A Novel Approach For Predicting Drug Response Similarity Using Machine Learning

M Supriya Menon[1], P Raja Rajeswari[2]

[1]Research Scholar, Department of computer science and Engineering
Koneru Lakshmaiah Education Foundation, Vaddesvaram, AP, India
[2]Professor, Department of computer science and Engineering
Koneru Lakshmaiah Education Foundation Vaddesvaram, AP, India

Email: [1]supriyamenon05@gmail.com, [2]rajilikhitha@gmail.com

Abstract: Medical domain is revolutionized in terms of Diseases, Diagnosis, and Treatment Prediction thereby undergoing immense pressure due to the high dimensionality of numerable multivariate attributes, suppressing the quality of the analysis. Many techniques like Clustering and Classification have ruled over despite, rendering few hairline gaps towards attaining maximum efficiency. Our Machine Learning-based approach heads towards filling these gaps by adopting advanced K-Means in anticipating Drug likelihood in core attributes of Patients. The proposed Methodology focuses on determining Drug Response similarity by enhanced clustering technique concerning sensitive attributes of Patients. We successfully demonstrated its performance on the UCI Patient dataset reflecting enhanced results concerning Quality Parameters.

Keywords: k-Means; Euclidean distance; KNN algorithm, Drug responses; clustering;

1. INTRODUCTION

Uncovered sensitive and relevant information that cannot be analyzed at first sight might be examined and consolidated for better treatment using data mining techniques like Associations mining [13], correlations, Classification, and Clustering. These techniques play a vital role in the medical field with vast amounts of patient data being poured into repositories. Medical data is prone to discrepancies like redundancy, complexity, and privacy [3] that can be handled with ease using mining strategies owing path to Medical Data Mining. The mined medical data helps doctors in taking crucial decisions in crucial situations. Multi variant attributes inpatient medical records provoke us towards the applicability of well-known mining mechanisms [17]like classification and Clustering.

Classification, one of the major functionality of data mining is a supervised technique performed when class labels are available. It classifies data records of patients based on attributes like disease, symptoms, age, treatment, and further. Unlike in scenarios where class labels or target classes are not directly defined, Clustering takes its place by grouping similar data points in n-dimensional space. Clustering works on its principle property of maximum similarity within-cluster, and minimum similarity between clusters. Many Data mining techniques are prevailing for disease prediction and treatment [4]. Similar approaches in a

combination considered for drug prediction and analysis offer new scope for research in the Drug response compendium[1].

Clustering works on both categorical and continuous data and forming quality clusters enhances the performance of analysis with cluster lifetime also impacting the clustering results in dynamic environments [8]. The dataset considered for our work consists of multi-variant attributes like Ethnicity, F-score, M-score, O-score, E-score, A-score, C-score, Impulsive, etc nominally numeric in our scenario, which perform well for identifying drug response similarity on patient data. There are 5 well-known personality traits defined as building blocks of human personality namely Ethnicity, Neuroticism, Extraversion, Agreeableness, conscientiousness.

Ethnicity a commonly preferred parameter in studies of health disparities and one of the major personality traits deal with a person's reaction about the place, culture, race, food, climate, and so on. It puts its mark on the patient's response to drugs. Neuroticism tends towards the impact of anxiety, depression, moodiness, self-doubt, threat, and frustration of personal feelings. Extraversion is mainly a character dealing with a person's excitability, sociability, emotional expressiveness, assertive, talkativeness that participates in responding to drugs. Openness has properties like imagination, insight, curiosity, and higher interests. Agreeableness includes features like Kindness, affection, trust, and pro socio behaviors. People with these high behaviour are very co-operative. Conscientiousness enlightens higher thoughtfulness, well organized, goal-oriented, and planned behavior. All these factors play a silent vital role in responding to a drug with variable variations. These factors are considered for the experimental demonstration of Drug consumption similarity prediction [12]with an advanced dynamic clustering approach. Feature extraction [24] in classification adds up to quality classes but poses a challenge in the Big data scenario[26], entailing network analysis features [22] for large outsourced data.

Several dreadful diseases like Brain tumors are severely affected by these factors, and these scenarios are well handled using convolution neural networks[10]. The above-mentioned personality traits impact a wide range of disease- treatment cases like cuff-less blood pressure [9] to malaria, cancer[5], amnesia, etc which are well studied using machine learning approaches[11]. Fluid mechanics also render good support in medical diagnosis [21] with Machine Learning approaches [20] well suited for securing patient sensitive data by avoiding Intrusion even in the private cloud [23] with mechanisms like honey encryption.

 Our paper focuses on the similarity aspect of clustering by considering drug similarities of patients for various diseases[18], which help in improving the efficiency of treatments by using advanced K-Means with neighborhood concept.

2. RELATED WORKS

Saad Haider and Pal in 2014 contributed a detailed analysis of drug sensitivity and modeling multivariate distribution concerning drug sensitivities. Li-Yu Hu, Huang, Ke, and Tsai in 2016 discussed the importance of Distance function like Euclidean distance, Manhattan, chi-square, etc in KNN Classification considering medical data. Jeongsu Park and Lee in 2018 came up with Privacy issues in e-cloud for k-Nearest Neighbour Classification. In 2019 Rashid, Yousuf, Ram, and Goyal proposed a novel approach for predicting drugs in medical datasets. Ehsan Ullah et. al., in 2017 discussed in recognizing cancer drug sensitivity based on associated genomic features. Wang et. al., in 2019 contributed Association identification among Drug –Disease related to Neighbourhood information in Neural Networks. Wen Chao Xing and Yilin Bei, in 2017 gave the classification of Medical, data with the KNN algorithm.

Ganesan. T and Rajeswari P, in 2019 proposed a genetic algorithm related to improving Cluster lifetime by optimal sensor placement thereby increasing the performance utilization of clusters. Sajana T and Narasingarao, in 2018 gave a detailed study on comparisons of Malaria Disease using machine learning techniques. Shinde A and Rajeswari P, in 2020 contributed a novel hybrid framework for health care related to Blood Pressure with the Machine Learning approach. Sowjanya, Divyambica, Gopinath, Vamsidhar, and Vijay Babu, in 2019 gave an impressive prediction model for diabetics disease based on Glucose levels in blood using Data science algorithms. Meghana, Manisha, and Rajeswari P, in 2019 proposed a deep Learning approach related to Brain Tumor disease with a convolution Neural Network approach. Supriya M and Rajeswari P, in 2017 reviewed different data mining techniques related to Association Rules for Privacy capabilities.

Jianping Gon, Xiong, and Kuang,2011 came up with a dual weighted voting function to detonate the effect of an outlier in K-nearest neighbor classification. Sandeep Kaur and Sheeta, in 2016 contributed a hybrid K-means in support of SVM for enhancing the efficiency in disease prediction. Sabthami, Thirumoorthy, and Munnswaran in 2016 proposed a multi-view clustering of medical records for their drug responses and hypothetical conditions. Alsayat and Sayed, in 2016 discussed existing clustering techniques and proposed an enhanced K-Means clustering with the SOM technique to overcome the centroid selection problem. Jadhav and Vijaya babu, in 2019 discussed the diverse features of network analysis. Srinivas et. al., in 2018 proposed honey encryption for private cloud. AmudhavelJ, Srikanth, Babu Karthik, and Sambasivam G contributed to the analysis of Fluid dynamics in the medical domain. Rajeswari P and Supriyamenon M, in 2018 discussed the privacy aspects in mining techniques based on the context and environment. Rama Rao, Sivakannan Subramani, Prasad, in 2017 gave a detailed study of technical challenges in Big data. Vidhullatha in 2019 spoke about intrusion detection in higher perspectiveSurlakar, Araiyo, and Sundaram, in 2016 contributed an appreciated comparative analysis of k-Means and K-Nearest Neighbour techniques to Image segmentation.

3. BASIC PRELIMINARIES

K- means: An unsupervised learning approach defined for clustering, where k in k-Means specifies the number of clusters the algorithm is intended to project. The resulting cluster may be of arbitrary shapes. K-Means works in 2 phases iterating till convergence. The initial is the assignment step continued by iteratively updating. k- Means takes into consideration the sum of squared means between data points and all centroids. It solves the problem of Expectation-Maximization, E step in assigning points to a nearer cluster, and M for calculating the centroid of Cluster. It evolves in several variations [15] like in combination with SVM, classification, etc define enhancement [16] and stands the most opted choice of researchers.

KNN: One of the supervised Machine Learning algorithm which ranks to be best voted for researchers and polls for solving Classification and Regression problems [14]. In KNN classification the result is a class membership whereas n Regression [7] the result is the property value of an object. KNN is a lazy learner and works by computing Euclidean distance between data points to find the nearest neighbor. The Accuracy of KNN downtrends with noise and performance detonates with large volumes of data [3]. A variation of KNN available is by assigning weights to the neighbors depending on the consistency of neighbors. Comparatively [19] K-Means is an eager learner, at times a fusion of KNN and k-Means

would result in an efficient model enhancing performance compared to individual contributions.

4. PROPOSED APPROACH

The Proposed approach relies on dynamic K-Means clustering for Drug response similarity Prediction, which helps doctors to adopt the right decisions at right time.

The Dataset considered for our work consists of 3500 records entailing 32 attributes holding both categorical and numeric values. Among them, numerical attributes like Ethnicity, E-score, N-score, O-score, C-score are taken up for identifying the associations aiming at improved similarity identification. As a fact of the word, the intended work mainly revolves around Drug similarities, and disclosing similarities is a basic amenity of Clustering.

Clustering algorithms are capable of generating clusters of variable sizes and shapes irrespective of volume constraints and hence preferred for our work. Many Clustering Techniques fill the Bag, among which K-Means stands to be the researcher's choice due to its scope of extension and performance accelerating parameters [2]. Our approach holds novelty in the aspect of measuring distance i.e. voting for nullifying the effect of spurious classes which paves the path for considering weighted inverse Euclidean distance in dynamic k-Means.

The dynamic K-Means starts with projecting the data points of attributes on to an N-dimensional space and initializing k value. Generating of clusters begins by computing weighted Euclidean distance between data points as discussed below.

1. Compute the distances between data points

$$ED = \sqrt{((x2 - x1)^2 + (y2 - y1)^2)}$$

2. Compute the inverse of each distance

$$I_i = 1 / ED_i \quad i=1 \text{ to } N$$

3. Find the sum of the Inverse distance

$$S = \sum_{i=1}^{N} Ii$$

4. Calculate weight for each data point based on their attribute values
$$W_i = I_i/S$$

5. Based on the resulting W_i values, data points are clustered.

Algorithm Procedure for the nearest drug similarity index :
1. Initialization phase

a.   Allocate N data values ( ) to N clusters with each cluster having one data point. Set N_t = N.

b.   Determine the corresponding dynamic k nearest neighbors DK[l] for each cluster I (I = 1 to N), thereby arrange DK[I] in the ascending order of their cluster distances to I.

c.   Generate IDK[l] for clusters I(l = 1 to N).

2. Merging phase:

a.   Consider DK list, to find clusters c1 and c2, with minimum cluster distance, and merge clusters into cluster c1c2. Set DK[c1c2] = ; and IDK[c1c2] = ;, where ; is an empty set.

b.   If Dc1,DK[c1][m] P Dc2,DK[c2][n], for each j 2 DK[c1], set DK[c1c2] = DK[c1c2] [{j} in the case of (Dc1c2,j + Dc1,c2) < Dc1,DK[c1][k], where m = jDK[c1]j and n = jDK[c2]j. Otherwise for each j 2 DK[c2], set DK[c1c2] = DK[c1c2] [{j} if (Dc1c2,j + Dc1,c2) < Dc2,DK[c2][k].

c.   DK[c1c2] = ;, determine DK[c1c2] from all clusters.

d.   Use the above index to determine nc1c2 and Cc1c2, respectively.

e.   Determine Su and delete cluster c2, where Su = IDK[c1]& [IDK[c2]n{c1,c2}.

3) Updating phase:

f.   Update IDK[j] = IDK[j]n{c1} for each cluster j 2 DK[c1], and update IDK[j] = IDK[j]n{c2} for each cluster j 2 DK[c2].

g.   Update DK[j] = DK[j]n{c1,c2} for each cluster j 2 Su.

h.   Replace DK[c1], nc1 and Cc1 by DK[c1c2], nc1c2, and Cc1c2, respectively, and update IDK[c1] = ;.

i.   Set IDK[i] = IDK[i] [{c1} for j = 1 to jDK[c1]j, where i = DK[c1][j].

j.   For each cluster j 2 Su, if Dc1,j < Dj,DK[j][m], update DK[j] = DK[j] [{c1} and IDK[c1] = IDK[c1] [{j}, where m = jDK[j]j.

k.   For each cluster j 2 Su, if DK[j] = ;, determine DK[j] from all clusters and for I = 1 to k, set IDK[i] = IDK[i] [{j}, where i = DK[j][I].

(4)   Set $N_t = N_t$ 1. Repeat steps (2)–(3) until $N_t < M$.

5. PERFORMANCE EVALUATION AND ANALYSIS:

To Implement the Proposed approach using Jsim, the Simulation parameters are initialized as shown in for Drug response similarity.

Table 1 : Simulation Parameters

PARAMETERS	VALUES
Simulator	JSIM

Simulator Time	100 s
Simulation Area	1000*1000 m
Proposed Protocol	Dynamic K-Means
No of datasets	3500
Number of attributes	32
No. of attributes considered for simulation	7

Figure 1: Dataset Selected for Clustering
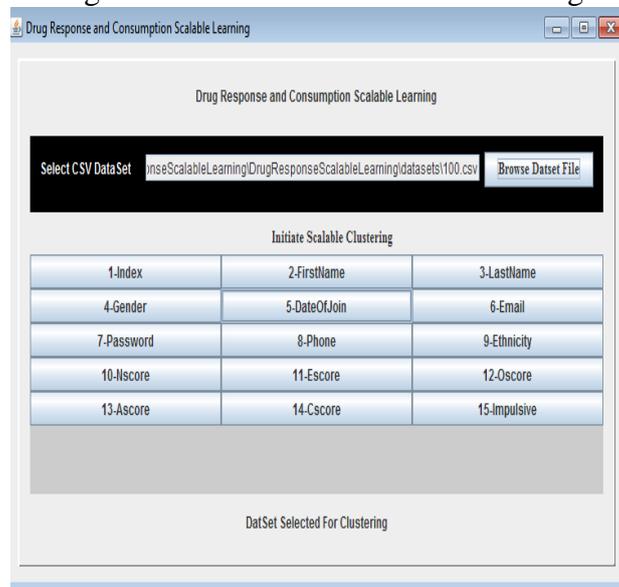
In figure 1 we visualize Drug Response and consumption Scalable learning by considering the above attributes from the selected dataset for clustering by emphasizing only on numeric attributes.
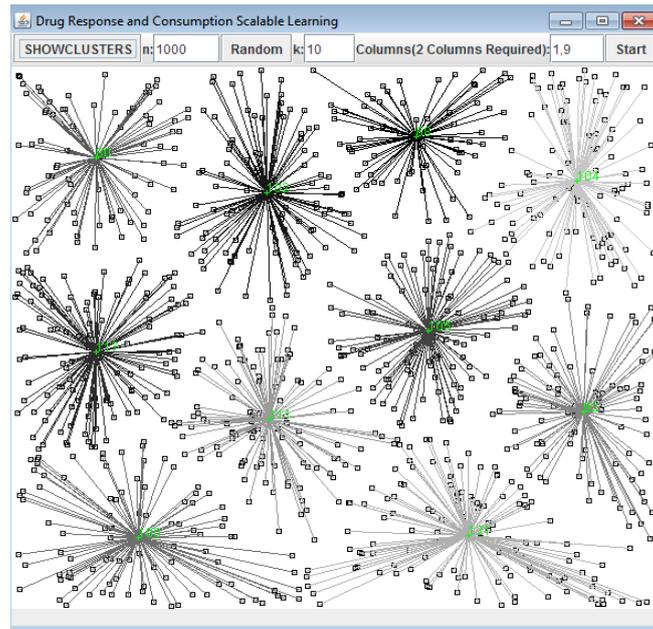
Figure 2: Clustered data sets

Figure 2 depicts the formation of clusters when k value is set to 10 for experimental evaluation of the proposed approach with k nearest values based on inverse Euclidean distance from the centroid.
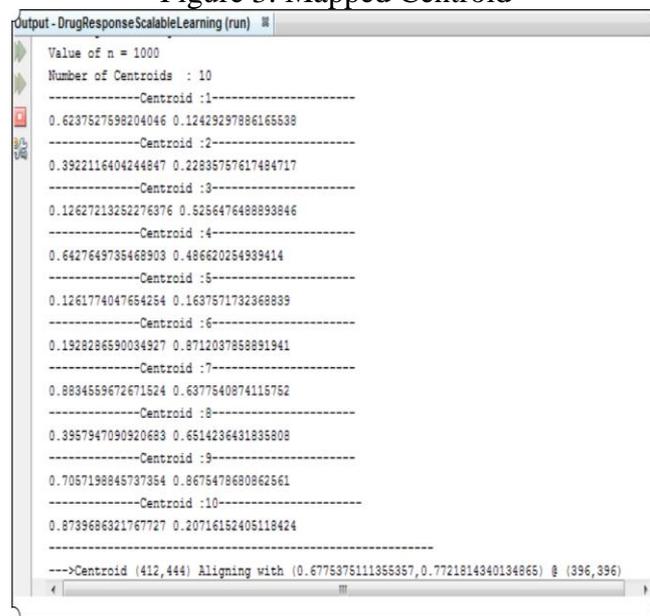
Figure 3: Mapped Centroid

Figure3 displays the coordinates of the 10 centroids selected for clustering datasets with its positioning parameters aligned.

Figure 4: Consolidated clustering values

```
Edge Cluster Walk

                [1115,Cluster : 8 @ (91,131) clusters count are : 123]
        [857,Cluster : 2 @ (477,73) clusters count are : 95]
                        [767,Cluster : 7 @ (90,438) clusters count are : 95]
                            [640,Cluster : 10 @ (298,443) clusters count are : 105]
            [360,Cluster : 4 @ (268,54) clusters count are : 89]
                        [282,Cluster : 9 @ (315,170) clusters count are : 91]
                    [266,Cluster : 5 @ (496,218) clusters count are : 87]
[211,Cluster : 1 @ (146,311) clusters count are : 105]
            [191,Cluster : 3 @ (347,296) clusters count are : 94]
                    [122,Cluster : 6 @ (498,418) clusters count are : 116]
Cluster Ensemble runs in 1.514 secs
```

The above figure 4 elucidates the number of data points in each cluster and coordinates of the respective centroid.

6. RESULTS

The Performance of our Proposed dynamic Approach is evaluated against k-means and KNN with metrics like f-measure, Accuracy, Recall, and Time Efficiency. These metrics contribute to evaluating and presenting improved results.

Accuracy: A metric for classification and clustering which well performs for categorical, numeric, and multiclass.

Recall An important measure for focusing actual positives among scenarios where our evaluation choice is to capture more positives from datasets.

Time Efficiency: These metric measures the time slice at which efficient results are achieved by various algorithms. In our experimental setup, we have taken the UCI dataset containing records about several attributes.

Table 2

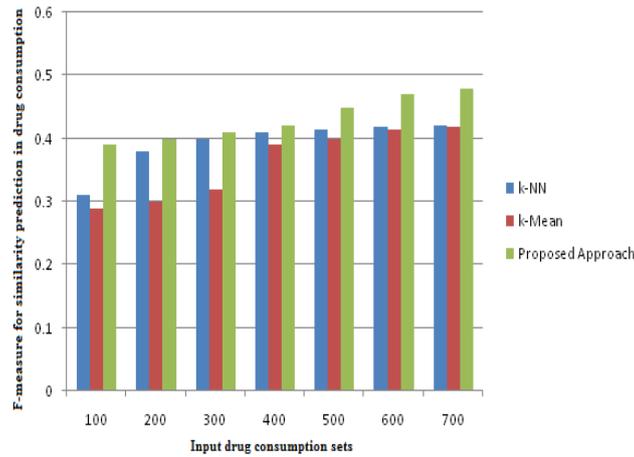| Input Drug Consumption data sets	k-NN	k-Mean	Proposed Approach DK
100	0.31	0.29	0.39
200	0.38	0.3	0.4
300	0.4	0.32	0.41
400	0.41	0.39	0.42
500	0.415	0.4	0.45
600	0.419	0.415	0.47
700	0.421	0.419	0.48

Figure 5: Performance of f-measure with drug consumption data sets.

Table 3

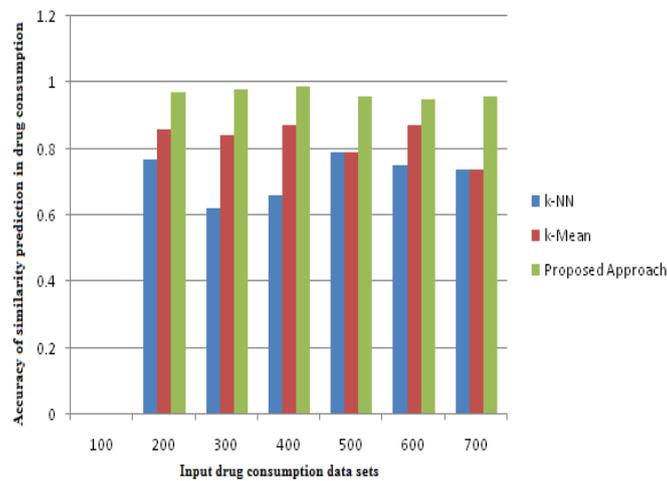| Input Drug Consumption data sets	k-NN	k-Mean	Proposed Approach DK
100	0.77	0.86	0.97
200	0.62	0.84	0.98
300	0.66	0.87	0.99
400	0.79	0.79	0.96
500	0.75	0.87	0.95
600	0.74	0.74	0.96
700	0.69	0.82	0.96

Figure 6: Performance of accuracy with processing drug consumption data sets.

Table 4

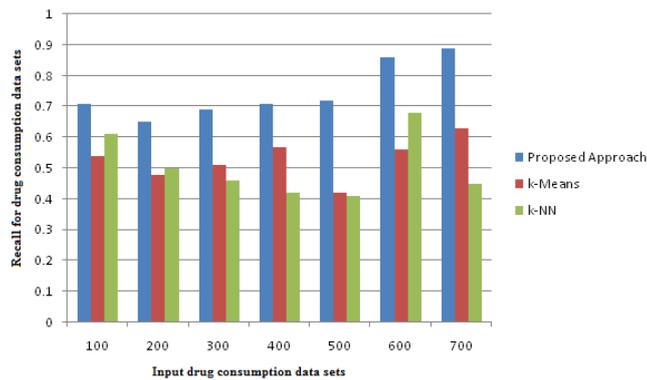| Input drug consumption data sets	Proposed Approach DK	k-Mean	k-NN
100	0.71	0.541	0.61
200	0.65	0.48	0.50
300	0.69	0.51	0.46
400	0.71	0.57	0.42
500	0.72	0.42	0.41
600	0.86	0.56	0.68
700	0.89	0.63	0.45

Figure 7: Performance of recall concerning the consumption of drug-related data sets

Table 5

Drug consumption data sets	k-NN	k-Means	Proposed Approach DK
100	0.61	0.49	0.3
200	0.73	0.54	0.31
300	0.71	0.59	0.32
400	0.85	0.61	0.33
500	0.92	0.71	0.34
600	0.95	0.62	0.36

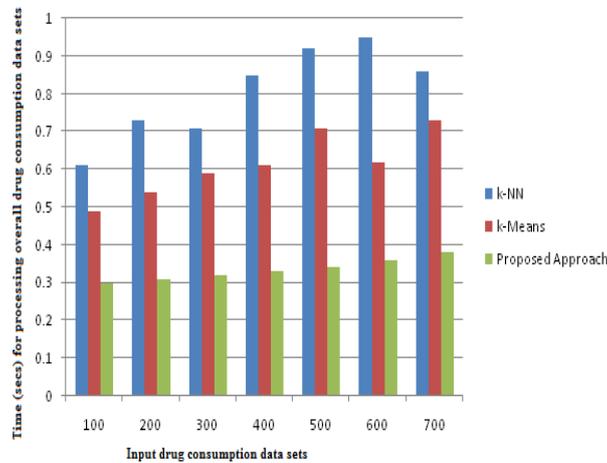| 700	0.86	0.73	0.38

Figure 9: Performance of time in processing drug consumption data sets

7. CONCLUSION

Identifying Disease treatment relationships continues to be a burning problem to numerous aspects taken into consideration. Researchers are striving to achieve maximum attainment in predicting optimal treatments for disease, still leaving a few coins unturned. Our present work extends predicting profound treatment for diseases by introspecting parameters like the cure, side effects [6] using classification strategies to identifying similarities of drug responses of patients depending on their behavioral traits using enhanced dynamic K-Means clustering. The proposed approach promised to generate improved results. Further identifying the co-relations between similarity clusters and optimizing the clusters for better results using optimization algorithms may turn out to be beneficial in building optimized predictive models in the medical domain.

8. REFERENCES

[1] Haider, S., & Pal, R. (2014). Analysis of multivariate drug sensitivity dependence structure using copulas. 2014 IEEE Global Conference on Signal and Information Processing globalSIP). doi:10.1109/globalsip.2014.7032345

[2] Wang, Y., Deng, G., Zeng, N., Song, X., & Zhuang, Y. (2019). Drug-Disease Association Prediction Based on Neighborhood Information Aggregation in Neural Networks. IEEE Access, 7, 50581–0587. doi:10.1109/access.2019.2907522.

[3] Park, J., & Lee, D. H. (2018). Privacy-Preserving k-Nearest Neighbor for Medical Diagnosis in e-Health Cloud. Journal of Healthcare Engineering, 2018, 1–11. doi:10.1155/2018/4073103.

[4] Rashid, M., Yousuf, M. M., Ram, B., & Goyal, V. (2019). Novel Big Data Approach for Drug Prediction in Health Care Systems. 2019 International Conference on Automation,

Computational, and Technology Management (ICACTM). DOI:10.1109/icactm.2019.8776823.

[5] Ullah, E., Mall, R., Bensmail, H., Rawi, R., Shama, S., Muftah, N. A., & Thmpson, I. R. (2017). Identification of cancer drug sensitivity biomarkers. 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). doi:10.1109/bibm.2017.8218043.

[6] Wang, Y., Deng, G., Zeng, N., Song, X., & Zhuang, Y. (2019). Drug-Disease Association Prediction Based on Neighbourhood Information Aggregation in Neural Networks. IEEE Access, 7, 50581–0587. doi:10.1109/access.2019.2907522.

[7] Xing, W., & Bei, Y. (2019). Medical Health Big Data Classification Based on KNN Classification Algorithm. IEEE Access, 1–1. doi:10.1109/access.2019.2955754.

[8] T.Ganesan, Pothuraju Raja Rajeswari, "Genetic Algorithm Based Optimization to Improve the Cluster Lifetime by Optimal Sensor Placement in WSN's", International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-8 Issue-8, June, 2019.

[9] A. Shinde, P. Raja Rajeswari, Santosh, "A Novel Hybrid Framework for Cuff-Less Blood Pressure Estimation based On Vital Bio Signals processing using Machine earning ", , International Journal of Advanced Trends in Computer Science and Engineering, Volume 9 No.2, March -April 2020.

[10] Amulya P, Sai Meghana S, Manisha A, Rajarajeswari P,"A Deep Learning Approach For Brain Tumor Segmentation using Convolution Neural Network", International Journal of Scientific and Technology Research, December 2019.

[11] Sajana T and Narasingarao, "A comparative study on imbalanced malaria disease diagnosis using machine learning techniques. Journal of Advanced Research in Dynamical and Control Systems, 2018, 10,552-56.

[12] Sowjanya V, Divyambica CH, Gopinath P, Vamsidhar M, B.Vijaya Babu, International Journal Of Engineering and Advanced Technology (IJEAT) ISSN: 2249-8958, Volume-8 Issue-4, April 2019, Improved Prediction of Diabetes based on Glucose Levels in blood using Data Science Algorithms.

[13] Supriyamenon M, Rajarajeswari P, A Review on Association Rule Mining Techniques with Respect to their Privacy-Preserving Capabilities (KLEF), International Journal of Applied Engineering Research July 2017, ISSN 0973-4562 Volume 12, Number 24 (2017) pp. 15484- 15488.

[14] Taisong Xiong and Yin Kuang , "A Novel Weighted Voting for K-Nearest Neighbor Rule" JOURNAL OF COMPUTERS, VOL. 6, NO. 5, MAY 2011,pp-833-838.

[15] Kaur, S., & Kalra, S. (2016), " Disease prediction using hybrid K-means and support vector machine", 2016 1st India International Conference on Information Processing (IICIP). doi:10.1109/iicip.2016.7975367.

[16] Alsayat, A., & El-Sayed, H. (2016). Efficient genetic K-Means clustering for health care knowledge discovery. 2016 IEEE 14th International Conference on Software Engineering Research, Management, and Applications (SERA). doi:10.1109/sera.2016.7516127

[17] Rajeswari P, Supriyamenon.M, "A contemporary way for enhanced modeling of context-aware privacy system in PPDM". Journal of Advanced Research in Dynamic and Control Systems,Vol.10,01-issue, July 2018.

[18] Sabthami, J., Thirumoorthy, K., & Muneeswaran, K. (2016), " Multi-view clustering of clinical documents based on conditions and medical responses of patients",2016 10th

International Conference on Intelligent Systems and Control (ISCO). doi:10.1109/isco.2016.7726951 .

[19] Surlakar, P., Araujo, S., & Sundaram, K. M. (2016). Comparative Analysis of K-Means and K-Nearest Neighbor Image Segmentation Techniques. 2016 IEEE 6th International Conference on Advanced Computing (IACC). doi:10.1109/iacc.2016.27 .

[20] vidyullatha pellakuri, "Performance analysis of machine learning techniques for intrusion detection system", 2019Proceedings – 2019, 5th International Conference on Computing, Communication Control and Automation, ICCUBEA 2019.

[21] Amudhavel, J.; Srikanth, V.; Babukarthik, R. G.; Sambasivam, G, "Fluid Dynamics in advanced medical Diagnostic Technologies: An in-depth Analysis", Bioscience Biotechnology Research Communications, 2018, volume 11, issue 1, pp: 34-38.

[22] Pranavati Jadhav and Dr. Burra Vijaya Babu, "Detection of Community within Social Networks with Diverse Features of Network Analysis", Journal of Advanced Research in Dynamical and Control Systems ISSN: 1943-023X Volume 11 | 12-Special Issue Pages: 366-371.2019.

[23] Srinivasu N, Sahil M, Francis J, et. al., "Security enhanced using honey encryption for private data sharing in cloud", International Journal of Engineering and Technology(UAE) (2018)

[24] B. Subramanian, V. Saravanan, R.K. Nayak, T. Gunasekaran and S. Hariprasath, "Diabetic Retinopathy – Feature Extraction and Classification using Adaptive Super Pixel Algorithm", International Journal of Engineering and Advanced Technology (IJEAT), Vol. 9, No. 2, pp. 618-627, 2019, ISSN: 2249-8958, B-Impact Factor-5.97, DOI: 10.35940/ijeat.B2656.129219.

[25] Kumar S.A, Vidyullatha P, "A comparative analysis of parallel and distributed FSM approaches on large-scale graph data", International Journal of Recent Technology and Engineering, Volume 7, Issue 6, April 2019, Pages 103-109.

[26] Kvsn RamaRao, Sivakannan Subramani, M.A. Prasad, "Technical challenges and perspectives in batch and stream big data machine learning", December 2017, International Journal of Engineering & Technology 7(1-3):48.