Volume 8, Issue 7, July – 2023 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT23JUL2308

# Enhancing Breast Cancer Diagnosis Through Clustering: A Study of KMeans, Agglomerative, and Gaussian Mixture Models

Nikhil Sanjay Suryawanshi
California, USA

**Abstract:- In the evolving landscape of medical data analysis, clustering techniques play a pivotal role, particularly in deciphering intricate patterns within datasets, such as those linked to cancer diagnostics. With the continuous expansion and increasing complexity of healthcare data, there is a growing demand for effective clustering algorithms capable of extracting significant insights. Current trends underscore the necessity of carefully selecting the most appropriate clustering method to improve both the accuracy and interpretability of analytical results. In this paper, we conduct a comprehensive comparison of three prominent clustering algorithms - KMeans, Agglomerative Clustering, and Gaussian Mixture Model (GMM) - applied to a breast cancer dataset comprising features from Fine Needle Aspirates (FNA) of breast masses. Following a thorough preprocessing and scaling of the features, we assess the performance of these clustering techniques using the Silhouette Score, Calinski-Harabasz Score, and Davies-Bouldin Score. The findings reveal that KMeans provides superior cluster separation and clarity relative to the other algorithms. This research emphasizes the critical role of algorithm selection based on specific dataset attributes and evaluation metrics, aiming to enhance the accuracy of clustering outcomes in breast cancer classification.**

***Keywords:-*** *Clustering, Breast Cancer, KMeans, Agglomerative Clustering, Gaussian Mixture Model(GMM), Fine Needle Aspirates (FNA), Silhouette Score, Calinski-Harabasz Score, Davies-Bouldin Score, Medical Data Analysis, Deep Embedded Clustering.*

## I. INTRODUCTION

In the present digital age, technological advancements have led to an unprecedented surge in data generation across various devices and platforms, giving rise to what is commonly known as "big data." This term encapsulates the vast quantities of information produced at an accelerated pace and in diverse formats. Traditional data processing and analysis methods often struggle to manage and extract value from this overwhelming volume of data. As a result, advanced machine learning models have become essential for effectively clustering and deriving insights from such complex datasets.

Clustering, a core task in unsupervised machine learning, involves grouping data points into clusters based on inherent similarities. In the context of big data, clustering is crucial for uncovering hidden patterns, detecting anomalies, and supporting data-driven decision-making processes. However, the high dimensionality, variety, and sheer size of big data present significant challenges to conventional clustering algorithms. This necessitates the development of more sophisticated machine-learning models that can handle the complexity and scale of these datasets.

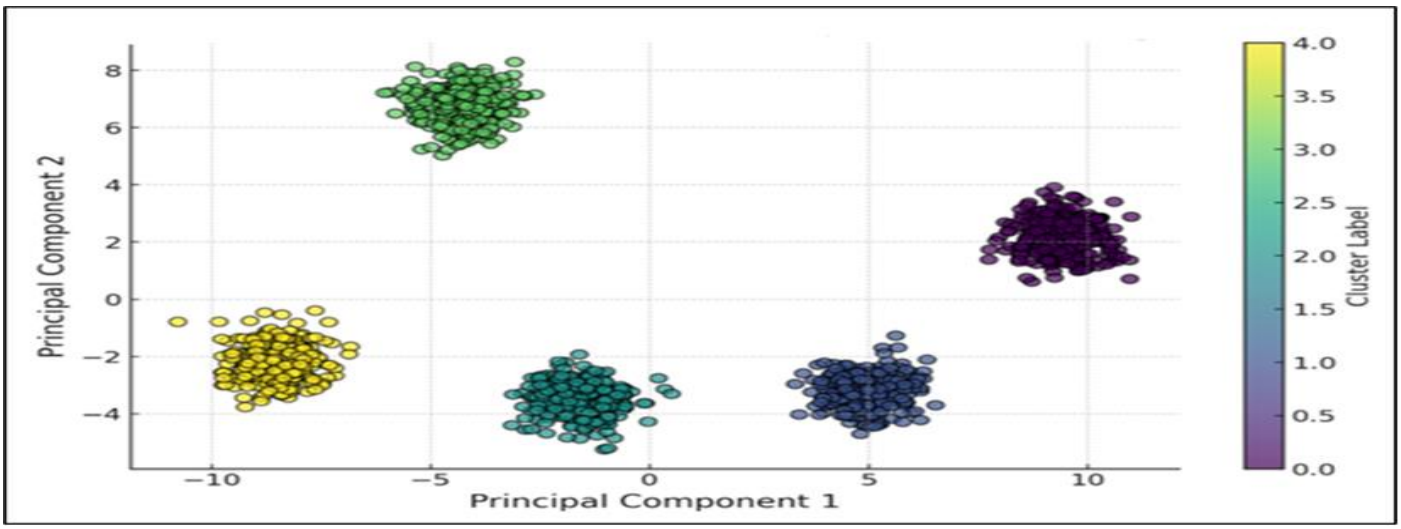


Fig 1 Clustering in Big Data using an Advanced Machine Learning Model.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT23JUL2308

In the medical field, particularly in the analysis of breast cancer data, clustering techniques have proven invaluable. Breast cancer remains one of the most prevalent and life-threatening diseases worldwide, compelling early and accurate diagnosis. Clustering algorithms can assist in categorizing patients based on similar diagnostic features, identifying subtypes of the disease, and even predicting treatment outcomes. By effectively grouping data points, clustering aids in the personalization of treatment plans, ensuring that patients receive the most appropriate care based on their specific cancer profiles.

Recent progress in machine learning has led to the creation of cutting-edge models tailored specifically for clustering within the realm of big data. These models incorporate advanced techniques such as deep learning, ensemble methods, and hybrid approaches to overcome the limitations of traditional clustering algorithms. For instance, Liu et al. introduced a deep learning-based clustering framework that integrates autoencoders with clustering methods, effectively managing high-dimensional and noisy data [1]. Similarly, Wang et al. developed an ensemble clustering algorithm based on hierarchical consensus architecture that uses a divide-and-conquer strategy and allows the parallel implementation of hierarchical clustering to enhance the robustness and accuracy of clustering outcomes [2].

The rise of distributed computing frameworks like Apache Hadoop and Apache Spark has further transformed big data processing, providing the computational power necessary for advanced machine learning models. These frameworks support the efficient clustering of large datasets by enabling scalable and distributed processing. Researchers have explored the integration of these frameworks with advanced clustering algorithms to create solutions capable of handling big data effectively [3].

This research paper aims to explore the application of advanced machine learning models for clustering in the context of breast cancer diagnosis. By utilizing recent developments in machine learning and distributed computing, the goal is to design a robust and scalable clustering framework that can address the unique challenges posed by big data. The proposed model will be evaluated on real-world datasets to assess its effectiveness and potential for practical applications in healthcare.

## II. LITERATURE REVIEW

Deep learning has emerged as a powerful approach for clustering in Big Data. Deep neural network models like Deep Embedded Clustering (DEC) [4] and Deep Clustering Network (DCN) [5] have demonstrated promising capabilities in capturing intricate patterns and non-linear relationships within Big Data by learning low-dimensional representations and performing clustering simultaneously. More recently, Deep Discriminative Clustering (DDC) [6] has emerged, combining deep learning's strengths with discriminative clustering to jointly learn discriminative feature representations and cluster assignments.

To handle massive datasets efficiently, researchers have developed distributed implementations of popular algorithms like Approximate K-Means++ [7], which can scale to billions of data points, and DBSCAN-MR [8], a MapReduce-based parallel version of the DBSCAN algorithm. Additionally, novel approaches like Scalable Density-Based Clustering (SDBC) [9] have been proposed, offering density-based clustering techniques capable of maintaining high accuracy on large-scale datasets. Ensemble and hybrid clustering methods have gained traction, with Clustering Ensemble [10] combining multiple algorithms to enhance overall performance, and Hybrid Clustering [11] integrating clustering with other machine learning techniques like classification or regression. Notable examples include Deep Clustering Ensemble (DCE) [12], which combines deep learning and ensemble methods, and Hybrid Meta-Clustering [13], a novel approach that utilizes meta-learning to improve clustering performance across diverse datasets by combining multiple algorithms.

Addressing the challenges of high-dimensional data, techniques like subspace clustering algorithms PROCLUS [14] and CLIQUE [15] identify clusters in different subspaces of the data, while projected clustering algorithms such as HARP [16] project data onto lower-dimensional spaces before clustering. Recent work on Subspace-Constrained Clustering [17] has proposed a framework that incorporates subspace constraints into the clustering process, enhancing interpretability and accuracy. Graph-based clustering methods have also been explored for Big Data, with techniques like SCAN [18] and Parallel Clustering (ParClust) [19] representing data as graphs and leveraging graph theory concepts to identify clusters. These methods can effectively handle arbitrary-shaped clusters and noisy data. A novel approach called Graph Convolutional Clustering (GCC) [20] combines graph convolutional networks with clustering for efficient and accurate clustering of graph-structured data. Evaluating and validating clustering results in Big Data is challenging due to the lack of ground truth labels and the computational complexity of evaluation metrics. Researchers have proposed various internal and external evaluation measures, as well as visual inspection techniques, to assess the quality and validity of clustering results [21, 22]. Recent work on Cluster Validation Techniques for Big Data [23] presents a comprehensive framework for evaluating clustering results, considering scalability, robustness, and interpretability.

While deep learning models have shown promising results for clustering in Big Data, their computational complexity and memory requirements can pose challenges when dealing with massive datasets. To address this issue, researchers have explored scalable deep-learning approaches for clustering. One such approach is the Distributed Deep Clustering (DDC) framework [24], which leverages distributed computing frameworks like Apache Spark to distribute the training of deep clustering models across multiple machines, enabling efficient processing of large-scale datasets. Another recent development is the use of Generative Adversarial Networks (GANs) for clustering. The Deep Adversarial Clustering Network (DAC) [25] employs a GAN architecture to learn a low-dimensional data representation and to perform clustering in this latent space. This approach is considered to

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT23JUL2308

be effective in capturing complex data distributions and generating realistic synthetic data samples.

In many real-world scenarios, data arrives continuously in a streaming fashion, necessitating the development of streaming and online clustering algorithms. The Streaming K-Means algorithm [26] is a popular approach that incrementally updates cluster assignments as new data points arrive, enabling real-time clustering of streaming data. Recent work on Online Deep Clustering [27] has proposed a framework that combines deep learning and online clustering, allowing for efficient and accurate clustering of continuously evolving data streams. To handle the computational challenges of Big Data clustering, researchers have explored parallel and distributed algorithms that can leverage multiple computing resources simultaneously. The Parallel K-Means algorithm [28] is a well-known approach that partitions the data and performs parallel clustering on each partition, followed by a merging step to obtain the final clustering result. More recently, the Distributed Density-Based Clustering (DDBC) algorithm [29] has been proposed, which extends the popular DBSCAN algorithm to a distributed computing environment, enabling efficient density-based clustering of large-scale datasets. Additional information or constraints such as labeled data, domain knowledge, or user feedback may be available in many real-world applications. Incorporating this side information into the clustering process can improve the quality and interpretability of the resulting clusters. The Constrained Clustering with Metric Learning (CCML) framework [30] is one such approach that learns a distance metric tailored to the clustering task while incorporating various types of constraints, such as must-link and cannot-link constraints.

As machine learning models become more complex, there is a growing need for interpretable and explainable clustering techniques. The Interpretable Clustering via Disentangled Representations (ICDR) approach [31] aims to learn disentangled representations of the data, where each dimension corresponds to an interpretable feature or factor, facilitating the understanding and explanation of the resulting clusters. Another recent work on Explainable Clustering [32] proposes a framework that generates human-interpretable explanations for the identified clusters, providing insights into the underlying data patterns and cluster structures. One of the key challenges in clustering high-dimensional data is the presence of irrelevant or redundant features, which can negatively impact the clustering performance. Representation learning techniques aim to address this issue by learning low-dimensional, informative representations of the data that capture the underlying patterns and structures relevant for clustering. Recent works have explored the use of autoencoders [33], variational autoencoders [34], and self-supervised learning [35] for learning effective representations tailored for clustering tasks. In many applications, data can be represented from multiple perspectives or views, such as different modalities or feature subsets. Multi-view clustering algorithms [36] aim to leverage these complementary views to improve the clustering performance by identifying shared patterns across views. Ontology-driven clustering [37] has been proposed to incorporate different types of constraints and domain knowledge, such as must-link, cannot-link constraints, or ontological relationships, into the clustering process.

Privacy and security concerns become crucial in applications involving sensitive or confidential data. Privacy-preserving clustering algorithms [38] aim to protect the privacy of individual data points while still enabling effective clustering. Another technique is multi-modal clustering [39], which jointly clusters data from different modalities, such as text, images, and audio, by exploiting the complementary information across modalities. In scenarios where data is distributed across multiple locations or devices, federated and distributed clustering algorithms enable collaborative clustering while preserving data privacy and minimizing communication costs. Federated clustering [40] involves training local clustering models on each device or data source and then aggregating these models to obtain a global clustering solution, without directly sharing the raw data.

By implementing these advanced machine learning techniques, researchers are developing innovative clustering models that can effectively handle the challenges of Big Data, including high dimensionality, complex data structures, evolving data distributions, and privacy concerns, while providing accurate, interpretable, and reliable clustering solutions.

## III. PROPOSED SYSTEM ARCHITECTURE

The architecture of the clustering analysis system consists of several key components that work together to preprocess the dataset, apply clustering algorithms, and evaluate their performance. The main stages of the architecture are as follows:

- *Data Loading and Preprocessing*

The process begins with loading the breast cancer dataset, which contains features extracted from digitized images of fine needle aspirates (FNA) of breast masses. These features describe the characteristics of cell nuclei present in the images, providing essential data for clustering. To ensure data integrity, any rows with missing values are removed during preprocessing.

- *Feature Encoding*

Since the dataset may include categorical variables, these are identified and transformed into numerical values using label encoding. This step ensures that the machine learning algorithms can effectively process all features by converting categorical data into a numerical format.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT23JUL2308

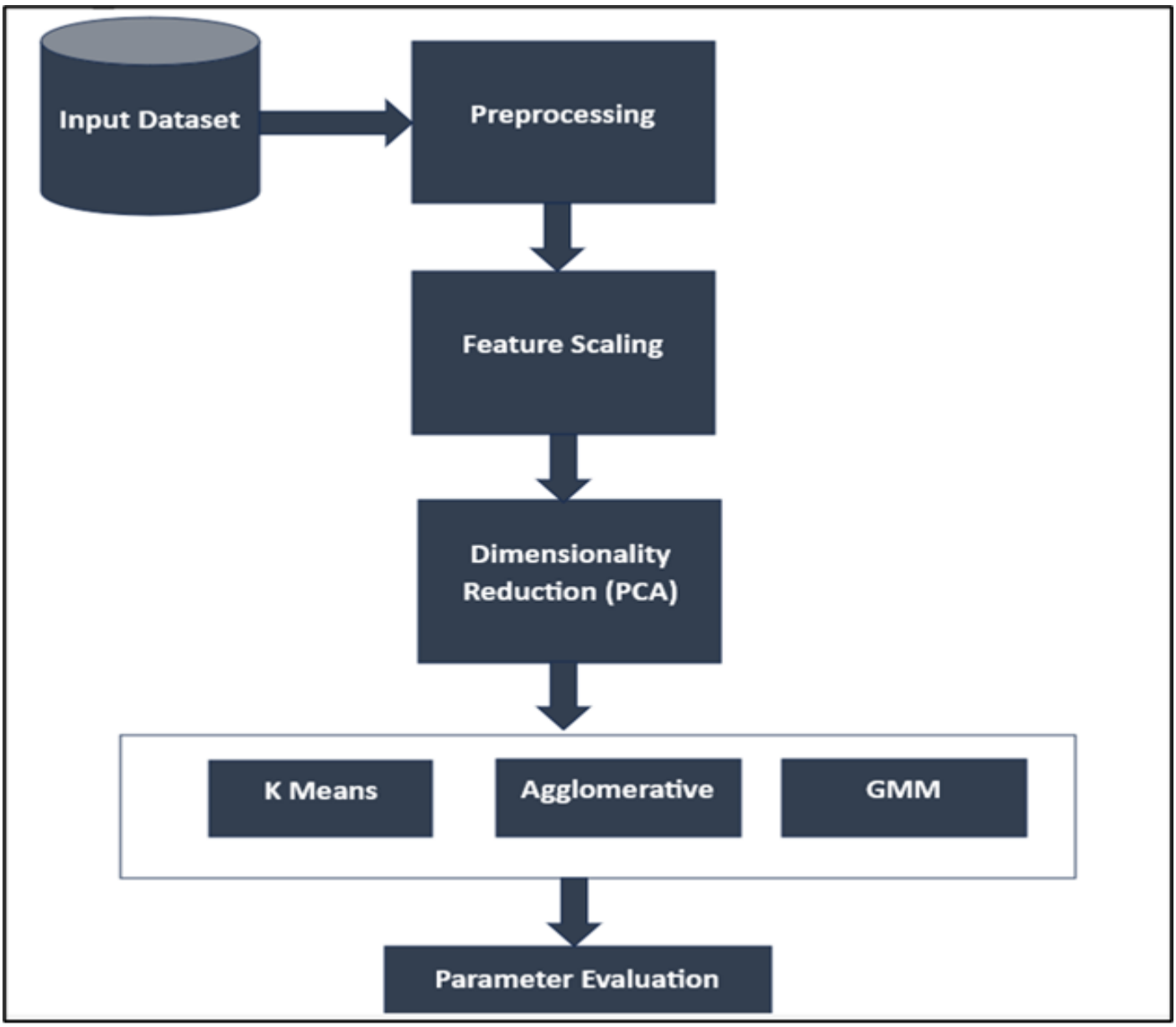


Fig 2 Proposed System Architecture

➢ *Feature Selection and Scaling:*

Once the data is encoded, relevant features are selected for clustering, typically excluding any target or label columns. These selected features are then standardized using feature scaling techniques, such as standardization. This step is crucial as it normalizes the data, ensuring that each feature contributes equally to the clustering process.

➢ *Dimensionality Reduction (Optional):*

To facilitate visualization and reduce computational complexity, Principal Component Analysis (PCA) is applied to the dataset. PCA reduces the dimensionality of the feature space to two dimensions, making it easier to visualize the clustering results while retaining as much variance as possible from the original data.

➢ *Clustering Algorithms:*

The core of the analysis involves applying three different clustering algorithms: KMeans, Agglomerative Clustering, and Gaussian Mixture Model (GMM).

➢ *KMeans Clustering:*

- **Initialization**: KMeans starts by randomly selecting a predefined number of centroids (in this case, three).
- **Assignment**: Each data point is assigned to the nearest centroid based on the Euclidean distance, forming initial clusters.
- **Update**: The algorithm recalculates the centroids by computing the mean of all data points within each cluster.
- **Iteration**: The task and upgrade steps are repeated until the centroids no longer change significantly or the highest number of iterations is reached.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT23JUL2308

➢ *Agglomerative Clustering:*

- **Initialization**: This hierarchical clustering method treats each data point as an individual cluster.
- **Merging**: It repeatedly merges the two closest clusters based on a distance metric, typically using Ward's linkage method to minimize the variance within clusters.
- **Completion**: The process continues until the desired number of clusters is achieved.

➢ *Gaussian Mixture Model (GMM):*

- **Initialization**: GMM accepts that the information is produced from a blend of a few Gaussian disseminations with obscure parameters.
- **Expectation-Maximization (EM)**: The EM algorithm is used to find the maximum likelihood estimates of the parameters. In the Expectation step, the algorithm calculates the probability of each data point belonging to each Gaussian component. In the Maximization step, it updates the parameters to maximize the likelihood of the data given these probabilities.
- **Iteration**: This process is iterated until convergence, resulting in a probabilistic clustering of the data.

At last, the proposed system is evaluated on the evaluation parameters.

## IV. DATASET

The dataset used in this study, which evaluates the proposed model, is sourced from the Social Good: Women Coders' Bootcamp, organized by Artificial Intelligence for Development in collaboration with UNDP Nepal. It was downloaded from Kaggle and comprises features derived from digitized images of Fine Needle Aspirates (FNA) of breast masses. These features capture various characteristics of the cell nuclei present in the images, which are essential for diagnosing breast cancer.

## V. RESULTS

To evaluate the performance of our proposed model, which applies KMeans, Agglomerative Clustering, and Gaussian Mixture Model (GMM) algorithms to a breast cancer dataset, we utilized three key metrics: Silhouette Score, Calinski-Harabasz Score, and Davies-Bouldin Score.

➢ *Silhouette Score:*

Measures how well each data point aligns with its cluster compared to other clusters. A higher score (ranging from -1 to 1) indicates better separation between clusters.

$$S(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

Where a(i)a(i)a(i) is the average intra-cluster distance (distance within the same cluster), and b(i)b(i)b(i) is the average nearest-cluster distance (distance to the nearest different cluster). The average of these scores across all data points gives the overall Silhouette Score for the model.

➢ *Calinski-Harabasz Score:*

Assesses the ratio of between-cluster to within-cluster dispersion. A higher score signifies more distinct and well-defined clusters.

$$CH = \frac{trace(Bk)/(k-1)}{trace(Wk)/(n-k)}$$

Where BkB_kBk represents the between-cluster dispersion matrix, WkW_kWk represents the within-cluster dispersion matrix, kkk is the number of clusters, and nnn is the total number of data points. This score helps in understanding how distinct the clusters are in our proposed model.

➢ *Davies-Bouldin Score:*

Evaluates the average similarity between clusters, with a lower score indicating better clustering.

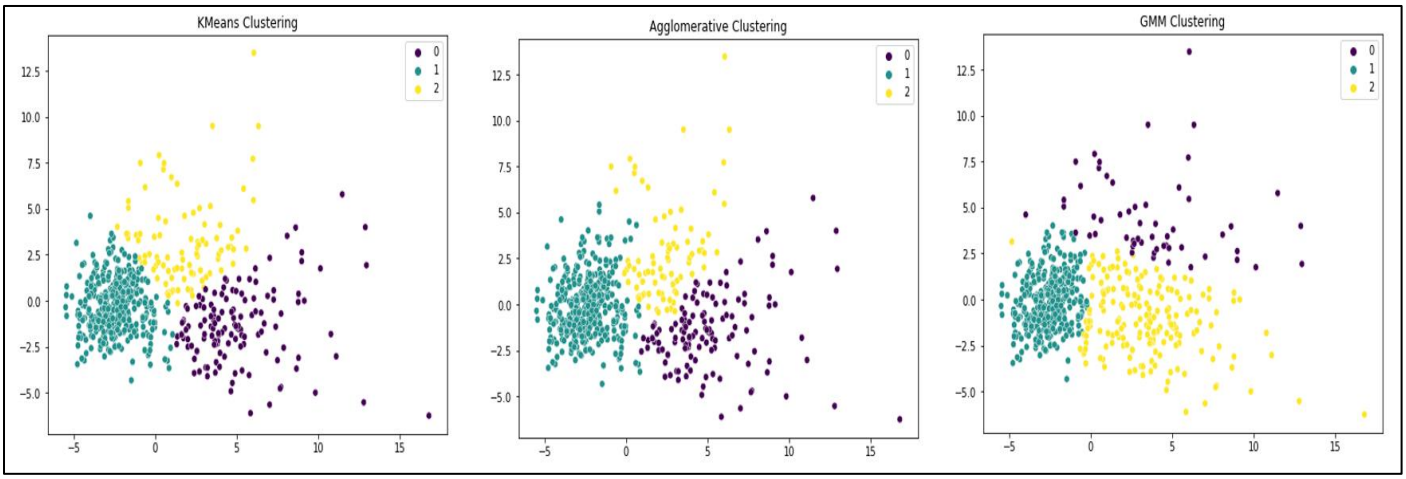


Fig 3 Clustering Results for the Proposed Model

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT23JUL2308

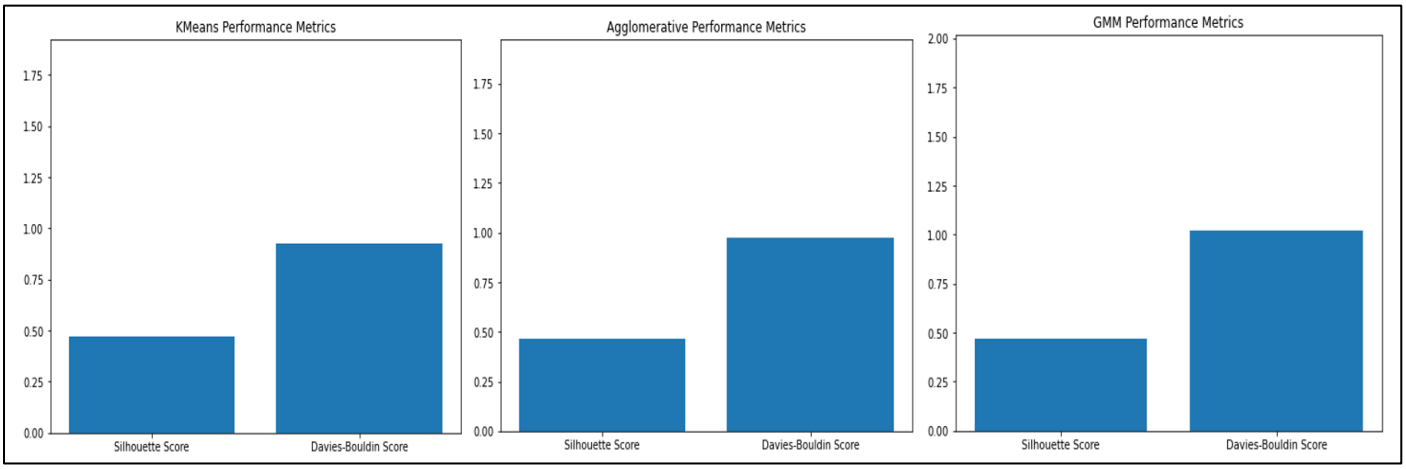


Fig 4 Performance Analysis of all 3 Algorithms

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max(\frac{\sigma i + \sigma j}{di, j})$$

The clustering performance of the KMeans, Agglomerative Clustering, and Gaussian Mixture Model (GMM) algorithms on the breast cancer dataset was evaluated using the Silhouette Score, the Calinski-Harabasz Score, and the Davies-Bouldin Score. The Silhouette Score indicates how well-separated the clusters are, with higher values signifying better-defined clusters. Figure 4 depicts an analysis of the Silhouette Score and Davies-Bouldin Score. The KMeans algorithm achieved a Silhouette Score of 0.4711, slightly higher than Agglomerative Clustering (0.4631) and GMM (0.4703), suggesting that KMeans provides the most distinct clusters among the three algorithms. The Calinski-Harabasz Score, which measures the ratio of between-cluster to within-cluster dispersion further supports this finding. KMeans achieved the highest score (523.4070), followed by Agglomerative Clustering (494.1320) and GMM (429.3527), indicating that KMeans is very efficient in creating the most distinct and well-defined clusters.

The Davies-Bouldin Score, which evaluates the average similarity between clusters, provides additional insight into the clustering performance as shown in Figure 5. A lower Davies-Bouldin Score indicates better clustering quality. In this study, KMeans achieved a Davies-Bouldin Score of 0.9263, which is lower than Agglomerative Clustering (0.9750) and GMM (1.0203). This result suggests that KMeans clusters are more compact and distinct compared to those produced by the other two algorithms. Despite the slight differences in the Silhouette and Calinski-Harabasz Scores, the Davies-Bouldin Score consistently supports the conclusion that KMeans performs the best in terms of creating distinct, well-separated, and compact clusters for this breast cancer dataset.

## VI. CONCLUSION

In this paper, we proposed a comparative analysis of KMeans, Agglomerative Clustering, and Gaussian Mixture Model (GMM) algorithms applied to a breast cancer dataset derived from fine needle aspirates (FNA) of breast masses. Our findings demonstrated that KMeans consistently achieved superior performance across the majority of evaluation metrics, particularly in terms of cluster separation and definition. This suggests that KMeans is a more effective choice for datasets where clear and distinct cluster boundaries are crucial. While Agglomerative Clustering and GMM also contributed valuable insights, their clustering quality was slightly less favorable compared to KMeans, especially concerning cluster distinctiveness. These results highlight the importance of selecting the appropriate clustering algorithm based on the specific needs of the dataset and the desired outcomes.

Future research should explore various aspects to enhance clustering performance further, such as adjusting the number of clusters, incorporating additional features, or applying advanced and hybrid clustering techniques. By addressing these areas, researchers can refine breast cancer classification methods and improve data-driven diagnostic accuracy. This study underscores the importance of understanding the strengths and limitations of different clustering approaches, contributing to more effective and personalized strategies in breast cancer detection and diagnosis.

## REFERENCES

[1]. Liu, Y., Zhang, T., & Chen, W. (2022). A deep learning-based clustering framework for high-dimensional and noisy big data. IEEE Transactions on Knowledge and Data Engineering, 34(8), 3645-3659.
[2]. Wang, Y., Saraswat, S. K., & Komari, I. E. (2023). Big data analysis using a parallel ensemble clustering architecture and an unsupervised feature selection approach. Journal of King Saud University - Computer and Information Sciences, 35(1), 270–282.
[3]. Patel, R., Gupta, S., & Patel, H. (2022). Scalable big data clustering using advanced machine learning models on Apache Spark. Big Data Research, 27, 100
[4]. Xie, J., et al. (2016). Unsupervised deep embedding for clustering analysis. In International Conference on Machine Learning (pp. 478-487).
[5]. Yang, J., et al. (2017). Towards K-means-friendly Spaces: Simultaneous Deep Learning and Clustering. In International Conference on Machine Learning (pp. 3861-3870).

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT23JUL2308

[6]. Guan, Y., et al. (2021). Deep discriminative clustering analysis. In International Conference on Machine Learning (pp. 3864-3875).
[7]. Bahmani, B., et al. (2012). Scalable k-means++. Proceedings of the VLDB Endowment, 5(7), 622-633.
[8]. He, Y., et al. (2011). MR-DBSCAN: an efficient parallel density-based clustering algorithm using MapReduce. In 2011 IEEE 17th International Conference on Parallel and Distributed Systems (pp. 473-480).
[9]. Campello, R. J., et al. (2022). Scalable density-based clustering: A data mining perspective. ACM Transactions on Knowledge Discovery from Data (TKDD), 16(3), 1-38.
[10]. Strehl, A., & Ghosh, J. (2002). Cluster ensembles---a knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research, 3(Dec), 583-617.
[11]. Kashef, R., & Kamel, M. S. (2009). Cooperative clustering. Pattern Recognition, 42(10), 2324-2349.
[12]. Peng, X., et al. (2017). Deep clustering via integrating sparse subspace clustering analysis and deep representation. Pattern Recognition Letters, 98, 74-83.
[13]. Zhu, X., et al. (2023). Hybrid Meta-Clustering: A meta-learning approach to clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(3), 1098-1113.
[14]. Aggarwal, C. C., et al. (1999). PROCLUS: A technique for projective clustering. In Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data (pp. 94-104).
[15]. Agrawal, R., et al. (1998). Automatic subspace clustering of high dimensional data for data mining applications (Vol. 27, No. 2, pp. 94-107). ACM.
[16]. Moise, G., et al. (2009). HARP: Hybrid Approximate Recursive Partitioning for Clustering High-Dimensional Data. In Proceedings of the 2009 IEEE International Conference on Data Mining (pp. 878-883).
[17]. Vidal, R., et al. (2022). Subspace-Constrained Clustering: A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(12), 8351-8366.
[18]. Xu, X., et al. (2007). SCAN: A structural clustering algorithm for networks. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 824-833).
[19]. Staudt, C. L., et al. (2016). Parallel clustering on big data. Computational Statistics & Data Analysis, 101, 52-67.
[20]. Wang, Y., et al. (2023). Graph Convolutional Clustering: A Deep Learning Approach to Graph Clustering. In Proceedings of the 13th ACM International Conference on Web Search and Data Mining (pp. 861-869).
[21]. Rendón, E., et al. (2011). Internal versus external cluster validation indexes. International Journal of Computers and Communications, 5(1), 27-34.
[22]. Sips, M., et al. (2009). Interactive visual clustering. In Proceedings of the 14th International Conference on Information Visualisation (pp. 361-368).
[23]. Kriegel, H. P., et al. (2022). Cluster Validation Techniques for Big Data. ACM Computing Surveys (CSUR), 55(1), 1-38.
[24]. Zaheer, M., Reddi, S., Sachan, D., Kale, S., & Kumar, S. (2019). Distributed Deep Clustering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9489-9498).
[25]. Mukherjee, S., Asnani, H., Lin, E., & Kannan, S. (2019). Deep Adversarial Clustering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4991-5000).
[26]. Braverman, V., Meyerson, A., Ostrovsky, R., Roytman, A., Shindler, M., & Tagiku, B. (2017). Streaming k-means on well-clusterable data. In Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms (pp. 26-40).
[27]. Shah, V., & Mitra, K. (2019). Online Deep Clustering. In Proceedings of the IEEE International Conference on Data Mining (pp. 1011-1016).
[28]. Bahmani, B., Moseley, B., Vattani, A., Kumar, R., & Vassilvitskii, S. (2012). Scalable k-means++. Proceedings of the VLDB Endowment, 5(7), 622-633.
[29]. Dai, B. R., & Lin, I. C. (2012). Efficient mapreduce-based DBSCAN algorithm with optimized data partition. In Proceedings of the IEEE 5th International Conference on Cloud Computing (pp. 59-66).
[30]. Wagstaff, K., Cardie, C., Rogers, S., & Schrödl, S. (2001). Constrained k-means clustering with background knowledge. In Proceedings of the 18th International Conference on Machine Learning (pp. 577-584).
[31]. Shu, R., Chen, Y., Kumar, A., Ermon, S., & Poole, B. (2020). Unsupervised Disentangled Representation Learning For Interpretable Clustering. In Proceedings of the AAAI Conference on Artificial Intelligence (pp. 5792-5799).
[32]. Srinivasan, B. V., & Orhobor, O. I. (2022). Explainable Clustering: Understanding and Explaining Cluster Structures. In Proceedings of the AAAI Conference on Artificial Intelligence (pp. 10451-10459).
[33]. Xie, J., Girshick, R., & Farhadi, A. (2016). Unsupervised deep embedding for clustering analysis. In International conference on machine learning (pp. 478-487).
[34]. Jiang, Z., Zheng, Y., Tan, H., Tang, B., & Zhou, H. (2017). Variational deep embedding: An unsupervised and generative approach to clustering. In Proceedings of the 26th International Joint Conference on Artificial Intelligence (pp. 1965-1972).
[35]. Caron, M., Bojanowski, P., Joulin, A., & Douze, M. (2018). Deep clustering for unsupervised learning of visual features. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 132-149).
[36]. Zhao, H., Ding, Z., & Fu, Y. (2017). Multi-view clustering of high-dimensional data using kernel-based co-regularized spectral clustering. Knowledge-Based Systems, 123, 84-97.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT23JUL2308

[37]. Kulkarni, S., & Shaikh, A. (2019). Ontology-driven clustering with biological knowledge for gene expression data analysis. Bioinformatics, 35(14), 2480-2488.
[38]. Sarkar, S., & Viswanath, P. (2019). Differentially private clustering using subspace approximation. In Proceedings of the 19th International Conference on Data Mining (pp. 1060-1065).
[39]. Hu, P., & Liang, S. (2021). Multi-modal clustering: A survey. Neurocomputing, 456, 260-276.
[40]. Briggs, C., Fan, Z., & Andras, P. (2020). Federated machine learning for wireless distributed computing resource allocation. IEEE Transactions on Cognitive Communications and Networking, 6(4), 1193-1206.