ISSN: 2230-9926

Available online at http://www.journalijdr.com

IJDR

International Journal of Development Research
Vol. 15, Issue, 02, pp. 67764-67768, February, 2025
https://doi.org/10.37118/ijdr.29189.02.2025

RESEARCH ARTICLE OPEN ACCESS

COMBINING MOTIF INFORMATION AND K-NEAREST-NEIGHBORS FOR FORECASTING HOSPITAL OUTPATIENT VISITS FLOW

*1Duong Tuan Anh and 2Nguyen Nhut Tan

1Department of Information Technology, Ho Chi Minh University of Foreign Languages and Information Technology, 828 Su Van Hanh Street, District 10, Ho Chi Minh City, Vietnam; 2Ho Chi Minh City Hospital of Dermato-Venereology, 2 Nguyen Thong Street, District 3, Ho Chi Minh City, Vietnam

ARTICLE INFO

Article History:
Received 11th December, 2024
Received in revised form
16th December, 2024
Accepted 19th January, 2025
Published online 28th February, 2025

Key Words:
Hospitaloutpatient visits flow, Time series motif, Motif discovery algorithm, K-nearest-neighbors, Artificial Neural Network (ANN).

Corresponding author: Duong Tuan Anh

ABSTRACT

Effective hospital outpatient visits flow forecasting is an important task for modern hospitals to implement intelligent management of medical resources. Since outpatient visits flow may be nonlinear and dynamic, we propose a hybrid model, which combines motif information and k-nearest-neighbors regressor. Time series motif is a previously unknown pattern appearing frequently in a time series. In the proposed approach, we first discover time series motif by using a segmentation-based method and then exploit motif information for forecasting in combination with a k-nearest-neighbors (kNN) model. The proposed approach is called kNN+Motif. To demonstrate that our kNN+Motif method is robust, we applied the new approach to forecast the outpatient visits flow in Ho Chi Minh City Hospital of Dermato-Venereology. The experiment was implemented to compare the proposed forecasting model against the single k-nearest-neighbors model and artificial neural network (ANN) model. The experimental results demonstrate that the proposed kNN+Motif model is more effective than the single k-nearest-neighbors method and ANN model. Besides, the kNN+Motif can run much faster than thesingle kNN model.

Citation*: *Duong Tuan Anh, Nguyen Nhut Tan. 2025. *"*Combining Motif Information and k-nearest-neighborsfor Forecasting Hospital Outpatient Visits Flow*". International Journal of Development Research*, 15, (02), 67764-67768.

INTRODUCTION

Accurate forecasting the healthcare demand and resource availability becomes more important and critical. Outpatient departments which play an important role in hospital service experience increasing pressure year by year (Lou *et al.*, 2017). Effective hospital outpatient visits forecasting is beneficial for the suitable planning and allocation of healthcare to meet the medical demands. Due to multiple characteristics of daily outpatient visits, such as randomness, cyclicity and trend (Lou *et al.*, 2017), in the past, several researchers have studied various methods for this prediction problem such as statistics analysis and machine learning methods. Some typical research works can be listed as follows. Li *et al.* (2014) proposed ARIMA (AutoRegressive Integrated Moving Average) model for forecasting monthly outpatient visits flow at a grade 3 hospital in China. Kim *et al.* (2020) compared the performances of two methods: ARIMA and SARIMA in forecasting outpatient visits flow on the datasets of some hospitals in Gangnam-gu, Seoul and the results show that SARIMA brings out the better prediction performance than ARIMA. Sumitra and Basri (2020) compared the performances of three methods: ARIMA, Simple Exponential Smoothing and Holt-Winters in forecasting outpatient visits flow at a Community Health Center in Indonesia and the results show that Holt-Winters brings out the best prediction performance.Guan and Elgelhardt (2019)compared the performances of four methods: linear regression, ANN, Recurrent Neural Network (RNN) and Long-Short-Term-Memory (LSTM) neural networks in forecasting pediatric outpatient volume. Thapa and Timalsina (2023) proposed Gated Recurrent Unit (GRU), an improved variant of LSTM, in forecasting hospital outpatient volume. There have been several hybrid methods proposed for forecasting outpatient visits flow which are listed as follows. Wang *et al.* (2015) proposed a hybrid method which combines a time series decomposition technique (EEMDAN) with multiple local predictor fusion for forecasting diarrhoea outpatient visits at some hospitals in Shanghai, China. Lou *et al.* (2017) used a combination model based on ARIMA and Simple Exponential Smoothing (SES) for forecasting outpatient visits flow at some large hospitals in China. Huang and Wu (2017) proposed a hybrid method which combined empirical mode decomposition (EMD)technique with multiple artificial neural networks in forecasting hospital outpatient volume. Jiang *et al.* (2019) proposed a combination method which couples deep neural networks and classical regression models for forecasting outpatient volume in Hong Kong. Deng *et al.* (2023) proposed a hybrid model which combines ARIMA with deep neural network LSTM model for forecasting outpatient visits flow at some departments at the First

Hospital of Shanxi Medical University in China. The hybrid model by Deng *et al.* (2023) aims to combine the ability of capturing the linear components of the time series by ARIMA model with the ability of capturing the nonlinear components of the time series by LSTM model.

Truong *et al.* (2012) proposed a hybrid framework for time series forecasting which combined motif information with a conventional prediction method. Motifs are frequently occurring subsequences in a long time series. In the work by Truong *et al.*, (2012), the proposed approach first discovers time series motif by using a segmentation-based method, called EP-C and then exploit the obtained motif information in combination with an artificial neural network (ANN) model for time series forecasting.

In this paper, we modify the framework in the work by Truong *et al.* (2012) with the main idea: we combine motif information with k-nearest-neighbor regressor in forecasting hospital outpatient visits flow. The newapproach is called kNN+Motif. In kNN+Motif, motif information and the k-nearest-neighbor regressor are complementary. With the new approach, our main contributions are as follows:

- We combine motif information with k-nearest-neighbor predictor in seasonal time series forecasting.
- We apply the proposed method for forecasting outpatient visits flow at Hospital of Dermato-Venereology in Ho Chi Minh City.
- We compare empirically the effectiveness of our proposed time series forecasting method with those of the two other typical methods: kNN and ANN model. Experimental results reveal that kNN+Motif performs better than the single kNN and ANN model in forecasting hospital outpatient visits flow. Besides, the kNN+Motif can run much faster than the single kNN model.

METHODOLOGY

Some Definitions: A time series $T = t_1, t_2, \ldots, t_m$ is an ordered set of *m* real values measured at equal intervals. Given a time series *T* of length *m*, a subsequence *C* is a subsection of length $n<m$ of contiguous positions from *T*, i.e., $C = t_p, t_{p+1}, \ldots, t_{p+n-1}$, for $1 \leq p \leq m-n+1$.

Definition 1. *Distance function*: *Dist*(*C, M*) is a positive value used to measure the difference between two time series *C* and *M*, based on some measure method.

Definition 2. *Match*: Given a positive real number *r* (called *range*) and a time series *T* containing a subsequence *C* beginning at position *p* and a subsequence *M* beginning at *q*, if the distance $Dist(C, M) \leq r$, then *M* is called a matching subsequence of *C*.

Definition 3. *Trivial Match*: Given a time series *T*, its two subsequences *P* of length *n* starting at position *p* and *Q* starting at position *q*, we say that *Q* is a trivial-match to *P*, if either $p = q$ or there does not exist a subsequence *Q'* beginning at *q'* such that $Dist(P, Q') \geq r$ and either $q<q'<p$ or $p<q'<q$.

Definition 4. *1-motif*: Given a time series *T*, a subsequence length *n* and a range *r*, the most significant motif in *T* (called the 1-motif) is the subsequence C_1 that has the highest count of non-trivial matches.

All subsequences that are similarto the 1-motif are called *instances* of that 1-motif. Figure 1 illustrates a motif consisting of three instances with length 40 in an electrocardiography (ECG) time series.

The *k*-th most significant motif in *T* (called the *k-motif*) is the subsequence C_k that has the highest count of non-trivial matches and satisfies $D(C_k, C_i) > 2r$ for all $1 \leq i \leq k$. All these above definitions are from the first formal definition of time series motif given by Lin *et al.* (2002).

Discovering Time Series Motifs by EP-C Algorithm: The first segmentation-based method for time series motif discovery is the work proposed by Gruber *et al.* (2006). This method is based on the concept of significant extreme points that was proposed by Pratt and Fink (2002). Figure 2 illustrates the definition of significant extreme point which can besignificant minimum (a) and significantmaximum (b). Given a time series, starting at the beginning of the time series, all significant minima and maxima of the time series are computed by using the algorithm given in Pratt and Fink (2002).

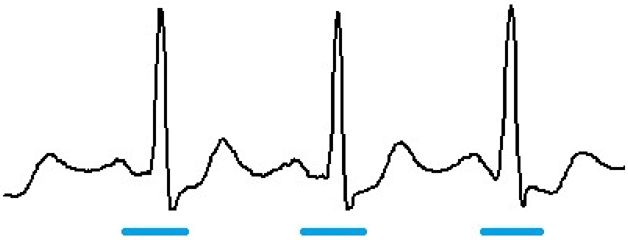

Figure 1. An electrocardiograph time series with three motif instances with length 40

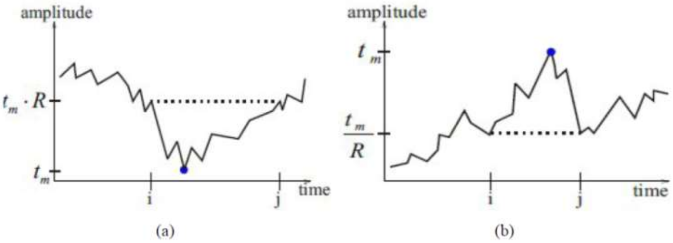

Figure 2. Illustration of significant extreme points, (a) minimum (b) maximum

The algorithm proposed by Gruber *et al.* (2006) for finding time series motifs consists of three steps: extracting significant extreme points, determining motif candidates from the extracted significant extreme points and clustering the motif candidates. Each motif candidate is the subsequence in the time series that is bounded by two important extreme points (the *i*-th and the (*i*+2)-th). Motif candidates are the subsequences that may have different lengths. To enable the computation of distances between them, the authors bring them to the same length by using homothety transform or spline interpolation. After clustering, the largest cluster contains the instances of the 1-motif in the time series. This algorithm is called EP-C (Extreme Points and Clustering). Through the experiments reported in Truong *et al.* (2012), EP-C is much more effective than Random Projection (Chiu *et al.*, 2003) in terms of time efficiency and motif accuracy. One interesting property of EP-C is that EP-C can discover not only the 1-motif but also all *i*-motif for all $1 \leq i \leq k$ since each cluster (among *k* clusters) found after clustering corresponding with an *i*-motif.

```
EP-C(T, motif_length)
// T is the time series, motif_length is a user-defined length for //motif candidates
1. SIG = Significant-Points (T)
// SIG is the array containing significant extreme points
// identified from time series T
2. fori = 1 to length(SIG) -2 do
3.  Subs[i] = T(SIG(i) to SIG(i+ 2))
// Subs is the array containing all motif candidates
4.  Subs = Homothety (Subs, motif_length)
// Convert all the extracted subsequences to the same length,
//i.e., motif_length; motif_length can be computed as the average
// length of all the extracted subsequences
5. (c_1, c_2,...,c_k) = Clustering (Subs,k)
//k: the number of clusters
// c_1, c_2,...,c_k are sorted in decreasing order of the number of
// subsequences in each cluster
```

// The most significant motif is represented by c_1, the cluster
// with the highest number of subsequences.

The k-Nearest-Neighbors Regressor: In this study, we use k-nearest neighbors method as a predictor in forecasting hospital outpatient visits flow. This method is a nonparametric method in time series forecasting. Nonparametric regression does not require any prior knowledge about the process to be modeled. Among nonparametric approaches the k-nearest-neighbors (kNN) method has shown to be promising and has been successfully applied in various prediction studies because of its ability to tolerate high-dimension and incomplete data (Lin *et al.*, 2012; Martinez *et al.*, 2019). The forecasting ability and simplicity of kNN makes its adaptation to hospital outpatient visits data suitable. Due to all these reasons, kNN is selected to be used as a machine learning predictor in this study. The main idea of applying kNN for time series forecasting is straighforward; given a series $T = (t_1, t_2, \ldots, t_n)$, the problem is to forecast t_{n+h}, where h is the forecast horizon. For predicting a pattern Q with target t_{n+1}, which contains $(t_{n-l}, \ldots, t_{n-1}, t_n)$, the kNN regressor method searches for the k most similar subsequences to Q. When the k most similar subsequences are found, the target of Q is obtained by averaging the targets of the k found subsequences. The l parameter in kNN is called the *pattern size*. Figure 3 illustrates an example of kNN algorithm (with k=2) in which there are three subsequences idenfified a stock time series. The three subsequences are highlighted. The third subsequence (C) is the pattern in question; the first (A) and the second subsequence (B) are the two nearest-neighbors which are found by the 2-NN algorithm for the pattern.

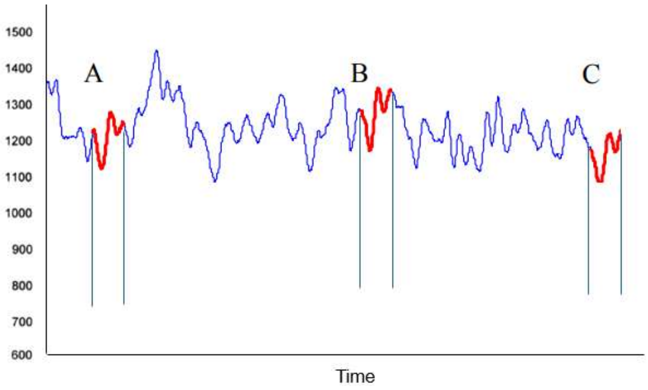

Figure 3. An example of k-nearest-neighbors with k = 2

The Proposed Method for Forecasting Hospital Outpatient Visits Flow: The process of finding the most significant motif in a time series can be seen as a training phase of the proposed forecasting method. The prediction phase after this training step is described as follows.

1. After finding the most significant motif in a time series T, we divide the whole motif into two parts: (a) the *prefix* that is the subsequence from the extreme point $ep_i$to the extreme pointep_{i+1} and (b) the *suffix* that is the subsequence from the extreme point $ep_{i+1}$to the extreme pointep_{i+2}. Notice that the prefix and suffix of all instances of the motif may have different lengths.
2. We determine the average length of the prefixes of all the instances of the motif, *av_prefix_length*. Then we apply homothety to transform all these prefixes to the subsequences of the same length (*av_prefix_length*). We determine the largest distance *R_prefix_max* between the transformed prefixes.
3. Given the current state (pattern) of the time series that we have to predict the value of the next time step, we determine the two last extreme points in this pattern and extract the subsequence between these two extreme points. Then we apply homothety to transform this subsequence to the length *av_prefix_length*. Wecompute the largest distance between this subsequence with the prefixes of all the instances of the motif.During this process we also determine the smallest distance between the subsequence in question with some prefix in the instances of the motif, we call this motif instance *most_similar_motif.* If this smallest distance is less than *R_prefix_max* then the subsequence in question matches with the prefix of a motif instance and we call this subsequence *prefix_predictable_motif,* and compute the prediction at the next time step using the suffix of *most_similar_motif.* Otherwise, we pass the current state to the kNN model to tackle the prediction.

Figure 4 illustrates the flow chart of the proposed kNN+Motif method.

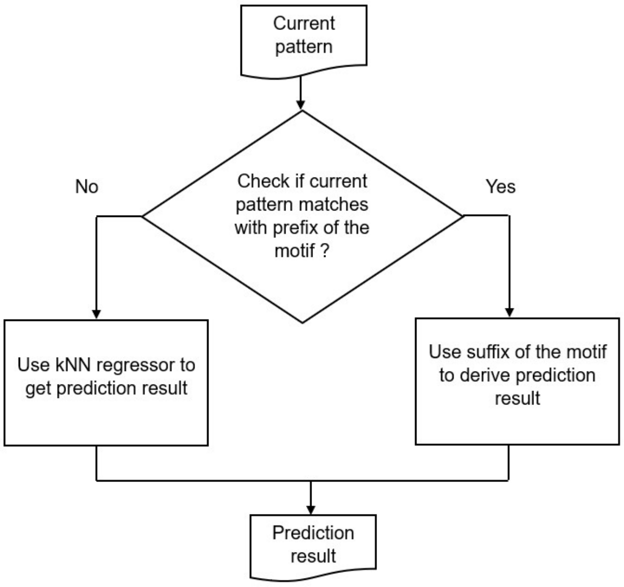
Current pattern
Check if current pattern matches with prefix of the motif ?
No
Yes
Use kNN regressor to get prediction result
Use suffix of the motif to derive prediction result
Prediction result

Figure 4. The flow chart of the kNN+Motif forecasting method

RESULTS AND DISCUSSION

To evaluate the effectiveness of the proposed approach kNN+Motif, we implemented the following experiment in which kNN+Motif is compared to the kNN model and the ANN model in hospital outpatient visits flow forecasting. We implemented k-nearest-neighbors repressor and ANN model by using the Scikit-learnlibrary (Hacketing, 2017). As for the motif discovery algorithm, EP-C, we implemented it by using Python programming language. We conducted the experiment on an Intel(R) Core™ i7-1081 CPU@ 1.61GHz RAM32GB PC.

Finally, only one-day ahead forecasting (a form of short-term forecating) is considered in this research.

The Dataset

The data for our study come from the Hospital of Dermato-Venereology in Ho Chi Minh City during the period from January 2016 to December 2023 with the total number of 5.666.002 outpatient visits. Thus, the dataset contains 2785 observations in time series. Figure 5 provides the curve of the daily outpatient arrivals during the eight years from 2016 to 2023. From Figure 5 we can intuitively see the nonstationary patterns of the time series. Figure 6 provides the curve of the daily emergency patient arrivals in one month from January 1st, 2018 to January 30th, 2018. From Figure 5 and Figure 6, it can be seen that the daily outpatient visits fluctuate greatly, especially on weekends and holidays. Moreover, a seasonal variation pattern is found over the period of one week. The dataset is divided into two sub-datasets: the training dataset (80%) and the testing dataset (20%).

Parameter Setting: After applying EP-C algorithm to find the 1-motif in the time series of hospital outpatient visits flow, we obtain the 1-motif with the length 7. Through experiment, we can determine the suitable values for the parameters of kNN+Motif, the single kNN method and ANN model. The best fit parameter values of the three method for the test dataset are given in Table 1.

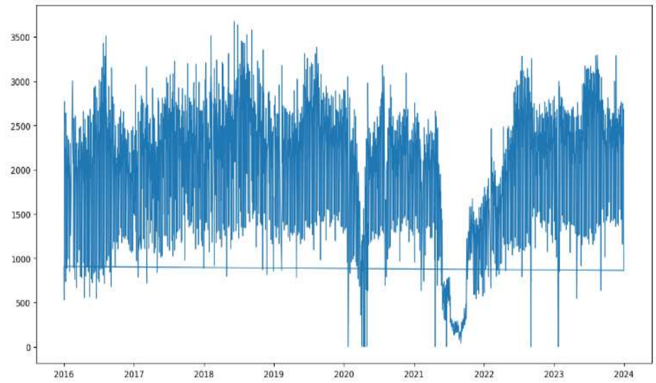

Figure 5. The daily outpatient visits from January 2016 to December 2023

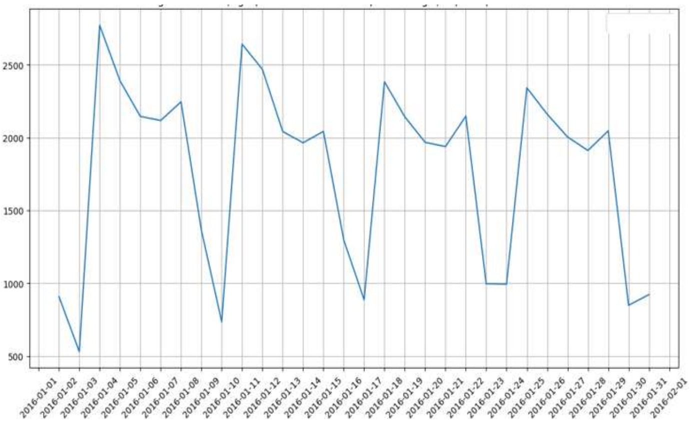

Figure 6. The daily outpatient visits in one month from January 1st , 2016 to January 31th , 2016.

Table 1. Parameter estimation results of the three methods

Method	Parameter values
kNN+Motif	Pattern size = 4, k = 3, Motif length = 7 Distance= Euclid
kNN	Pattern size = 4, k = 3 Distance= Euclid
ANN	Input layer: number of units = 128 Hidden layer: number of units = 64 Batch_size = 64 Number of epochs = 500 Optimizer = Adam

Prediction Evaluation Measures: In this study, the root mean squared error (RMSE), and the mean absolute percentage error (MAPE) are used as evaluation criteria. The formula for RMSE, and MAPE are given as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{2}(y_i - \hat{y}_i)^2} \qquad (1)$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|\hat{y}_t - y_t|}{y_t} \times 100 \qquad (2)$$

where n is the number of observations, y_t is the actual value for time point$t$, and $\hat{y}_t$ is the forecast value for time point$t$.

The use of these measures represents different aspects to evaluate forecasting models. The first is absolute performance measure while the last one (MAPE) is a relative measure.The MAPE is a scale-invariant statistic that expresses error as a percentage. The model has higher predictive accuracy when RMSE or MAPE is much closer to 0.

EXPERIMENTAL RESULTS

In the experiment, we compare kNN+Motif with the two other methods: the kNN regressor model and the ANN model. The MAPE and RMSE error results of the three forecasting methods are given in Table 2. From the experimental results in Table 2, we can see that:

- Based on both the MAPE and RMSE indicators, the kNN+Motif approach performs better than the k-NN approach. As for kNN+Motif, the improvement based on MAPE in comparison to the kNN on the test dataset is 5.47% and the improvement based on RMSE in comparison to the kNN on the test dataset is 49.66%.

- Both kNN+Motif and kNN approaches perform better than ANN model. This research is the first work which compares the performances of two typical forecasting methods: kNN and ANN in outpatient volume forecasting and finds out that kNN is better than ANN in this special forecasting problem.

- The order of the three comparative methods in terms of prediction accuracy on the test dataset can be listed as follows: kNN+Motif > kNN > ANN.

Besides prediction accuracy, we also assessed the three methods kNN+Motif, kNN and ANN in terms of running time. The running times of the three methods in seconds are given in Table 3.

Table 2. Forecasting errors of the three methods on the daily on hospital outpatient visits dataset

Method	MAPE	RMSE
kNN+Motif	16.24%	256.4
kNN	17.18%	509.37
ANN	17.97%	514.16

From the experimental results in Table 3, we can see that kNN+Motif improves the running time remarkably in comparison to kNN. The methos kNN+Motif is about 79.5 times faster than kNN. This finding indicates that by using Motif information, the combined method, kNN+Motif, can reduce remarkably the work load of the kNN predictor on seasonal time series. Furthermore, kNN+Motif can run much faster than ANN.

Table 3. Execution times (in seconds) of each method

Method	Execution time
kNN+Motif	0.88
kNN	70
ANN	38.53

CONCLUSIONS

This paper contributes to exploration of prediction method to forecast outpatient visits flow at Hospital of Dermato-Venereology in Ho Chi Minh City. We have proposed a method which combines motif information with kNN regressor for short-term forecasting on the above-mentioned dataset. In our proposed method, we used a segmentation-based method, called EP-C, for motif discovery in time series. Experimental results on the test dataset reveal that the performance of kNN+Motif is better than those of the single kNN and ANN model. Besides, the kNN+Motif method can run much faster than the single kNN method. Furthermore, the proposed method is by no means limited to application in outpatient visits flow prediction and can easily adapted to other applications. As for future work, we plan to forecast outpatient volume on the dataset collected from one other hospital in Ho Chi Minh City, Vietnam. Besides, we intend to combine motif-information with some deep learning neural networks, such as LSTM or GRU (Thapa and Timalsina, 2023), in hospital outpatient visit flow forecasting.

Acknowledgement: The authors are grateful to the Ho Chi Minh City Hospital of Dermato-Venereology for providing the data.

REFERENCES

Chiu, B., Keogh, E. and Lonardi, S. (2003) Probabilistic discovery of time series motifs. In:*Proc. of9th Int. Conf. on Knowledge Discovery and Data Mining* (KDD 2003), pp.493–498.

Deng, Y., Fan, H., Wu, S. (2023) A hybrid ARIMA-LSTM model optimized by BP in the forecast of outpatient visits.*Journal of Ambient Intelligence and Humanized Computing*, vol. 14, pp.5517-5524.

Gruber, C., Coduro, M., Sick, B.(2006) Signature verification with dynamic RBF network and time series motifs. In: *Proc. of 10th International Workshop on Frontiers in Hand Writing Recognition*,Université de Rennes 1, October, La Baule, France.

Guan, G.,Engelhardt, B. E. (2019) Predicting Sick Patient Volume in a Pediatric Outpatient Setting using Time Series Analysis, *Machine Learning for Healthcare*vol. 106, pp.1-16.

Hacketing, G. (2017) Mastering Machine Learning with scikit-learn, PACKT-Publishing, Birmingham-Mumbai.

Huang, D., Wu, Z. (2017) Forecasting outpatient visits using empiricalmode decomposition coupled with backpropagation artificial neural networks optimized by particle swarm optimization. *PLOS ONE*, vol. 7, February 21.

Jiang, S., Xiao, R., Wang, L., Luo, X.,Huang, C.,Wang, J.H., Chin, K.S.,and Nie, X. (2019) Combining Deep Neural Networks and ClassicalTime Series Regression Models for Forecasting Patient Flows in Hong Kong, IEEE ACCESS2936550.

Kim, K.R., Park, J.E., Jang, I. T. (2020)Outpatient forecasting model in spine hospital using ARIMA and SARIMA methods, *Journal of Hospital Management and Health Policy*, vol. 4, September.

Li, Y., Wu. F., Zheng, C., Hou. K., Wang, K., Sun, N. (2014) Predictive analysis of outpatient visits to a grade 3, class a hospital using ARIMA model. In: *Proceedings of the 2014 International symposium on information technology* (ISIT 2014). Dalian: CRC Press, 2015, p 285.

Lin, A., Shang, P., Feng, G. and Zhong, B. (2012) Application of empirical mode decomposition combined with k-nearest neighbors approach in financial time series forecasting, *Fluctuation and Noise Letters* 11 (02), p. 1250018.

Lin, J., Keogh, E., Patel, P., Lonardi, S. (2002) Finding Motifs in Time Series. In: *Proceedings of the 2nd Workshop on Temporal Data Mining,* at the *8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*,July 23-26, Edmonton, Alberta, Canada.

Lou, L., Lou, L., Zhang, X., He, X.(2017) Hospital daily outpatient visits forecasting using a combinatory model based on ARIM and SES models.*BMC Health Services Research*vol. 17, no.469.

Martínez, F., Frías, M. P., Pérez, M. D., Rivera, A. D. (2019) A methodology for applying *k*-nearestneighbor to timeseries forecasting.*Artificial Intelligence Review*, vol. 52, pp. 2019-2037.

Pratt, K.B., Fink, E.: (2002) Search for patterns in compressed time series. *International Journal of Image and Graphics* 2(1), pp. 89-106.

Sumitra, I. D., Basri, I. (2020) Forecasting the Number of Outpatient Patient Visits Using the ARIMA, SES And Holt-Winters Methods at XYZ Community Health Center, *IOP Conf. Ser.: Mater. Sci. Eng.* vol. 879,012060.

Thapa, K., Timalsina, A. K. (2023) Hospital Outpatient Visit Forecasting using Gated Recurrent Unit, In: *Proc. of 6th Int. Conf. on Information Systems and Computer Networks* (ISCON), Mathura, India, 03-04 March.

Truong, C. D., Tin, H. N., Anh, D. T. (2012) Combining motif information and neural network for time series prediction. *International Journal of Business Intelligence and Data Mining***7(**4) pp. 318-339.

Wang, Y., Gu, J., Zhou, Z., Wang, Z. (2015) Diarrhoea outpatient visits prediction based on time series decomposition and multi-local predictor fusion. *Knowledge Base Systems*, vol. 88 pp. 12-23.

****