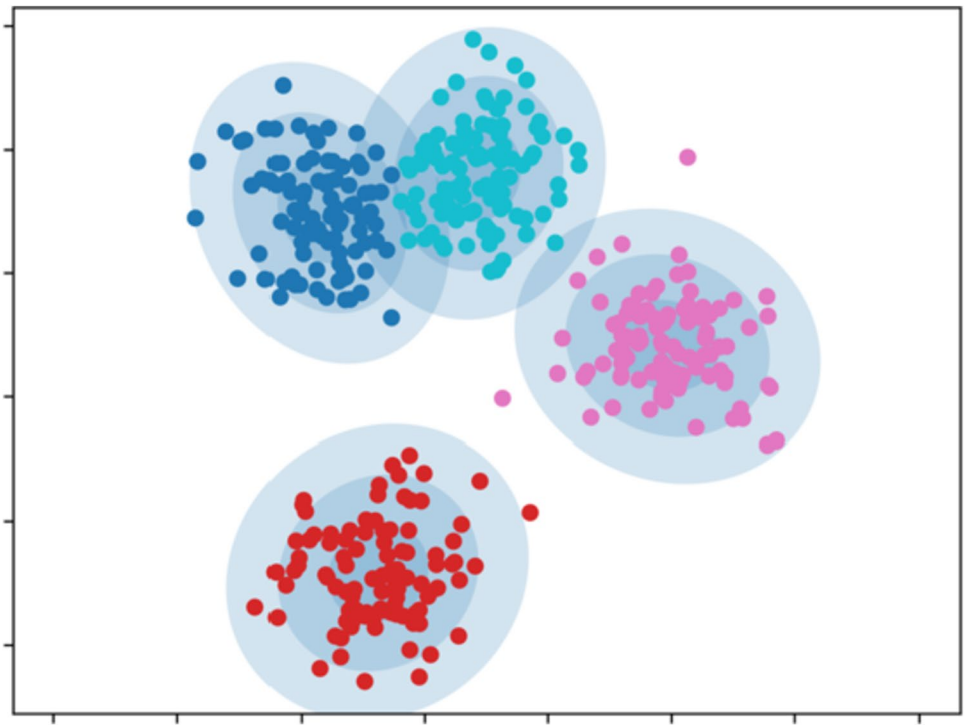

Fig. 3 Demonstration of the Gaussian Mixture Model

an example of a dendrogram for the hierarchical clustering. There are two main types of hierarchical clustering: agglomerative and divisive. Agglomerative clustering is a method that starts with each point as its cluster. As the process progresses, the nearest pair of clusters is merged in each step. This merging continues until it culminates in a single cluster or a specific number of clusters, depending on the parameters set at the outset of the process [22]. On the other hand, Divisive clustering is a method that begins with a single, all-encompassing cluster. As the process evolves, a cluster is

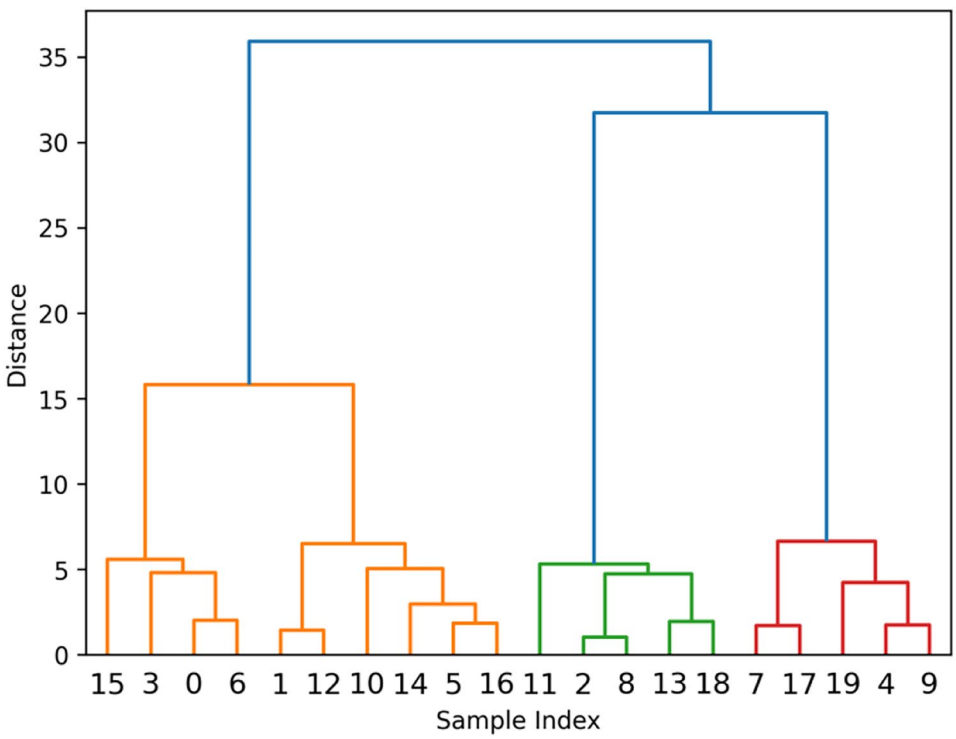

Fig. 4 Dendrogram for Hierarchical Clustering

split at each step. This splitting continues until every cluster contains only an individual point or a predetermined number of clusters are achieved, depending on the initial setup of the procedure [22]. This study uses both Agglomerative clustering and Divisive clustering for comparison.

2.1.4 Density-based

The density-based method relies on density as the local cluster criterion, such as points connected by density. Characteristics and features of density-based clustering include identifying clusters of any shape. It also effectively handles noise within the data. It requires only a single scan, examining the local region to validate the density. However, it necessitates the specification of density parameters as a condition for termination [22]. Density-based spatial clustering of applications with noise (DBSCAN) is a famous example of a density-based method [23]. This method labels high-density areas as clusters and low-density areas as outliers. It helps discover clusters of varied forms and deal with noise without requiring a set number of clusters [23]. Figure 5 visualises the clusters using the DBSCAN method.

2.2 Performance comparison measures

The performance of various unsupervised ML methods is assessed using different evaluation techniques, such as Adjusted Rand Index (ARI), Adjusted Mutual Information (AMI), Homogeneity, Completeness, V-measure, and Silhouette Coefficient. These are applied to establish comparative performance metrics in this study.

2.2.1 Adjusted Rand Index

Adjusted Rand Index (ARI) is a modification to the Rand index. It calculates a similarity metric between two clusters by considering all sample pairs and then counting those pairs that are either similarly or differently assigned in the predicted and actual clusters [24]. The formular for ARI is

$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathrm{Expecte\,RI}}{\mathrm{Max(RI)} - \mathrm{Expected\,RI}}$$

where *RI* is the Rand Index, *Expected RI* is the Expected Rand Index, and *Max* (*RI*) is the maximum possible Rand Index.

The ARI value lies between -1 and 1, where 1 means identical clustering and –1 means dissimilar clustering. If the ARI is equal to 0, it indicates random labelling.

2.2.2 Adjusted Mutual Information

Adjusted mutual information (AMI) modifies the Mutual information (MI) score to account for chance [25]. It acknowledges that MI tends to increase with larger clusters, independent of the actual amount of shared information between them. The formula for AMI is

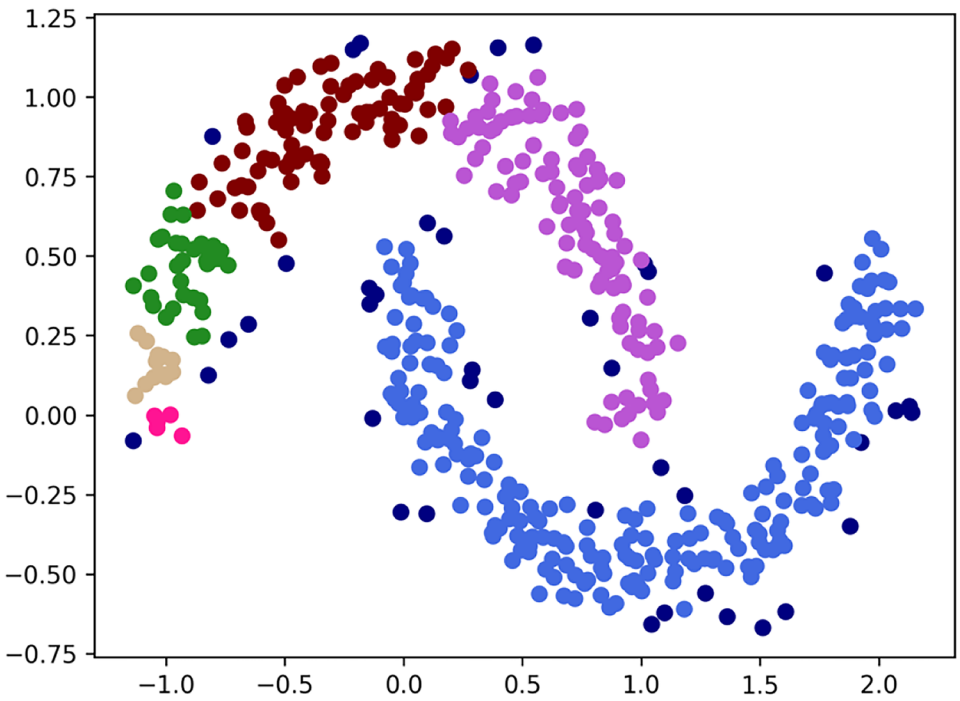

Fig. 5 DBSCAN Clustering

$$\text{AMI(X, Y)} = \frac{\text{MI} - \text{Expected MI}}{\text{Max(H(X), H(Y))} - \text{Expected MI}}$$

where *MI* is the mutual information, $H(X)$ and $H(Y)$ are the entropies of X and Y, and *Expected MI* is the expected mutual information. AMI ranges from 0 to 1, where a score of 1 indicates perfect agreement between two clustering. A score close to 0 suggests largely independent clustering or a result no better than random chance.

2.2.3 Homogeneity

Homogeneity is a clustering measure that compares the outcomes to a ground truth. It denotes that a cluster is homogenous if made up entirely of data points from a single class [26]. The formula for Homogeneity is

$$\text{Homogeneity} = 1 - \frac{\text{H}(\text{Y}_{\text{true}}|Y_{\text{predict}})}{\text{H}(Y_{true})}$$

where Y_{true} is the ground truth, $Y_{predict}$ is the predicted clusters. $H(Y_{true}|Y_{predict})$ is the conditional entropy of the ground truth given the cluster predictions. $H(Y_{true})$ is the entropy of the ground truth. Homogeneity is a metric that also varies between 0 and 1. A score of 1 means that each cluster contains only members of a single class, signifying perfect Homogeneity. A score of 0 indicates that the clusters are randomly assigned, lacking any homogeneity.

2.2.4 Completeness

Completeness is another clustering evaluation metric determining whether all data points in a given class are clustered. The clustering result is deemed complete when each class is contained inside a single cluster [26]. The formula for this measure is

$$\text{Completeness} = 1 - \frac{H(Y_{predict}|Y_{true})}{H(Y_{predict})}$$

where Y_{true} is the ground truth, $Y_{predict}$ is the predicted clusters. $H(Y_{predict}|Y_{true})$ is the conditional entropy of the cluster predictions given the ground truth. $H(Y_{predict})$ is the entropy of the cluster predictions. Completeness ranges from 0 to 1. A score of 1 is achieved when all class members are assigned to the same cluster, indicating complete capture of all classes within the clusters. A score of 0 would imply that the clustering assignments are completely scattered without capturing the essence of classes.

2.2.5 V-measure

The V-measure is the harmonic mean between Homogeneity and Completeness [26], and the formula is

$$V - measure = \frac{2 \times (Homogeneity \times Completeness)}{Homogeneity + Completeness}$$

The V-measure score lies between 0 and 1, where 1 stands for perfectly complete and homogeneous labelling. V-measure ranged from 0 to 1. A score of 1 represents perfect clustering with both complete capture of all classes within clusters and each cluster containing only members of a single class. A score of 0 would indicate that the clustering fails on both homogeneity and completeness grounds.

2.2.6 Silhouette coefficient

The silhouette coefficient is used in cluster analysis to assess clustering quality. It computes the distance between each data point in one cluster and the points in neighbouring clusters, measuring how well each data point fits into its allocated cluster [27]. The formula is

$$\text{Silhouette} = \frac{b_i - a_i}{max(a_i, b_i)}$$

where a_i is the average distance inside the cluster, and b_i is the average distance nearest other clusters. Silhouette Coefficient values range from -1 to 1. A score of 1 denotes that the clusters are well apart from each other and clearly distinguished. A score of 0 indicates overlapping clusters. A negative value suggests that data points might have been assigned to the wrong clusters. This metric gives a perspective on the distance and separation between the formed clusters.

3 Research dataset

Table 1 presents the datasets utilised in this study, outlining their respective attributes, including the number of features and data size. These datasets were sourced from the UCI Machine Learning Repository [28] and Kaggle [29]. This research uses the original data without preprocessing to ensure an unbiased comparison. We drop any entries with missing values.

4 Results

We employed the default parameters provided by *Sklearn* for training our unsupervised ML models [44]. Tables 2, 3, 4, 5, 6 and 7 show various models' ARI, AMI, Homogeneity, Completeness, V-measure and Silhouette metrics. They have been trained on our research datasets.

Table 2 illustrates the ARI of 15 datasets. Based on these 15 datasets, the best-performing method is the Divisive clustering for D12 (0.8510), followed by BGM for D5

Table 1 A summary of the dataset

ID	Dataset name	Number of features	Data size	Reference
D1	Heart Disease	13	303	Detrano et al. [30]
D2	Heart failure clinical records	13	299	Chicco and Jurman [31]
D3	Pima Indians Diabetes	8	768	Smith et al. [32]
D4	Heart Disease Prediction	13	270	Detrano et al. [30]
D5	Breast Cancer Wisconsin (Diagnostic)	5	569	Mangasarian et al. [33]
D6	Cervical Cancer	19	72	Machmud and Wijaya [34]
D7	Indian Liver Patient Dataset	10	583	Ramana et al. [35]
D8	Lung Cancer	15	309	Hong and Yang [36]
D9	Thyroid Disease	5	7200	Quinlan [37]
D10	Chronic Kidney Disease	25	400	Soundarapandian et al. [38]
D11	Indian Liver Patient Dataset	10	583	Lichman [39]
D12	Autism Screening Adult	21	704	Thabtah et al. [40]
D13	Prostate Cancer	10	100	Mahmood [41]
D14	Breast Cancer Coimbra	10	116	Patrício et al. [42]
D15	Cervical cancer	36	858	Fernandes et al. [43]

(0.6413). Performance varied widely across methods and datasets, underlining the necessity of testing multiple techniques. For AMI, the highest-performing models changed by dataset (Table 3). For instance, the best performance was observed with Divisive clustering in dataset D12 (0.7504), followed by BGM in dataset D5 (0.5337). For Homogeneity (Table 4), DBSCAN performs remarkably well on eight datasets, while the Divisive clustering performs best on D1 and D12. Regarding Completeness (Table 5), DBSCAN performs best on seven datasets, while BGM and Divisive clustering showed strong results on three datasets. DBSCAN has revealed the best performance on eight datasets with the V-measure metric. Evaluating the Silhouette score, the Agglomerative clustering dominated four datasets.

Additionally, Table 8 illustrates how often each model scored the highest in any given measure. DBSCAN showed the best performance most times (31), followed by BGM (18), Divisive clustering (15) and Agglomerative

Table 2 Adjusted Rand Index (ARI) comparison among unsupervised machine learning models

Dataset	Partitioning clustering		Model-based clustering		Hierarchical clustering		Density-based clustering
	Classic k-means	*Mini Batch k-means*	*Gaussian Mixture*	*Bayesian Gaussian Mixture*	*Agglomerative Clustering*	*Divisive Clustering*	*DBSCAN*
D1	0.0205	0.0205	0.0012	0.0012	0.0117	**0.0271**	-0.002
D2	0.0174	-0.0270	0.0471	**0.0499**	0.0044	-0.0021	-0.0002
D3	0.0744	0.0697	0.0009	0.0021	**0.1003**	-0.0002	0.0001
D4	0.0302	0.0248	0.0471	**0.0494**	0.0103	0.0490	0.0001
D5	0.4457	0.4358	0.6359	**0.6413**	0.2745	0.4661	0.0752
D6	0.2961	0.2403	0.3271	0.3271	**0.4292**	0.1385	0.0034
D7	-0.0209	-0.0709	-0.0700	-0.0240	-0.0410	-0.0529	**0.0001**
D8	-0.0213	-0.0131	0.0609	0.0136	-0.0595	-0.0211	**0.0612**
D9	0.1867	0.0514	-0.0465	0.1383	0.1134	0.0266	**0.2459**
D10	**0.3234**	0.0859	**0.3234**	**0.3234**	0.2135	0.0456	0.0000
D11	-0.0209	-0.0709	-0.0041	-0.0081	-0.0041	-0.0529	**0.0001**
D12	0.7172	0.6842	0.4318	0.3849	0.3742	**0.8510**	0.0000
D13	0.0557	0.0448	0.0521	0.0083	**0.0743**	0.0018	0.0009
D14	-0.0084	-0.0085	**0.0677**	0.0339	-0.0015	-0.0077	-0.0103
D15	0.0016	0.0049	0.0669	**0.0807**	0.0008	-0.0061	-0.0087

Table 3 Adjusted Mutual Information (AMI) comparison among unsupervised machine learning models

Dataset	Partitioning clustering		Model-based clustering		Hierarchical clustering		Density-based clustering
	Classic k-means	*Mini Batch k-means*	*Gaussian Mixture*	*Bayesian Gaussian Mixture*	*Agglomerative Clustering*	*Divisive Clustering*	*DBSCAN*
D1	0.0114	0.0114	-0.0023	-0.0023	0.0086	**0.0272**	-0.002
D2	-0.0004	-0.0025	0.0094	**0.0104**	-0.0040	-0.0022	0.0042
D3	0.0285	0.0258	0.0003	0.0002	**0.0447**	-0.0002	0.0001
D4	0.0172	0.0132	0.0094	0.0104	0.0108	**0.0317**	0.0010
D5	0.4188	0.411	0.5215	**0.5337**	0.3076	0.3759	0.1191
D6	0.2192	0.2092	0.2382	0.2382	**0.3247**	0.2448	0.0010
D7	0.016	0.0703	0.0517	**0.0956**	0.0002	0.0548	0.0041
D8	0.0044	-0.0014	0.0073	0.0010	0.0166	0.0017	**0.0460**
D9	0.0840	0.0674	0.0058	0.0676	0.0339	0.0214	**0.1452**
D10	**0.3013**	0.0423	**0.3013**	**0.3013**	0.2084	0.0303	0.0000
D11	0.0160	**0.0703**	-0.0013	0.0007	0.0002	0.0548	0.0041
D12	0.6621	0.6355	0.2939	0.3400	0.4144	**0.7504**	0.0000
D13	0.0406	0.0338	0.0446	0.0178	**0.0544**	0.0363	0.0540
D14	-0.0063	-0.0054	**0.1058**	0.0844	0.0085	-0.0029	0.0285
D15	0.0114	0.0114	-0.0023	-0.0023	0.0086	**0.0272**	-0.002

clustering (14). For individual performance metrics, the DBSCAN is the top performer regarding Homogeneity, Completeness and V-measure. BGM did well against the ARI and AMI metrics. Unsupervised MLs based on *k-mean*s (Classic and Mini Batch) showed the most minor performance.

The best model to choose will depend on the particulars of a specific application and the performance indicators that are most important to the stakeholders. From above, the DBSCAN model received the highest score among 15 datasets, demonstrating the best overall performance. However, DBSCAN is sensitive to parameter settings and may struggle

Table 4 Homogeneity comparison among unsupervised machine learning models

Dataset	Partitioning clustering		Model-based clustering		Hierarchical clustering		Density-based clustering
	Classic k-means	*Mini Batch k-means*	*Gaussian Mixture*	*Bayesian Gaussian Mixture*	*Agglomerative Clustering*	*Divisive Clustering*	*DBSCAN*
D1	0.0135	0.0135	0.0006	0.0006	0.0111	**0.0185**	0.0058
D2	0.0023	0.001	0.0111	0.0118	0.0002	0.0003	**0.8948**
D3	0.0267	0.0247	0.0013	0.0012	0.0423	0.0008	**0.9944**
D4	**0.0196**	0.0156	0.0113	0.0118	0.0134	0.0341	0.001
D5	0.3764	0.368	0.5094	**0.5186**	0.2514	0.3842	0.4253
D6	0.2413	0.2235	0.2599	0.2599	**0.3247**	0.2448	0.001
D7	0.0108	0.0638	0.043	0.0977	0.0019	0.0537	**0.9772**
D8	0.0103	0.0024	0.0131	0.0058	0.0248	0.0066	**0.0379**
D9	0.0878	0.0917	0.0065	0.0868	0.0321	0.0114	**0.7913**
D10	0.2354	0.0505	0.2354	0.2354	0.1509	0.0383	**1.0000**
D11	0.0108	0.0638	0.0001	0.0021	0.0019	0.0537	**0.9772**
D12	0.6911	0.6661	0.2972	0.3629	0.4445	**0.7599**	0.0000
D13	0.0420	0.0349	0.0442	0.0152	**0.0557**	0.0224	0.0546
D14	0.0001	0.0010	**0.1039**	0.0807	0.0143	0.0036	0.0260
D15	0.0002	0.0009	0.0357	0.0396	0.0001	0.0019	**0.0467**

Table 5 Completeness comparison among unsupervised machine learning models

Dataset	Partitioning clustering		Model-based clustering		Hierarchical clustering		Density-based clustering
	Classic k-means	*Mini Batch k-means*	*Gaussian Mixture*	*Bayesian Gaussian Mixture*	*Agglomerative Clustering*	*Divisive Clustering*	*DBSCAN*
D1	0.0135	0.0135	0.0006	0.0006	0.0111	**0.0185**	0.0058
D2	0.0027	0.0001	0.014	0.0156	0.0008	0.0003	**0.1026**
D3	0.0331	0.0294	0.0012	0.0011	0.0499	0.0008	**0.0972**
D4	0.0203	0.0163	0.014	0.0156	0.0135	0.0345	**0.1228**
D5	0.4741	0.4676	0.5356	**0.5511**	0.3997	0.3695	0.1166
D6	0.2160	0.2044	0.2348	0.2348	**0.3035**	0.2162	0.1447
D7	0.0688	0.082	0.0707	**0.0960**	0.0507	0.0590	0.0928
D8	0.0059	0.0013	0.0094	0.0032	0.0168	0.004	**0.1741**
D9	0.0810	0.0535	0.0058	0.0557	0.0366	**0.3468**	0.1087
D10	**0.4391**	0.0443	**0.4391**	**0.4391**	0.3744	0.0325	0.1156
D11	0.0688	0.0820	0.0001	0.0019	0.0507	0.0590	**0.0928**
D12	0.6361	0.6085	0.2925	0.3214	0.3895	0.7417	**1.0000**
D13	0.0420	0.0355	0.0480	0.0276	0.0557	**0.1466**	0.0918
D14	0.0001	0.0011	**0.1213**	0.1043	0.0161	0.0039	0.0713
D15	0.0001	0.0003	0.0154	0.0181	0.0000	0.0007	**0.0330**

with clusters of varying densities, whereas Divisive clustering does not rely on specific parameter settings and is better at handling clusters with different densities. Additionally, unlike Divisive clustering, DBSCAN can face challenges in high-dimensional spaces and in preserving the global structure of data. A critical observation is the wide range of variance in model performance across different datasets, although DBSAN dominated in most cases. This could be reflective of the innate differences in data distribution, noise, and feature relevance. This variation underscores the need for a sophisticated and discerning approach to choosing the appropriate unsupervised ML model, carefully weighing the dataset's unique properties alongside each model's inherent advantages.

Table 6 V-measure comparison among unsupervised machine learning models

Dataset	Partitioning clustering		Model-based clustering		Hierarchical clustering		Density-based clustering
	Classic k-means	*Mini Batch k-means*	*Gaussian Mixture*	*Bayesian Gaussian Mixture*	*Agglomerative Clustering*	*Divisive Clustering*	*DBSCAN*
D1	0.0138	0.0138	0.0007	0.0007	0.0111	**0.0186**	0.0111
D2	0.0025	0.0001	0.0124	0.0134	0.0004	0.0004	**0.1842**
D3	0.0295	0.0269	0.0013	0.0012	0.0458	0.0008	**0.1771**
D4	0.0199	0.0159	0.0124	0.0134	0.0134	0.0343	**0.2188**
D5	0.4196	0.4119	0.5222	**0.5344**	0.3087	0.3767	0.2008
D6	0.2280	0.2180	0.2467	0.2467	**0.3137**	0.2296	0.2527
D7	0.0190	0.0718	0.0534	0.0970	0.0370	0.0563	**0.1700**
D8	0.0075	0.0017	0.0109	0.0041	0.0200	0.0049	**0.0623**
D9	0.0843	0.0676	0.0062	0.0678	0.0342	0.0220	**0.1912**
D10	**0.3065**	0.0472	**0.3065**	**0.3065**	0.2151	0.0352	0.2073
D11	0.0186	0.0718	0.0001	0.0020	0.0037	0.0563	**0.1696**
D12	0.6625	0.6360	0.2948	0.3409	0.4152	**0.7507**	0.0000
D13	0.0420	0.0352	0.0460	0.0196	0.0557	0.0388	**0.0684**
D14	0.0001	0.0010	**0.1119**	0.0910	0.0152	0.0037	0.0381
D15	0.0001	0.0005	0.0215	**0.0249**	0.0001	0.0010	0.0387

Table 7 Silhouette comparison among unsupervised machine learning models

Dataset	Partitioning clustering		Model-based clustering		Hierarchical clustering		Density-based clustering
	Classic k-means	*Mini Batch k-means*	*Gaussian Mixture*	*Bayesian Gaussian Mixture*	*Agglomerative Clustering*	*Divisive Clustering*	*DBSCAN*
D1	0.0389	0.0389	0.0022	0.0022	**0.3431**	0.2821	0.0906
D2	0.5829	0.4561	0.0174	0.1902	**0.6789**	0.4576	0.1604
D3	**0.5688**	0.562	0.392	0.3452	0.5533	0.408	0.0176
D4	0.3804	**0.3847**	0.1740	0.1902	0.3193	0.367	0.0024
D5	0.6991	**0.7001**	0.6107	0.617	0.6827	0.5102	-0.0511
D6	0.2801	0.2670	**0.2821**	**0.2821**	0.2704	0.2118	0.0566
D7	0.8573	0.6628	0.5752	0.4723	**0.9256**	0.5954	0.0624
D8	0.5088	0.4762	0.0210	0.0145	0.4723	**0.5139**	0.4364
D9	0.5509	0.3393	-0.0972	0.2782	0.6002	**0.8426**	-0.5019
D10	0.5970	**0.8155**	0.6820	0.6820	0.7250	0.3565	-0.0655
D11	0.8573	0.6628	0.0656	0.0923	**0.9256**	0.5954	0.0624
D12	0.4632	0.3596	0.1514	0.1516	0.3567	0.4235	**0.5000**
D13	0.2169	0.2109	0.1644	**0.2225**	0.2094	-0.0034	0.1692
D14	0.2033	0.2172	0.1911	0.2101	0.2144	0.1755	**0.4014**
D15	0.4050	**0.4108**	0.1994	0.2255	0.3822	0.2977	-0.4949

Furthermore, the fact that DBSCAN consistently exhibits high Homogeneity, Completeness and V-measure implies that this model is particularly well-suited for datasets where classes are separated by density. This insight could prove invaluable for practitioners dealing with such data characteristics. Conversely, the strong performance of BGM in the ARI and AMI metrics across various datasets indicates its potential as a versatile model capable of capturing the structure of the data with a reasonable balance between cluster purity and recovery.

The Python code used to implement the unsupervised machine learning models considered in this study is available at https://github.com/haohuilu/unsupervisedml/.

5 Discussion

This research compares unsupervised machine learning models applied to eight different health-related datasets. The datasets were sourced from the UCI Machine Learning Repository and encompass a variety of health issues, including heart disease, diabetes, and multiple forms of cancer. These datasets exhibit diverse numbers of features and sizes. The primary goal of this study was to contrast the performance of these models across multiple measures without undertaking any data preparation, ensuring an unbiased comparison.

Table 8 Comparison of unsupervised machine learning models showing the number of times they presented the highest measurement

Models	Adjusted Rand Index	Adjust Mutual Information	Homogeneity	Completeness	V-measure	Silhouette	Total
Classic k-means	1	1	1	1	1	1	6
Mini Batch k-means	0	1	0	0	0	4	5
Gaussian Mixture	2	2	1	2	2	1	10
Bayesian Gaussian Mixture	5	4	1	3	3	2	18
Agglomerative Clustering	3	3	2	1	1	4	14
Divisive Clustering	2	4	2	3	2	2	15
DBSCAN	4	2	8	7	8	2	31

Table 9 Basic principles, Pros and Cons of different unsupervised machine learning models

Model	Principle	Pros	Cons
Classic k-means	Separates data into K number of clusters based on the features	• Easy to understand • Efficient computational cost [19]	• Need to define the number of clusters [19] • Sensitive to the initial choice of centroids [19]
Mini Batch k-means	It uses a randomly selected subset of the data in each iteration	• More efficient than classic k-means [19] • Reduces noise and variance through averaging [19]	• It might produce worse results due to randomness [19]
Gaussian Mixture	A probabilistic model assumes the data points are generated from a mixture of a finite number of Gaussian distributions	• Allow for elliptical clusters [51]	• May converge to a local optimum [51]
Bayesian Gaussian Mixture	The extension of the Gaussian mixture model. It incorporates a prior distribution of the parameters	• No need to define the number of clusters • Allow for elliptical clusters [52]	• Computationally expensive [52]
Agglomerative Clustering	A hierarchical clustering method starts with each data point in a separate cluster and merges them iteratively	• Can produce a hierarchy of clusters • No need to define the number of clusters [53]	• Not suitable for large datasets [53]
Divisive Clustering	A top-down approach to hierarchical clustering, where you start with the whole dataset as a single cluster and then recursively divide it into smaller clusters	• Can identify broad data subgroups [54]	• Computationally demanding • Sensitive to split decisions [54]
DBSCAN	A density-based clustering algorithm separates high-density regions from low-density regions and may identify arbitrarily formed clusters while excluding noise points	• Can discover arbitrarily shaped clusters • Less sensitive to initialisation values [55]	• Not suitable for data with varying densities [55]

The performance of the models differed based on the dataset and the evaluation metrics employed: ARI, AMI, Homogeneity, Completeness, V-measure, and Silhouette. Each metric provides a distinct insight into clustering quality. ARI and AMI measure the clustering against the ground truth. Homogeneity evaluates if each cluster solely comprises members of a single class, while Completeness assesses if all members of a specific class are grouped into the same cluster. V-measure is a harmonic mean of these two, and Silhouette gauges cluster separation and cohesion. DBSCAN's excellent performance in Homogeneity suggests its robustness in capturing dense clusters, but it also flags potential shortcomings in handling data with varying densities or noise levels.

Meanwhile, the BGM model ranks second in overall performance across 15 datasets. It shows notable strength in ARI scores for five datasets and AMI scores for four. However, its high computational requirements might limit its use in large datasets or those needing immediate analysis. BGM models excel at autonomously determining cluster numbers in complex datasets and resist overfitting by integrating prior distributions. However, their high computational demand and reduced effectiveness with non-Gaussian data or inappropriate priors are notable drawbacks [45]. The third best performing model, Divisive clustering, designed for sequencing datasets like life-course histories, leverages Classification and Regression Tree analysis principles, including tree pruning, to predict cluster counts. It excels in hierarchical, large datasets by uncovering complex relationships but can struggle with overlapping or non-hierarchical data, leading to less accurate clustering [46]. Moreover, the consistent performance of Agglomerative Clustering in terms of the Silhouette score suggests its potential utility in datasets where clear separation between clusters is present. Nevertheless, Mini Batch *k*-means offers an alternative that might better manage noise while sacrificing some degree of performance due to its inherent randomness.

The selection of models in unsupervised learning tasks is nuanced and contextual. For instance, while hierarchical methods like agglomerative and divisive clustering do not require the specification of the number of clusters, their computational intensity and potential to create unbalanced hierarchies must be considered, especially for large datasets. In the literature, the application of unsupervised machine learning models in disease prediction must be judicious, considering the unique characteristics of healthcare data. For example, *k*-means is known for its efficiency and has been widely used in medical data analysis for its simplicity [47]. However, its performance can be hindered by the requirement to specify the number of clusters and its sensitivity to outliers [48]. DBSCAN is favoured for its ability to find clusters of arbitrary shapes and sizes, which is often suitable for the complex patterns present in medical datasets [49]. Yet, its performance can degrade with varying density clusters. The Gaussian Mixture Model offers flexibility due to its probabilistic nature and can accommodate the varied distribution of medical data [20], though it can be computationally intensive, which may not be optimal for all applications. Experts agree that there is no one-size-fits-all model, and the choice should depend on the specific requirements of the data and the task at hand [50].

To sum up, while DBSCAN frequently emerged as the top performer, no singular model consistently outshone others across every dataset and metric. The choice of model should be influenced by the unique attributes of the dataset and the relevance of the evaluation metrics for the particular research or application context. This study serves as a valuable reference for future unsupervised learning endeavours in health-related fields. It also emphasises the importance of continued exploration in model selection and optimisation techniques. The basic principles, pros and cons of various unsupervised models are detailed in Table 9.

6 Conclusion

This study comprehensively compared unsupervised learning models within the realm of disease prediction. The diversity of data types within this field, from heart disease to prostate cancer, demands a flexible approach to model selection. Based on the evaluated performance metrics, two models emerged as particularly promising: DBSCAN and BGM. The former demonstrated robust performance in the Homogeneity and V-measure. Conversely, BGM excelled in the ARI and AMI metric. This underscores DBSCAN's aptitude for discerning densely populated clusters of similarity, even across heterogeneous datasets. Such findings highlight the potential prowess of these models in disease prediction. Their consistently high performance across diverse datasets indicates their capability to transcend the inherent challenges posed by the varied scales and ranges typical of medical data. Despite the intricate nature of medical datasets, these models succeeded in effectively clustering the data. The findings from this study serve not only as a testament to the capabilities of these models in transcending the challenges posed by medical datasets but also as a caveat to the user to be mindful of the models' limitations. Future research directions could delve into applying deep learning models for predicting disease risks, drawing from an even broader pool of medical datasets. One of the most noteworthy attributes of unsupervised machine learning models is their flexible architecture, which facilitates adaptability and continuous enhancement. It is important to note that unsupervised learning is an evolving domain, and ongoing advancements in algorithm efficiency, model robustness and interpretability

are expected to enhance further their application in disease prediction and other medical applications.

Author's contribution H.L.: Writing, Data analysis and Research design; S.U.: Research design, Writing, Conceptualisation and Supervision.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions This research received no specific grant from the public, commercial, or not-for-profit funding agencies.

Data availability statement This research used open-access public datasets for research investigation.

Declarations

Conflict of interests The authors declare that they do not have any conflict of interest.

References

1. Alloghani M, Al-Jumeily D, Mustafina J, Hussain A, Aljaaf AJ. A systematic review on supervised and unsupervised machine learning algorithms for data science. In: Supervised and unsupervised learning for data science. Springer; 2020. p. 3–21.
2. Chen H, Wu L, Chen J, Lu W, Ding J. A comparative study of automated legal text classification using random forests and deep learning. Inf Process Manage. 2022;59(2):102798.
3. Uddin S, Ong S, Lu H. Machine learning in project analytics: a data-driven framework and case study. Sci Rep. 2022;12(1):15252.
4. Jáñez-Martino F, Alaiz-Rodríguez R, González-Castro V, Fidalgo E, Alegre E. A review of spam email detection: analysis of spammer strategies and the dataset shift problem. Artif Intell Rev. 2023;56(2):1145–73.
5. Miklosik A, Evans N. Impact of big data and machine learning on digital transformation in marketing: A literature review. Ieee Access. 2020;8:101284–92.
6. Lu H, Uddin S. A disease network-based recommender system framework for predictive risk modelling of chronic diseases and their comorbidities. Appl Intell. 2022;52(9):10330–40.
7. Singh A, Thakur N, Sharma A. A review of supervised machine learning algorithms. In: 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom). Ieee; 2016.
8. Hahne F, Huber W, Gentleman R, Falcon S, Gentleman R, Carey V. Unsupervised machine learning. In: Bioconductor case studies. Springer; 2008. p. 137–57.
9. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak. 2019;19(1):281.
10. Katarya R, Meena SK. Machine learning techniques for heart disease prediction: a comparative study and analysis. Heal Technol. 2021;11:87–97.
11. Rahman AS, Shamrat FJM, Tasnim Z, Roy J, Hossain SA. A comparative study on liver disease prediction using supervised machine learning algorithms. Int J Sci Technol Res. 2019;8(11):419–22.
12. Shamrat FJM, Asaduzzaman M, Rahman AS, Tusher RTH, Tasnim Z. A comparative analysis of parkinson disease prediction using machine learning approaches. Int J Sci Technol Res. 2019;8(11):2576–80.
13. Sinha P, Sinha P. Comparative study of chronic kidney disease prediction using KNN and SVM. Int J Eng Res Technol. 2015;4(12):608–12.
14. Uddin S, Haque I, Lu H, Moni MA, Gide E. Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. Sci Rep. 2022;12(1):1–11.
15. Vats V, Zhang L, Chatterjee S, Ahmed S, Enziama E, Tepe K. A comparative analysis of unsupervised machine techniques for liver disease prediction. In: 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT). IEEE; 2018.
16. Antony L, Azam S, Ignatious E, Quadir R, Beeravolu AR, Jonkman M, De Boer F. A comprehensive unsupervised framework for chronic kidney disease prediction. IEEE Access. 2021;9:126481–501.
17. Alashwal H, El Halaby M, Crouse JJ, Abdalla A, Moustafa AA. The application of unsupervised clustering methods to Alzheimer's disease. Front Comput Neurosci. 2019;13:31.
18. Hartigan JA, Wong MA. Algorithm AS 136: A k-means clustering algorithm. J R Stat Soc Ser C Appl Stat. 1979;28(1):100–8.
19. Sculley D. Web-scale k-means clustering. In: Proceedings of the 19th international conference on World wide web. 2010.
20. Reynolds DA. Gaussian mixture models. In: Encyclopedia of biometrics, vol. 741. Springer; 2009. p. 659–63.
21. Roberts SJ, Husmeier D, Rezek I, Penny W. Bayesian approaches to Gaussian mixture modeling. IEEE Trans Pattern Anal Mach Intell. 1998;20(11):1133–42.
22. Han J, Pei J, Tong H. Data mining: concepts and techniques. Morgan kaufmann; 2022.
23. Ester M, Kriegel H-P, Sander J, Xu X. Density-based spatial clustering of applications with noise. Int. Conf. knowledge discovery and data mining; 1996.
24. Steinley D. Properties of the hubert-arable adjusted rand index. Psychol Methods. 2004;9(3):386.
25. Vinh NX, Epps J. Bailey, J2738784: Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. J Mach Learn Res. 2010;11:2837–54.
26. Rosenberg A, Hirschberg J. V-measure: a conditional entropy-based external cluster evaluation measure. Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL); 2007.
27. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65.
28. Asuncion A, Newman D, UCI machine learning repository. Irvine. USA: CA; 2007.
29. Kaggle. Kaggle. 2023. www.kaggle.com. Cited 16 June 2023.
30. Detrano R, Janosi A, Steinbrunn W, Pfisterer M, Schmid J-J, Sandhu S, Guppy KH, Lee S, Froelicher V. International application of a new probability algorithm for the diagnosis of coronary artery disease. Am J Cardiol. 1989;64(5):304–10.
31. Chicco D, Jurman G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med Inform Decis Mak. 2020;20(1):1–16.

32. Smith JW, Everhart JE, Dickson W, Knowler WC, Johannes RS. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In: Proceedings of the annual symposium on computer application in medical care. American Medical Informatics Association; 1988.
33. Mangasarian OL, Street WN, Wolberg WH. Breast cancer diagnosis and prognosis via linear programming. Oper Res. 1995;43(4):570–7.
34. Machmud R, Wijaya A. Behavior determinant based cervical cancer early detection with machine learning algorithm. Adv Sci Lett. 2016;22(10):3120–3.
35. Ramana BV, Babu MSP, Venkateswarlu N. A critical study of selected classification algorithms for liver disease diagnosis. Int J Database Manag Syst. 2011;3(2):101–14.
36. Hong Z-Q, Yang J-Y. Optimal discriminant plane for a small number of samples and design method of classifier on the plane. Pattern Recognit. 1991;24(4):317–24.
37. Quinlan R. Thyroid disease data set. 1987. https://archive.ics.uci.edu/ml/datasets/thyroid+disease. Accessed 3 Jul 2022.
38. Soundarapandian P, Rubini L, Eswaran P. Chronic kidney disease data set. Irvine, CA, USA: UCI Mach. Learn. Repository, School Inf. Comput. Sci., Univ. California; 2015.
39. Lichman M, UCI machine learning repository. Irvine. USA: CA; 2013.
40. Thabtah F, Kamalov F, Rajab K. A new computational intelligence approach to detect autistic features for autism screening. Int J Med Informatics. 2018;117:112–24.
41. Mahmood S. Prostate cancer. 2023. https://www.kaggle.com/datasets/sajidsaifi/prostate-cancer. Cited 15 Jun 2023.
42. Patrício M, Pereira J, Crisóstomo J, Matafome P, Gomes M, Seiça R, Caramelo F. Using Resistin, glucose, age and BMI to predict the presence of breast cancer. BMC Cancer. 2018;18(1):1–8.
43. Fernandes K, Cardoso JS, Fernandes J. Transfer learning with partial observability applied to cervical cancer screening. In: Pattern Recognition and Image Analysis: 8th Iberian Conference, IbPRIA 2017, Faro, Portugal, June 20–23, 2017, Proceedings 8. Springer; 2017.
44. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
45. Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. PLoS Genet. 2015;11(4):e1004969.
46. Chander S, Vijaya P. Unsupervised learning methods for data clustering. In: Artificial Intelligence in Data Mining. Elsevier; 2021. p. 41–64.
47. Jain AK. Data clustering: 50 years beyond K-means. Pattern Recogn Lett. 2010;31(8):651–66.
48. Celebi ME, Kingravi HA, Vela PA. A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Syst Appl. 2013;40(1):200–10.
49. Ester M, Kriegel H-P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. kdd; 1996.
50. Bouveyron C, Brunet-Saumard C. Model-based clustering of high-dimensional data: A review. Comput Stat Data Anal. 2014;71:52–78.
51. McLachlan GJ, Lee SX, Rathnayake SI. Finite mixture models. Annual review of statistics and its application. 2019;6:355–78.
52. Ghahramani Z, Beal M. Variational inference for Bayesian mixtures of factor analysers. In: Advances in neural information processing systems, vol. 12. NeurIPS; 1999.
53. Ackermann MR, Blömer J, Kuntze D, Sohler C. Analysis of agglomerative clustering. Algorithmica. 2014;69:184–215.
54. Sonagara D, Badheka S. Comparison of basic clustering algorithms. Int J Comput Sci Mob Comput. 2014;3(10):58–61.
55. Khan K, Rehman SU, Aziz K, Fong S, Sarasvady S. DBSCAN: past, present and future. In: The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014). IEEE; 2014.