Clustering similar time series data for the prediction the patients with heart disease

3 authors, including:

Mohanad AL-Musaylh
University of Southern Queensland
14 PUBLICATIONS **565** CITATIONS

SEE PROFILE

Clustering similar time series data for the prediction the patients with heart disease

Raid Luaibi Lafta[1], Mohanad S. AL-Musaylh[2], Qahtan Makki Shallal[2]
[1]Department of Studies and Planning, Thi-Qar University, Nasiriyah, Iraq
[2]Department of Information Technologies, Management Technical College, Southern Technical University, Basrah, Iraq

Article Info

ABSTRACT

Developed intelligent technologies are become play a promising role in providing better decision-making and improving the medical services provided to the patients. A risk prediction task for short-term is big challenge task; however, it is a great importance for recommendation systems in health care field to provide patients with accurate and reliable recommendations. In this work, clustering method and least square support vector machine are used for prediction a short-term disease risk prediction. The clustering similar method is based on euclidean distance which used to identify the similar sliding windows. The proposed model is trained by using the slide windows samples. Finally, the appropriate recommendations are generated for heart diseases patients who need to take a medical test or not for following day using least square support vector machine. A real dataset which collected from heart diseases patient is used for evaluation. The proposed method yields a good results related by the recommendations accuracy generated to chronicle heart patients and reduce the risk of incorrect recommendations.

Corresponding Author:

Qahtan Makki Shallal
Department of Information Technologies, Management Technical College, Southern Technical University
Basrah, Iraq
Email: qahtan.makii@stu.edu.iq

1. INTRODUCTION

In the last few decades, may people worldwide have been death due to chronicle disease [1]. The quality of patients' lives of chronic disease patients may be effected due to the existing medical recommendation systems, which are limited in providing treatment and care for patients that require frequent medical attention. Therefore, many efforts were already prepared in this field to discover the powerful tools to solve this matter. Due to the heart disease, many cases of death appeared in the world. It was picked up as the top deadliest disease between non-infectious diseases which involves much cost as well as effort regards to treatment and protection [2]. The quality of patient that having chronic disease lives continues to be mainly influenced due to unavailability of effective medical suggestions which could be made to get a greater care and treatment.

An exact prediction of the short-term disease possibly be useful for the medical practitioners to make accurate decision, reduce the workload needed in medical and reduce the incorrect recommendations rates. Many telehealth applications are designed to deliver the medical information into the medical practitioners and patients [3], [4]. For example, the digital tools permit medical practitioners and patients to give each other the medical suggestions as well as personal reminders, upload detail from the devices like the monitors of blood glucose/pressure cuff, publish the information along with providers of health care and store health records [5], [6]. The providers of health care are able to collaborate in the real time and also face-to-

face with their patients employing the services of telehealth (such as the video tools or internet) to deliver suitable suggestions for the whole conditions of health. The devices like monitors of heart rate and monitors of blood pressure can easily link to some applications that are related to web-based in the environment of telehealth to exchange the medical data among patients and the providers of health care [7]. The system of telehealth is helpful for the patients that having chronic diseases, also individuals who live in far areas in which a lesser number of professionals are able to reach to the organizations of healthcare. As a consequence of the significance of disease danger prediction toward the life of the patient and seeking on additional active techniques of analytic for the disease danger prediction, inclusive efforts will need to increase the quality of medical recommendations and evidence-based decisions in the environment of telehealth. In fact, the patients of chronic disease are needed to undertake several medical tests on daily basis in order to control and observe on their entire conditions of chronic health via a telehealth system. This practice may bring plenty of difficulties to patients and negatively impacts their quality of life. Providing perfect medical suggestions to monitor their day-to-day medical test routine can successfully decrease the workload related with having those particular tests while maintaining the concerned health risk in notable low level [8]–[11].

In most of the cases, the essential function for the telehealth systems is to produce accurate recommendation, which can be frequently depending on the prediction related to the risk of short-term disease. In literature, various disease risk prediction models have been applied by the use of statistical analysis tools and data mining principle to handle many healthcare and medical concerns [12]–[17]. However, in previous literature, researchers have not especially interacted with the issues of chronic disease. Moreover, most researchers deal with recommendation considering the long-term disease risk prediction on recommendations. In this study, short-term prediction is tougher compare to the prediction of long-term as the conditions of patients might possibly experience a lot of sudden changes through a small timeframe. Additionally, the recommendations of short-term are patient's benefit as they deliver guidance of the patient's requirement to do for the upcoming few day [18]–[22]. The section of training data gets a huge influence on the results of the prediction method [23]. For example, the relationship that connecting both variables of input and output in our disease risk prediction model could be easily constructed if the sliding windows are similar as training samples. In this research, the similar sliding windows are clustered into two groups: either the patient needs to get the medical test or not needed to take it. A method of short-time disease risk prediction is suggested, and the important method contributions are summarized as shown [24]–[27]: i) The time series data is partitioned into smaller overlapped sliding windows depending on the sliding window size utilized in the data analysis; ii) A clustering method is carried out on all-time series sliding windows to identify the similar sliding windows. The clustering similar method is based on euclidean distance helps to recognize the similar sliding windows that are close in the distance of space; iii) A clustering similar sliding windows are dealt as training samples belong to suggested model; iv) Least square-support vector machine is applied to generate suitable recommendations for the patients which are having chronic heart diseases in regards to the requirement of taken the medical test or not toward the next upcoming day; v) A comparison has already been done among our suggested model and the researches that already established to solve the identical concern to prove that our technique is superior.

In order to achieve the experimental evaluations, many patients of heart disease were used to collect the dataset of real-life time series. The results acquired has presented that the introduced method was effected in delivering the perfect recommendations for the patients having heart disease and decreasing the workload required in their medical tests. In addition, it has significantly decreased the improper rates of recommendations for these type of patients. The other part of has been designed as follows. In section 2 deep explained about the approach of proposed recommendation that belongs to the patients of chronic disease. Further, in section 3 present the results that carried out from the experiments to accomplish the performance evaluation for the suggested method. Afterwards, in the section 5 the research has been concluded and illustrate the significant future work.

2. RESEARCH METHOD

The present study attempts to explore the overall impact of proposed method to provide the patients that having heart diseases with medical recommendations to fulfill the need of getting the medical test toward the next coming day. In fact, we first explain the architecture's overview for our suggested method. In the next stage, a discussion in details has been concluded on the clustering method and least square support vector machine (LS-SVM), the two technical parts of the suggested method, is explained in the current section.

2.1. Methodology overview

Figure 1 illustrates the structure of the suggested method that recommended to supply a proper medical recommendation to the patients in the environment of telehealth. First of all, a given time series data is partitioned to smaller overlapped sliding windows depend on slide window's size which is empirically determined. In second step, the technique of clustering is used to recognize the similar sliding windows. The clustering similar method is based on euclidean distance aim to recognize the similar sliding windows which are really close in distance of space [28], [29]. Finally, the clustering similar sliding windows are dealt as training samples of LS-SVM classifier in order to make a binary recommendation in regards to the specific condition of patient. The generated medical recommendations will definitely be made the decision to detect if the given patient required to get the medical test toward the next upcoming day or not. More details are presented in the following subsections [27], [30].
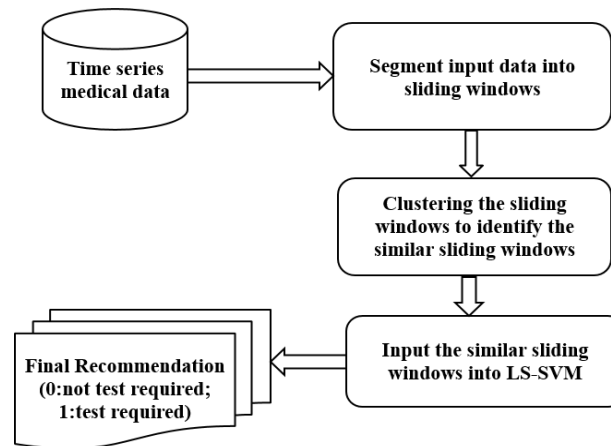

Figure 1. Architecture of the suggested method

2.2. Time series clustering method

For time series data, similarity measures usually include distance method. Euclidean Distance is well-known distance measure in the data mining problem [30], [31]. It is introduced as square root of the summation, which is calculated by using squares of amount of differences among the corresponding coordinates of two points; or as the distance of straight line located between two points in the space of euclidean [f][b]. Let $\{X = X_{ij}, i = 1, 2, 3, .., n; j = 1, 2, 3, ..., m\}$ be a set of data samples. The euclidean distance between the two samples can be explained as:

$$D(x_i, x_k) = \sqrt{\sum_{j=1}^{m}(x_{ij} - x_{kj})^2} \tag{1}$$

where n is the samples number and m is the samples dimension. Based on the above equation, the smaller value of equation means the two samples are more similar.

2.3. Least square support vector machine (LS-SVM)

It is a technique of machine learning which is developed by [32] from the latest version of a support vector machine. A set of linear equations is used for training. In the last years, least square support vector machine has been effectively used to solve the pre-diction and classification issues in medical domain. Due to its high performance to classify the time series data with a high classification accuracy and a minimum time execution [33], it is employed in various fields such as for prediction of muscle fatigue in electromyogram signals [32] and breast cancer prediction [34].

A LS-SVM is one of well-known forms of LS-SVM. It is built to categorize a given dataset into two classes introduced as 1, -1 [8]. In LS-SVM, the given data is mapped towards the space of high-dimensional. Thereafter, it utilized a hyper plane separating the two particular classes required by increasing the distance located between the support vectors and plane. Consider a set of data (x1, y1), (x2, y2), (x3, y3), ..., (xn, yn) Rn, m is the number of data. To isolate these classes, LS-SVM tries to get the most effective separating hyperplane with the highest possible margin. Based on the following rules, LS-SVM is solved a given problem:

$$y_i[(ux_i) + u_0] = 1 - \xi_i \quad i = 1,2,3,\ldots,n \tag{2}$$

$$\frac{1}{2} \parallel u \parallel^2 + \frac{c}{2} \sum_{i=1}^{n} \xi_i{}^2 \tag{3}$$

Based on those formals, a given problem can be drawn up as (4).

$$((u; b; a; \xi) = \frac{1}{2} \parallel u \parallel^2 + \frac{c}{2} \sum_{i=1}^{n} \xi_i{}^2 - \sum_{i=1}^{n} a_i\{y_i[(ux_i) + u_0] - 1 + \xi_i \tag{4}$$

3. EXPERIMENTAL RESULTS

To make performance evaluation for the proposed method, extensive experiments have already been designed and performed using a real-life dataset. First, the evaluations of experiment are discussed in this section, which includes the details of the performance metrics and dataset that utilized in the process of evaluation. Followed by the detailed experimental results.

3.1. Experimental setup

To achieve the test of practical applicability that belongs to suggested recommendation method, Tunstall Healthcare dataset found from industry collaborator are used. This dataset were acquired from an experimental study carried out on a number of patients that are having a disease chronic heart, and the data gathered contained the everyday medical reading of the patients which are dissimilar necessary medical measurements in the environment telehealth. This dataset is basically a time series reading that includes data extracted from 6 patients with 7,147 records of different time series obtained between the months of May to October 2012. Every individual dataset's record divide to a number of patient-related meta-data attributes, such as measurement value, measurement type, visit-id, measurement unit, patient id, date and received date. Table 1 is showing the characteristics of data attributes. The dataset of every day that belong to each patient are also contains a numerical reading that are collected from a number of critical medical measurement numerical readings over the time of study, including diastolic blood pressure (DBP), blood glucose and weight, heart rate, oxygen saturation (SO2), mean arterial pressure (MAP), of which the data related to heart. The Table 1 illustrates the dataset meta-data attributes.

In the current work, the dataset is split into two individual sets: the testing set and training set. The proposed model was trained utilizing the set of training and subsequently approved by the use of testing set. 75% of the dataset were divided as the training data and the rest of 25% were utilized in the study to testing the data. The generated recommendations were reasonably compared to a real readings of the test to evaluate the capability of the suggested model on creating recommendations of high quality. The issue of class-imbalance (this means that the quantity of normal data is bigger compared to abnormal data) in the historical medical data of patient is carefully addressed when training the classifiers. To solve this issue, two methods are used: over-sampling, and under-sampling [35].

Table 1. Data attributes of the dataset

Name of Attribute	Type of Attribute
id	Numeric
patient-id	Numeric
hcn	Numeric
visit-id	Numeric
measurement type	Nominal
measurement value	Numeric
measurement unit	Nominal
measurement question	Nominal
date-received	Numeric
date	Numeric

3.2. The performance metrics

In this paper, three metrics of performance were introduced to get the performance of the potential method in comparison with the benchmark models as follows:
−   Calculating the accuracy (5) that refers to the correctly recommended days in percentage (N_C) against the number of days ($|D|$) in dataset.

$$Accuracy = \frac{N_C}{|D|} * 100\% \tag{5}$$

- On the other hand, workload saving (5), which denotes to the total number of days (NNo) when medical recommendations are generated for only skipping tests against the total number of days, has also been used.

$$Saving = \frac{NN_O}{|D|} * 100\% \tag{6}$$

- Assessing the methods using the risk that has been calculating utilizing (7).

$$Risk = \frac{NR}{|D|} * 100\% \tag{7}$$

where NR denotes the percentage of number of days that having a risky medical recommendation which that predicted as a skipped test for a given medical measurement but they should be suggested as abnormal in the testing set. A correct recommendation is considered when the model produces a recommendation that which "test required" for the following day and a the reading is normal for that day in the dataset. Otherwise, the recommendations are considered incorrect. In this study, the proposed and benchmark methods were developed and evaluated using the MATLAB which running on an Intel i7 processor at 3.40 GHz with 8.00 GB RAM.

3.3. Results and analysis
3.3.1. Evaluating the proposed method
The medical data of patients are clustered to 2 classes according to the characteristics of slide windows. In the cluster 1, most of the slide windows are collected for patients who needs to take a medical test related with a medical measurement on that day. However, cluster 2 contains the slide windows that the patient is not required to take a medical measurement. The clustering similar slide windows are treated as training samples of the proposed model.

3.3.2. Proposed model performance
The size of sliding window has a significant influence on the performance of the model. Therefore, the prediction model is applied with various sliding windows (different sliding windows) in order to improve the performance of the proposed model. Table 2 presents the percentage of accuracy, saving and risk for different sizes of sliding windows. In general, when the value of k ranged from 3 to 5, the proposed model is achieved the best accuracy. In addition, when most recent days are consider, the risk prediction is more accurate. However, to investigate the effectiveness of the prediction model in generating the recommendation with different time period prediction, five different time periods of prediction containing respectively three, four, five, six, and seven days were selected to evaluate the prediction model. Table 3 shows the performance of the prediction model based on the time period of prediction. Based on the results, the prediction time period of 3, 4, and 5 days in advanced achieved high accuracy compared with those of other 6 and 7 days. It is clear that the proposed model yields the highest accuracy with a short time prediction i.e 3-5 days in advanced.

Table 2. The performance of the prediction model based on different sliding window sizes

Size of sliding window	Accuracy (%)	Saving (%)	Risk (%)
3 days	96.00	66.32	01.25
4 days	95.45	65.55	01.90
5 days	95.00	65.80	01.95
6 days	94.20	64.20	02.50
7 days	85.10	61.80	05.80

Table 3. The performance of the prediction model based on the time period of prediction

Time period of prediction	Accuracy (%)	Saving (%)	Risk (%)
3 days	95.00	65.32	02.00
4 days	95.00	65.00	02.00
5 days	96.00	64.80	01.10
6 days	90.20	63.50	04.00
7 days	87.10	62.00	05.00

3.4. Effectiveness comparison with previous methods
A compression of the proposed method with the previous work has been made in this section. For fair comparison, the same Tunstall dataset were employed in the prior work to process the same issue.

A basic heuristic tool was developed for generating suitable recommendations for patients with heart diseases in a telehealth environment [36]. This approach was combined with two methods which are a hybrid method and regression-based prediction algorithm [37] to process the same issue in this field. However, a fast Fourier transformation was used with a machine learning based ensemble model for generating appropriate medical recommendations to patients suffering from chronic diseases [38], [39].

The comparisons results obtained in Table 4 showed that the suggested technique yielded the best accuracy in compression with the benchmark approaches. It is clearly showed that, the accuracy percentage was improved from 94% to 96% while a little improvement from 63% to more than 65% using workload saving has been shown. Additionally, the recommendation risk of the method used in this study was also lower than the three procedures developed in this study.

Table 4. The prediction model performance comparison with previous methods

| Method | Techniques used | Tunstall medical dataset | | |
		Accuracy (%)	Saving (%)	Risk (%)
[36]	Basic generated heuristic algorithm	86	10	8
[37]	Basic generated heuristic algorithm Regression-based algorithm and Hybrid algorithm	91	15	5
[38]	Fast Fourier transformation coupled with ensemble model	94	63	3
Proposed method	Clustering method and a least square-support vector machine	96	65	1.25

4. CONCLUSIONS AND FUTURE WORK

In this pilot study, a new method was introduced utilizing the clustering method and a least square-support vector machine for predicting a short-term disease risk. The obtained results found that our new method can possibly utilized as a better effective tool of medical test recommendation for the environment of telehealth to patients with heart disease. The classification model utilized in the method requires the effective use of euclidean distance as a similarity measurement with least square-support vector machine to obtain whether or not a certain patient requires to have a test for his physical body at this moment using the facility of telehealth. Using 3-5 days frame time, the proposed method yielded a better predictive performance in comparison to the benchmark frame times. Our findings showed that using the most recent days (the sliding window size is ranging between 3-5 days) gives a higher achievement of the suggested method compared to the other sliding window sizes. Additionally, the suggested method is compared to with some of previous works implemented to find out the solution for the identical issue. According to the evaluations got from the experimental work, we are aware that the suggested method could be applied as an impactful tool to enhance the aspect of decision-making by which the cost and time related to the everyday medical test can easily be decreased. The study results proved that the suggested method is an efficient for enhancing the aspect of the medical decisions relying on the clinical evidence and decreasing the time costs caused by the patients of chronic diseases by getting their medical tests on daily basis, by which giving them a better generic lives.

As a future work, collect a set of techniques, like Adaboost and boosting, may be applied to generate accurate and perfect recommendations and executing a comparison study between different models of ensemble. Moreover, implement the current suggested method to the patients having another types of disease in support telehealth care to verify the suggested method is more comprehensive in regards to the data of medical time series.

REFERENCES

[1] N. Potischman, R. Troisi, and L. Vatten, "A life course approach to cancer epidemiology," in *A Life Course Approach to Chronic Disease Epidemiology*, Oxford University Press, 2004, pp. 260–280.

[2] L. Lavoie, H. Khoury, S. Welner, and J. B. Briere, "Burden and prevention of adverse cardiac events in patients with concomitant chronic heart failure and coronary artery disease: a literature review," *Cardiovascular Therapeutics*, vol. 34, no. 3, pp. 152–160, May 2016, doi: 10.1111/1755-5922.12180.

[3] A. R. Dewar, T. P. Bull, D. M. Malvey, and J. L. Szalma, "Developing a measure of engagement with telehealth systems: the mHealth technology engagement index," *Journal of Telemedicine and Telecare*, vol. 23, no. 2, pp. 248–255, Jul. 2016, doi: 10.1177/1357633X16640958.

[4] J. Wang, M. Qiu, and B. Guo, "Enabling real-time information service on telehealth system over cloud-based big data platform," *Journal of Systems Architecture*, vol. 72, pp. 69–79, Jan. 2017, doi: 10.1016/j.sysarc.2016.05.003.

[5] L. Van Ma, J. Kim, S. Park, J. Kim, and J. Jang, "An efficient session-weight load balancing and scheduling methodology for high-quality telehealth care service based on WebRTC," *Journal of Supercomputing*, vol. 72, no. 10, pp. 3909–3926, Jan. 2016, doi: 10.1007/s11227-016-1636-8.

[6] N. Kulkarni, "Support vector machine based alzheimer's disease diagnosis using synchrony features," *International Journal of Informatics and Communication Technology (IJ-ICT)*, vol. 9, no. 1, p. 57, Apr. 2020, doi: 10.11591/ijict.v9i1.pp57-62.

[7] R. Lafta, J. Zhang, X. Tao, Y. Li, M. Diykh, and J. C.-W. Lin, "A structural graph-coupled advanced machine learning ensemble model for disease risk prediction in a telehealthcare environment," in *Studies in Big Data*, Springer Singapore, 2018, pp. 363–384.

[8] K. Polat and S. Güneş, "Breast cancer diagnosis using least square support vector machine," *Digital Signal Processing: A Review Journal*, vol. 17, no. 4, pp. 694–701, Jul. 2007, doi: 10.1016/j.dsp.2006.10.008.

[9] J. G. Yang, J. K. Kim, U. G. Kang, and Y. H. Lee, "Coronary heart disease optimization system on adaptive-network-based fuzzy inference system and linear discriminant analysis (ANFIS-LDA)," *Personal and Ubiquitous Computing*, vol. 18, no. 6, pp. 1351–1362, Oct. 2014, doi: 10.1007/s00779-013-0737-0.

[10] A. S. Sánchez, F. J. I.-Rodríguez, P. R. Fernández, and F. J. d. C. Juez, "Applying the K-nearest neighbor technique to the classification of workers according to their risk of suffering musculoskeletal disorders," *International Journal of Industrial Ergonomics*, vol. 52, pp. 92–99, Mar. 2014, doi: 10.1016/j.ergon.2015.09.012.

[11] M. S. Mohktar *et al.*, "Predicting the risk of exacerbation in patients with chronic obstructive pulmonary disease using home telehealth measurement data," *Artificial Intelligence in Medicine*, vol. 63, no. 1, pp. 51–59, Jan. 2015, doi: 10.1016/j.artmed.2014.12.003.

[12] J. Zhang *et al.*, "Coupling a fast fourier transformation with a machine learning ensemble model to support recommendations for heart disease patients in a telehealth environment," *IEEE Access*, vol. 5, pp. 10674–10685, 2017, doi: 10.1109/ACCESS.2017.2706318.

[13] C. D. Chang, C. C. Wang, and B. C. Jiang, "Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors," *Expert Systems with Applications*, vol. 38, no. 5, pp. 5507–5513, May 2011, doi: 10.1016/j.eswa.2010.10.086.

[14] F. Huang, S. Wang, and C. C. Chan, "Predicting disease by using data mining based on healthcare information system," in *Proceedings - 2012 IEEE International Conference on Granular Computing, GrC 2012*, Aug. 2012, pp. 191–194, doi: 10.1109/GrC.2012.6468691.

[15] H. Y. Lu, C. Y. Huang, C. T. Su, and C. C. Lin, "Predicting rotator cuff tears using data mining and bayesian likelihood ratios," *PLoS ONE*, vol. 9, no. 4, p. e94917, Apr. 2014, doi: 10.1371/journal.pone.0094917.

[16] V. Krishnaiah, G. Narsimha, and N. S. Chandra, "Diagnosis of lung cancer prediction system using data mining classification techniques," *International Journal of Computer Science and Information Technologies*, vol. 4, no. 1, pp. 39–45, 2013.

[17] D. Y. Yeh, C. H. Cheng, and Y. W. Chen, "A predictive model for cerebrovascular disease using data mining," *Expert Systems with Applications*, vol. 38, no. 7, pp. 8970–8977, Jul. 2011, doi: 10.1016/j.eswa.2011.01.114.

[18] X. Zhu, S. Zhang, R. Hu, Y. Zhu, and J. Song, "Local and global structure preservation for robust unsupervised spectral feature selection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 3, pp. 517–529, Mar. 2018, doi: 10.1109/TKDE.2017.2763618.

[19] W. Zheng, X. Zhu, Y. Zhu, R. Hu, and C. Lei, "Dynamic graph learning for spectral feature selection," *Multimedia Tools and Applications*, vol. 77, no. 22, pp. 29739–29755, Oct. 2018, doi: 10.1007/s11042-017-5272-y.

[20] X. Zhu, X. Li, S. Zhang, Z. Xu, L. Yu, and C. Wang, "Graph PCA hashing for similarity search," *IEEE Transactions on Multimedia*, vol. 19, no. 9, pp. 2033–2044, Sep. 2017, doi: 10.1109/TMM.2017.2703636.

[21] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient kNN classification with different numbers of nearest neighbors," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 5, pp. 1774–1785, May 2018, doi: 10.1109/TNNLS.2017.2673241.

[22] R. Hu *et al.*, "Graph self-representation method for unsupervised feature selection," *Neurocomputing*, vol. 220, pp. 130–137, Jan. 2017, doi: 10.1016/j.neucom.2016.05.081.

[23] G. Sun *et al.*, "Short-term wind power forecasts by a synthetical similar time series data mining method," *Renewable Energy*, vol. 115, pp. 575–584, Jan. 2018, doi: 10.1016/j.renene.2017.08.071.

[24] R. Likhitha and A. Manjunatha, "Power quality disturbances classification using complex wavelet phasor space reconstruction and fully connected feed forward neural network," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 6, pp. 2980–2988, Dec. 2021, doi: 10.11591/eei.v10i6.3207.

[25] U. N. Wisesty and T. R. Mengko, "Comparison of dimensionality reduction and clustering methods for sars-cov-2 genome," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 4, pp. 2170–2180, Aug. 2021, doi: 10.11591/EEI.V10I4.2803.

[26] A. Alzu'Bi and M. Barham, "Automatic BIRCH thresholding with features transformation for hierarchical breast cancer clustering," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 2, pp. 1498–1507, Apr. 2022, doi: 10.11591/ijece.v12i2.pp1498-1507.

[27] Q. M. Shallal, Z. A. Hussien, and A. A. Abbood, "Method to implement K-NN machine learning to classify data privacy in IoT environment," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 20, no. 2, pp. 985–990, Nov. 2020, doi: 10.11591/ijeecs.v20.i2.pp985-990.

[28] Y. Pratama, I. G. E. Dirgayussa, P. F. Simarmata, and M. H. Tambunan, "Detection roasting level of lintong coffee beans by using euclidean distance," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 6, pp. 3072–3082, Dec. 2021, doi: 10.11591/eei.v10i6.3153.

[29] T. A. Salih, M. T. Ghazal, and Z. G. Mohammed, "Development of a dynamic intelligent recognition system for a real-time tracking robot," *IAES International Journal of Robotics and Automation (IJRA)*, vol. 10, no. 3, p. 161, Sep. 2021, doi: 10.11591/ijra.v10i3.pp161-169.

[30] M. G. C. P. and R. Hajare, "Space-time trellis codes: Field programmable gate array approach," *International Journal of Reconfigurable and Embedded Systems (IJRES)*, vol. 9, no. 3, p. 213, Nov. 2020, doi: 10.11591/ijres.v9.i3.pp213-223.

[31] E. San Segundo, A. Tsanas, and P. Gómez-Vilda, "Euclidean distances as measures of speaker similarity including identical twin pairs: A forensic investigation using source and filter voice characteristics," *Forensic Science International*, vol. 270, pp. 25–38, Jan. 2017, doi: 10.1016/j.forsciint.2016.11.020.

[32] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, no. 3, pp. 293–300, 1999, doi: 10.1023/A:1018628609742.

[33] Y. Bai, X. Han, T. Chen, and H. Yu, "Quadratic kernel-free least squares support vector machine for target diseases classification," *Journal of Combinatorial Optimization*, vol. 30, no. 4, pp. 850–870, Mar. 2015, doi: 10.1007/s10878-015-9848-z.

[34] N. S. A. Sharawardi, Y. H. Choo, S. H. Chong, A. K. Muda, and O. S. Goh, "Single channel sEMG muscle fatigue prediction: An implementation using least square support vector machine," in *2014 4th World Congress on Information and Communication Technologies, WICT 2014*, Dec. 2014, pp. 320–325, doi: 10.1109/WICT.2014.7077287.

[35] S. Wang, Z. Li, W. Chao, and Q. Cao, "Applying adaptive over-sampling technique based on data density and cost-sensitive SVM to imbalanced learning," Jun. 2012, doi: 10.1109/IJCNN.2012.6252696.

[36]   R. Lafta, J. Zhang, X. Tao, Y. Li, and V. S. Tseng, "An intelligent recommender system based on short-term risk prediction for heart disease patients," in *Proceedings - 2015 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2015*, Dec. 2016, pp. 102–105, doi: 10.1109/WI-IAT.2015.47.
[37]   R. Lafta *et al.*, "An intelligent recommender system based on predictive analysis in telehealthcare environment," *Web Intelligence*, vol. 14, no. 4, pp. 325–336, Nov. 2016, doi: 10.3233/WEB-160348.
[38]   R. Lafta *et al.*, "A fast fourier transform-coupled machine learning-based ensemble model for disease risk prediction using a real-life dataset," in *Lecture Notes in Computer Science,* vol. 10234 LNAI, Springer International Publishing, 2017, pp. 654–670.
[39]   A. A. Abbood, Q. M. Shallal, and H. K. Jabbar, "Intelligent hybrid technique to secure bluetooth communications," in *Advances in Intelligent Systems and Computing*, vol. 1333 AISC, Springer Singapore, 2021, pp. 135–144.

BIOGRAPHIES OF AUTHORS

Raid Luaibi Lafta received the B.Sc. degree in computer science from University of Thi-Qar, Iraq, in 2002, and the M.Sc. degree in information technology from Voronezh State University, Russia, in 2010. He obtained his PhD degree in 2018 from the University of Southern Queensland in Australia. His research interests include recommendation systems and predictive data mining. He can be contacted at email: raidluaibi.Lafta@utq.edu.iq.

Mohanad S. AL-Musaylh, received his bachelor's and master's degrees in mathematics from Basra University in south of Iraq. He obtained his PhD degree in 2020 from the University of Southern Queensland in Australia. Currently working as, the director of scientific and graduate studies in Management Technical College of Basra at Southern Technical University in Iraq. He has published more than 6 papers in journals that indexed in Scopus, and more than 3 papers in reputed journals and confernces. His research interests include data intelligence models, machine learning, deep learning, artificial intelligence models and data forecasting. He can be contacted at email: Mohanad.al-musaylh@stu.edu.iq.

Qahtan Makki Shallal, received his bachelor's degree in computer science from Basra University in south of Iraq. He obtained his master's degree from Jamia Hamdard at New Delhi in 2008. He received his PhD degree in 2018 from AMU in India. He has more than 10 years teaching experience and currently working a Head of Information technologies department in Management Technical College of Basra at Southern Technical University in Iraq. He has published more than 12 papers in journals that classified as Scopus, and more than 10 papers in reputed journals and confernces. His research interests include wireless networking, IoT, cloud computing security, and machine learning. He can be contacted at email: qahtan.makii@stu.edu.iq.