A Hybrid Machine Learning approach for Drug Repositioning

Supriya Menon M[1]
Research Scholar
Dept. of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
AP, India

Pothuraju Rajarajeswari[2]
Professor
Dept. Of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
AP, India

Abstract— Data Mining continues its battle against dense volumes of data poured in, to withstand the urge for knowledge discovery. Knowledge enhances the decision-making capability providing a peak edge in the competitive world. The medical domain is one such application, where a fast and wise decision gains to be a lifesaver. Several mining techniques already prevailed in the medical world with their essence reflected in their advancements. Among the discrete usage scenarios, Drug similarity prediction is evolving to be an attractive choice of research, as new drug development is expensive and time-consuming. Also, the approval rate of the FDA is stepping down, which paves attention towards drug repositioning based on similarities and a quench for optimized results. Our research addresses this for the mentioned complications by employing a hybrid approach. Our framework blends the features of KNN with ACO to attain enhanced Drug consumption similarities. The proposed hybrid computational method promises leveraged results, simulated with JSIM evaluating performance parameters like recall, accuracy, and time.

Keywords *—Ant Colony Optimization, K++ Means Clustering, KNN- approach.*

I. INTRODUCTION

Pharmaceutical Industry is playing a significant role in research and development by leveraging the pharmaceutical R & D expenditure drastically to 186 billion US dollars in 2019, while 126 billion in 2012. Most trending advancements in this area are outsourcing of R&D by drug manufacturers targeting cost reduction, enhancing predictive modeling in clinical research with big data usage, and grabbing technology companies with real-world evidence residing in data warehouses from heterogeneous sources. Regardless of all such contributions, pharmacy companies failed to maintain their pace with fewer numbers of novel drugs being approved and restrictions on prices pose a tough challenge to the drug industry. This struggle for existence forces the drug developers to enhance creative skills in finding new applications for existing drugs by proposing Drug Repositioning [8].

Drug repositioning is an attractive innovative stream, benefitting drug manufacturers with safer prescriptions for patient's proceeds by selecting a drug for therapeutic effects and carrying out clinical tests. Several approaches like computational, biological [9, 3], experimental, and fusion-based have marked their success in drug repositioning. With the availability of huge amounts of medical data over the digital platform, computational approaches are gaining concern in new directions for approved drugs. Dynamically growing medical data is questioning the technology to drill down the enhancements benefitting the society.

Data mining resolves this quench for a deep analysis of high dimensional medical datasets with numerous techniques like classifications, Associations, Clustering, and Outlier analysis by deriving profound conclusions targeting better decision making [21]. Among fore mentioned mining techniques Association mining aims at discovering association patterns resulting in Association rules. The classification process categorizes datasets into classes with rules defined in the learning process. Clustering emphasizes unsupervised learning with similarity and dissimilarity proximities. Few well accepted clustering algorithms are K-Means, Optics, dbscan and so on [20]. Outliers serve remedies for suspicious uncommon behavior, encouraging fraudulent analysis.

Drug repositioning, basically characterized with similarity measure owes outstanding results with clustering techniques whose basic functionality is grouping similar data items into one group or cluster [13]. They work on the property of minimum inter-cluster similarities and maximum intra cluster similarities. Feature based similarity identification and creation help prediction of drug and disease similarity [11] to yield better results [5, 6]. It uses statistical computations to uncover hidden structures, in scenarios of exploring and evaluating data analysis by researchers to disclose features existing without prior knowledge.

Our approach focuses on optimization of drug similarity prediction for diseases, lending support towards drug repositioning with the help of clustering mechanism for similarity and optimization carried out using Ant Colony Technique.

Manuscript received December 5, 2020
Manuscript revised December 25, 2020
https://doi.org/10.22937/IJCSNS.2020.20.12.24

II. RELATED WORK

Martens, Backer, and Haesen, 2007, proposed an Ant Miner technique with a Max-Min ant system by including class variables to avoid the effect of a Multi-class problem. Kentzoglanakis and pook, 2012, investigated the problem of gene regulatory networks and proposed an alternative with particle swarm optimization. Ibrahim et. al., 2020, introduced an Adaptive neurological fuzzy inference system infusion with PSO and Grey wolf optimization to enhance disease prediction with their characteristic features. Ding et. al., 2019, came up with a novel prediction approach for ensuring improved results by considering feature-based prediction of drug similarities. Liu et. al., 2020 contributed a new framework termed TS-SVD, by considering drug-disease and drug proton to analyze drug-disease associations which achieved better results. Mavrovouniohs and Yang, 2013, discussed in detail about DOP problem in ACO and resolved it by proposing ACO for dynamic environments. Chen and Zhang, 2018, reviewed Micro RNA's that drugs target to control expression levels and repositioning.

Huang et. al., 2019, proposed a computational environment to analyze similarity in drugs for Drug repositioning resulting in reduced cost. Reddy and Supreethi, 2017, proposed K-Means with ACO to beat the adverse effects of the Local Minima problem in K-Means. Zheng, Shameek, and Jinyar, 2017, developed an optimal Drug similarity framework to enhance the performance of Side-Effect prediction by picking 917 drugs from the drug bank as a dataset. Celebi et. al., 2015, employed a Page rank algorithm in Drug-Drug interaction to identify new DDI considering the weight approach to analyze the similarities of drugs. Jiao and Zhiyong proposed a novel bi-partite graph technique by the inclusion of drug structure information for computation of Drug similarity Index. Surlakar, Araujo, and Sundaram, 2016 talked in detail about K-Means & KNN, in Pathology to identify various issues. Premalatha and Subhasree, 2017, discussed three clustering algorithms like Medical Storage Platform for DM, Homogeneity similarity-based Hierarchical, and K- Harmonic Means overlapped K-means and hierarchical algorithms for clustering by evaluating results. Altaseva et. al., 2016, proposed a disease prediction system for heart diseases using Naïve Bayes and K-Means Clustering algorithms. Vikram, Negi, 2019, came forward with Classification Model considering a structured biological knowledge graph. Cenaroglu , 2019, proposed K-Means Clustering in combination with envelopment analysis to understand the technical efficiencies of public hospitals in turkey by considering validity index measure. Yanchun et. al., 2019, developed a technique called weighted-KNN considering multi label linear discriminate for Weight calculation to enhance the accuracy of calculation with 3 different types of data sets. Xia, Stilgic, and Wang, 2018, talked about role of data mining technique to handle noise and missing values in large volumes of medical data, resulting in a combined technique of k-Means and clustering enhancing performance. Gupta and Chandra, 2020, reviewed data mining techniques, tasks, challenges, issues, and their applications in real world. Supriya and Rajeswari, 2020, came up with a novel approach for predicting drug response similarity using machine learning with dynamic K-means approach.

III. EXISTING APPROACH

Ant Colony Optimization

The optimization technique aims at resolving computational issues in a probabilistic manner by disclosing optimal paths. Other optimization algorithms like PSO focuses on iterative improvement of solutions until an optimal is found [2]. Conversely Grey Wolf Optimization technique, a meta heuristic technique follows grey wolf nature [4] to hunt optimal solution varying in structure from others. ACO is modeled concerning ant colony actions by optimal solutions for exploring artificial ants [7]. The simulation agents designated as artificial ants move over the parameter range projecting all acceptable solutions. The path over the accepted range is recorded with their quality ranked, To compare and evolve better solution in the further iteration [1].

The issues related to optimization problems are grabbing interest both in scientific and industrial domains. Few areas urging the need for such optimization are vehicle routing salesman, scheduling problems, Image processing, Problem Assignment problems, set problems, and so on. Many optimization algorithms are available to address the aforementioned problems. Few algorithms are differential evolutions, Generic approaches, Evolutionary, dynamic, hill climbing.

The algorithm initiates with an artificial ant moving in search of the best solutions for the considered optimization problem. At every step, the ant constructs a solution and evaluates it against other solutions, until the optimal one is attained. The step by step algorithm is as follows:

The procedure of ACO is

While not terminated
Do
Construct solutions ()
Perform relevant actions ()
Update solutions ()
Repeat

End procedure.

K –Means

A well known unsupervised clustering algorithm for grouping the data based on similarities undergoes iteration to partition the dataset into k clusters [14]. The iterations compute the Euclidean distance [22] from the initial center and keep updating until the final clusters are shaped. All the points within the cluster are at different distances belonging to the same cluster [18]. The lower the variation within clusters, the maximum quality is the cluster. This approach of K-Means is termed Expectation-Maximization, with assigning data points to near cluster and calculating centroid for each cluster [16]. Many research discussions even revolve around other clustering algorithms[15] like Hierarchical, emphasizing on levels to form clusters.

.

KNN Classification

It is a well-opted classification algorithm that succeeds in handling both Classification [17] and Regression problems relying on supervised Learning contributed by Thomas Cover. KNN a lazy learner algorithm by definition locally approximates the function and postpones the computation to function evaluation. KNN finds the nearest neighbors to the k and classifies them to fall into the same class [14]. All Neighbors within a specific range give the same effect of belonging to the same class. Weights [12] can be assigned to the contributions of neighbors to prioritize them. WKNN works similarly but assigns weights to its neighbors with the highest to nearer one and lowest to farther one [19]. This task of weighing is handled in 2 phases, like first finding distance and second transforming them into weight using a kernel function. Several kernel functions for KNN are Quartic function, Cosine function, Triweight function, Gaussian function, and Inverse function. These weights help in finding the most optimal neighbors with all benefits. Few kernel functions used in our approach are discussed in the proposed work.

IV. PROPOSED APPROACH

The information of diseases, drugs related to diseases, classified psychological parameter values of patients from UCI repository are contributed for our experimental implementation. A detailed explanation regarding the proposed hybrid approach, a blend of ACO and dynamic K Means approach is given below. The K-Means considered is designated - Dynamic, as they are clustered by considering weights based on the distance to the nearest neighbors.

Phase 1:

The Dataset considered for hybrid framework initiates by co-relating different activities to different attributes, thereby attaining training data set and test data set.

Phase 2:

In this phase training data set is subjected to Dynamic K-Means to extract the most relevant nearest neighbors using Euclidean distance and later transforming distance into weights with the three kernel functions like weighted quadric, cosine, tri weights are calculated. The respective Distance function D_f and Kernel functions are formulated below.

$$D_f = D(x, x_j) = \frac{d(x, x_i)}{d(x, x_{\lambda+1})}$$

Where $x_1..x_n$ be learning set of Observations

Distance to Weight function given as

$$W_j = T(D_j)$$

Kernel functions :

$$\text{Quadric function} = \frac{15}{16}(I - D^2)^2 - I(|D| <= 1)$$

$$\text{Tri weight} = \frac{33}{22}(I-D^2)^3 + (|D| <= 1)$$

$$I = \begin{cases} 1 \text{ if } |D| <= 1 \\ 0 \text{ if } |D| > 1 \end{cases}$$

Class Membership function

$$\hat{Y} = \max_r \left[\sum_{j=1}^{k} W(j) I(y(j) = r \right]$$

Phase 3:

Input the clusters from dynamic k-means to ACO for i=1to n do

- Ant selects an item by a random walk
- Compute probability for selection and dropping
- Compute the Fitness function to evaluate the efficiency of the opted solution. The fitness is determined by the following equation.

$$\frac{R_p}{\sigma_p}$$

Fitness function =

(Where Rp is a portfolio return value and σp the Covariance)

If fitness function (threshold range)
{
Accept
}
Else
{
Drop
}
Repeat

Phase 4: The optimized results help in identifying the drug similarity co-relation with improvement in F-measure

The proposed approach attains optimized co-relations among attributes that help in uncovering drug similarities for enduring Drug Repositioning which involves an examination of pre-existing drugs for new therapeutic impetus.

PERFORMANCE EVALUATION AND ANALYSIS:

Our Proposed Approach optimizes the existing Dynamic K Means approach by considering datasets in variable numbers with metrics like Accuracy, Recall, Precision, and Time Efficiency. The results of the fore mentioned parameters vary with the input count of datasets.

Accuracy: A quintessential metric of Classification well performs for binary and multiclass is the number of true results for total cases being classified.

Recall : An important measure for focusing actual positives among scenarios where our evaluation choice is to capture more positives from datasets. It defines the measure of completeness.

Precision: A choice metric to answer the question of true positives divided by true positives and true negatives i.e. Correctness of the predictions made during the classification process. It defines the measure of exactness.

F-Measure: It specifies the balance between the Precision and the Recall using Harmonic mean between them.

Time Efficiency: This metric measures the time at which efficient results are dropped out by the classification algorithms.

To Implement the Proposed approach using Jsim, the Simulation parameters are initialized as shown in Table 1.

Table 1 : Simulation Parameters

PARAMETERS	VALUES
Simulator	JSIM
Simulator Time	100 s
Simulation Area	1000*1000 m
Proposed Protocol	Hybrid Approach
No of datasets	2500
Number of attributes	32
No. of attributes considered for simulation	7

Input Drug Consumption data sets	K- Means	Proposed Hybrid Approach
100	0.31	0.89
200	0.38	0.83
300	0.4	0.72
400	0.41	0.89
500	0.415	0.83
600	0.419	0.86
700	0.42	0.79

Table 2. Comparison of f-measure with variable drug consumption data sets.

The medical drug dataset consisting of 2500 records with 32 attributes are considered for our experimental work, out of which variable data sets are chosen and their performances are analyzed against various parameters.

According to the simulation results shown in Table 2, the performance of the hybrid approach is efficient than that of conventional k-Means with different data sets.

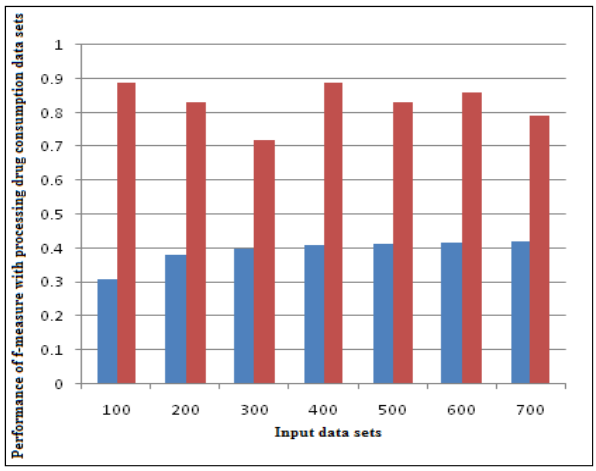

Figure 1. Performance of f-measure with variable drug consumption data sets.

Figure 1 depicts the increase in performance concerning f- measure with a red bar for the proposed model and a blue bar for existing K – Means.

Input Drug Consumption data sets	K- Means	Proposed Hybrid Approach
100	0.86	0.98
200	0.84	0.98.6
300	0.87	0.99
400	0.79	0.99.89
500	0.87	0.98.67
600	0.74	0.99.6
700	0.82	0.98.9

Table 3. Comparison of Accuracy measure with variable drug consumption data sets.

In Table 3 we acquired higher Accuracy values for the proposed hybrid approach, analyzing different data sets among available.

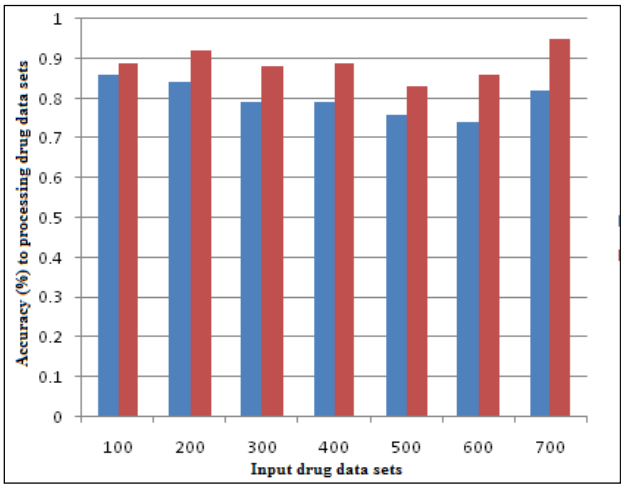

Figure 2. Performance of Accuracy with variable drug consumption data sets.

Figure 2 plots the improved Accuracy of variable drug consumption data sets using the ACO approach with a red spike when compared to a blue spike for the existing model.

Input Drug Consumption data sets	K- Means	Proposed Hybrid Approach
100	0.61	0.49
200	0.73	0.54
300	0.71	0.59
400	0.85	0.61
500	0.92	0.71
600	0.95	0.62
700	0.86	0.73

Table 4. Comparison of Time efficiency with variable drug consumption data sets.

Table 3 compares the runtime in milliseconds of Traditional k-Means and Hybrid approach. It is evident from the table that the time efficiency of the proposed is higher.

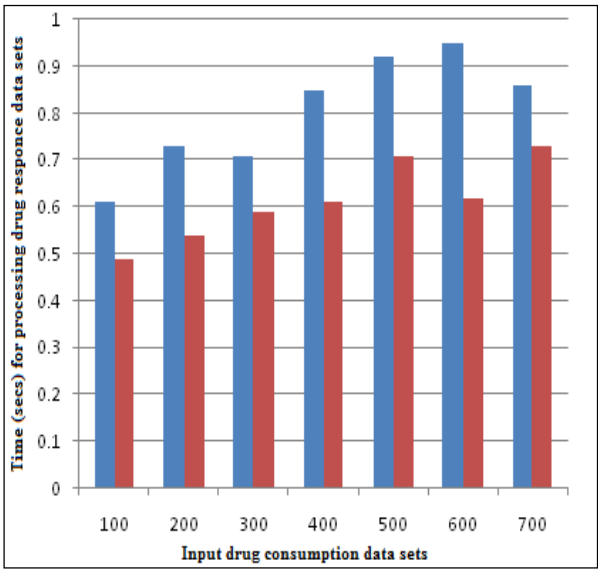

Figure 3. Performance of Time efficiency with variable drug consumption data sets.

Figure 3 projects the leveraged time efficiency of the proposed hybrid approach over the traditional approach.

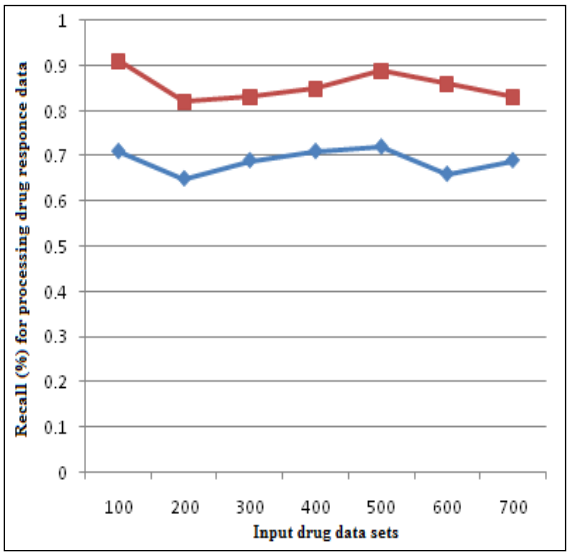

Figure 4: Performance of Recall with variable drug consumption data sets.

Figure 4 depicts an improved Recall measure indicated with a red spike when compared to the existing technique.

Input drug consumption data sets	Proposed Approach	Proposed Hybrid Approach
100	0.71	0.91
200	0.65	0.82
300	0.69	0.83
400	0.71	0.85
500	0.72	0.89
600	0.66	0.86
700	0.69	0.83

Table 5. Comparison of Recall with variable drug consumption data sets.

Table 5 shows improvement in recall from traditional k-means to hybrid approach considering variable size drug data sets.

VI CONCLUSION

This paper considers and analyzes the concept of applying Ant colony optimization to optimize the clusters formed. It initiates by selecting k initial clusters then upgrades with K nearest neighbors considering weights using kernel functions. This technique considered against drug consumption dataset improves the efficiency of Drug repositioning by optimizing drug response similarities. Applying Ant colony optimization to these Clusters leverages various performance factors like Accuracy, Recall, f-measure, and Time efficiency of the clustered data sets. The Optimization techniques promise enhanced results by overriding the adverse effects of traditional K-Means paving the scope for better Drug repositioning. The knowledge gained from such an analysis would enable the pharmaceutical domain to serve society with advanced treatments.

REFERENCES

[1] Martens, D., De Backer, M., Haesen, R., Vanthienen, J., Snoeck, M., & Baesens, B. (2007). Classification With Ant Colony Optimization. IEEE Transactions on Evolutionary Computation, 11(5), 651–665. doi:10.1109/tevc.2006.890229 .

[2] Kentzoglanakis, K., & Poole, M. (2012). A Swarm Intelligence Framework for Reconstructing Gene Networks: Searching for Biologically Plausible Architectures. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9(2), 358–371. doi:10.1109/tcbb.2011.87

[3] Ruskin, H. J., & Roznovat, I. A. (2015). Computational Models & Methods in Systems Biology & Medicine. IET Systems Biology, 9(6), 217–217. doi:10.1049/iet-syb.2015.0078 .

[4] El-Hasnony, I. M., Barakat, S. I., & Mostafa, R. R. (2020). Optimized ANFIS Model Using Hybrid Metaheuristic Algorithms for Parkinson's Disease Prediction in IoT Environment. IEEE Access, 8, 119252–119270. doi:10.1109/access.2020.3005614 .

[5] Ding, P., Yin, R., Luo, J., & Kwoh, C. K. (2018). Ensemble Prediction of Synergistic Drug Combinations Incorporating Biological, Chemical, Pharmacological and Network Knowledge. IEEE Journal of Biomedical and Health Informatics, 1–1. doi:10.1109/jbhi.2018.2852274.

[6] Liu, J., Zuo, Z., & Wu, G. (2020). Link Prediction Only with Interaction Data and Its Application on Drug Repositioning. IEEE Transactions on NanoBioscience, 1–1. doi:10.1109/tnb.2020.2990291

[7] Ping, G., Chunbo, X., Yi, C., Jing, L., & Yanqing, L. (2014). Adaptive ant colony optimization algorithm. 2014 International Conference on Mechatronics and Control (ICMC). doi:10.1109/icmc.2014.7231524 .

[8] Prediction of drug-disease associations for drug repositioning through drug-miRNA-disease heterogeneous network. (2018). IEEE Access, 1–1. doi:10.1109/access.2018.2860632 .

[9] Huang, L., Luo, H., Yang, M., Wu, F.-X., & Wang, J. (2019). Drug and disease similarity calculation platform for drug repositioning. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). doi:10.1109/bibm47256.2019.8983401 .

[10] [10] Reddy, T. N., & Supreethi, K. P. (2017). Optimization of K-means algorithm: Ant colony optimization. 2017 International Conference on Computing Methodologies and Communication (ICCMC). doi:10.1109/iccmc.2017.8282522 .

[11] An Optimized Drug Similarity Framework for Side-effect Prediction Yi Zheng, Shameek Ghosh and Jinyan Li, ISSN: 2325-887X DOI:10.22489/CinC.2017.128-068, Computing in Cardiology 2017; VOL 44.

[12] Celebi, R., Mostafapour, V., Yasar, E., Gumus, O., & Dikenelli, O. (2015). Prediction of Drug-Drug Interactions Using Pharmacological Similarities of Drugs. 2015 26th International Workshop on Database and Expert Systems Applications (DEXA). doi:10.1109/dexa.2015.23 .

[13] [14] Li, J., & Lu, Z. (2012). A new method for computational drug repositioning using drug pairwise similarity. 2012 IEEE International Conference on Bioinformatics and Biomedicine. doi:10.1109/bibm.2012.6392722 .

[14] 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

[15] Premalatha, P., & Subasree, S. (2017). Performance analysis of clustering algorithms in medical datasets. 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT). Doi:10.1109/icecct.2017.8117894

[16] Altayeva, A., Zharas, S., & Cho, Y. I. (2016). Medical decision making diagnosis system integrating k-means and Naïve Bayes algorithms. 2016 16th International Conference on Control, Automation and Systems (ICCAS). doi:10.1109/iccas.2016.7832446 .

[17] Apoorva Silchar, Atul Negi ,Towards Better Drug Repositioning Using Joint Learning ,2019 IEEE https://sci-hub.se/10.1109/indicon47234.2019.9029004.

[18] Cinaroglu, S. (2019). Integrated k-means clustering with data envelopment analysis of public hospital efficiency. Health Care Management Science. doi:10.1007/s10729-019-09491-3

[19] Ma1 ,Qing Xie1 , Yongjian Liu , Shengwu Xiong A weighted KNN-based automatic image annotation method Yanchun Springer-Verlag London Ltd., part of Springer Nature 2019

[20] Hu, X., Štiglic, G., & Wang, F. (2018). Special Issue on Data Mining in Health Informatics. Journal of Healthcare Informatics Research, 2(4), 367–369. doi:10.1007/s41666-018-0039-4

[21] Manoj Kumar Gupta1 • Pravin Chandra1 A comprehensive survey of data mining Manoj Kumar, January 2020 Bharati Vidyapeeth's Institute of Computer Applications and Management https://doi.org/10.1007/s41870-020-00427-7

[22] Supriya Menon M, Rajeswari P, A Novel Approach for predicting Drug response Similarity using Machine Learning, European journal of Molecular & Clinical Medicine,2020, Volume 7, Issue 8, pp 796-808.