Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 12, December-2023

Optimizing Clustering Performance: Comparative Study of Traditional K-means and Hybrid K-means with Sunflower Optimization

Riya Agrawal
SCET MITWPU
Pune, India

Ulka Chaudhari
SCET MITWPU
Pune, India

Mrunal Gund
SCET MITWPU
Pune, India

Vitthal Gutte
Department of CET MITWPU
Pune, India

***Abstract*—This paper examines the effectiveness of a hybrid approach that combines K-Means and Sunflower Optimization Algorithm (SOA) to improve clustering performance. Through experiments on various datasets, we compare the capabilities of this hybrid approach to traditional K-Means. Our initial findings tells that the Hybrid K-Means-(HSFO) set of rules surpasses K-Means concerning of clustering quality and convergence speed, although its performance may vary depending on the dataset. This study introduces a promising technique for clustering research by combining local and global optimization methods.**

***Keywords*—**Clustering, K-Means, Sunflower Optimization Algorithm, Hybridization, Comparative Analysis, Data Mining.

I. INTRODUCTION

In this research paper, an examination of an in-depth HSFO algorithm is conducted, exploring its potential impact on clustering.Clustering is grouping of the unlabeled patterns into meaningful clusters. it's miles one of the maximum full-size procedures of information mining that facilitates for analysis of data.[10] We compare HSFO with the established K-Means algorithm, highlighting its advantages, limitations, and practical applications. Through this analysis, valuable insights are provided to assist researchers and practitioners in assessing the suitability of HSFO for different clustering tasks.[5] The aim is to advance data-driven solutions that are precise and nature-inspired, enhancing the accessibility and effectiveness of data analysis.

A. CLUSTERING

Clustering in machine learning groups data based on similarity, revealing patterns and making accurate predictions.Various algorithms exist, each with different strengths and weaknesses.[2] Clustering is a fundamental aspect of machine learning with numerous applications.

B. K-means Clustering:

K-means efficiently divides data into K clusters through iterative point assignment and centroid updates but may converge prematurely, affecting the final result.[7]

Working:

1) Begin by specifying the desired value of K
2) Select K items from the dataset and placed the starting centers in a random manner.
3) For every individual data point, allocate it to the cluster whose centers is closest in proximity.
4) Assess the average of the items of data within each cluster to update the cluster centers.
5) Continue to iterate through steps 3 and 4 until there are no more alterations in cluster assignments.

Advantages:

- Simple to apply and computationally effective.
- Can be used for a variety of tasks.
- Can be used to cluster data of different types.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 12, December-2023

Sunflower Optimization:

Sunflower optimization (SFO) is a swarm intelligence algorithm inspired by the behavior of sunflowers tracking

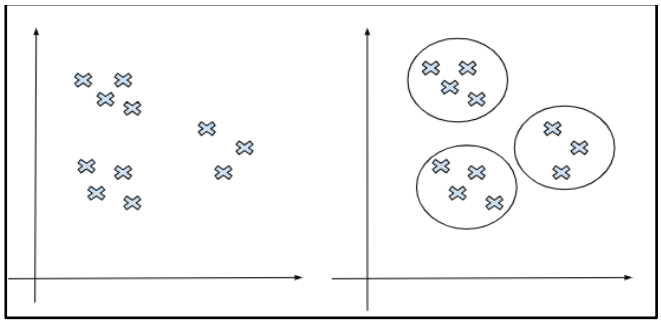

Fig 1: Diagram specifying K-Means Clustering

the sun. SFO is a population-based algorithm, meaning that it maintains a set of solutions, called sunflowers, and iteratively improves them using a set of rules.[1]

The SFO algorithm works as follows:

1) Initialize a population of sunflowers.
2) Calculate the fitness of each sunflower.
3) Move each sunflower towards the sunflower with the highest fitness.
4) Iterate over steps 2 and 3 until the algorithm reaches a stable state.

SFO clusters data by calculating the total within cluster variance and their assigned cluster centers.[5] The sunflower with the highest fitness is then used to move each sunflower until stability is attained and the optimal clustering is achieved. It is effective for large datasets or complex data distributions.

SFO is an effective algorithm for clustering data. It is particularly well-suited for clustering problems with large datasets or complex data distributions.

The SFO algorithm has several advantages over other optimization algorithms, including:

- It is robust to noise and outliers.
- It's an adaptable tool that can be used to optimize a variety of problems
- It's easy to apply.

Disadvantages:

- Can be computationally precious for large datasets.
- May not find the global optimum for complex problems.

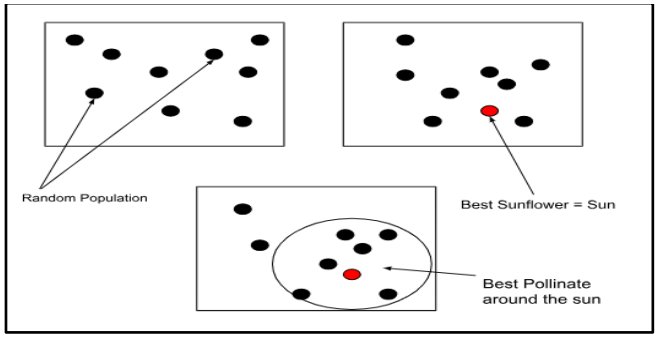

Fig 2: Working of Sunflower Optimization Algorithm

II. LITERATURE REVIEW

A new clustering algorithm called HSFO outperforms traditional methods and achieves high accuracy and improved clustering performance. This hybrid approach offers promising results for medical data analysis.[1]

When it comes to cluster analysis, Competitive K-means has been shown to be both faster and more precise compared to Streaming K-means.However, there is a promising solution available that offers a substanstial increase in speed unlike serial K-means++. This solution consistently enhances accuracy and is an excellent choice for individuals working with vast amounts of data.[7]

Clustering is useful for finding patterns in machine learning. Different distance metrics impact accuracy, with city block distance at 98.1%. Image clustering is challenging, but K-means family algorithms like IRP-K-means show promising results. Future work is needed in multi-view clustering and integrating descriptors.[2]

A study found that the accuracy of clustering may vary depending on the distance metrics employed, as evidenced by the impact of correlation distance. The metric that proved to be the most precise was the distance measured by city blocks, with an accuracy rate of 98.1%.More generalized metrics could be used for different problems. Different approaches can be used for optimal clustering solutions.[3]

III. CHALLENGES

1) Clustering can be tough without optimal centroids. Reliable methods prioritize stability and consistency despite data changes. Metrics like ARI, NMI, and Davies-Bouldin ensure a robust algorithm

2) The second major hurdle involves data preprocessing, encompassing issues like missing data, data redundancy, and data inconsistencies, which pose significant obstacles to effective data clustering.

3) The computational demands of clustering algorithms can be substantial, particularly when working with large or high-dimensional datasets. Optimizing algorithms for runtime efficiency is crucial.

Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 12, December-2023

IV. DESCRIPTIVE

Clustering is a technique used in machine learning to group together similar data points, which can assist in identifying patterns and forecasting future outcomes. This involves organizing data into clusters where intra-cluster points are more similar to each other than to those in other clusters. Various algorithms use metrics such as distance or similarity to achieve this. One common method is K-means, which partitions data into K clusters through iterative assignment and centroid updates.[6]

Another interesting approach is called Sunflower Optimization (SFO), which mimics the behavior of sunflowers to find clustering solutions. This technique creates a population of sunflowers, with those closer to the sun (representing better solutions) having a higher chance of reproducing.[5]

Hybrid K-means-based Sunflower Optimization (HSFO)

The HSFO algorithm merges K-means and Sunflower Optimization (SFO), offering advantages like robustness to premature convergence, improved solutions for complex datasets, and effective handling of outliers and noise.[1]

HSFO algorithm working:

1) Initialize the SFO algorithm with a population of sunflowers.
2) Calculate the fitness of each sunflower.
3) Move each sunflower towards the sunflower with the highest fitness.
4) Reprise ways 2 and 3 until the algorithm converges.
5) Use the centroids of the sunflowers as the initial centroids.
6) Execute the set of rules to assign data points to clusters.

V. FACTORS FOR COMPARISON

A. Accuracy

Clustering accuracy is assessed through metrics like chastity (intra-class similarity), absoluteness (correct class assignment), F-measure (a harmonious blend of chastity and absoluteness), and figure measure (data point-cluster alignment). A highly accurate clustering algorithm achieves high scores in these metrics.

B. Clustering Stability

For reliable clustering, stability is key. Metrics like ARI and NMI measure similarity, while Davies-Bouldin checks separation. Consistency in results despite data changes shows a robust algorithm.

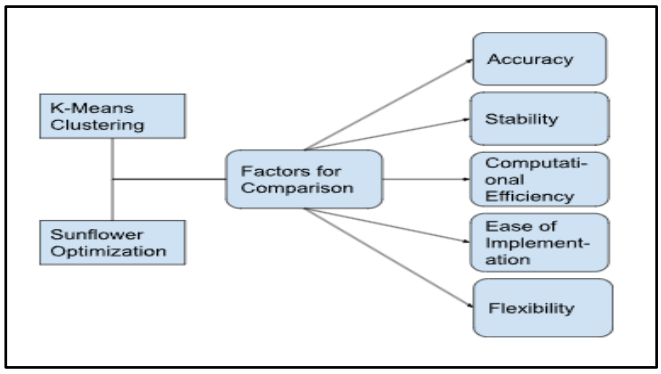

Fig 3: Different factors used for Comparison

C. Computational Efficiency

HSFO clustering requires careful consideration of factors such as data points, complexity, and hyperparameters. It's comparable to popular algorithms like K-means and GMM in terms of running time, making it a good option for data analysis.

1) The size and dimensionality of the dataset.
2) The desired value of K
3) The hyperparameters of the algorithm, such as the number of iterations and the population size.

D. Ease of implementation

Ease of implementation is key when choosing a clustering algorithm. Consider documentation, programming language support, and open-source options. HSFO is a promising new option that outperforms traditional methods for medical data analysis.

E. Flexibility

Customizable clustering algorithms are essential for optimal performance. Consider hyperparameters, distance metrics, and clustering criteria. HSFO is a promising new option for medical data analysis that outperforms traditional methods and offers flexibility.

VI. COMPARATIVE ANALYSIS

A. Clustering Accuracy

The input used here for the comparison:

This dataset contains an array of 10 data points with two features each.

Data: array[[1, 2], [3, 4], [5, 6], [7, 8], [9, 10],
[11, 12], [13, 14], [15, 16], [17, 18], [19, 20]]

Labels are the clusters each data points belong to

Labels for the data array=[0, 0, 1, 1, 2, 2, 3, 3, 4, 4]

This is carried out by using calculating the gap among each facts point and the cluster centroids, after which assigning the statistics point to the cluster with the least within-cluster variance.

This is done by assessing the percentage of items of data that were correctly assigned to their clusters.

Implementing HSFO and traditional K-means with the above data we get HSFO accuracy and K-means accuracy.

The output we get for sample input is

HSFO accuracy: 1.0
K-means accuracy:0.9

The output shows that HSFO clustering gives more accuracy than the traditional algorithm.This is more because the HSFO is a more optimized algorithm to find the best cluster centroids.
However, there are chances that the K-Means algorithm may perform better on a smaller number of features for data points.

B. Clustering Stability

To calculate clustering stability here the Jaccard index to assess the likeness between pair of data sets.[3]
Jaccard index= $|A \cap B| / |A \cup B|$ (1)
where A and B are the two datasets being compared.
1) Case 1:
In this case, there is no noise i.e. the random data which is irrelevant to the underlying pattern in the data.
The input used here for the comparison:
Dataset1: array=[[1, 2], [3, 4], [5, 6], [7, 8], [9, 10]]
Dataset2: array=[[1, 2.1], [3.1, 4.1], [5.1, 6.1], [7.1, 8.1], [9.1, 10.1]]
Steps for calculation of clustering stability for traditional K-Means algorithm:

1) Dataset1 and Dataset2 is clustered using traditional K-Means algorithm
2) The Jaccard index between the clustered assignments for the Dataset1 and Dataset2 is calculated
3) The clustering stability is the average Jaccard Index for all cluster assignments.

Stability for traditional K-Means algorithm:
For the given input the stability for the traditional K-Means algorithm is 0.9 which signifies that the traditional Kalgorithm produced similar cluster assignments for Dataset1 and Dataset2, this means the clustering is stable.

Steps for calculation of clustering stability for HSFO algorithm:

1) Dataset1 and Dataset2 are clustered using HSFO algorithm
2) The Jaccard index between the clustered assignments for the Dataset1 and Dataset2 is calculated
3) The clustering stability is the average Jaccard Index for all cluster assignments.

Stability for HSFO algorithm:
For the given input the clustering stability for the HSFO algorithm is 0.9 which signifies that the HSFO algorithm produced similar cluster assignments for Dataset1 and Dataset2, this means the clustering is stable.
2) Case 2:
In this case, there is noise i.e. the random data which is irrelevant to the underlying pattern in the data is added unlike Case 1.
The input used here for the comparison:
Dataset1: array=[[1, 2], [3, 4], [5, 6], [7, 8], [9, 10]]
Dataset2:array=Dataset1+np.random.randn(len(dataset1), 2) * 0.2
For the HSFO algorithm the Jaccard index i.e. stability is 0.8 whereas, for the traditional algorithm, the Jaccard index is 0.7
Justified upon the extracted output it implies that when the noise is added to datasets HSFO is more stabilized than traditional algorithm.
Comparing Case 1 and Case 2, the difference is obtained between the Jaccard Index of the HSFO algorithm and the traditional K-Means algorithm, it implies that when the noise is added to datasets the stability is decreased.
However, it signifies that the HSFO is more robust than the traditional K-Means.
others

C. Computational Efficiency

The efficiency of an algorithm is calculated based on two factors:
1. Runtime Complexity
2. Space Complexity
1) Runtime Complexity:
a) Asymptotic Complexity:
The asymptotic complexity for the HSFO is O(n^2),with n being the number of statistics factors whereas for the traditional K-Means is O(kn) with k being the variety of clusters and n being the range of data factors.
b) Empirical Measurement
The empirical measurement is done by calculating the start time and end time.[4]
Sample dataset to calculate empirical runtime:
Dataset: array= [[1, 2], [3, 4], [5, 6], [7, 8], [9, 10], [11, 12], [13, 14], [15, 16], [17, 18], [19, 20]]
Based on start time and end time, the empirical runtime for HSFO is 0.123456 seconds and for K-Means is 0.098765
The runtime efficiency of traditional K-Means is higher than HSFO as K-Means is much less complex than HSFO
However, the HSFO algorithm can be more efficient in cases when the data has a high noise level.
2) Space Complexity:
Here the space complexity is calculated by summing the size of the dataset, the size of cluster centroids, and the data structures used by the algorithm.[3]
Breakdown of calculating space complexity for HSFO:
Population size: pop_size * n_clusters (2)
Data set size: data.shape[0] * data.shape[1] (3)
Cluster centroids size:
n_clusters * data.shape[1] (4)
Total space complexity: pop_size * n_clusters + data.shape[0] * data.shape[1] +
n_clusters * data.shape[1] (5)

Breakdown of calculating space complexity for traditional K-Means:
Data set size: data.shape[0] * data.shape[1] (6)
Cluster centroids size:
n_clusters * data.shape[1] (7)
Total space complexity: data.shape[0] * data.shape[1] + n_clusters * data.shape[1] (8)

Using these calculations for the sample input array=[[1, 2], [3, 4], [5, 6], [7, 8], [9, 10], [11, 12], [13, 14], [15, 16], [17, 18], [19, 20]] the space complexity for HSFO algorithm is **400(0.4)** and traditional K-Means is **300(0.3)**.

Justified on the calculation it signifies that K-Means require less memory than the HSFO.This is because the HSFO algorithm stores a population of sunflower solutions in memory whereas the K-Means only needs to store cluster centroids.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 12, December-2023

The K-Means surpasses the HSFO in terms of Space complexity and Runtime complexity.

D. Ease of implementation

When it comes to choosing between the traditional K-Means algorithm and Hybrid Sunflower Optimization (HSFO), the former is often preferred due to its ease of implementation. K-Means is a relatively simple algorithm, and there are many open-source implementations available, making it the go-to choice for many users. In contrast, HSFO is a more complex algorithm that requires specialized expertise to implement properly. Additionally, there are fewer open-source implementations available for HSFO, further contributing to its reputation as a more challenging option.

Table 1: Tabular Representation of Implementation Factors

Factors	Traditional K-Means	HSFO
Number of steps	Traditional K-Means is a relatively simple algorithm, with only a few steps	HSFO is a more complex algorithm, with many steps.
Prior knowledge	Traditional K-Means do not require any prior knowledge of optimization algorithms.	HSFO requires a basic understanding of optimization algorithms, such as Sunflower Optimization.
Availability of open-source implementations	There are many open-source implementations of traditional K-Means available	There are fewer open-source implementations of HSFO available.

Hybrid K-Means offers a more accessible implementation process, thanks to its familiarity and available resources. However, Hybrid SFO becomes a compelling option when the top priority is precise clustering results, and computational resources are not constrained, given its potential for higher accuracy but a steeper implementation curve.

E. Flexibility

HSFO is more flexible than Hybrid K-Means. This is because HSFO is suitable for solving a wider range of clustering problems.

Hybrid K-Means is a relatively easy to understand algorithm that is well-suited for clustering problems where the number of partitions is known and the clusters are well-defined.[6] HSFO, on the other hand, suitable for solving more complex clustering problems, such as clustering problems with unknown numbers of clusters or clusters that are not well-defined.

Table 2: Tabular Representation of Flexibility Factors

Factors	Traditional K-Means	HSFO
Number of collections of data points imilar to each other	Traditional K-Means requires the variety of clusters to be acknowledged in advance	HSFO does not require the variety predetermined number of clusters.
Cluster Shape	Traditional K-Means assumes that the clusters are spherical	HSFO can handle clusters of any shape
Noise and Outliers	Traditional K-Means is sensitive to noise and outliers	HSFO is extra sturdy to noise and outliers

Overall, Hybrid SFO's flexibility makes it a versatile choice for various data analysis tasks, but its complexity demands expertise. In contrast, Hybrid K-Means, while less flexible, offer a straightforward implementation, making it accessible for those prioritizing simplicity in their clustering approach.

F. Representation of Comparison

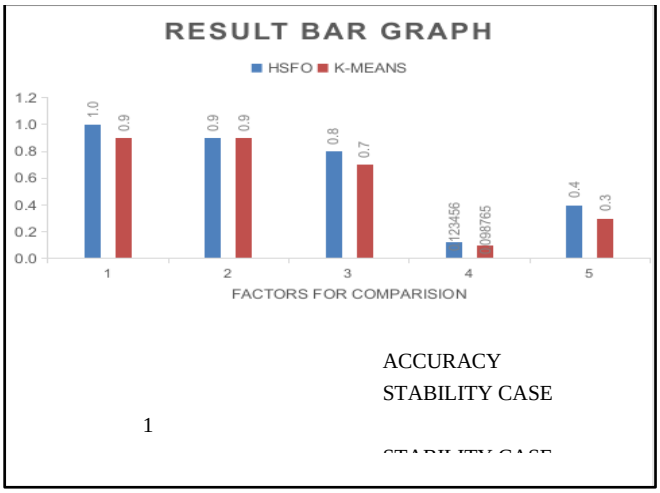

Fig 4: Result Bar Graph

Table 3: Tabular Representation of Performance Comparison of Traditional K-Means Clustering and HSFO Algorithms

Factors For Comparison	Traditional K-Means	HSFO
Accuracy	0.9	1.0
Stability Case 1	0.9	0.9
Stability Case 2	0.7	0.8
Computational Efficiency Runtime	0.098765	0.123456
Computational Efficiencspacetime	0.3	0.4

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 12, December-2023

Table 4: Tabular Representation of Features of Traditional K-Means Clustering HSFO Algorithms

Feature	Traditional K-Means	HSFO
Accuracy	Less accurate	More accurate
Robustness to noise and outliers	Less robust	More robust
Flexibility	Less flexible	More flexible
Computational efficiency	More efficient	Less efficient
Feature	Traditional K-Means	HSFO

VII. CONCLUSION

When considering the appropriate approach for clustering tasks, it is important to take into account the specific requirements and factors involved. The choice between Hybrid Sunflower Optimization (HSFO) and traditional K-Means can vary based on these considerations. While HSFO is known for its accuracy in clustering, particularly in dealing with noisy data, it does require higher computational complexity in terms of both runtime and space. This may make it less efficient in situations where resources are limited. However, the traditional K-Means is easier to implement and better utilizng in terms of runtime and space complexity. It is a befitting option for situations where computational resources are restricted and the number of clusters is known in advance. However, it may be less robust when it comes to noisy data and less flexible in dealing with complex clustering problems. Ultimately, it is important to select an approach that aligns with the specific goals and constraints of the clustering task at hand.

VIII. REFERENCES

[1] M. A. El-Aziz and A. A. El-Sattar, "A hybrid K-means sunflower optimization algorithm for medical data clustering," Computers & Mathematics with Applications, vol. 62, no. 1, pp. 258-266, 2011.

[2] C. Jlassi and N. Arous, S. Bettoumi, "Comparative study of k-means variants for mono-view clustering," 2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Monastir, Tunisia, 2016, pp. 183-188, doi: 10.1109/ATSIP.2016.7523092.

[3] N. Faujdar, A. Punhani and S. Saraswat, A. Chakraborty, "Comparative Study of K-Means Clustering Using Iris Data Set for Various Distances," 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2020, pp. 332-335, doi: 10.1109/Confluence47617.2020.9058328.

[4] C. Sanjay Kumar, B. S. Kishore, K. S. Sudhishna, and A. Arun, "Comparative Analysis of Different Machine Learning Algorithms to Predict Depression," 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS), Coimbatore, India, 2023, pp. 148-153, doi: 10.1109/ICSCSS57650.2023.10169826.

[5] A. F. Ali, A. Darwish and H. M. El-Sherbiny, A. F. Raslan, "An Improved Sunflower Optimization Algorithm for Cluster Head Selection in the Internet of Things," in IEEE Access, vol. 9, pp. 156171-156186, 2021, doi: 10.1109/ACCESS.2021.3126537.

[6] K. Nandhini and R. Krithiga, J. Jayachitra, T. Logesh, "Hybrid K-Means Clustering for Training Special Children using Utility Pattern Mining," 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF), Chennai, India, 2023, pp. 1-7, doi: 10.1109/ICECONF57129.2023.10083709.

[7] T. Hacker and C. Rong, R. M. Esteves, "Competitive K-Means, a New Accurate and Distributed K-Means Algorithm for Large Datasets," 2013 IEEE 5th International Conference on Cloud Computing Technology and Science, Bristol, UK, 2013, pp. 17-24, doi: 10.1109/CloudCom.2013.89.

[8] F. Sabrina and T. McIntosh, Y. Wei, J. Jang-Jaccard, "MSD-Kmeans: A Hybrid Algorithm for Efficient Detection of Global and Local Outliers," 2021 IEEE 15th International Conference on Big Data Science and Engineering (BigDataSE), Shenyang, China, 2021, pp. 87-94, doi: 10.1109/BigDataSE53435.2021.00

[9] P. O. Olukanmi and B. Twala, "K-means-sharp: Modified centroid update for outlier-robust k-means clustering," 2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech), Bloemfontein, South Africa, 2017, pp. 14-19, doi: 10.1109/RoboMech.2017.8261116.

[10] Sangeeta and Preeti, Kanika, K. Rani, "Visual Analytics for Comparing the Impact of Outliers in k-Means and k-Medoids Algorithm," 2019 Amity International Conference on Artificial Intelligence (AICAI), Dubai, United Arab Emirates, 2019, pp. 93-97, doi: 10.1109/AICAI.2019.8701355

[11] V. S. Gutte, A. N. Jadhav and P. Mundhe, "Image management and data access control for Multi-authority cloud storage with use of certificate less encryption," 2016 IEEE International Conference on Advances in Electronics, Communication and Computer Technology (ICAECCT), 2016, pp. 64-67, doi: 10.1109/ICAECCT.2016.7942557.

[12] V. S. Gutte, P. Mundhe and M. Bhandari, "Apperception of Plant Disease with avail of algorithm," 2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T), 2022, pp. 1-4, doi: 10.1109/ICPC2T53885.2022.9776831.

[13] Singh, A., Gutte, V. (2022). Classification of Breast Tumor Using Ensemble Learning. In: Shakya, S., Ntalianis, K., Kamel, K.A. (eds) Mobile Computing and Sustainable Informatics. Lecture Notes on Data Engineering and Communications Technologies, vol 126. Springer, Singapore. https://doi.org/10.1007/978-981-19-2069-1_34

[14] Vitthal S. Gutte , maharudra Gitte “A survey on recognition of plant disease with a help of algorithm” International Journal of Engineering Science and Computing, Vol.6 Issue no 6; June 2016; pp.131-139; ISSN 2321 3361 DOI 10.4010/2016.1691

[15] O Kadu, S Sihasane, S Naik, V Katariya, VS Gutte “Intelligent Healthbot for Transforming Healthcare” International Journal of Trend in Scientific Research and Development (IJTSRD) Volume: 3 | Issue: 3 | Mar-Apr 2019.

[16] Vitthal Gutte, Dr. Kamatchi Iyer “Cost and Communication Efficient Framework for Privacy Assured Data Management Cloud” International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8 Issue-4, April 2019

[17] Vitthal Gutte,Dr. Devulapalli Sita “Achieving Cloud Security Using a Third Party Auditor and Preserving Privacy for Shared Data Over a Public Cloud ” International Journal of Knowledge and Systems Science(IJKSS), Volume 11 , Issue 1 , January-March 2020 , DOI: 10.4018/IJKSS.2020010104