International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Deep Belief Network Model for Detection of an Outlier in Healthcare Data

[1]Dharmesh Dhabliya, [2]Ankur Gupta, [3]Dr Abhijit Dandavate, [4]Dushyant Kaushik, [5]Swati Vitthal Khidse, [6]Shrinivas R. Zanwar, [7]Dr. Jambi Ratna Raja Kumar,

Submitted: 26/05/2023 **Revised**: 16/07/2023 **Accepted**: 29/07/2023

Abstract: Sports wristbands provide a rich source of information for a thorough understanding of people's physical conditions in the light of the popularisation of intelligent wearable gadgets. Outlier detection is still important since there are unknown outliers in the multi-dimensional activity data it supplies. Traditional methods of density estimation are hindered by the "curse of dimensionality," resulting in poor detection results. A Gaussian mixture generative model (GMGM) health data detection method is employed to address this issue. To begin, the model trains the original data with a variational autoencoder (VAE) and recovers latent features by lowering the reconstruction error. The latent distribution and extracted attributes are then utilised to forecast the varied membership of the samples using a deep belief network (DBN). Then, to prevent the effects of model decoupling, the variational autoencoder, deep belief network, and Gaussian mixture model (GMM) are optimised together. The Gaussian mixture model predicts the sample density of each data set and considers samples with densities more than the threshold as anomalies during the training phase. On the ODDS standard dataset, the model's performance is tested. The results reveal that the AUC index of GMGM is enhanced by 5.5 percent points on average when compared to the deep autoencoder Gaussian mixture model (DAGMM). Finally, the method's usefulness is demonstrated by the experimental findings on real datasets.

Keywords: *Curse of Dimensionality, Deep Belief Network, Gaussian Mixture Model, Variational Autoencoder, Healthcare*

1. Introduction

People have been paying more and more attention to healthy lifestyles in recent years. Sports bracelets are becoming increasingly popular as a way to track one's health. Hands that are athletic Rings may track people's activities and behaviours, such as how much they sleep, how long they sleep, their heart rate, and how many steps they take at the gym. A illness was discovered in the literature [1]. There are considerable disparities between sick and healthy bracelet wearers in the bracelet data, and various indications are more strongly related with specific circumstances, such as activity. Cardiovascular illness and metabolic disorders are linked to both steps and resting heart rate. The scarcity of information for bracelet wearers, in terms of practical analysis, relying solely on the data provided by the bracelet does not provide an accurate picture of their physical condition. For bracelets collected data, outliers are variations from indicators connected with specific conditions. As a result, the difference in the bracelet data must be determined. Constant value used to predict whether the user's body contains any hidden risks.

Because the distance between typical points is great, calculate the spread (or average distance) between each sample point and compare it to the distance threshold; if it is greater than the threshold, values are considered outliers. When dealing with high-dimensional data, however, the correlation distance becomes more important.

Distance and nearest neighbour lose their meaning, and the effect of anomaly detection deteriorates. The "dimension disaster" problem is prone to occur when doing anomaly detection in this era of big data since data has high-dimensional properties. To overcome this challenge, many studies have concentrated on the Constant value detection approach. A two-step strategy is used in the classic technique [2-4]. This is the first drop dimension, after which anomaly detection is performed. Both of these steps are taught separately.

1Professor, Department of Information Technology, Vishwakarma Institute of Information Technology, Pune, Maharashtra, India Email: dharmesh.dhabliya@viit.ac.in
orcid.org/0000-0002-6340-2993
2Assistant Professor, Department of Computer Science and Engineering, Vaish College of Engineering, Rohtak, Haryana, India ankurdujana@gmail.com
3Associate Professor, Automobile engineering, Dhole Patil college of Engineering Pune.
Email abhidandavate@gmail.com
4Assistant Professor, Department of Computer Science and Engineering, MERI College of Engineering and Technology, Sampla, Rohtak, Haryana, India
dushyant.kaushik@meri.edu.in
5Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra,swatikhidse@gmail.com
6Dept of Artificial Intelligence and Data Science, CSMSS, Chh. Shahu College of Engineering, Aurangabad, MS, India.
shrinivas.zanwar@gmail.com
Orcid : 0000-0002-0719-0260
7Associate Professor, Department of Computer Engineering, Genba Sopanrao Moze College of Engineering, Balewadi, Pune, Maharashtra, India Email: ratnaraj.jambi@gmail.com

Anomaly detection is used to guide dimension reduction training, which makes it easy to miss outliers. Detection of critical information is common. Literature [5] neural network with deep learning (deep neural network, DNN) K-means and dimensionality reduction (K-means) Clustering approaches are integrated to allow for simultaneous optimization of both tasks, which saves time and money. To improve the detection effect, decouple the influence of learning.

A variety of anomaly detection methods have been proposed by deep learning researchers. To improve detection performance, many techniques are employed. Literature [6] DAMM To combine low-dimensional feature representation with reconstruction error characteristics, the method first utilises a deep autoencoder to transform the original data into Row latent space representation. For density estimation, the attributes are fed into GMM (Gaussian mixture model).

The amount of numbers exceeding the density threshold is recorded as outliers by selecting an acceptable density threshold. This technique, on the other hand, assumes that exceptions are uncompressible. As a result, the input data cannot be reconstructed efficiently from the low-dimensional latent space. Mutually The reconstruction error is insufficient when compared to VAE employing reconstruction probability to recreate the original data[7-8].

Due to a lack of objectivity, the DAGMM method's detection performance is poor. GMGM (Gaussian mixture generative model) is comparable; literature [7] suggested using VAE (variational autoencoder DL-GMM approach paired with GMM, which employs a hybrid). The Gaussian distribution approximates the posterior of the VAE, therefore enhancing the capacity of the original VAE. It is not, however, appropriate for unsupervised outlier detection. [8] Literature A proposed approach for anomaly detection using a multi-view topic model is anomaly detection based on a multi-view subject model. The features are modelled to get the corresponding relationship, which reduces the detection time dramatically. This approach, however, has low detection accuracy.

GMGM is used in this paper to detect anomalies in human activity data. Use the VAE from the generative model in this model. To train DBN (deep brief network) to forecast the diverse membership of the sample, generate data latent distribution and reconstruction error. As the sample density is abnormal, Gaussian, the mixture model, acquires each sampling density of the data; the density is higher than the threshold of the training phase. Starting with Avoiding the impact of model decoupling, GMGM optimises VAE, DBN, and GMM together.

The following are the three primary contributions of this paper:

(1) In order to preserve as much of the original data's features as feasible, the generative network For real-world samples, use VAE to produce latent distributions and recover error characteristics.

(2) To avoid calculating the sample density owing to the matrix during the calculation, The covariance matrix's singularity problem cannot be solved; GMGM constructs the covariance matrix by combining probability, mean, and covariance of samples. To calculate the sample density, use the Cholesky decomposition.

(3) Due to the traditional two-step approach, important information is lost while doing anomaly detection; GMGM collaborates end-to-end. Optimize VAE, DBN, and GMM to maintain the original properties of the data. This approach is used throughout the book to discover anomalies in health data, and tests are done on actual data sets to illustrate the algorithm's efficiency. This approach may be used to discover anomalies in health data.

2. Related work

2.1 Variational Autoencoder

Variational autoencoders are proposed to solve the problems of traditional algorithms. It can solve the complicated and expensive issues of inference and training in complex scenarios the s—the ability to generate low-dimensional representations of latent variables of the input data. Variational self-editing of the encoder can be regarded as a feature constructed to generate its probability distribution to reconstruct the data based on the original sample distribution. Compared to deep autoencoders, the reconstruction error is used to reconstruct the data, and the reconstruction probability is a probability measurement. It considers the variability of variable distributions and is more moral than reconstruction error Sexuality and Objectivity [9]. Therefore, this paper selects VAE for feature extraction, Solving the "Curse of Dimensionality" while preserving the multimodality of the original data feature. In recent years, various autoencoders have gradually been combined with deep neural networks, connectedby stacking hidden layers in an unsupervised manner, Parameter optimisation. Assuming that x ∈RD is a vector of dimension D, that z ∈Rd′ is the corresponding latent representation of dimension d′, and that P (.) is the probability rate distribution function, the probability distribution can be made by:

$$p(x) = \int p(x \mid z)p(z)dz \qquad \text{.....(1)}$$

2.2 Gaussian Mixture Model

The Gaussian mixture model breaks the distribution of variables into a number of statistical models. The predicted maximum is used in the Gaussian probability density ion distribution. The expectation-maximization (EM) algorithm estimates the probability density parameters of the degree function [10]. When doing anomaly detection, GMM finds the probability density of the data, t. The higher the probability density, the sample is outlier, the more significant the possibility, and the more its goal is compared to binary classification. The calculation will show up, though, when GMM is used to estimate the density of high-dimensional data. It is a hard problem to solve with computers, so this article starts by showing the raw data in the space mentioned above, and then uses GMM to estimate a density. When fitting data with more than one dimension, GMM has three parameters: the mixture probability, the mean, and the covariance. If GMM has K parts, the probability of the mixture, the mean, and the covariance variance of the kth part are by ϕ(k), μ(k) and σ(k), and ∑(i)= 1 K ϕ(i)= 1, I a 1D GMM. For example, the probability density function is shown. In formula (2) and formula (3):

$$p(x) = \sum_{i=1}^{k} \Phi_i \, N(x_i|\mu_i) \qquad \text{……..(2)}$$

$$N(x \mid \mu_i, \sigma_i) = \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right) \qquad \text{……(3)}$$

In the training phase of the GMM model, the EM algorithm is used to maximise the likelihood. The optimal parameters of the model are solved using a natural function, namely the mixture probability ϕ(k), the average value μ(k) and covariance σ(k) until the model converges.

3. Proposed Algorithm

This paper uses GMM for Anomaly detection when estimating density on high-dimensional data for problems that take a long time to solve. Figure 1 shows that the model is mostly made up of the generative model and the Gaussian mixture model. The basic idea behind GMM The following explains why: First, the generative model reduces the input samples through VAE dimensional processing to make latent space representations of sample points. Then, the DBN uses the feed to predict the diverse membership of the sample point. Finally, using the mixed membership attribute, GMM shows the sample density of each data, and data above the threshold in the training phase is considered abnormal.

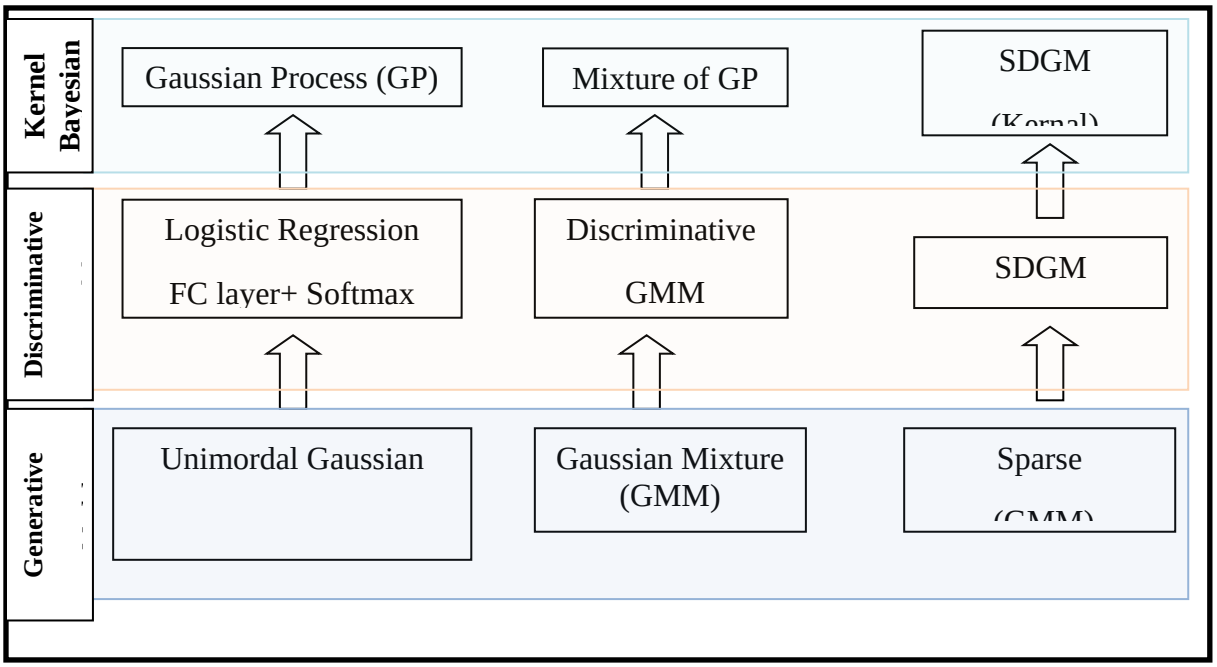

Fig 1: GMM Anomaly Detection Model

In high-dimensional space, there will be a thing called a "dimensional disaster." As the number of data dimensions goes up, the amount of time it takes to do density prediction will go up, and performance will go down. To solve this problem, the generative model is the data input X = [x1, x2,..., xN] ∈RD× N for reconstruction processing to get the sample points' potential space. Between the representation Z and the reconstructed feature, , ω is used to keep the sample information's multimodality and feed it into the DBN.

$$Z = f_o(X) \qquad \text{…… …(4)}$$

$\hat{X} = g_o(Z)$ ……..(5)

$\omega = h(X, \hat{X})$ ……..(6)

Where Z is the input sample X's latent representation and X is the sample's reconstruction error features, fθ(·) and gϕ(·) are the encoding and decoding functions, respectively. is the reconstruction error eigenvector, and h(.) is the A function used to calculate the error vector. H (.) can be found using the relative Euclidean distance, the absolute Euclidean distance, or the root mean square error representation in GMGM. In writing, the relative Euclidean distance is used [6] the square root of the error.

For density estimation in Gaussian mixture models, the membership of the mixture members of each sample should be considered. Existing methods either randomise the membership of the mixture during the initialisation phase or take an average probability for each component, which are problematic. Reference [6] uses a multi-layer perceptron, and GMGM uses a deep belief network to adaptively calculate the mixing probability of each component [11]

Solve this problem. In the output layer of DBN, an M-dimensional vector Γ̂ i =[γ̂ i1,γ̂ i2,…, γ̂ im,…,γ̂ iM] is generated for each sample xi using the Softmax function to estimate its mixed membership:

Where bλ(·) denotes DBN, γ̂it represents the ith in the Gaussian mixture model the sample is the probability generated by the kth component.

Inspired by the M steps in the expectation-maximization (EM) algorithm, GMGM utilises a sample of size N and a mixed membership Γ̂ I estimate Parameters of the Gaussian mixture model:

$\widehat{\emptyset}_{\mathrm{k}} = \sum_{i=1}^{N} \frac{\widehat{\Gamma_{ik}}}{N}$ ……(8)

$\hat{\mu}_{\mathrm{k}} = \frac{\sum_{i=1}^{N} \frac{\widehat{\Gamma_{ik}}}{N} Z_i}{\sum_{i=1}^{N} \frac{\widehat{\Gamma_{ik}}}{N}}$ ……(9)

$\hat{\sigma}_{\mathrm{k}} = \frac{\sum_{i=1}^{N} \frac{\widehat{\Gamma_{ik}}}{N} (z_i - \widehat{\mu_k})^2}{\sum_{i=1}^{N} \frac{\widehat{\Gamma_{ik}}}{N}}$ ……(10)

Among them, ∀1 ≤ k ≤ K, K is the number of Gaussian mixture model components; ϕ̂ k is the mixing probability for component k; μ̂ k and σ̂k are the mean of component k, respectively, value and covariance. Then, using the estimated parameters, we can further steps to infer the sample density.

To fit a Gaussian distribution to a single data using a maximum likelihood point, you will get a very sharp Gaussian distribution, and the whole model will "collapse". When the above situation occurs in a Gaussian mixture model, there will be. As a result, the inversion of the covariance matrix cannot be achieved, a phenomenon called the singularity Question [12]. Reference [6] uses the multivariate height with covariance matrix inversion. The direct expression of the s distribution calculates the energy of each sample. But, Due to the singularity problem of the matrix, the covariance matrix Σ̂ The inverse of k may have no method to solve. Therefore, GMGM utilises the mixed probability ϕ̂ (k), mean μ̂ (k) and covariance σ̂(k)to compute the Cholesky decomposition of the covariance matrix, and compute Sample Density:

$ln|\widehat{\sigma_i}| = 2 * \sum_{i=1}^{N} ln\, diag(L)$ …….(11)

$\Delta = \ln\widehat{\sigma_k} - .5 * \sum_{i=1}^{N} \vartheta^2 + d' * ln2\pi + ln|\widehat{\sigma_k}|$ ……..(12)

$E(z) = -\ln\sum_{i=1}^{k} e^{\Delta}$ ……..(13)

Where L is the covariance matrix σ̂k, the compensation term is divided by Cholesky. The lower triangular matrix of the solution; v is the solution of the linear system of equations; d' is the generating module the dimensionality of the low-dimensional representation is provided by the type.

When doing anomaly detection with the traditional two-step method, you lose the key. Important data, so the dimensionality reduction process needs to be combined with the density estimation process training and mutual optimisation [13].GMM is using the EM algorithm to model. During exercise, first, calculate the mixed composition of each data according to the current parameters. Member membership degree, then use the obtained hybrid membership degree to estimate the model type parameters until convergence. Therefore, in this paper, GMGM will expect the maximum probability replacement that the samples in the E step of the algorithm belong to each sub-distribution. Generate the output of the model in an end-to-end structure in an end-to-end, the generative model and GMM are jointly trained in the same way; then, the EM algorithm is used for The M step in the method to estimate the parameters of the mean, covariance, etc. in the GMM. Then maximise the likelihood function, which is more accessible than traditional training methods, to achieve the ideal detection effect.

In the testing phase, GMGM can predict samples according to Eq. (13). The density of the sample density is higher than the threshold of the training phase as the abnormal.

3.1 Objective function

Decoupled learning in GMM doesn't work very well, so VAE, DBN, and GMM are combined to train models.

Given N The objective function for a sample set of data points is the following:

$$\min \mathcal{J}(f_\theta, g_\emptyset, b_\lambda) = \Omega + \frac{\lambda_1}{N}\sum_{i=1}^{N} E(z) + \lambda_2 b_\lambda(z+\omega) \quad \text{……..(14)}$$

$$\Omega = D_{KL}\left[q(z,\Gamma \mid x) \parallel p(z,\Gamma \mid x)\right] \quad \text{…….(15)}$$

Equation (15) represents the posterior distribution q(z, Γ|x) and the maximum likelihood distribution KL divergence of p(z, Γ|x). By minimising the posterior distribution with maximum likelihood, The KL divergence of the distribution maximises the possibility of multidimensional inputs. E(z) simulates the probability that an input sample can be observed. The minimum densifying of the sample density maximises the likelihood of observing the input sample; the optimal combination of VAE, DBN and GMM parameters is obtained

λ1 and λ2 are hyperparameters used to standardise the objective function; in the experiment, λ1 = 0.1 and λ2 = 0.001 usually give better results. Minimum J(fθ,gϕ,bλ) can provide the best parameters for generative models and GMMs. Number combination.

3.2 Algorithm Complexity Analysis

Assume X ∈ RD × N is a primitive of size N and dimension D in the input data; the GMGM technique must recover the original data. Set the number of hidden layers to three, which corresponds to the three-layer encoder and three-layer decoder layer; D ′ is the number of nodes in each hidden layer (i.e., the output dimension of each layer). The temporal complexity of this portion reaches a maximum of O(NKD' 3); DBN Predict the probability that each sample belongs to each of the K components separately; this part includes the Back-propagation and Softmax processes, and its time complexity is O(NKD′ 3; density estimation using GMM has a time complexity of O((K + 1) 3, so the time complexity of GMGM is O((K + 1) 3 +(K + 1)ND′ 3). SOS algorithm (stochastic outlier selection) [14] To approximate the affinity, use the dissimilarity matrix. The time complexity of the relationship between two points is O(N3), which is substantially larger than the time complexity of the connection between two points. The time complexity of VAE [15] c, the traditional anomaly detection method in this study, is O(NKD′ 3; the time complexity of DAGMM is O((K + 1) 3 + N(K + 2)D′ 3).

4. Experimental evaluation

The experimental platform is set up using Windows 10, an Intel Core i7-7700HQ CPU processor running at 2.80 GHz, and 20 GB of RAM; all algorithms are written in Python. Anomaly classes are found in 5 datasets from the ODDS database, which are classified based on sample labels. The data indicated with a 0 represents the normal class, whereas the data marked with a 1 represents the abnormal class. The dataset's data properties are shown in Table 1.

The approach in this study was compared to SOS, variation-based Anomaly Detection Algorithm for Encoders [14-15], and Deep Autoencoder Gaussian Mixture Model (deep autoencoding Gaussian mixture model, DAGMM) [6] in order to test its performance. The rationale for the choice is because the SOS method calculates the outlier probability for each data point using the related probability Read, which is consistent with the predictions in this study. Each sample point has a comparable density, and the method in this research is based on that variation. The improvement of the autoencoder's anomaly detection technique is chosen as the ratio; DAGMM employs a deep autoencoder to recover the original data's properties, and the membership degree of the mixed members of the sample is calculated by the multilayer perceptron and the most. GMM then calculates the energy of each sample point in order to discover anomalies. Because the detection effect is excellent and the structure is comparable to that of the method in this work, use it as a comparison algorithm.

TABLE 1. Dataset information

Dataset	No. of data	Dimension	Number of outliers (proportion)
Ionosphere	342	29	135
Arrhythmia	436	256	59
Musk	3062	166	97
Speech	3686	400	61
Shuttle	49097	9	3511

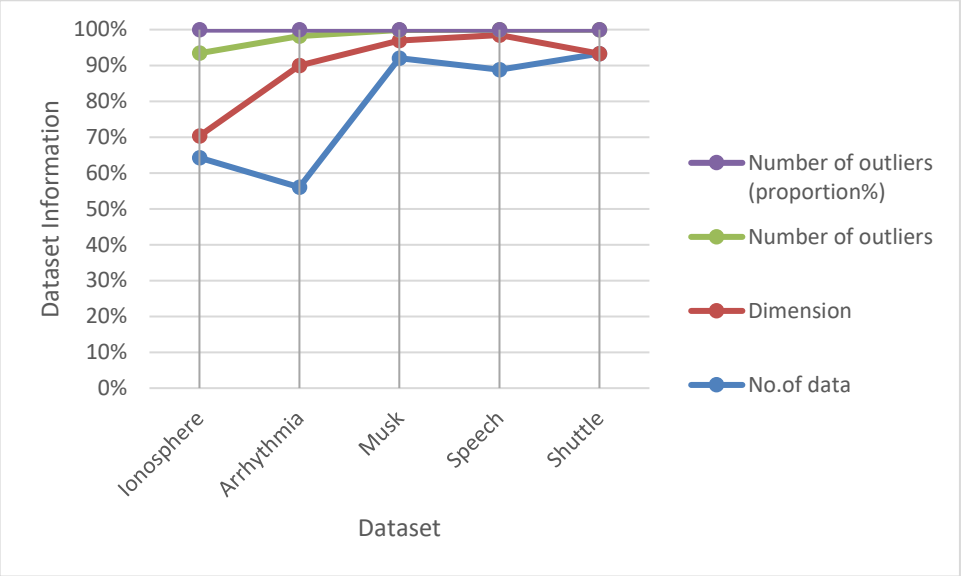

Fig 1: Dataset information

The approach in this study was compared to SOS, variation-based Anomaly Detection Algorithm for Encoders [15], and Deep Autoencoder Gaussian Mixture Model (deep autoencoding Gaussian mixture model, DAGMM) [16] in order to test its performance. The rationale for the choice is because the SOS method calculates the outlier probability for each data point using the related probability Read, which is consistent with the predictions in this study. Each sample point has a comparable density, and the method in this research is based on that variation. The improvement of the autoencoder's anomaly detection technique is chosen as the ratio; DAGMM employs a deep autoencoder to recover the original data's properties, and the membership degree of the mixed members of the sample is calculated by the multilayer perceptron and the most. GMM then calculates the energy of each sample point in order to discover anomalies. The detection effect is good, and the structure is similar to the algorithm in this paper, so choose It to act as a comparison algorithm.

Recall, F1 - Score, Accuracy (ACC), and Receiver operating curve are the performance measures utilised in this article to assess the anomaly detection method (area under curve, AUC). A superior exception Recall, F1-Score, ACC, and AUC should all be high in the detection method.

3.1 Experimental comparison results and analysis

For each sample set, the parameters of GMGM are set as follows: Data set Dives for Ionosphere, Arrhythmia, Musk, Speech and Shuttle. The spatial representation dimensions are 3, 4, 4, 4, and 2; to determine the GMM, The optimal number of components, some analytical criteria need to be used to evaluate the type of possibility. This paper refers to the literature [6] and literature [7] and found that the Mainly using the Bayesian information criterion (Bayesian information criterion) criterion, BIC) [16] evaluation method to determine the number of components, the model's The lower the BIC value, the better the performance of the GMM in predicting the sample density of the sample data it is good. For all datasets in this paper, the number of GMM components is 3, and the model's BIC value is the smallest, so for all datasets, GMM The number of members is set to 3.

To verify the optimal performance of GMGM for high dimensional data detection, we selected the Speech dataset with a larger dimension and adopted a qualitative method. Formula, with SOS, VAE and DAGMM algorithms for the ROC curve for comparison, the comparison results are shown in Figure 2. As can be seen from the figure, compared to The AUC values of the area under the ROC curve of the SOS, VAE and DAGMM algorithms, The GMGM anomaly detection method has the largest size and the highest AUC value. Among them, the VAE algorithm has the worst detection effect, probably because VAE is when the data is represented in the latent space, the original sample is compared with the anomaly. The critical information was erroneously removed, leading to the detection of AUC. The value is lower; while GMGM adopts end-to-end joint training, which can simultaneously train VAE, DBN and GMM to make the three model parameters reach Optimal, the detection effect is ideal.

As can be seen from Figure 3, for different datasets, this paper calculates when the method achieves the best detection effect, the corresponding VAE encoder. The number of layers o is different. When the value of o

increases, the corresponding AUC of each dataset always increases and then decreases. This is because increasing the value of o first can make the encoder performs data reconstruction well and learns the original sample well. Characteristic of this, the AUC value increases; but then continues to increase as o is large, which leads to over fitting of training and reduces the AUC value of the algorithm. After careful consideration, the o values for the five datasets in Figure 3 are 4 (33- 16-8-3), 5 (274-136-64-16-5), 5 (166-84-42-12-5), 5 (400- 200-100-50-5), 2 (9-2).

To verify the advantage of GMGM in time complexity, it is Average detection with SOS algorithm, VAE algorithm and DAGMM algorithm. The time is compared, and the comparison results are shown in Table 2.

Table 2. Comparison of average detection

ALGORITHM	AVERAGE DETECTION TIME
SOS	2.48
VAE	0.48
DAGMM	1.06
ALGORITHM	0.64

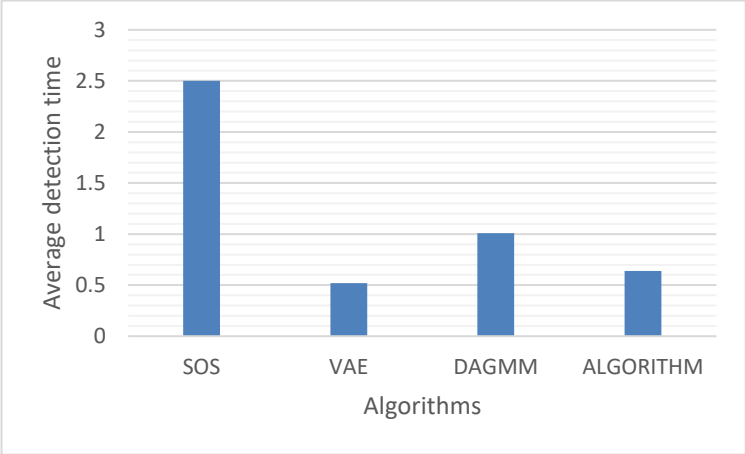

Fig 2: Algorithms and their average detection time

Although the average detection time of the method in this study is not the shortest, it is the VAE algorithm with the shortest average detection time alone, as demonstrated in Table 2. The difference is 0.12 seconds, and its average detection time is faster than that of The DAGMM method has been improved by 37%, as seen by the algorithm's detection in this work.

The Table 3 shows the experimental findings in comparison to the independently trained model. The GMGM with end-to-end training has greater metrics than separately trained models, as shown in the table.

Table 3. Comparison of experimental results of different model structures

MODEL STRUCTURE	DATA SET	ACC	RECALL	F1-SCORE	AUC
model of this paper	Ionosphere	0.88842	0.49062	0.51102	0.90066
	Arrhythmia	0.83742	0.43044	0.50694	0.84864
	Musk	1.0149	1.00776	1.0047	1.01796
	Speech	0.9384	0.98838	0.97716	0.93636
	Shuttle	0.99144	0.43554	0.47124	0.99348
ne d ind ep en de ntl y	Ionosphere	0.83742	0.40902	0.40698	0.82416

	Arrhythmia	0.71196	0.31824	0.39372	0.71502
	Musk	0.79152	0.84966	0.91494	0.81702
	Speech	0.8517	0.88842	0.85068	0.89862
	Shuttle	0.92922	0.93636	0.95574	0.95574

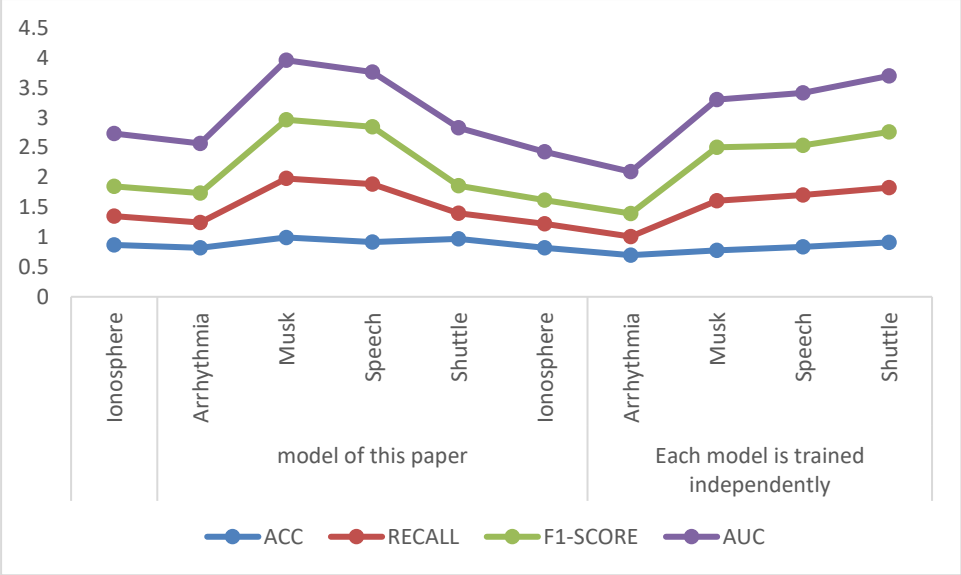

Fig 3: Comparison of experimental results of different model structure

In order to verify the advantages of the performance of the algorithm in this paper, the algorithm in this paper is compared with VAE algorithm; SOS algorithm and DAGMM algorithm are compared and calculated. The performance indicators of each anomaly detection algorithm ACC, Recall, F1-Score and The AUC values are listed in Table 4.

Table 4. Comparison of experimental results of different Algorithms

DATA SET	ALGORITHM	ACC	RECALL	F1-SCORE	AUC
IONOSPHERE	SOS	0.7415	0.3631	0.4009	0.7783
	VAE	0.8231	0.3458	0.3927	0.7732
	DAGMM	0.8507	0.4447	0.4549	0.8548
	ALGORITHM	0.8884	0.4906	0.5110	0.9007
Arrhythmia	SOS	0.6885	0.2315	0.3213	0.5885
	VAE	0.5110	0.3478	0.3805	0.5131
	DAGMM	0.8201	0.2591	0.4192	0.6344
	ALGORITHM	0.8374	0.4304	0.5069	0.8486
MUSK	SOS	0.6905	0.7579	0.7120	0.7548
	VAE	0.7415	0.7558	0.8007	0.7568
	DAGMM	0.8303	0.8425	0.8690	0.8446
	ALGORITHM	1.0149	1.0078	1.0047	1.0180
	SOS	0.7742	0.7987	0.7721	0.7415

SPEECH	VAE	0.7364	0.7721	0.7936	0.8252
	DAGMM	0.9058	0.9078	0.9415	0.9394
	ALGORITHM	0.9384	0.9884	0.9772	0.9364
SHUTTLE	SOS	0.9017	0.4508	0.2744	0.7885
	VAE	0.8558	0.2662	0.3080	0.8884
	DAGMM	1.0108	0.3131	0.4009	0.7752
	ALGORITHM	0.9914	0.4355	0.4723	0.9935

Among them, the number of hidden layers of the VAE algorithm and the number of nodes in each layer is the same as the VAE in the generative network; DAGMM and Reference [6] have the same parameter setting. From the comparative experimental results in Table 4, it can be seen that the quasi-accuracy rate is only slightly lower than the DAGMM algorithm on the extensive data set Shuttle.

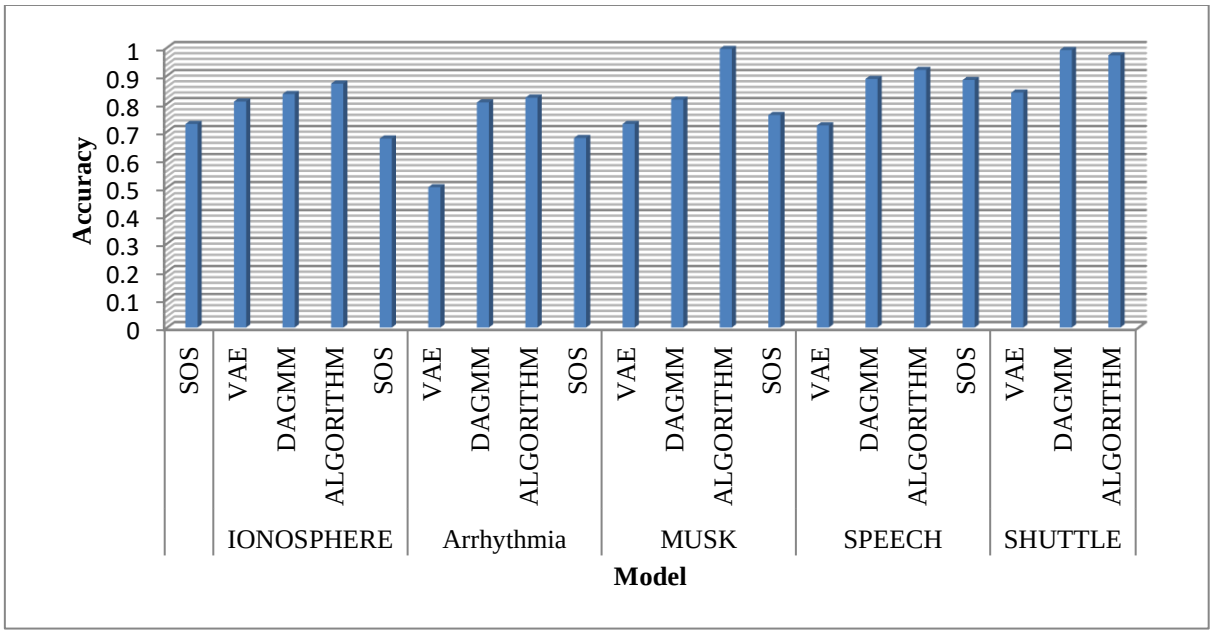

Fig 4: Comparative analysis of Accuracy of proposed algorithm with other

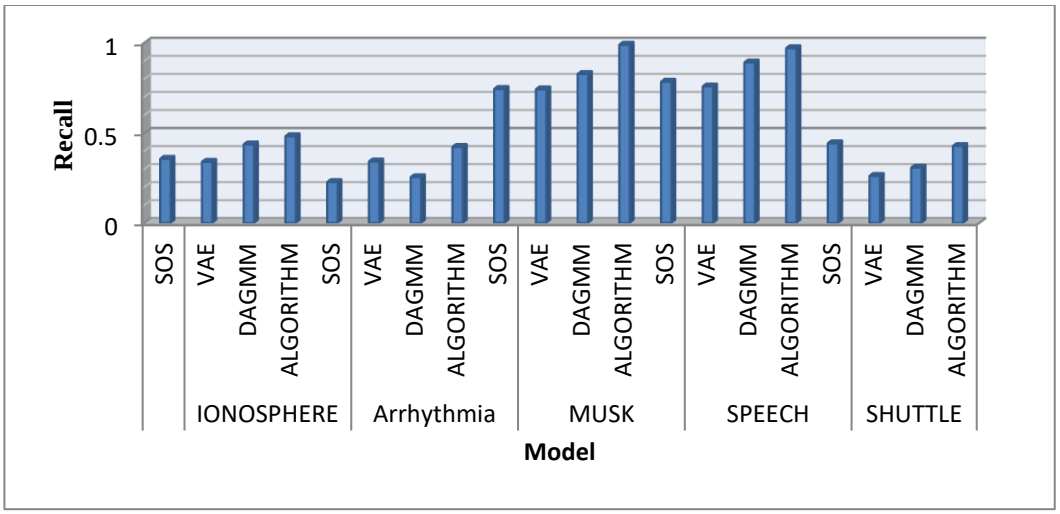

Fig 5: Comparative analysis of Recall of proposed algorithm with other

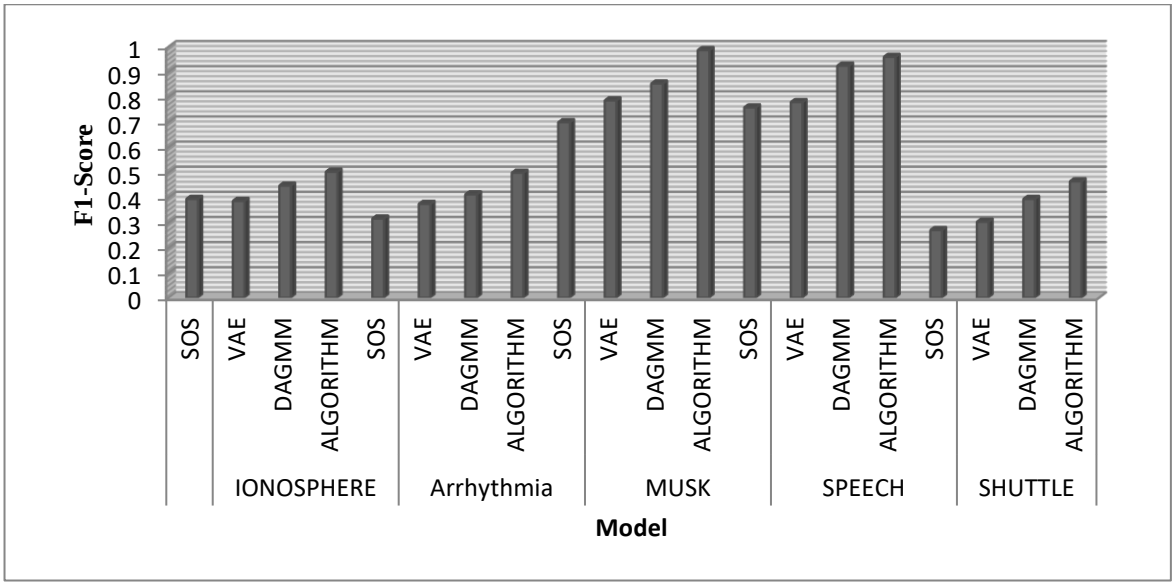

Fig 6: Comparative analysis of F1-Score of proposed algorithm with other

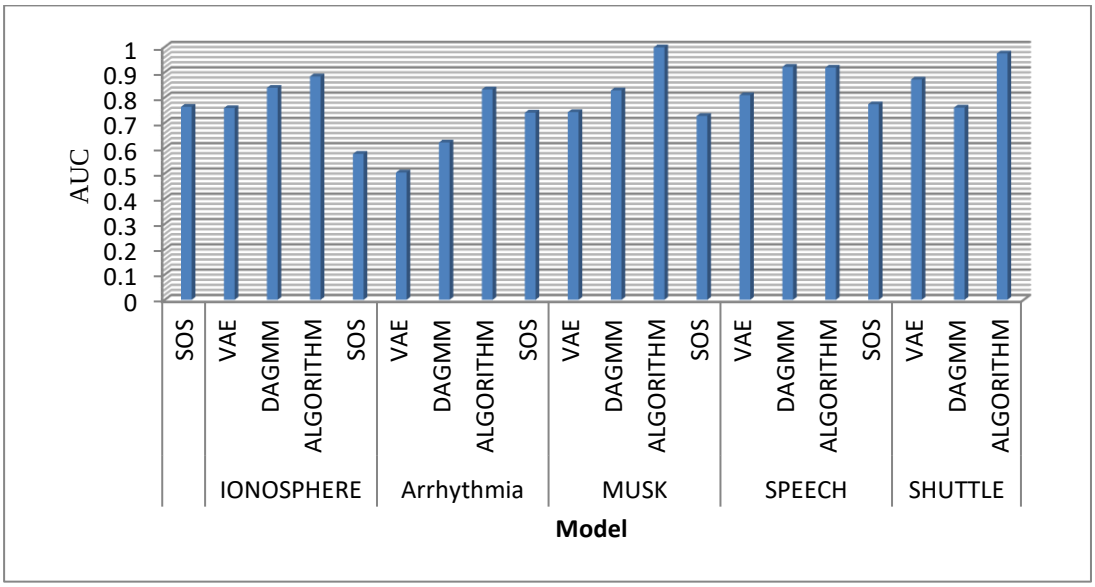

Fig 7: Comparative analysis of AUC of proposed algorithm with other

Its AUC value is also only marginally lower than DAGMM on the Speech dataset Algorithm; the Recall value on the large dataset Shuttle is not the highest, but about the same as the highest value; accurate on the high-dimensional dataset Musk The accuracy rate reaches 0.995, which is much higher than the 0.677 of the SOS algorithm. The Arrhythmia dataset with a relatively high and large amount of data also shows an ideal detection effect; on the Shuttle dataset, although this paper calculates The ACC and Recall of the law are slightly lower, but the F1 - Score and AUC values increased by 7 and 1.4 percentage points, respectively. This situation the reason for this may be that the latent space representation in the algorithm can compare. It captures the overall characteristics of the data well and improves the local structure of the data. The construction ability reduces the time complexity of the algorithm. Still, at the same time, the VAE is in the When representing the latent space of the large dataset Shuttle, due to a large amount of data will inevitably lead to over fitting, which also This is where the algorithm of this paper needs to be improved.

3.2 Health data anomaly detection results

After the algorithm's performance has been verified, the algorithm is used to perform on the collected health data to detect outliers. Figure 4 is the result of anomaly detection visualisation using the algorithm in this paper. Black dots represent normal data, and red dots represent abnormal data**.**

To highlight the advantages of this algorithm, the detection effect with the same good DAGMM algorithm performs on the same health in the same experimental environment.

To highlight the advantages of this algorithm, the detection effect with the same good DAGMM algorithm performs on the same health in the same experimental environment. The experiment was carried out on the Kang data, and the results are shown in Figure 5.

Comparing Figure 4 and Figure 5, it can be seen that in the two detection methods, the more obvious abnormal sample points can be detected. Still DAGMM algorithm has missed judgments at the edge of the data. Labelled 1 and 3, sample 2 are missed judgments, and the sample points marked with two are misjudged. And this article, when the algorithm detects edge anomalies, only 3 sample points is missed. Overall performance is good.

5. Conclusion

The activity data collected by the sports wristband contains unknown oddities. GMGM is used to discover anomalies in data, according to the problem. To train the DBN to estimate the membership of the mixed members of each sample, use the sample latent distribution and reconstructed features from the generative model; then, use GMM to forecast the density of each sample to discover outlier's measurement. To avoid model decoupling influences, the generative network and GMM are optimised together. A representative number of anomalous saws were used in the experiment. On the dataset, experiments are conducted, and the results reveal that the approach has an optimum detection effect fruit. Finally, the approach is applied to real-world datasets to visualise abnormalities. The detection findings reveal that the DAGMM algorithm has a lower false negative and false positive rate.

References

[1] L. F. M. Carvalho, C. H. C. Teixeira, W. Meira, M. Ester, O. Carvalho and M. H. Brandao, "Provider-Consumer Anomaly Detection for Healthcare Systems," 2017 IEEE International Conference on Healthcare Informatics (ICHI), 2017, pp. 229-238, doi: 10.1109/ICHI.2017.75.

[2] S. V. Georgakopoulos, P. Gallos and V. P. Plagianakos, "Using Big Data Analytics to Detect Fraud in Healthcare Provision," 2020 IEEE 5th Middle East and Africa Conference on Biomedical Engineering (MECBME), 2020, pp. 1-3, doi: 10.1109/MECBME47393.2020.9265118

[3] J. Pereira and M. Silveira, "Learning Representations from Healthcare Time Series Data for Unsupervised Anomaly Detection," 2019 IEEE International Conference on Big Data and Smart Computing (BigComp), 2019, pp. 1-7, doi: 10.1109/BIGCOMP.2019.8679157.

[4] F. Ahamed and F. Farid, "Applying Internet of Things and Machine-Learning for Personalized Healthcare: Issues and Challenges," 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), 2018, pp. 19-21, doi: 10.1109/iCMLDE.2018.00014

[5] N. I. Haque, A. A. Khalil, M. A. Rahman, M. H. Amini and S. I. Ahamed, "BIOCAD: Bio-Inspired Optimization for Classification and Anomaly Detection in Digital Healthcare Systems," 2021 IEEE International Conference on Digital Health (ICDH), 2021, pp. 48-58, doi: 10.1109/ICDH52753.2021.00017

[6] F. A. Bellini, J. G. Gutierrez-Zorrilla, L. E. Anza, E. D. Ferreira, L. G. Deneault and G. Vanerio, "MDi: Acquisition, analysis and data visualization system in healthcare," 2017 IEEE URUCON, 2017, pp. 1-4, doi: 10.1109/URUCON.2017.8171879.

[7] J. Fiaidhi, "Envisioning Insight-Driven Learning Based on Thick Data Analytics With Focus on Healthcare," in IEEE Access, vol. 8, pp. 114998-115004, 2020, doi: 10.1109/ACCESS.2020.2995763

[8] M. Kavitha, P. V. V. S. Srinivas, P. S. L. Kalyampudi, C. S. F and S. Srinivasulu, "Machine Learning Techniques for Anomaly Detection in Smart Healthcare," 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), 2021, pp. 1350-1356, doi: 10.1109/ICIRCA51532.2021.9544795

[9] L. Servi, R. Paffenroth, M. Jutras and D. Burchett, "Reducing Reporting Burden of Healthcare Data Using Robust Principal Component Analysis," 2020 IEEE International Conference on Big Data (Big Data), 2020, pp. 3827-3836, doi: 10.1109/BigData50022.2020.9378410.

[10] A. Biwalkar, R. Gupta and S. Dharadhar, "An Empirical Study of Data Mining Techniques in the Healthcare Sector," 2021 2nd International Conference for Emerging Technology (INCET), 2021, pp. 1-8, doi: 10.1109/INCET51464.2021.9456157

[11] W. Yao, K. Zhang, C. Yu and H. Zhao, "Exploiting Ensemble Learning for Edge-assisted Anomaly Detection Scheme in e-healthcare System," 2021 IEEE Global Communications Conference (GLOBECOM), 2021, pp. 1-7, doi: 10.1109/GLOBECOM46510.2021.9685745

[12] J. Seo and O. Mendelevitch, "Identifying frauds and anomalies in Medicare-B dataset," 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2017, pp. 3664-3667, doi: 10.1109/EMBC.2017.8037652

[13] M. Nawaz, J. Ahmed, G. Abbas and M. Ur Rehman, "Signal Analysis and Anomaly Detection of IoT-Based Healthcare Framework," 2020 Global Conference on Wireless and Optical Technologies (GCWOT), 2020, pp. 1-6, doi: 10.1109/GCWOT49901.2020.9391621

[14] F. Hounaida, B. d. Wided, M. -M. Amel and Z. Faouzi, "A Learning based Secure Anomaly Detection for Healthcare Applications," 2020 IEEE 29th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 2020, pp. 124-130, doi: 10.1109/WETICE49692.2020.00032.

[15] M. Nawaz, J. Ahmed, G. Abbas and M. Ur Rehman, "Signal Analysis and Anomaly Detection of IoT-Based Healthcare Framework," 2020 Global Conference on Wireless and Optical Technologies (GCWOT), 2020, pp. 1-6, doi: 10.1109/GCWOT49901.2020.9391621.

[16] A. A. Sathio, M. Ali Dootio, A. Lakhan, M. u. Rehman, A. Orangzeb Pnhwar and M. A. Sahito, "Pervasive Futuristic Healthcare and Blockchain enabled Digital Identities-Challenges and Future Intensions," 2021 International Conference on Computing, Electronics & Communications Engineering (iCCECE), 2021, pp. 30-35, doi: 10.1109/iCCECE52344.2021.9534846

[17] Mr. Dharmesh Dhabliya, Dr.S.A.Sivakumar. (2019). Analysis and Design of Universal Shift Register Using Pulsed Latches . International Journal of New Practices in Management and Engineering, 8(03), 10 - 16. https://doi.org/10.17762/ijnpme.v8i03.78

[18] Kumar, P. ., Gupta, M. K. ., Rao, C. R. S. ., Bhavsingh, M. ., & Srilakshmi, M. (2023). A Comparative Analysis of Collaborative Filtering Similarity Measurements for Recommendation Systems. International Journal on Recent and Innovation Trends in Computing and Communication, 11(3s), 184–192. https://doi.org/10.17762/ijritcc.v11i3s.6180

[19] Kumar, P. ., Gupta, M. K. ., Rao, C. R. S. ., Bhavsingh, M. ., & Srilakshmi, M. (2023). A Comparative Analysis of Collaborative Filtering Similarity Measurements for Recommendation Systems. International Journal on Recent and Innovation Trends in Computing and Communication, 11(3s), 184–192. https://doi.org/10.17762/ijritcc.v11i3s.6180