IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

CNN-based Health Model for Regular Health Factors Analysis in Internet-of-Medical Things Environment

WALAA N. ISMAIL[1], MOHAMMAD MEHEDI HASSAN[2], (Senior Member, IEEE), HESSAH A. ALSALAMAH[3,4], GIANCARLO FORTINO[5],(Senior Member, IEEE)

[1]College of Computer and Information Sciences, Minia University, Egypt.
[2]Research Chair of Smart Technologies, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia
[3]College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia
[4]College of Engineering and Architecture, Al Yamamah University, Riyadh 11512, Saudi Arabia
[5]Department of Informatics, Modeling, Electronics, and Systems,University of Calabria, Rende, 87036, Italy

Corresponding author: Mohammad Mehedi Hassan (e-mail: mmhassan@ksu.edu.sa).

This work is supported by the Deanship of Scientific Research at King Saud University through the Vice Deanship of Scientific Research Chairs. Chair of Smart Technologies. The authors also thank the RSSU at King Saud University for their technical support.

ABSTRACT Remote health monitoring applications with the advent of Internet of Things (IoT) technologies have changed traditional healthcare services. Additionally, in terms of personalized healthcare and disease prevention services, these depend primarily on the strategy used to derive knowledge from the analysis of lifestyle factors and activities. Through the use of intelligent data retrieval and classification models, it is possible to study disease, or even predict any abnormal health conditions. To predict such abnormality, the Convolutional neural network (CNN) model is used, which can detect the knowledge related to disease prediction accurately from unstructured medical health records. However, CNN uses a large amount of memory if it uses a fully connected network structure. Moreover, the increase in the number of layers can lead to an increase in the complexity analysis of the model. Therefore, to overcome these limitations of the CNN-model, we propose a CNN-regular target detection and recognition model based on the Pearson Correlation Coefficient and regular pattern behavior, where the term regular denotes objects that generally appear in similar contexts and have structures with low variability. In this framework, we develop a CNN-regular pattern discovery model for data classification. First, the most important health-related factors are selected in the first hidden layer, then in the second layer, a correlation coefficient analysis is conducted to classify the positively and negatively correlated health factors. Moreover, regular patterns behaviors are discovered through mining the regular pattern occurrence among the classified health factors. The output of the model is subdivided into regular-correlated parameters related to obesity, high blood pressure, and diabetes. Two distinct datasets are adopted to mitigate the effects of the CNN-regular knowledge discovery model. The experimental results show that the proposed model has better accuracy, and low computational load, compared with three different machine learning techniques methods.

INDEX TERMS Internet-of-Medical-Things, Convolutional neural network, Regular health pattern discovery, Context management

I. INTRODUCTION

In modern society, monitoring human daily life using Internet of Things (IoT) technology includes the activities, vital signs and physiological parameters, stress, and sleep of a user in the health industry. IoT is in the spotlight because of the expected improvement it can deliver in medical services, disease predication and prevention, and cost reduction [1]–[3] Meanwhile, the number of patients suffering from chronic diseases is increasing because of the increase in unhealthy dietary habits. Such chronic diseases require a continuous monitoring process and management through healthcare and disease management services. Furthermore, patients should gain monitoring knowledge related to their health status [4]–[6] as it is possible to support life care based on living pattern analysis. IoT-based healthcare services feature a continuous generation of data, and thereby analyzing such data

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2980938, IEEE Access

to generate new knowledge can improve the quality of life [7]–[9]. Deep learning (DL) has become one of the leading paradigms, which can provide accurate pattern prediction and classification services [10], [11].

Due to DL technology, the CNN model is widely used for knowledge extraction from both structured and unstructured data. In the medical health field, a CNN is widely used as a representation tool for unstructured medical data analysis, and implementation of complex models [12]–[14]. One drawback of the CNN model is that it can cause an overfitting problem, an increase in computation load and have huge memory requirements. With the advent of IoMT environment, explosive volumes of vital sign data describing the human daily life became available. Additionally, DL models facilitates the opportunity of understanding these data through learning, extracting, and analyzing features and patterns from the collected biological data. for instance, health care users and caregivers can understand human normal health status, life routine, or sudden disease symptoms through the analysis of the collected patterns. Moreover, health care institutions can make critical health care decisions and treatments plan from the collected data. One example of such data-driven judgment is the regular pattern behavior.

Human health factors follow a regular behavior so that when the patient presents with a chronic disease, he will experience an increase in more than one health parameter (e.g high temperature) at the same time. Additionally, he may experience a specific decrease in another vital health parameter (e.g blood pressure). Therefore, studying such regular behavior knowledge [15]–[17]from the collected data facilitate the exploration of more characteristics regarding health-related parameters management and analysis.

To this end, the contributions of this work can be summarized as follows:

1) A new CNN learning model for knowledge discovery of regular correlated health-related factors to reveal regular co-occurring disease and symptoms relationships.
2) The model classifies the data using a double-layer CNN structure. In the first layer, the model input data are processed to extract potential parameters through multivariate analysis, then in the second layer, the selected data are classified based on the degree of correlation between each other.
3) Regular pattern behavior is studied to reveal only regular co-occurring health parameters. Furthermore, regular co-occurring patterns are discovered through regular pattern mining algorithm which extracts the periodic behavior of the collected factors within a specific period defined by the user. Then, the output of the model are classified according to the collected health-related parameters for obesity, high blood pressure, and diabetes.

II. RELATED WORK

Artificial intelligence has revolutionized remote health monitoring and decision support systems through the use of deep learning technology. Traditional statistical techniques and feature extraction techniques use a supervised learning model, such as logistic regression [18], support vector machines (SVM) [19], and random forest [20]. Such traditional techniques have been used efficiently for pattern analyses and mining of EHR data [21]–[23]. Despite the simplicity and interoperability of such statistical models, such models have restricted ability in dealing with high-dimensional input data, and their need for labelled data restricts their usage for comprehensive analyses of EHR data making the process impractical [24]–[26]. To overcome these limitations, recently, the Intelligent Information Society has advanced the analysis and modelling of EHR using DL technology. DL structure learns the dataset first and classifies it. Subsequently, it will make self-predications for similar data cases [27], [28]. Additionally, it can learn and discover knowledge from the unstructured medical context data. The key DL architectures are feed-forward neural networks (FFNN), convolutional neural networks (CNN), and recurrent neural networks (RNN).

An automatic diabetes detection method was developed in [29] using CNN model. Additionally, they used a combined network of a CNN-LSTM for abnormality detection of diabetes. In [30], CNN-brain tumor classification system is developed for automatic learning of tumor regions. The system efficiently solves the problem of insufficient data availability through the use of MRI images, by learning tumor regions from MR images and classifying this. The work presented in [28] used a deep generative learning model. The proposed model achieves 87.26% accuracy in detecting the use of traditional Chinese medicine using electronic health records (EHR).

RNN-based model is utilized in [31] for the detection of heart failure with an accuracy of 88.3%. Despite the good validation results achieved through those models in detecting different diseases, their methods have some scalability limitations as the data must be transformed in the preprocessing step which decreases the models accuracy. Electronic medical records-driven non-negative restricted Boltzmann machines (eNRBM) learning model is presented in [32]. The eNRBM uses an automatic feature extraction model from complete HER data. The proposed model performance outperforms the manual feature selection and engineering models. Moreover, the model can predict and classify the characteristics of patients and diseases accurately.

Miatto et al [29]introduced deep stacked diagnosing auto encoder (SDA) for training a universal feature extractor for clinical risk-prediction tasks. SDA can predict chronic health diseases such as diabetes mellitus, cancer of rectum and anus, cancer of liver and intrahepatic bile duct, congestive heart failure (non-hypertensive), among others. However, supervised learning model was needed to map the representation of each predicted risk. A deep record-learning model was introduced in [28]. The proposed model uses a CNN architec-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2980938, IEEE Access

ture to predict unplanned re-admission after discharge. The model maps the HER of each patients history to the predicted risk directly. Despite the great ability of deep learning in handling huge amounts of data, the learning models require a large amount of memory and computation. Machine learning models have been highly effective in different fields, but they still have limited application for clinical decision support systems. However, employing deep learning models on Electronic Health Records (HER) for personalized prediction of risks and abnormal health statuses may help in building effective solutions for caregivers in many situations.

Pham et al. [24] introduced an RNN architecture (called DeepCare) to predict the future medical risks based on patients health status and abnormalities. RNNs have a powerful ability in learning new representations from the collected HER. DeepCare model uses a modified short-term memory (LSTM) unit [33] to handle irregular patient intervisit periods. The performance of the model was tested for the prediction of unplanned re-admission within 12 months, where an Fscore of 0.791 was obtained, an improvement over traditional machine learning techniques such as SVM (F-score of 0.667) and random forests (Fscore of 0.714).

The CNN structure can affect the efficiency and accuracy of the classification model, therefore, in this study, we build a double layer CNN-model to extract and classify regular health-related data [34], [35]. The proposed method handles all the valid information available in the training dataset without transformation and then it is passed to the deep learning system for feature extraction and classification using multivariate analysis. Additionally, regular knowledge in the collected information is discovered accurately. The important health factors which are important for health status analysis are extracted through multivariate analysis. Subsequently, the correlation of the collected factors are discovered to identify the main causes of chronic diseases related to obesity, high blood pressure, and diabetes. Finally, regular behavior of each disease factors are studied and the knowledge related to regular co-occurred health factors are analyzed.

In [36] a similar work was presented using CNN-based health model for chronic diseases prediction. The model uses a CNN double layer structure for factors classifications. In the first layer, the model selects the significant health factors, however, the second layer conducts the analysis of the selected factor. Using PearsonCorrelation coefficient, the factors with a positive correlation above 0.5 are selected as positive significant factors; factors with a correlation of less than 0.5 are classified as negative correlated factors. Finally, associated rules are defined to classify and find the frequently occurring rules which may help to identify new knowledge from the classified dataset parameters. The output factors are classified into three chronic diseases, such as obesity, high blood pressure, and diabetes. The major differences between our work and the work presented in [36] are as follows:

1) The proposed model initiates a CNN-based model for the classification of positively correlated and negatively correlated health-related factors. Despite the great ability of such a model in discovering the association between unstructured data as in the case of EHRs, using it alone for attribute classification could hinder the clinical decision process as some parameters may have a negative correlation with each other and we want to study such abnormalities accurately.
2) They use association rule mining to discover more hidden knowledge; however, we are studying regular pattern behavior, regular mean patterns that co-occur within repeated time periods which could reveal more interesting knowledge related to pattern occurrence and its regular behavior. Exploring the more frequent association rule, however, just reveals the association between more frequent occurred patterns rather than the correlated regular behavior. Therefore, we decrease the time complexity of exploring a huge search space for all frequently occurred factors. Moreover, regular knowledge mining adds more insight in terms of learning human health factors and could help the caregiver to specify the right service for the right patient at the right time accurately.

III. CNN BASED REGULAR HEALTH DATA ANALYSIS MODEL

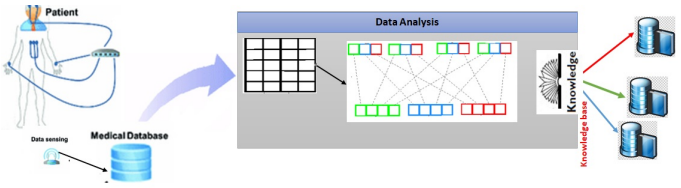

FIGURE 1: CNN-based regular factors discovery model.

The IoT devices can collect the health information at any time. Those data represent the health conditions and lifestyle parameters related to the selected patients. The proposed work as shown in Figure 1, is used for the discovery of regular correlated health factors that may affect health by using CNN-model. First, the medical data are collected from the patients and then preprocessed. From those processed data, the most important factors are selected to detect the relations among specific diseases, such as obesity, high blood pressure, and diabetes. Additionally, the productive correlation among the factors are analyzed to identify the positive correlated factors and negative correlated factors. Finally, the regular co-occurred parameters are selected to derive the required knowledge. In the next subsections, the details of each step are explained.

A. EXTRACTION OF DISEASE AND SYMPTOM CONCEPTS

From the collected unistructural medical health conditions, multivariate analysis is used to extract the correct factors and from the collected data, the model classifies them into input and output parameters.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2980938, IEEE Access

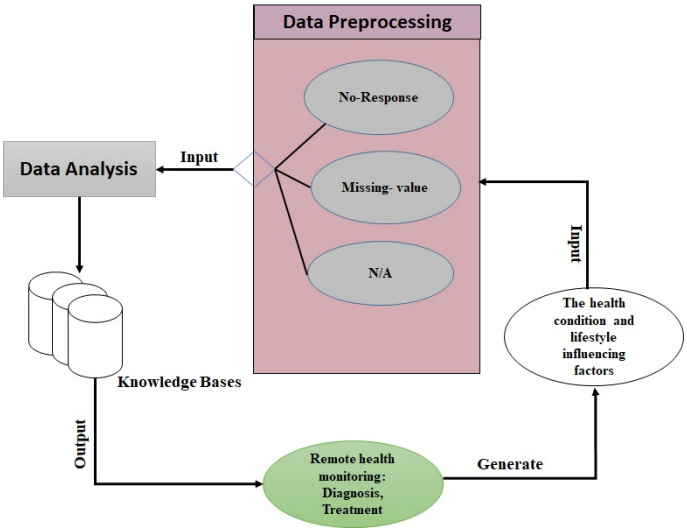

FIGURE 2: Data Preprocessing architecture for EHR data.

For our experiments, the National Health and Nutrition Examination Survey data [37] is preprocessed and used as shown in Figure 2,. The datasets were first preprocessed in order to extract potential characteristics related to health conditions and lifestyle parameters, and the feature was divided into input and output parameters. Factors related to chronic diseases are extracted and represented as input parameters, and the existence of obesity, diabetes, high blood pressure are the output parameters. The input is health conditions and lifestyle factors and the output is the patient status.

The complexity of the neural network will be increased with the increase in the number of inputs to the model. Therefore, Dimensionality reduction techniques should be applied for efficient analysis of linear and non-linear data. Principal Components Analysis and factor analysis are two of the most popular techniques for this purpose. For data preprocessing, we start by reducing the high dimension of the collected dataset as it is a large dataset with 49 attributes. Table 1 shows the input data and Table 2 shows the output data obtained. [Attribute] indicates the variable names. [Acronym] represents a simple abbreviation of the variables. The range of data is represented as [Range]. The data values can be [Num] as numbered values, not applicable [N/A], or there is no response for this attribute [Non-response]. For each attribute in the dataset, we ignore the irrelevant attributes having [N/A] or [Non-response] values. That way, the dataset attributes are reduced from high dimension to low dimension.

The 20 input variables are represented in Tables 1 and 2; output attributes are given in Table 2. For the selected 23 attributes, we preprocess the data to remove unnecessary items such as no response, and missing value. The dataset consists of 187,473 records and after preprocessing we extract 171,299 records, from which the relevant factors are analyzed and extracted to help remote monitoring application provide the right treatment for the right person at the right time. Figure 2 shows the complete process for collecting, preprocessing and finally the analysis of the collected input data and creation of the knowledge base.

TABLE 1: Input context data

Attribute	Acronym	Range
Water intake(g)	V_WT	[Num]
Protein intake (g)	V_PT	[Num]
Fat intake (g)	V_FT	[Num]
Energy intake (g)	V_EG	[Num]
Carbohydrate intake (g)	V_CB	[Num]
60-s pulse	B_PL	[Regular, irregular]
Systolic blood pressure (mmHg)	R_SBP	1) Low $\leq$ 120 2) Normal: 120 -140 3) Above $\leq$ 140
Diastolic blood pressure (mmHg)	R_DBP	1) Low $\leq$ 80 2) Normal: 80 - 90 3) Above: $\leq$ 90]
Body mass index (kg/m2)	R_BMI	1) Low $\leq$ 18.5 2) Normal: 18.5 - 25.0 3) Above: $\leq$ 25.0
Fasting blood sugar (mg/dL)	R_FSUG	1) Low $\leq$ 100 2) Normal: 100 - 126 3) Above: $\leq$ 126
Average sleep time/weekday (min)	B_ASLW	[Num, N/A, non-response]
Average sleep time/weekend (min)	B_ASLD	[Num, N/A, non-response]
Family history chronic disease	B_CH	[Yes, no, non-response]
Average smoking per day	B_SM	[Num, N/A, non-response]
Physical activity time (min)	B_PHC	[Num, N/A, non-response]
Time to sit and spend (h)	B_TSS	[Num, N/A, non-response]
Walk duration (min)	B_WD	[Num, N/A, non-response]
Drinking	B_DR	[1 2,3 4,5 6,7 9,10, N/A, non-response]
Sex	B_SEX	[Male, female]
Cholesterol intake (mg)	V_CHK	[Num]

TABLE 2: Output Classified data

Attribute	Acronym	Range
Presence of obesity	DS_ weight	[Low, normal, obesity]
Presence of diabetes	DS_IFG	[Normal, moderate, diabetes]
Presence of hypertension	DS_HYP	[Normal, pre, high]

B. APPLYING CORRELATION COEFFICIENT METHOD FOR INPUT DATA CLASSIFICATION

Clinical health records are not always correlated with each other. This may lead to overfitting of the learning model in which case the model generates accurate results with the training data rather than the testing data. For clinical health records knowledge generation, it is important to test the correlation between the collected attributes to prevent the overfitting in recognizing the test and training of health influencing parameters.

The Pearson correlation coefficient [38], [39] is used in different work for managing attributes relationships by means

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2980938, IEEE Access

of positive or negative correlation coefficient values. Positive correlation coefficient value means high correlation between attributes, and negative value means low correlation between the collected attributes [40]. Despite the efficiency of such a classification example, there are still some missing values which can hinder the models efficiency and accuracy.

The present study determines the significant relationships between the selected attributes by using the Pearson correlation coefficient (PCF)) presented in [38]. PCF Can show strength of relationship between two variables and study regular correlation behaviour between the selected health factors efficiently. An item has productive self-sufficient attributes if its significant level is greater than 0.1 [41]. This way, we ensure that the selected item is correlated and eliminate the overfitting problems. In order to test the correlation degree between attribute F_1 and F_2; the PCF can be calculated as follows:

$$\rho_{F_1,F_2} = \frac{Cov(F_1,F_2)}{\sigma F_1 \sigma F_2} = \frac{\sum |(F_1 - \mu F_1)(F_2 - \mu F_2)|}{\sigma F_1 \sigma F_2}$$

Two variables F_1, F_2 are correlated if their correlation coefficient (PCF) is greater than 0.1 which implies that F_1, F_2 are positively correlated. The coverage of F_1, F_2 are defined by $Cov(F_1, F_2)$. The deviation of the two parameters F_1, F_2 is represented as σF_1 and σF_2 respectively, while their respective means are μF_1 and μF_2. Further, if $PCF < 0.1$ then we said that F_1, F_2 do not have a strongly negative correlation to each other.

From the collected disease and symptom dataset in the pre-processing step, we test the correlation between potential health diseases (given in Table 3), whereas the correlation between health risk factors are given in Table 4 which demonstrates the three most relevant diseases of given diseases based on their association scores. To test the correlation between the input and output attributes, factors with PCF level of 0.5 are extracted. From the results of analyzing the health risk factors and their correlation, new methods could be employed to analyze the health conditions and the health factors related to each patient. For example, in Table 4, we can explore a negative CFP between carbohydrate intake ($V_C B$) and fat intake ($V_F T$). Therefore, carbohydrate and fat should not be ingested at the same time [42].

In this way, the health related attributes can be classified into positive and negative attributes based on the CFP using classification mining. For each health related record, Algorithm 1 is used to determine whether the parameters are positively or negatively correlated.

C. CNN-BASED HEALTH KNOWLEDGE MODEL

From what has been discussed above, we proposed a new model for the discovery of regular correlated health-related factors to detect any abnormality in health status. This model can extract all the available regular factor behavior to create the implicit knowledge related to remote monitoring and managing human lifestyle.

TABLE 3: Examples of diseases and associated diseases based on the association scores

	DS_OB	DS_DBT	DS_HYP
V_WT	0.654	0.099	0.765
V_EG	0.721	0.921	0.001
V_PT	0.900	0.711	0.028
V_FT	0.856	0.753	0.013
V_CB	0.989	0.765	0.040
B_PL	0.000	0.911	0.002
R_SBP	0.073	0.000	0.543
R_DBP	0.073	0.000	0.000
R_BMI	0.000	0.879	0.654
R_FSUG	0.654	0.002	0.000
B_ASLW	0.733	0.911	0.821
B_ASLD	0.023	0.732	0.991
B_CH	0.674	0.654	0.654
B_SM	0.059	0.711	0.551
B_PHC	0.811	0.609	0.655
B_TSS	0.653	0.666	0.732
B_WD	0.908	0.571	0.809
B_DR	0.765	0.901	0.503
B_SEX	0.760	0.811	0.666
V_CHK	0.633	0.712	0.671

We aim to enable care providers to provide a continual remote monitoring of patient health status by using the correlation results of the health attributes. Thus, the clinical workload would be reduced efficiently and the process of exploring different symptoms correlation at the same time could provide a robust CNN-based diagnosis support system. The CNN model is used to subdivide the selected correlated attributes discovered in the hidden layer. Additionally, undetected regular correlated readings are found by extracting the regular behavior of the correlated health factors [9], [12], [16].

From the collected Electronic health record factors, we found that there are some regularity-related health statuses . Regularity means that a certain disease may result in a change in specific health-related parameters for a specific time period and this may provide an important insight into a persons health status. Therefore, exploring regular factor behavior is a major issue in analyzing health records. For example, supposing disease X results in increases in body temperature and heart rate three times within one month, recording and analyzing this notable occurrence may help in preventing some sort of heart attack related to this patient.

Therefore, detecting the strong correlation between health factors and understanding the regular characteristics of the collected data would help in exploring more knowledge that is not due to random occurrences [9], [16]. To enable reporting only the regular behavior of health parameters, we use Algorithm 2 to calculate the periodicity of each pattern based on the user-defined regularity threshold value.

The input to the algorithm is the set of selected strong correlated factors and the output is the set of regular correlated attributes. CL denotes a candidate set of n-sized health fac-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2980938, IEEE Access

TABLE 4: Examples of symptoms and associated symptoms based on the association scores

	V_EG	V_WT	V_PT	V_FT	V_CB	B_PL	R_SBP	R_DBP	R_FSUG
V_EG	1	0.663	0.663	0.643	0.707	-0.211	-0.028	0.073	-0.765
V_WT	0.663	1	0.499	0.541	0.622	-0.034	-0.102	0.193	-0.034
V_PT	0.663	0.499	1	0.544	0.400	-0.145	-0.055	0.321	-0.091
V_FT	0.643	0.541	0.544	1	0.013	0.632	-0.099	0.923	-0.069
V_CB	0.707	0.622	0.400	-0.013	1	-0.0754	-0.171	0.044	-0.135
B_PL	-0.211	-0.034	-0.145	0.632	-0.0754	1	-0.011	0.500	-0.054
R_SBP	-0.028	-0.102	-0.055	-0.099	-0.171	- 0.011	1	0.779	0.911
R_DBP	0.073	0.193	0.321	0.923	0.044	0.500	0.779	1	0.781
R_BMI	0.677	0.213	0.311	0.114	0.231	0.159	0.234	0.453	0.104
R_FSUG	-0.765	-0.034	-0.069	-0.069	-0.135	-0.054	0.911	0.781	1

Algorithm 1: Algorithm for the Classification of correlated factors based on PCF value

Input : HDS: health-data set, Threshold value
Output: $PPPs$: Positive Productive Parameters, $NPPs$: Negative Productive Parameters.

Define Number PCF ;
if *(new updated transaction* T *in* HDS*)* **then**
  foreach *(item* e *in* T*)* **do**
    PCF = Pearson Correlation Coeffieent (e);
    if *(*$PCF \geq$ *Threshold)* **then**
      Call Update-class();
      return;
    end
  end
end
Update-class(input:factor e);
PCF = get Pearson Correlation Coeffieent (e);
if *(*$PCF \geq 0$*)* **then**
  Add e to $PPPs$;
end
else if *(*$PCF \leq 0$*)* **then**
  Add e to $NPPs$;
end
else
  Add e to $NULL - class$;
end

Algorithm 2: Knowledge mining for regular co-occurred influencing factors

Input : Cl: class of health-data set record; Database D, $MinSup$: minimum support count(0.5), Reg: regularity threshold value (0.1)
Output: regular co-occurred item list (RCL)

Define Frequent Item List $FSet$;
foreach *item* I *in class* CL **do**
  $R(I)$ = Call Check_item_regularity (I);
  if *Count* $I \geq MinSup$ *AND* $R(I) \leq Reg$*)* **then**
    Add I to Regular-Set RSS;
  end
end
Define Function Check_item_regularity(input item I, Database D);
foreach *(item* I *in* D*)* **do**
  Increment the count of all items in FSet that contain item I ;
  Update item regularity occurrences;
end
Return RCL;

tors. RCL denotes a set of regular factors. For each database transaction item e in class C in a given database, we calculate the the regularity of attribute e. If e satisfy the defined regularity threshold value given by the user, it will be added to the subset of the regular candidate set. In Algorithm 2, regular correlated factors that satisfy a minimum periodicity of 0.5 are found. The discovery of regular factors, which co-occurred regularly in the patients health lifestyle data, would help in providing remote assistance to the patients and predict any health risks accurately prior to any such health risks occurring [15], [16]. A knowledge base is then created by deriving such regular health knowledge. In terms of updating the regular health knowledge analysis model, the user can specify a specific time period continuously then the system would gather the factors, analyze regularity then send and save them in knowledge databases.

IV. PERFORMANCE EVALUATION

For performance evaluation, the experiment was carried out on a 64-bit Core i5 processor running Windows 10 pro, and with 12 GB of free RAM, SPSS. In this study, for the analysis of the proposed method, we use the data provided by the Korea National Health and Nutrition Examination Survey [43]. The context data are health -related factors and lifestyle attributes. They are the results of real-time health examinations of 10,806 citizens. Each citizen respond to a health survey with 768 items as part of the 7th Korea National Health and Nutrition Examination Survey. Multivariate analysis is used in the SPSS to extract the health-related factors which helps in reducing the computation complexity and decreases the number of attributes efficiently without missing any important values. For the analysis of model efficiency, various attributes are explored which affect

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2980938, IEEE Access

TABLE 5: Accuracy of regular CNN-based health model

	RMSE		Calculation speed		Complexity	
	Apply	None	Apply	None	Apply	None
Presence of obesity	0.0919	0.7681	15.525	36.681	0.452	1.100
Presence of high blood pressure	0.1793	0.2027	16.452	52.27	0.454	0.956
Presence of diabetes	0.2562	0.5976	16.964	62.13	0.654	1.923

obesity, high blood pressure, and diabetes. The selection of the health parameters is based on the correlation between attributes and we calculate the correlation coefficient of each attribute, and a correlation co-efficient of 0.5 or more are extracted. As a result, the attributes which are significant for obesity, high blood pressure, and diabetes disease prediction are selected first. Furthermore, the correlation between the selected attributes is calculated and classified into positive and negative relationships. Identifying the positively related factors could help in improving the persons daily life activity by monitoring such important positive correlated data between them. Additionally, any abnormality can be predicted through monitoring the negatively correlated parameters values. CNN is used to subdivide them. Additionally, regular factor behavior is analyzed to discover any additional factors behavior among the selected correlated factors. This reveals which attributes feature a regular correlation or an abnormal one; Therefore, the characteristics of the collected parameters are analyzed and classified into obesity, high blood pressure, and diabetes. This could significantly improve health status through providing more knowledge regarding obesity, high blood pressure, and diabetes diseases and their causes prior to their occurrence.

For the experiments, among the context data selected, we used 4,759,777 records in appropriate data formats that have no response and missing values are removed, including 1,499,423 records. We divide the collected records into 70% train data and 30% test data [44]. 10-fold cross validation algorithm was used to optimize hyper parameters (such as early stopping for training). Moreover, the root mean square error RMSE is evaluated to measure the CNN models accuracy. For each model, the RMSE, calculation speed, and complexity are calculated if we apply the proposed CNN-based health model. Additionally, we record the same evaluation criteria in the general CNN model for the prediction of obesity, diabetes, and high blood pressure. RMSE is used to measure the difference between the predicted and observed value [44]. The value of RMSE is evaluated as follows:

$$RMSE = \frac{\sqrt{\sum_1^n (P_K - O_K)^2)}}{n}$$

The smaller value of RMSE means high accuracy of the model prediction, however, the increase in RMSE value means low prediction capability of the model. For the identified three diseases; if obesity is predicted and the value is 3, and applying our model, the actual observed value is 2, the error is 1. Table 5 represents the RMSE results of the CNN-model. The prediction RMSE of our model is compared versus the general CNN model for the prediction of obesity, diabetes, and high blood pressure. The calculation speed of the proposed model is calculated with 2-hidden layer and the complexity is also calculated. For each diagnosed disease, the general CNN-model has high RMSE value, which indicates low predication accuracy with high calculation speed and complexity. For example, the RMSE of the general CNN model is about 0.87 with 1.1 complexity. However, the maximum RMSE of the proposed model prediction is 0.2562 for the presence of diabetes.

Furthermore, to compare our algorithm performance, we trained the following general models:

1) M1: our proposed model,
2) M2 : a LSTM model which is identified by its efficiency in finding correlated data,
3) M3: SVM
4) M4: traditional neural network.

The performance of each model is shown in Table 6. In this study, we propose a CNN-learning algorithm for the diagnosis and referral of three common diseases through discovering the regular behavior of correlated health parameters. By exploiting such knowledge of regular correlated algorithm, our model demonstrated competitive analysis performance on 4,759,777 medical records. Table 6 shows the performance analysis of the model. The accuracy of diagnosis and referral of our model reached 80.43%; 80.85%; 91.49%; 82.61%; 95.60% with a testing dataset, respectively. Regarding the other model, M2 (LSTM model), when trained with the collected data after preprocessing it and removing only irrelevant data, this proved not as effective as our M1. Moreover, the model represents a potential accuracy when compared with other traditional machine learning algorithms (M3, M4) learning model. Therefore, effective knowledge mining and analysis of medical data is of great significance in the discovery and diagnosis of health status and medical conditions. The model we proposed can extract features from the collected data which enables it to deliver accurate and robust results for the presence of obesity, high blood pressure,

TABLE 6: Different learning model accuracy comparison.

Disease	Learning Model			
	M1	M2	M3	M4
Presence of obesity	91.3	89	76	73
Presence of high blood pressure	93.5	90.01	91.3	73
Presence of diabetes	95	91	60	82
Four comprehensive	95	90.1	83	67

or diabetes.

V. CONCLUSION

In the medical big data field, critical health decisions are required to help patients attend to their health status. This work presents a CNN-based regular pattern mining model for the discovery of knowledge related to regularity in health conditions. The proposed method uses the health conditions and lifestyle patterns related to chronic diseases collected through IoT-devices. For the experiment, medical Korea National Health and Nutrition Examination Survey context data are used. A double-layer fully connected CNN structure is used to select and classify the collected data. The process begins with multivariate analysis to select the significant health factors, then the classification of such factors is conducted in the second layer. Finally, the regularity of the classified factors are analyzed and the most important regular-occurred health characteristics are selected. The model can detect either the regular positive correlated factors which can be maintained for healthcare, or the regular negative correlated factors which provide knowledge for improving unhealthy lifestyles and daily life activities. Regarding the performance study of the proposed model, it provides knowledge related to regular-correlated health parameters of obesity, high blood pressure, and diabetes. For future work, more types of diseases need further investigation with more vital sign collected data. Data quantification methods will affect the CNN model learning accuracy and performance, therefore, we plan to use different raw data preprocessing methods.

REFERENCES

[1] R. Indrakumari, T. Poongodi, P. Suresh, and B. Balamurugan, "The growing role of internet of things in healthcare wearables," in Emergence of Pharmaceutical Industry Growth with Industrial IoT Approach. Elsevier, 2020, pp. 163–194.

[2] B. Farahani, F. Firouzi, and K. Chakrabarty, "Healthcare iot," in Intelligent Internet of Things. Springer, 2020, pp. 515–545.

[3] L. Chen, G. Xu, S. Zhang, W. Yan, and Q. Wu, "Health indicator construction of machinery based on end-to-end trainable convolution recurrent neural networks," Journal of Manufacturing Systems, vol. 54, pp. 1–11, 2020.

[4] S. Kang, "A study on smart homecare for daily living ability and safety management of the elderly," in Information Science and Applications. Springer, 2020, pp. 707–710.

[5] M. M. Hassan, M. G. R. Alam, M. Z. Uddin, S. Huda, A. Almogren, and G. Fortino, "Human emotion recognition using deep belief network architecture," Information Fusion, vol. 51, pp. 10–18, 2019.

[6] Y. Zhang, R. Gravina, H. Lu, M. Villari, and G. Fortino, "Pea: Parallel electrocardiogram-based authentication for smart healthcare systems," Journal of Network and Computer Applications, vol. 117, pp. 10–16, 2018.

[7] V. Puntambekar, S. Agarwal, and P. Mahalakshmi, "Dynamic monitoring of health using smart health band," in Soft Computing for Problem Solving. Springer, 2020, pp. 453–462.

[8] A. Subasi, K. Khateeb, T. Brahimi, and A. Sarirete, "Human activity recognition using machine learning methods in a smart healthcare environment," in Innovation in Health Informatics. Elsevier, 2020, pp. 123–144.

[9] W. N. Ismail, M. M. Hassan, and H. A. Alsalamah, "Context-enriched regular human behavioral pattern detection from body sensors data," IEEE Access, vol. 7, pp. 33 834–33 850, 2019.

[10] S. Sengupta, S. Basak, P. Saikia, S. Paul, V. Tsalavoutis, F. Atiah, V. Ravi, and A. Peters, "A review of deep learning with special emphasis on architectures, applications and recent trends," Knowledge-Based Systems, p. 105596, 2020.

[11] G. Fortino, S. Galzarano, R. Gravina, and W. Li, "A framework for collaborative computing and multi-sensor data fusion in body sensor networks," Information Fusion, vol. 22, pp. 50–70, 2015.

[12] W. N. Ismail and M. M. Hassan, "Mining productive-associated periodic-frequent patterns in body sensor data for smart home care," Sensors, vol. 17, no. 5, p. 952, 2017.

[13] M. M. Hassan, A. Gumaei, A. Alsanad, M. Alrubaian, and G. Fortino, "A hybrid deep learning model for efficient intrusion detection in big data environment," Information Sciences, vol. 513, pp. 386 – 396, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020025519310382

[14] A. K. Sahoo, C. Pradhan, R. K. Barik, and H. Dubey, "Deepreco: deep learning based health recommender system using collaborative filtering," Computation, vol. 7, no. 2, p. 25, 2019.

[15] W. N. Ismail, M. M. Hassan, and H. A. Alsalamah, "Mining of productive periodic-frequent patterns for iot data analytics," Future Generation Computer Systems, vol. 88, pp. 512 – 523, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X17326481

[16] W. N. Ismail and M. M. Hassan, "Mining productive-associated periodic-frequent patterns in body sensor data for smart home care," Sensors, vol. 17, no. 5, p. 952, 2017.

[17] G. Fortino, D. Parisi, V. Pirrone, and G. Di Fatta, "Bodycloud: A saas approach for community body sensor networks," Future Generation Computer Systems, vol. 35, pp. 62–79, 2014.

[18] P. Ranganathan, C. Pramesh, and R. Aggarwal, "Common pitfalls in statistical analysis: logistic regression," Perspectives in clinical research, vol. 8, no. 3, p. 148, 2017.

[19] M. Malvoni, M. G. De Giorgi, and P. M. Congedo, "Data on support vector machines (svm) model to forecast photovoltaic power," Data in brief, vol. 9, pp. 13–16, 2016.

[20] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random forests and decision trees," International Journal of Computer Science Issues (IJCSI), vol. 9, no. 5, p. 272, 2012.

[21] C.-W. Song, H. Jung, and K. Chung, "Development of a medical big-data mining process using topic modeling," Cluster Computing, vol. 22, no. 1, pp. 1949–1958, 2019.

[22] A. Ismail, A. Shehab, and I. El-Henawy, "Healthcare analysis in smart big data analytics: Reviews, challenges and recommendations," in Security in Smart Cities: Models, Applications, and Challenges. Springer, 2019, pp. 27–45.

[23] P. Pace, G. Aloi, R. Gravina, G. Caliciuri, G. Fortino, and A. Liotta, "An edge-based architecture to support efficient applications for healthcare industry 4.0," IEEE Transactions on Industrial Informatics, vol. 15, no. 1, pp. 481–489, 2018.

[24] N. Wickramasinghe, "Deepr: a convolutional net for medical records." 2017.

[25] T. Pham, T. Tran, D. Phung, and S. Venkatesh, "Predicting healthcare trajectories from medical records: A deep learning approach," Journal of biomedical informatics, vol. 69, pp. 218–229, 2017.

[26] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun et al., "Scalable and accurate deep learning with electronic health records," NPJ Digital Medicine, vol. 1, no. 1, p. 18, 2018.

[27] F. Doshi-Velez, Y. Ge, and I. Kohane, "Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis," Pediatrics, vol. 133, no. 1, pp. e54–e63, 2014.

[28] J. R. A. Solares, F. E. D. Raimondi, Y. Zhu, F. Rahimian, D. Canoy, J. Tran, A. C. P. Gomes, A. H. Payberah, M. Zottoli, M. Nazarzadeh, N. Conrad, K. Rahimi, and G. Salimi-Khorshidi, "Deep learning for electronic health records: A comparative review of multiple deep neural architectures," Journal of Biomedical Informatics, vol. 101, p. 103337, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046419302564

[29] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: an unsupervised representation to predict the future of patients from the electronic health records," Scientific reports, vol. 6, no. 1, pp. 1–10, 2016.

[30] S. Ebadollahi, J. Sun, D. Gotz, J. Hu, D. Sow, and C. Neti, "Predicting patients trajectory of physiological data using temporal trends in similar patients: a system for near-term prognostics," in AMIA annual symposium proceedings, vol. 2010. American Medical Informatics Association, 2010, p. 192.

[31] E. Gawehn, J. A. Hiss, and G. Schneider, "Deep learning in drug discovery," Molecular informatics, vol. 35, no. 1, pp. 3–14, 2016.

[32] T. Tran, T. D. Nguyen, D. Phung, and S. Venkatesh, "Learning vector representation of medical objects via emr-driven nonnegative restricted

boltzmann machines (enrbm)," Journal of biomedical informatics, vol. 54, pp. 96–105, 2015.

[33] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Learning to diagnose with lstm recurrent neural networks," arXiv preprint arXiv:1511.03677, 2015.

[34] Y. Liu, Q. Zhang, G. Zhao, Z. Qu, G. Liu, Z. Liu, and Y. An, "Detecting diseases by human-physiological-parameter-based deep learning," IEEE Access, vol. 7, pp. 22 002–22 010, 2019.

[35] S. Shah, X. Luo, S. Kanakasabai, R. Tuason, and G. Klopper, "Neural networks for mining the associations between diseases and symptoms in clinical notes," Health information science and systems, vol. 7, no. 1, p. 1, 2019.

[36] J.-W. Baek and K. Chung, "Cnn-based health model using knowledge mining of influencing factors," Personal and Ubiquitous Computing, pp. 1–11, 2019.

[37] S. Kweon, Y. Kim, M.-j. Jang, Y. Kim, K. Kim, S. Choi, C. Chun, Y.-H. Khang, and K. Oh, "Data resource profile: the korea national health and nutrition examination survey (knhanes)," International journal of epidemiology, vol. 43, no. 1, pp. 69–77, 2014.

[38] Z. Xuanxuan, "Multivariate linear regression analysis on online image study for iot," Cognitive Systems Research, vol. 52, pp. 312–316, 2018.

[39] R. Gravina and G. Fortino, "Automatic methods for the detection of accelerative cardiac defense response," IEEE Transactions on Affective Computing, vol. 7, no. 3, pp. 286–298, 2016.

[40] J. Deshmukh and U. Bhosle, "Image mining using association rule for medical image dataset," Procedia Computer Science, vol. 85, pp. 117–124, 2016.

[41] J. Kang, S. J. Shin, J. Park, and S. Bang, "Hierarchically penalized quantile regression with multiple responses," Journal of the Korean Statistical Society, vol. 47, no. 4, pp. 471–481, 2018.

[42] Q. Xu, M. Zhang, Z. Gu, and G. Pan, "Overfitting remedy by sparsifying regularization on fully-connected layers of cnns," Neurocomputing, vol. 328, pp. 69–74, 2019.

[43] H. Yoo and K. Chung, "Heart rate variability based stress index service model using bio-sensor," Cluster Computing, vol. 21, no. 1, pp. 1139–1149, 2018.

[44] Z. Trunkvalterova, M. Javorka, I. Tonhajzerova, J. Javorkova, Z. Lazarova, K. Javorka, and M. Baumert, "Reduced short-term complexity of heart rate and blood pressure dynamics in patients with diabetes mellitus type 1: multiscale entropy analysis," Physiological measurement, vol. 29, no. 7, p. 817, 2008.