Enhanced Vaccine Recommender System to prevent COVID-19 based on Clustering and Classification

1st Bahaulddin Nabhan Adday
Directorate General of Education Diyala
Ministry of Education
Baghdad, Iraq
b.aladday@gmail.com

2nd Dr. Faris Ali Jasim Shaban
dept. of Electronic Engineering
Alnukhba University College
Baghdad, Iraq
faris2005@gmail.com

3rd Mohammed Rasool jawad
dept. of Network technology and info. systems
University of Babylon
Babylon, Iraq
mohammed.rasool@uobabylon.edu.iq

4th Refed Adnan Jaleel
dept. of Info. and comm. Engineering
Al-Nahrain University
Baghdad, Iraq
iraq_it_2010@yahoo.com

5th a,b Musadaq Mahir Abdel Zahra
[5a] dept. of Computer Tech. Engineering
[5a]Al-Mustaqbal University College
[5b]dept. of Electrical Engineering
[5b]University of Babylon, Iraq
musaddaqmahir@mustaqbal-college.edu.iq

2021 International Conference on Engineering and Emerging Technologies (ICEET) | 978-1-6654-2714-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICEET53442.2021.9659629

***Abstract*— Due to the fact that countries are presently dealing with the third wave of COVID-19 pandemic and in present time, the data of vaccines for preventing COVID-19 has triggered massive information, it is vital to create a system that can assist decision-makers and health care practitioners in combating COVID-19 and to combat the problem of vaccine information overload is to provide patients with personalized vaccine recommendations. Because of the ability of recommender systems (RSs) that use Collaborative Filtering (CF) to interpret decision-maker expectations, methodologies, it widely used and direct them towards linked tools that are acceptable to recommend the suitable vaccine for the persons. In this paper, we adopted an Enhanced Vaccine RSs for preventing COVID-19, which is called EVRSs-19. EVRSs-19 face some problems such as sparsity and diversity of vaccines data. To overcome these problems, we adopted two proposals. First, use clustering of K-Means to cluster the persons in several groups according to vaccine types to cope with sparsity of vaccines data. Second, use the K-Nearest Neighbors classifier-depend model of CF to discover neighbors in each vaccine cluster to increase diversity. Evaluating the EVRSs-19 system implemented on vaccines data in two testing using some metrics and the findings of these metrics after running the clustering and classification show that the system of EVRSs-19 has a perfect structure. Such as recall (0.92), precision (0.89), diversity score (9). As the vaccines recommendation list progressed, NDCG and DCG for persons are decreased.**

Keywords*— *Vaccine, COVID-19, RSs, K-Means, K-NN

I. INTRODUCTION

2.89 million verified deaths and 133 million confirmed instances of infection with the COVID-19 virus have been reported as of April 7, 2021. Virus mortality and transmission have been lowered since the start of the pandemic thanks to a variety of strategies: people' preventive steps, such as social distance, hand hygiene, and wearing face-masks; identify individuals infected with the virus through broad testing; and responses from governments about non-pharmaceutical policy, including bans on public gatherings, workplace and school closures, stay-at-home orders, and travel restrictions. Recently, with the successful evaluation, development, and production of various vaccines, governments are focusing on vaccination as a critical solution to the COVID-19 pandemic [1] [2].

Vaccines are one of the public health interventions that are cost-effective and most reliable, it's saving every year millions of lives. Pharmaceutical companies and scientists are racing against time to develop vaccines following the declaration of the pandemic by WHO in March 2020 and deciphering the genome sequence of COVID-19 in early 2020. Moderna (mRNA-1273) mRNA and Pfizer-BioNTech's (BNT162b2) vaccines in the United States it is approved for emergency usage [3]. With the encouraging news that the COVID-19 vaccine has been approved, there is growing hope that the epidemic will be ended through herd immunity [4]. According to estimates, the threshold for COVID-19 herd immunity ranges from 50 to 67%. Vaccine hesitation and skepticism among the global population are thought to be a major impediment to achieving such a goal. The WHO Strategic Advisory Group of Experts described vaccine hesitancy as a "delay in accepting or refusing immunization notwithstanding the availability of vaccination services."[5].

The RSs helps decision-makers to increase the desire of people to take the vaccines. RSs are filtering information systems applied to predict the user's preference to afford an item [6] [7]. Thus in this paper, we implement Enhanced RSs for the vaccine to prevent COVID-19 pandemic defined as ECRSs-19. When we build EVRSs-19, we faced different problems like sparsity and diversity of coronavirus data; in this paper, we shall handle these issues. Sparsity problem, in which sparse caused from big data of vaccines, this makes an obstacle on the work of EVRSs-19 and the Diversity problem, in which the weak in the diversity of results caused by the over fitting problem that is coming from memory-based CF, the system will stick in most available vaccines this will prevent other good less known vaccines from suggested. So, we have to find some way to provide diverse recommendations for a person.

The suggested EVRSs-19 predicts the suitable vaccine of the person on the foundation of various factors. The EVRSs-19 works depend on k-means clustering and K-Nearest classifier for three objects. The first object is to handle the sparsity of vaccine data and diversity of vaccines data. The second object to obtain a suitable vaccine type for the persons depend on neighbors. The third object is to increase the desire of the person to take a vaccine.

978-1-6654-3897-1/21/$31.00 ©2021 IEEE

The rest of the paper, as in section 2, describes the theoretical background and last works about RSs. K-Means and K-NN were done by authors with the specifics of techniques used by various authors. Section 3 describe the performance measuring, Section 4 shows the research methodology of the proposed EVRSs-19. Section 5 shows the result discussion of the suggested EVRSs-19. Section 6 describes the conclusions.

II. Literature Review

A. Recommender System

In the healthcare industry, the use of RSs has the potential to develop, as it may provide useful information on patients' health data and make recommendations for further medical treatment [8]. Because countries are presently dealing with the third wave of the COVID-19 pandemic, it is vital to create a framework that can assist decision-makers and healthcare practitioners to combat COVID-19 [9].

Recently, in response to the spreading of the COVID-19 pandemic, few researchers have thoroughly examined the use of RSs technology such as Reference [10] suggested a system for the improvement of the health RS, built to cater to COVID-19 symptoms' monitoring and self-assessment as well as to provide recommendations for medical and self-care treatments in Malaysia.

Reference [11] discussed the influence of the COVID-19 pandemic on the RSs environment. The authors described how using online machine learning algorithms that can identify and react to consumer profiles and behaviors changes also described how to make accurate and timely predictions about this rapidly expanding consumer sector.

Reference [12] described the recommending devices and the role of recommendation agents (RSs) in online retailing websites. Defensive flaws in present recommendation agents are also revealed, and solutions are provided for their effectiveness in COVID-19 outbreaks.

B. K-Means Clustering

K-means is unsupervised machine learning. It begins by randomly selecting k as a number of clusters and randomly selecting data points as centers of clusters. Then every point is inserted in the cluster whose center is closest to it. The closest between centers x and y computed depends on Euclidian distance, represented in equation (1).

$$d(x,y) = \sqrt{(x_1 - y_n)^2 + \cdots + (x_n - y_n)^2} \qquad (1)$$

By means of the points of the cluster, the centers of the clusters are updated and replaced. The process is iterated until reaching the stability of the clusters [13]. K-means can handle a lot of data. The simplicity of its implementation is also an advantage of the algorithm. K-means, on the other hand, has several drawbacks. The selection of the integer k can alter the outcomes [14]. The average of the points within the cluster causes early convergence and makes the method susceptible to outliers and noise. It also imposes a limitation on non-numerical data types like ordinal and categorical data [15].

Many studies used k-mean for COVID-19 pandemic such as Reference [16] attempted to identify the affected Union Territories and Indian states by COVID-19 using the K-means clustering method based on the secondary sources of data collected from the Indian Health and Family welfare Organization until March 24, 2020. It produced an effective result and visualized their work.

Reference [17] discussed the grouping of COVID-19 cases and deaths using the K-Means clustering in Southeast Asia. And utilizing Rapid Miner tools, the data were clustered into various clusters using the K-Means Clustering Process. Data that have been used are 2020 deaths from WHO for April of the year 2020, statistics of the country, and confirmed cases of COVID-19.

Reference [18] intended to map COVID-19 cases in Indonesia's provinces using the K-means clustering method. It had been an initial attempt to inform the public and raise awareness of the disease spread. It had been hoped that this study might assist towards optimal handling of the pandemic in Indonesia.

C. K-Nearest Neighbors Classifier

One of the most basic supervised machine learning algorithms is K-Nearest Neighbors (K-NN). Because similarity measurements such as Euclidean distance are used to assign test samples to the K closest training sample classes, this method is computationally simple [19][20].

K-NN is called sometimes instance-depend learner, there is no training phase, and because it postpones training until a test sample needs to be categorized [21] [22]. K-NN is also utilized to resolve nonlinear problems, such as customer rankings for banks and credit rating agencies, when the data gathered does not always conform to the linear theory. Therefore, when there is little or no prior knowledge about the distribution data, K-NN should be one of the first choices. [23].

Different prediction and classification algorithms are used to study the possibility of spreading COVID-19, such as Reference [24] studied the various classification algorithms of machine learning to forecast the COVID-19 deceased and recovered cases. The authors discovered that the K-NN classification algorithm outperforms other algorithms in terms of accuracy.

Reference [25] implemented a new COVID-19 diagnosis strategy based on an enhanced K-NN classifier and hybrid feature selection Methodology.

Reference [26] using an actual dataset acquired from multiple Iraqi hospitals through a questionnaire produced and distributed for 290 patients, the K-NN algorithm was used to build a classification model for predicting patients' state.

The difference between this paper & the related works is that we propose RS for vaccine to fight COVID-19 depend on integrating clustering (K-Means) and classification machine learning (K-NN) algorithms, which consider the main contribution of this paper.

III. Performance Measuring

A. Precision and Recall

In decision support, each aspect of the system is evaluated to determine if it was right or wrong. If you look at each recommended item in an RS and compare it to a user's actual consumption, you can get one of four results: It is possible for an item to be suggested or not, and for the consumer to have consumed it or not. If the item is recommended by the recommender, it is deemed positive. We say it was the proper decision if the user consumes the item. [27]. The item is

considered positive if it has been recommended, and true if both the recommender and the user agree on its value, so you get these outcomes [28]:

- True Positive (TP): item consumed and recommended by the person.
- False Positive (FP): item didn't consume and was recommended.
- False Negative (FN): the person consumed it and the recommender didn't include the item in a recommendation.
- True Negative (TN): the person didn't consume it and it wasn't recommended.

In Fig. 1, this is shown as a table and show how the precision and recall have been computed. Precision defined as what fraction of the recommended items the person consumed as in the following equation (1) [29]:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive + False Positive}} \quad (1)$$

While Recall defined as what, out of all the items that the user consumed, was recommended as in the following equation (2) [29]:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive + False Negative}} \quad (2)$$

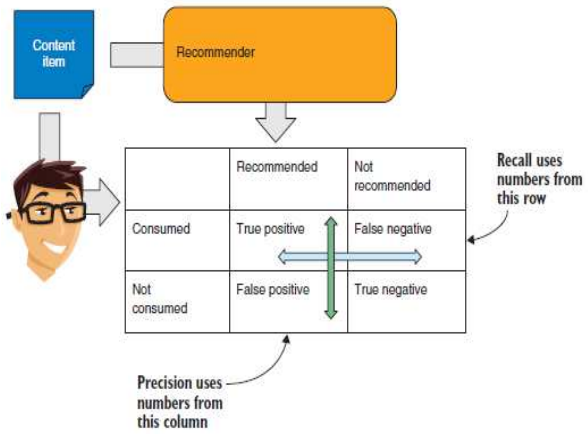

Fig. 1. How precision and recall are calculated

B. DCG and NDCG

Discounted cumulative gain (DCG) and Normalized DCG (NDCG) measures the quality of recommendations. DCG represents the relevancy of each recommendation to its position on the list. So, the DCG gain involves discounting the relevance score by dividing it with the log of the corresponding position, according to that DCG take into consideration the position of each item. If the DCG is high that's mean the list is better in quality. The DCG for each recommended item has to be decreased as we are moving on the list. Every person may receive a different number of recommendations. As a result, the DCG will change. We need a score with correct upper and lower bounds to average all of the suggestions scores to come up with a final score. DCG can be calculated using the equation (4) [30] [31]:

$$\text{DCG@k} = \sum_{i=1}^{k} \frac{\text{relevance}}{\log2(i+1)} \quad (3)$$

The NDCG brings about this normalization. DCG of the suggested order and DCG of the ideal order (IDCG) must be computed for each recommendation set to determine NDCG. NDCG is then computed as the ratio of DCG of recommended order to DCG of ideal order as shown in formula (5); this ratio will always be in the range [0, 1] [32].

$$\text{NDCG} = \frac{\text{DCG}}{\text{IDCG}} \quad (4)$$

IV. EVRSs-19 Methodology

The EVRSs-19 with the CF technique has been proposed to result in the most linked recommendations about vaccines by tackling significant challenging problems such as diversity and sparsity. We are taken data of seven vaccine types, which are Sputnik V, AstraZeneca-Oxford, Pfizer, Moderna, Novavax, Johnson & Johnson, and Sinopharm from [1]. The flowchart of the suggested EVRSs-19 algorithm is shown in Fig. 1 that boils down to the following steps:

1. In the beginning, the EVRSs-19 call K-Means clustering algorithm; Set the random number of clusters (k) for grouping persons using vaccine types data, set centroid person randomly from all the persons found in vaccine data for each vaccine cluster, set or join each person inside the nearest vaccine cluster, the closest between centroid and person computed depend on Euclidean distance. The centroid of each vaccine cluster is updated by taking the mean of all persons in each cluster, updated continuously until the vaccine cluster becomes stable.
2. Finally, the EVRSs-19 system call K-NN algorithm for each vaccine cluster; set a number of nearest neighbours (k) initially, calculate Euclidean distance between the new person and each person in vaccine clusters, order the distance to get the k nearest neighbors and then do majority voting about the suitable vaccine type for the new person.

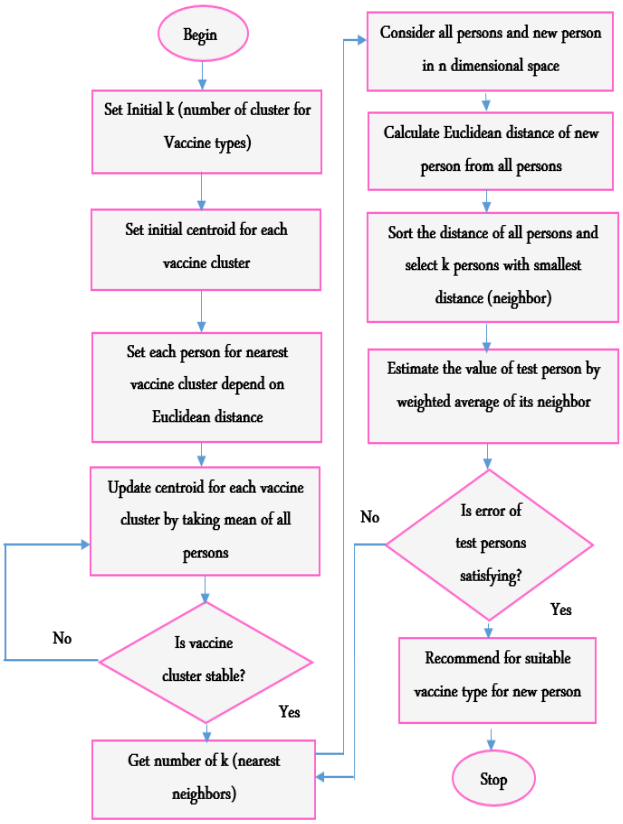

Fig. 2. Flowchart of EVRSs-19

V. RESULTS ANALYSIS

In this section, some of the results from implementing suggested EVRSs-19 and evaluating postposed systems will be shown, which have implemented using Java packages.

A. Elbow and K-Means Clustering

Before implementing the K- Mean clustering, which is solved for the sparsity problem, there is a way to check the optimum cluster number that must be recommended. This testing technique is known as the technique of the elbow. The elbow technique specifies the optimal number of clusters to be adopted with the highest variance between data that belong to the different clusters of vaccine types. So the system needs a seven k-mean clustering process, one for each vaccine type. The first elbow calculation plot is to specify the optimal number of clusters for vaccine types. All of those elbow plots are referring to that the best number of clusters is 7. The elbow and graph of clustering the users under all vaccine types are shown in Fig. 3 and Fig. 4, respectively.

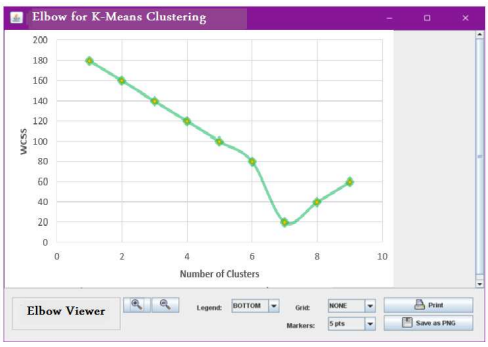

Fig. 3. Elbow method for Vaccine types clustering.

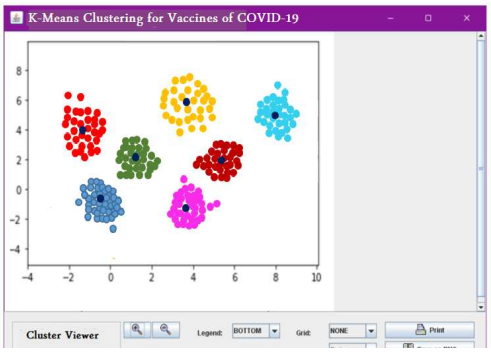

Fig. 4. K-Means Clustering for Vaccines of COVID-19

B. Evaluating of EVRSs-19

Finally, two tests after generating the EVRSs-19 recommendation list of vaccines for evaluating the proposed system have been implemented by using some factors such as Precision, Recall, Diversity Score, DCG, and NDCG. The first test contains 25 closest and 25 furthest, and the second test contains 50 closest and 50 furthest. In the first test, the Recall and Precision equal 0.89 and 0.92, respectively, while in the second test, the Recall and Precision equal to 0.91, as shown in Fig. 5. The diversity score in the first test and second test equal 8.79 and 9, respectively, as shown in Fig. 6.

DCG score is essential to understand the quality of RS. In the test operation for DCG, one persons has been taken as a test sample, DCG has been registered for all seven vaccines in the recommendations list see Fig. 7. it's clear that person1 got

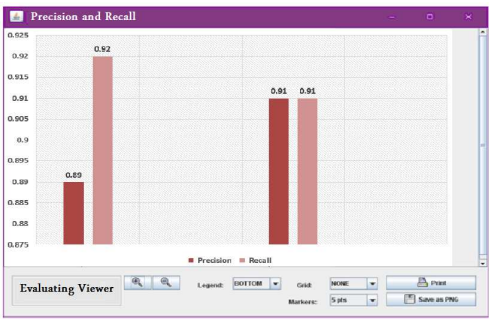

Fig. 5. Precision and Recall

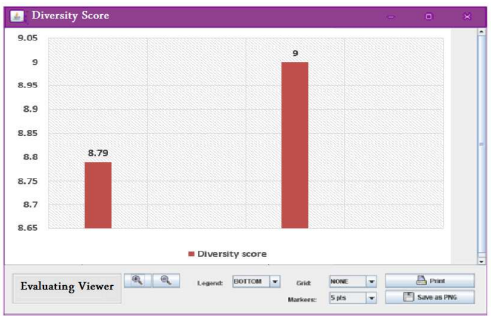

Fig. 6. Diversity Score

the most relevant high ranked vaccines in the list while the scores of the vaccines are decreased gradually. In the second test, when to increase the number of neighbor persons to 50, DCG not differs from the previous one it's like it precisely, and this is one of the system advantages.

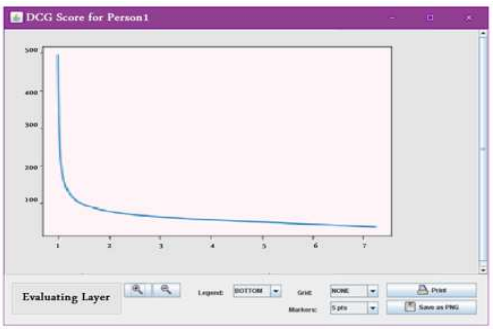

Fig. 7. DCG Score of Person1

Now to compute the NDCG, the results shown in Fig. 8 show that the NDCG for person1 is decreased as the recommendation list of vaccines progressed; the first three vaccines are ranked as necessary between 0.89 and 0.97, which is good relevancy, the relevancy decreased when the list progressed, and the person will show the first four vaccines for example.

In the second test, NDCG is decreasing as the recommended list progressed, which refer to the decreased of relevancy as the recommended list progressed as shown in Fig. 9 for person1, NDCG decreasing from 1.00 to 0.87 for the first three vaccines results in the recommendation list, the x-axis represents the index of the suggested vaccines. In contrast, the y-axis represents the NDCG score for each vaccine. Results showing that the good order for the recommendation list given to the person by the proposed system.

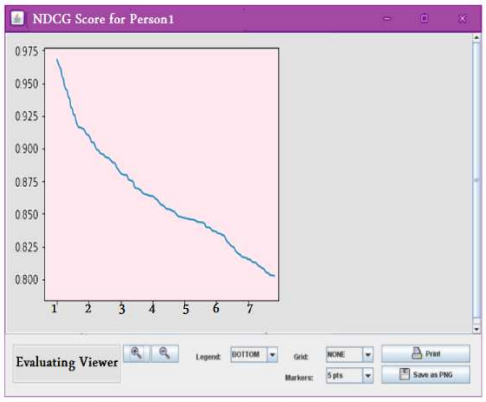

Fig. 8. Test 1 for NDCG Score of Person1

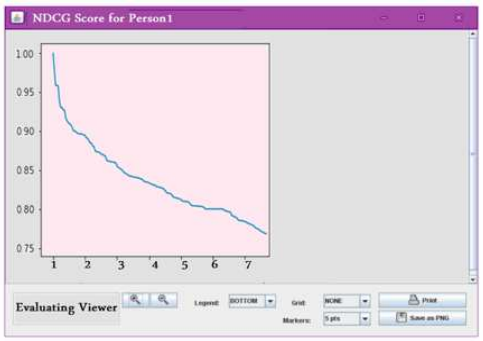

Fig. 9. Test 2 for NDCG Score of Person1

VI. CONCLUSIONS

The proposed EVRSs-19 shows improved efficiency by conducting the k-means on all persons, then implementing K-NN (CF) on this vaccine clusters to handle diversity without affecting accuracy. Sparsity has been neutralized using clustering without any imputation techniques. The decreasing of the NDCG and DCG is a good indicator because it shows that the high relevance suggestion located in the first part of the vaccines recommendation list and the relevance decreased when the list of vaccines progress, also the little decreasing in the NDCG represent the closeness between DCG and IDCG this means that the vaccines recommendation list has a good organization.

REFERENCES

[1] E. Mathieu, H. Ritchie, E. Ortiz-Ospina, et al., "A global database of COVID-19 vaccinations", *Nature human behaviour, vol. 5, pp. 947-953, July 2021.*

[2] O. Sharma, A. A.Sultan, H. Ding, and C. R.Triggle, "A Review of the Progress and Challenges of Developing a Vaccine for COVID-19", Frontiers in immunology, vol. 11, 2413, 14 October 2020.

[3] A. A.Malik, S. M.McFadden, J. Elharake, and S. B.Omer, "Determinants of COVID-19 vaccine acceptance in the US", EClinicalMedecine, vol. 26, 100495, September 2020.

[4] B. F.Haynes, et al., "Prospects for a safe COVID-19 vaccine", Science Translational Medicine, vol. 12, issue 568, eabe0948, Novmber 2020.

[5] B. S.Graham, "Rapid COVID-19 vaccine development", Science, vol. 368, Issue 6494, pp. 945-946, May 2020.

[6] F. Pajuelo-Holguera, J. A.Gómez-Pulido, F. Ortega, and J. M. Granado-Criado, "Recommender system implementations for embedded collaborative filtering applications", Microprocessors and Microsystems, vol. 73, 102997, March 2020.

[7] B. Walek and V. Fojtik , "A hybrid recommender system for recommending relevant movies using an expert system", Expert Systems with Applications, vol. 158, 113452, November 2020.

[8] H. Kaur, N. Kumar, and S. Batra, "An efficient multi-party scheme for privacy preserving collaborative filtering for healthcare recommender system", Future Generation Computer Systems, vol. 86, pp. 297-307, September 2018.

[9] S. Jayita, C. Chandreyee and B. Suparna, "Review of machine learning and deep learning based recommender systems for health informatics", Deep Learning Techniques for Biomedical and Health Informatic, Studies in Big Data, vol. 68, pp. 101-126, 2020.

[10] M. Othman , N. Muhd Zain , Z. Paidi, and F. Amir Pauzi, "Framework of Health Recommender System for COVID-19 Self-assessment and Treatments: A Case Study in Malaysia", IJCSNS International Journal of Computer Science and Network Security, vol.21, No.1, pp. , 12-18, January 2021.

[11] R. Abdulrahman and H. L.Viktor, "Personalised Recommendation Systems and the Impact of COVID-19: Perspectives, Opportunities and Challenges", KDIR, pp. 295-301, 2020.

[12] M. Nilashi, S. Asadi, R. Ali Abumalloh, S. Samad, and O. Ibrahim, "Intelligent Recommender Systems in the COVID-19 Outbreak: The Case of Wearable Healthcare Devices", Journal of Soft Computing and Decision Support Systems, vol.7, No.4, pp. 8-12, June 2020.

[13] S. Hussein Toman, M. Hamzah Abed, and Z. Hussein, "Cluster-based information retrieval by using (K-means)- hierarchical parallel genetic algorithms approach", TELKOMNIKA Telecommunication, Computing, Electronics and Control, vol. 19, No. 1, pp. 349~356. February 2021.

[14] P. Govender and V. Sivakumar, "Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019) ,"Atmospheric Pollution Research, vol. 11, pp. 40–56, 2020.

[15] A. Poompaavai and G. Manimannan, "Clustering study of Indian states and union territories affected by coronavirus (COVID-19) using k-means algorithm" International Journal of Data Mining And Emerging Technologies, vol. 9, pp. 43-51, 2019.

[16] Poompaavai A. and Manimannan G, "Clustering study of Indian states and union territories affected by coronavirus (COVID-19) using k-means algorithm", International Journal of Data Mining And Emerging Technologies, vol. 9(2), pp. 43-51, 2019.

[17] J. Hutagalung, et al., "COVID-19 Cases and Deaths in Southeast Asia Clustering using K-Mean Algorithm", Journal of Physics: Conference Series, Annual Conference on Science and Technology Research (ACOSTER), vol. 1783, No. 1, pp. 012027, 2020.

[18] F. Virgantari and Y. Erika Faridhan , "K-Means Clustering of COVID-19 Cases in Indonesia's Provinces", Proceedings of the International Conference on Global Optimization and Its Applications Jakarta, Indonesia, pp. 21-22, November 2020.

[19] S. Zhi-gang, et al. "A distributed rough evidential K-NN classifier: integrating feature reduction and classification." IEEE Transactions on Fuzzy Systems , 2020.

[20] A. mit Kumar, et al. "A Novel K-Means Clustering and Weighted K-NN Regression Based Fast Transmission Line Protection." IEEE Transactions on Industrial Informatics , 2020.

[21] X. Truong, et al. "An efficient sampling algorithm with a K-NN expanding operator for depth data acquisition in a LiDAR system." IEEE Transactions on Circuits and Systems for Video Technology 30.12 , 2020: 4700-4714.

[22] I. Hakim, et al. "Sentimen Analisis Stay Home menggunakan metode klasifikasi Naive Bayes, Support Vector Machine, dan k-Nearest Neighbor." Paradigma-Jurnal Komputer dan Informatika 22.2, 2020: 169-174.

[23] M. Nikooghadam, et al. "COVID-19 Prediction Classifier Model Using Hybrid Algorithms in Data Mining." International Journal of Pediatrics 9.1 , 2021: 12723-12737.

[24] P. Theerthagiri, I. Jeena Jacob, A. Usha Ruby, and Y. Vamsidhar, "Prediction of COVID-19 Possibilities using KNN Classification Algorithm", International Journal of Current Research and Review, vol. 13, 156, 2021.

[25] W. M. Shaban, A. H. Rabie, A. I. Saleh, and M.A. Abo-Elsoud, "A new COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier", Knowledge-Based Systems, 205: 106270., July, 2020.

[26] R. Adnan Jaleel, I. Mohammed Burhan, and A. Mohaisen Jalookh, "A Proposed Model for Prediction of COVID-19 Depend on K-Nearest Neighbors Classifier:Iraq Case Study", In Press, Proc. of the 3rd

International Conference on Electrical, Communication and Computer Engineering (ICECCE) , 12-13 June 2021, Kuala Lumpur, Malaysia.

[27] K. Haruna, et al., "Research paper recommender system based on public contextual metadata, Scientometrics, vol. 125, pp. 101-114, 2020.

[28] S. Tiwari, "Implicit Preferences Discovery for Biography Recommender System Using Twitter", Procedia Computer Science, vol. 167, pp. 1411-1420, 2020.

[29] M. Sharma , L. Ahuja, and V. Kumar, "A Novel Rule based Data Mining Approach towards Movie Recommender System", Journal of Information and Organizational Sciences, vol. 44, No. 1, 2020.

[30] M. Chen and P. Liu,"Performance Evaluation of Recommender Systems", International Journal of Performability Engineering, 2017, vol. 13, Issue (8), pp. 1246-1256, 2017.

[31] M. S Hegde, G. Krishna , and Dr. R. Srinath, "An Ensemble Stock Predictor and Recommender System", International Conference on Advances in Computing, Communications and Informatics (ICACCI), . IEEE, pp. 1981-1985, September 2018.

[32] S. Forouzandeh, K. Berahmand, and M. Rostami "Presentation of a recommender system with ensemble learning and graph embedding: a case on MovieLens", Multimedia Tools and Applications, vol. 80, pp. 7805–7832, 2021.