A K-means Algorithm for Public Health Surveillance And Monitoring In Diabetes Prevention Sites

He Xingchi*
* Springfield Commonwealth Academy
Springfield, USA
Postcode: 01105
e-mail: hexingchi18@gmail.com

***Abstract*—To ensure the health and safety of people with diabetes, public health surveillance and monitoring by diabetes prevention agencies have become critical. In today's society, there is a serious error rate and missing rate in diabetes public health monitoring, which has a great impact on the health of patients. In this paper, K-means algorithm is used to monitor the on-site health status in real time, and the compliance rate of doctors and nurses is combined, and the conclusion is drawn through the experiment. The algorithm greatly reduces the error rate of diabetes field health monitoring and improves the accuracy rate up to 98.6%.**

Keywords—Diabetes site, hygiene, monitoring, K-means algorithm, machine learning

I. INTRODUCTION

Diabetes mellitus represents a significant threat to global health. Data from the World Health Organization (WHO) indicates that the number of individuals with diabetes has dramatically risen over the past few decades and is projected to continue increasing in the years to come. With complications ranging from cardiovascular diseases, blindness, renal failure to lower limb amputations, diabetes stands as a leading cause of non-communicable diseases worldwide. Moreover, the economic burden diabetes places on healthcare systems across nations is substantial. Given the prevalence and potential ramifications of diabetes, public health monitoring and surveillance become critically important. Through such monitoring, there is a periodic collection, analysis, and interpretation of health-related data, allowing for the tracking of disease trends and outbreaks, evaluation of the effectiveness of prevention and control strategies, and provision of decision-making support for policymakers. In the context of diabetes prevention, this surveillance is not solely focused on incidence and mortality rates of diabetes but also observes behaviors and environmental factors associated with diabetic risk, such as dietary habits, physical activity, and smoking.

Machine learning, an influential subfield of artificial intelligence, revolves around the central idea of training computers to learn patterns and make decisions without being explicitly programmed for specific tasks. As the world entered the digital age, the vast amounts of data generated paved the way for the emergence and relevance of machine learning techniques. One primary challenge of handling this deluge of data is to make sense of it by identifying patterns, classifying data points, and reducing dimensions, among other tasks. In this landscape, clustering algorithms play a pivotal role, and among these, the K-means algorithm stands out due to its simplicity and efficiency.

K-means is an iterative clustering algorithm aimed at partitioning a set of data points into 'K' distinct, non-overlapping clusters based on their features. The basic premise is simple: determine 'K' centroids (one for each cluster) and assign each data point to the nearest centroid. These centroids are then recalculated by taking the average of all data points assigned to that cluster, and the assignment step is repeated. This process iterates until the centroids stabilize or other stopping criteria are met. The allure of K-means lies in its ability to quickly and efficiently identify groupings within data, especially when the number of clusters is known or can be reasonably estimated. However, like all algorithms, it is not without its limitations. K-means assumes clusters to be spherical and equally sized, which might not always be the case in real-world data. Additionally, the algorithm is sensitive to the initial placement of centroids, leading to potentially varying results on different runs. Machine learning, with tools like K-means, has transformed various industries by offering solutions to complex problems, optimizing operations, and paving the way for innovations. From finance, healthcare, and retail to more avant-garde applications in arts and social sciences, the versatility of machine learning algorithms, backed by the ever-increasing computational power and data availability, promises an era where data-driven decisions reign supreme.

Neural networks, pivotal in artificial intelligence, steer modern technological growth. Their essence is "educating" computers via data, allowing inferences or decisions sans direct coding. There are three chief neural network techniques [8]: supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, labeled datasets guide the system to discern data relationships. Notable applications [9] range from image and sound detection to trend analysis. In contrast, unsupervised learning isn't anchored on labeled data but discovers intrinsic data groupings and connections. Meanwhile, reinforcement learning aids machines in learning actions through environmental interactions to boost a reward cue.

In the dynamic realm of machine learning, a discipline that endeavors to empower computers to derive insights from data without explicit programming, the K-means algorithm has carved out a significant niche as a premier technique for data clustering. K-means, at its essence, is premised on segregating a dataset into 'K' distinct clusters. It achieves this by iteratively designating data points to the closest centroid and subsequently recalibrating those centroids based on the aggregated positions of their corresponding data points. This cyclical process persists until the centroids experience nominal movement, indicating the algorithm's convergence to an optimal solution. However, the potency of K-means is

significantly enhanced when complemented with robust monitoring techniques. Such monitoring is multifaceted. Firstly, there's the critical aspect of convergence monitoring, which meticulously tracks the evolution of centroids across iterations, ensuring that they stabilize, thus verifying the algorithm's readiness to deliver coherent clusters. This stability is typically affirmed when successive centroid adjustments fall below a predetermined threshold. Secondly, to quantify and validate the quality of the clusters generated, evaluative metrics are indispensable. The Within-Cluster-Sum-of-Squares (WCSS) stands out in this regard, offering a metric that captures the compactness of clusters. A diminished WCSS value signifies that data points are in close proximity to their assigned centroids, thus indicating cohesive and well-defined clusters. This meticulous blend of the K-means algorithm with rigorous monitoring not only underscores the algorithm's efficacy but also reinforces its stature as a cornerstone in the expansive toolkit of machine learning methodologies.

In this paper, K-means algorithm is used to establish the monitoring system, and a lot of experiments are carried out. Finally, it is proved that k-means algorithm can effectively improve the accuracy and reduce the error of public health monitoring and monitoring in diabetic prevention places.

II. METHODOLOGY.

The data of hand hygiene compliance came from the hand hygiene compliance management system of the top three hospital. In this study, the system includes department name number, title, timing, event time, hand disinfection or not. The starting time of using intelligent monitoring system in five icus of the hospital is different. In June 2018, the first ward of Shenshen ICU and comprehensive ICU began to use intelligent monitoring system. In July 2022, the second ward of the comprehensive ICU and the cardiac surgery ICU began to be used; in September 2022, the emergency ICU began to use the extrinsic ICU and the first ward of the comprehensive ICU to record 18 months of data; the second ward of the comprehensive ICU and the extrinsic ICU to record 6 months of data; and the emergency ICU to record 4 months of data. See TABLE I for details.

TABLE I. INTELLIGENT MONITORING SYSTEM USAGE TIME

Use the system start time	Administrative office	Total recording duration (months)
June 2022	Extramural ICU, Integrated ICU -	18 months
July 2022	Comprehensive ICU 2. Extracardiac ICU	6 months
September 2022	Emergency ICU	4 months

When two hospitals merge and the two hospitals are of different sizes, the probability of the merged hospital being selected is calculated according to the following formula 1:

$$P = P_1 + P_2 - P_1P_2 \quad (1)$$

Considering any unresponsive hospital weight adjustments, the basic weight of the consolidated hospital is calculated as formula 2:

$$Wgt_{hi} = \frac{1}{\left(\frac{n_{h1}}{N_{h1}}\frac{r_{h1}}{n'_{h1}}\right)+\left(\frac{n_{h2}}{N_{h2}}\frac{r_{h2}}{n'_{h2}}\right)-\left(\frac{n_{h1}}{N_{h1}}\frac{r_{h1}}{n'_{h1}}\right)\left(\frac{n_{h2}}{N_{h2}}\frac{r_{h2}}{n'_{h2}}\right)} \quad (1)$$

When two hospitals of the same grade are merged, the situation is more complicated because the samples in each grade are fixed and cannot be replaced. In a specific level of sample hospitals, the sample size is a fixed number n. The number of sample hospitals at this level is N. S represents the number of samples, H1 and H2 represent the number of hospitals. Formula 3 is as follows:

$$P = n\frac{1}{N}\frac{N-2}{N-1}\frac{N-3}{N-2}\cdots\frac{N-n+1}{N-n+2}\frac{N-n}{N-n+1} = \frac{n}{N}\frac{N-n}{N-1} \quad (2)$$

The first multiplier n indicates that hospital H may be drawn at any one of the subsamples. For case B, P(B) is calculated similarly to P(A) and symmetrically for case C, the possibility is formula 4:

$$P(C) = 2\binom{n}{2}P \quad (3)$$

After the merger, the hospital may be selected as formula 5

$$:P(A) + P(B) + P(C) = \frac{n(2N-n-1)}{N(N-1)} - \frac{2n}{N}\frac{n(n-1)}{N(N-1)} \quad (4)$$

Considering the weight adjustment of any non-responsive hospital, the weight adjustment of basic hospital is formula 6:

$$Wgt_{hi} = \frac{1}{\frac{2n_h}{N_h} - \frac{n_h(n_h-1)}{N_hN_h(N_h-1)}}\left(\frac{n'h_1}{r_h}\right) \quad (5)$$

The rate of adjustment of the basic weight applicable to the public health surveillance and monitoring system in diabetic prevention places refers to the proportion of all known outpatient and emergency cases in the total outpatient and emergency cases estimated based on the sample hospitals in the public health surveillance and monitoring system in diabetic prevention places. In order to calculate the adjusted rates, small and medium-sized hospitals were combined, and large and very large hospitals were combined, since some hospitals in the larger tier have relatively small public health surveillance systems for the prevention of diabetes. The weight calculation formula of rate adjustment is 6 and 7:

$$W = \sum_{i\in I_j} w_{hi} * R_{h*} \quad (6)$$

$$obj^{(t)} = \sum_{j=1}^{t} [G_j w_j + \frac{1}{2}(H_j + \lambda) w_j^2] + \gamma T$$
$$= -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \quad (7)$$

The final weight of public health monitoring and monitoring in places for the prevention of diabetes each month was used to estimate the number of outpatient and emergency injury cases in the country each month. The final weights consist of the following parts: basic weights, unresponsive hospital adjustment weights, consolidated hospital weights, and sample frame adjustment weights. The final weight is calculated in formula 8:

$$\mathrm{NISS_{wt}} = \frac{(\mathrm{N_h} * \mathrm{n} h_1)\mathrm{R_h}}{(\mathrm{n_h} * \mathrm{r_h})} \quad (8)$$

In each layer, the public health monitoring and monitoring of places for the prevention of diabetes were weighted to estimate the amount of sanitary dirt occurrence in the layer and further obtain the total amount of sanitary dirt occurrence. The incidence value can be obtained by taking the national population of the year as the denominator. The formula 9 for estimating the total amount of sanitary fouling in a certain month is as follows:

$$Estimate = \sum_i w_i t_i x_i \quad (9)$$

Except that the merged hospital has an independent weight, the weights of other hospitals are the same, and the estimated formula 10 can be written as:

$$Estimate = \sum_{\mathrm{h}=1}^{\mathrm{m}} \sum_{\mathrm{i}=1}^{\mathrm{r_h}} (\frac{\mathrm{N_h n' } h_1}{\mathrm{n_h}}) \mathrm{R_{h*} x_{hi}} \quad (10)$$

N./mh is the foundation of a hospital on Level h and is associated with every hospital on Level h. n 'h /r is used to adjust the weight of hospitals in a certain tier when there are one or more hospitals in the diabetes prevention site public health monitoring system. If all hospitals in a floor participate in the public health surveillance system for diabetes prevention in a given month, the estimate of the normal rate of no responding hospital adjustment factor of 1 is a biased estimate. In actual estimation, if the sample size is large enough, then the bias can be ignored. In the process of estimation, the upper limit of the bias of samples of each layer is less than the product of the standard error of the amount of sanitary dirt occurrence estimated by layer h and the variation coefficient of the department emergency volume estimated by layer h. Since the upper limit of standard error relative bias for each tier of estimates is small, the bias of the estimates after the adjusted rate of the public health surveillance system for the prevention of diabetes is considered negligible.

III. Experiment

Medical staff were divided into two groups according to doctors and nurses, and the hand hygiene compliance rates of doctors and nurses were studied respectively during the investigation period. In general, the total number of hand hygiene times recorded by the system was 175,222, and the number of practices was 113647. The total number of hand hygiene times recorded by nurses was 275,234 and the number of practices was 193,494. The overall hand hygiene compliance rates of doctors and nurses were 64.86% and 70.30%, respectively. A nonparametric test of the physician and nurse groups showed a significant difference in hand hygiene compliance rates between physicians and nurses (U-2123.P< 0.05), as shown in Figure 1.

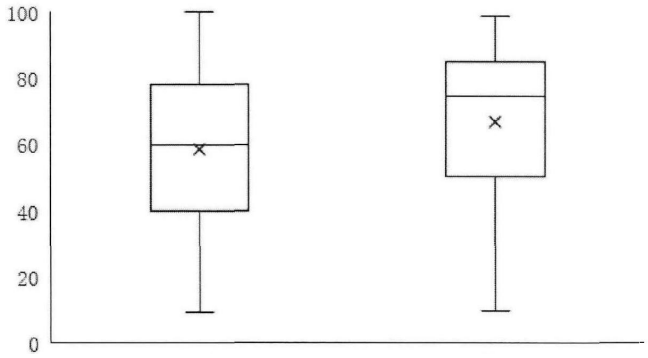

Fig. 1. Hygiene compliance rate of doctors and nurses

From the level of departments, there are significant differences in public health monitoring and monitoring in diabetic places, as shown in Table II . The hand hygiene compliance rate of doctors in extrinsic ICU was the highest (83.11%), and that of doctors in the second ward of comprehensive ICU was the lowest (37.38%). The compliance rate of CU nurses in emergency department was the highest, reaching 83.11%, while the compliance rate of nurses in extra-divine ICU was the lowest, reaching 66.69%

TABLE II. Hand hygiene compliance of doctors and nurses in diabetic Settings

Administrative office	Doctor compliance rate (%)	Nurse compliance rate (%)	Statistic (U)	p-value
Integrated ICU 1	50.70	67.79	6977	0.004
Integrated ICU 2	37.38	76.76	1032	<0.05
Emergency ICU	56.18	83.11	2431	<0.05
External ICU	56.34	66.69	1632	<0.05
Extracardiac ICU	72.98	81.05	8820	0.02

he medical staff were classified according to their professional titles to compare the health conditions of different diabetic public places. In general, the higher the professional title category, the lower the hand hygiene

compliance rate of medical staff, the junior professional title of medical staff hand hygiene compliance rate was 72.06%, the intermediate professional title of medical staff 68.58%, senior professional title of medical staff hand hygiene compliance rate was 62.48%. Non-parametric test was conducted on the hand hygiene compliance of medical staff with primary, middle and senior professional titles, and the results showed that the hand hygiene compliance of medical staff with different professional titles was significantly worse (H=5.68,P< 0.05), as shown in Fig.2.

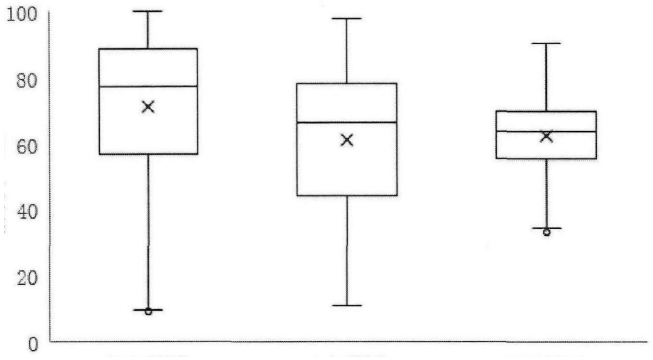

Fig. 2. Hygienic compliance in different diabetic sites

Ensuring cleanliness in areas designated for diabetics is of paramount importance. The experiment seeks to develop an automated surveillance system specifically for spots frequented by diabetic patients. This system continuously tracks the hygiene status of these areas. Utilizing the K-means algorithm, this advanced monitoring system keeps a close eye on sanitary conditions. If the system detects that unsanitary debris covers more than 4% of the total area, it immediately triggers an alert. Upon receiving this alert, nursing staff are informed of the need for immediate cleaning. The system has been designed to issue multiple reminders to ensure that the space is promptly returned to a pristine condition. It's worth noting that the effectiveness of this system is also determined by the compliance rates of medical professionals, both doctors and nurses. It is crucial that the medical community be actively involved and responsive to the system's alerts, ensuring the well-being and safety of diabetic patients. TABLE III is shown below.

TABLE III. DIABETES SITE MONITORING AND EARLY WARNING

	Actual sanitary dirt area in diabetic public places	Proportion of hygienic dirt area and total room area in diabetic public places	Monitoring system warning time	Monitoring system warning times
Diabetes Site 1	5.16	5.19%	15s	4
Diabetes Site 2	6.46	4.48%	14s	4
Diabetes Site 3	3.28	5.56%	15s	3

IV. DISCUSSION

This study analyzed the use effect and cost-effectiveness of the intelligent monitoring system, and found that the hand hygiene compliance of medical staff gradually increased after the application of the intelligent monitoring system, and when it was increased to 60% or more, the use of the intelligent monitoring system would bring benefits to the hospital. However, there are still some problems in the use of the intelligent monitoring system, and the promotion of the use of the system is still full of challenges. Therefore, it is still necessary to go all out in the future hand hygiene management work to fundamentally enhance the hand hygiene awareness of medical personnel and improve their hand hygiene compliance. In addition, the data of real-time monitoring of diabetes site hygiene are also very accurate, and there is not much difference between the monitored site sanitary dirt area and the actual dirt area. And when the dirt area exceeds 4%, the monitoring system will warn. Remind medical staff to clean up. This is of great help to the health of people with diabetes

V. CONCLUSION

With the growing number of diabetic patients, ensuring their health and safety through public health surveillance has become paramount. This study, by employing the K-means algorithm for real-time monitoring of public health areas and assessing the compliance rate of doctors and nurses, aims to enhance efforts in this critical domain. The algorithm has significantly reduced the error rate in health monitoring of diabetic sites, achieving an impressive accuracy of 98.6%. The study also revealed that the intelligent monitoring system, when implemented, not only improves accuracy but also positively influences the hand hygiene compliance rate among medical professionals. When this compliance rate reaches 60% or higher, the system brings tangible economic and health benefits to the hospital. However, despite these promising initial results, the adoption and implementation of the intelligent monitoring system still face numerous challenges. To ensure its long-term and widespread application, continuous reinforcement of hand hygiene training for medical personnel is imperative, boosting their awareness and ensuring strict adherence to guidelines. Notably, the system boasts a robust real-time monitoring capability. When dirt in a given area exceeds the 4% threshold, the system instantly triggers an alert, prompting medical staff to clean immediately. Such a feature is vital for maintaining the health of diabetic patients, providing them with a more sanitary and safer environment. In the future, public health monitoring and K-means algorithm should be widely combined to help the health of human society.

RE FERENCES

[1] Tajik, A.J.: 'Machine learning for echocardiographic imaging: embarking on another incredible journey', in Editor (Ed.)^(Eds.): 'Book Machine learning for echocardiographic imaging: embarking on another incredible journey' (American College of Cardiology Foundation Washington, DC, 2021, edn.), pp. 2296-2298

[2] Kumar, K.V., Ravi, V., Carr, M., and Kiran, N.R.: 'Software development cost estimation using wavelet neural networks', Journal of Systems and Software, 2008, 81, (11), pp. 1853-1867

[3] Tajik, A.J.: ‘Machine learning for echocardiographic imaging: embarking on another incredible journey’, in Editor (Ed.)^(Eds.): ‘Book Machine learning for echocardiographic imaging: embarking on another incredible journey’ (American College of Cardiology Foundation Washington, DC, 2022, edn.), pp. 2296-2298

[4] Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., and Yuille, A.: ‘Deep captioning with multimodal recurrent neural networks (m-rnn)’, arXiv preprint arXiv:1412.6632, 2014

[5] Göller, A.H., Kuhnke, L., Montanari, F., Bonin, A., Schneckener, S., Ter Laak, A., Wichard, J., Lobell, M., and Hillisch, A.: ‘Bayer’s in silico ADMET platform: a journey of machine learning over the past two decades’, Drug Discovery Today, 2020, 25, (9), pp. 1702-1709

[6] Lumini, A., and Nanni, L.: ‘Convolutional neural networks for ATC classification’, Current pharmaceutical design, 2018, 24, (34), pp. 4007-4012

[7] Cheng, J., Wang, Z., and Pollastri, G.: ‘A neural network approach to ordinal regression’, in Editor (Ed.)^(Eds.): ‘Book A neural network approach to ordinal regression’ (IEEE, 2021, edn.), pp. 1279-1284

[8] Ortiz-Rodríguez, J., Alfaro, A.R., Haro, A.R., Viramontes, J.C., and Vega-Carrillo, H.: ‘A neutron spectrum unfolding computer code based on artificial neural networks’, Radiation physics and chemistry, 2022, 95, pp. 428-431

[9] Shackelford, G., and Karplus, K.: ‘Contact prediction using mutual information and neural nets’, Proteins: Structure, Function, and Bioinformatics, 2023, 69, (S8), pp. 159-164

[10] D. Sleeman, "The challenges of teaching computer programming," Communications of the ACM, vol. 29, no. 9, pp. 840-841, 2022

[11] M. N. Ismail, N. A. Ngah, and I. N. Umar, "Instructional strategy in the teaching of computer programming: a need assessment analyses," The Turkish Online Journal of Educational Technology, vol. 9, no. 2, pp. 125-131, 2020.

[12] Roccetti, M., Delnevo, G., Casini, L., and Cappiello, G.: ‘Is bigger always better? A controversial journey to the center of machine learning design, with uses and misuses of big data for predicting water meter failures’, Journal of Big Data, 2022, 6, (1), pp. 1-23

[13] Kanika, S. Chakraverty, and P. Chakraborty, "Tools and techniques for teaching computer programming: A review," Journal of Educational Technology Systems, vol. 49, no. 2, pp. 170-198, 2020.