SN Computer Science (2024) 5:596
https://doi.org/10.1007/s42979-024-02886-2

ORIGINAL RESEARCH

Indexing ECG for Integrated Health Social Networks Predicting Keywords from ECG to Access Online Information

Yi Huang[1] · Insu Song[1]

Received: 14 May 2022 / Accepted: 9 April 2024

Abstract

Health Social Networks (HSN) provide rich medical knowledge bases that are scalable and sustainable, while IoT provides non-invasive, pervasive, and low-cost methods to collect patient data. However, receiving relevant information from HSN is time consuming and challenging for users, such as searching for the right relevant information using keywords and filtering. On the other hand, healthcare IoT has limited access to the vast medical knowledge bases, such as HSN, to interpret the collected data. To address these challenges, we propose Keyword-based Integrated HSN of Things (KIHoT), an approach that combines the strengths of both HSNs and IoT to overcome their limitations. In this method, data (biosignals) collected via IoT devices are converted to human readable keywords using word embedding vector features and CNN (Convolutional Neural Network) predictors. The CNN predictors are trained to predict keywords that individuals search within an HSN to extract relevant information of the given biosignals. Those keywords are encoded as word embedding for searching relevant information. KIHoT utilizes contrast learning techniques to extract latent feature representations of electrocardiogram (ECG) signals, which are then used to predict disease-related keywords. The proposed method was evaluated using 11,936 ECG signals from patients with heart disease and achieved an accuracy of 98% for disease prediction. Our results suggest that KIHoT can effectively extract relevant information from HSN portals, making it easier for researchers and clinicians to access valuable medical knowledge.

Keywords Health social networks · Internet of things · Remote diagnosis · Electrocardiogram · Word embedding

Introduction

Common chronic medical conditions, such as heart and respiratory diseases, are the leading causes of global death [1, 2]. Persistent care and monitoring are required to prevent these deaths. However, rising cost of healthcare in the aging population remains a significant challenge to those essential healthcare services [3]. The training of medical professionals is responsible for the rising medical cost. For example, in the U.S., training a General Practitioner (GP) costs more than US$300,000 [4]. Health Social Networks (HSNs) are the potential solutions to the low access to healthcare.

HSN comprise patient-driven healthcare that provides rich medical information, as social media allows millions of users to upload their data, such as status updates and images [3]. Health portals in the U.S. alone have more than 40,000 active members and 1.5 million unique monthly visits [3].

Healthcare Social Networks (HSNs) have the potential to be an invaluable source of medical information for researchers, clinicians, and patients alike. However, the large amounts of data contained in HSNs can make it difficult for users to find relevant information. Furthermore, current automated diagnosis tools based on machine learning are limited in their ability to provide detailed diagnoses and are often trained using expensive, time-consuming labeled data. As a result, these tools are only able to treat a limited number of diseases, limiting their usefulness to patients and clinicians.

On the others hand, the Health Internet of Things (IoT) provides low-cost, pervasive, and objective health monitoring [5]. The current methods for automated diagnosing heart disease are based on heart sounds or electrocardiogram

✉ Yi Huang
yi.huang3@my.jcu.edu.au

Insu Song
insu.song@jcu.edu.au

[1] College of Science and Engineering, James Cook University, Singapore, Singapore

Published online: 28 May 2024

(ECG). ECG provides information about heart function, such as heart rhythm [6]. Furthermore, some approaches also utilize IoT for heart diagnosis using ECG [7]. However, healthcare IoT has limited access to the vast medical knowledge bases, such as HSN, to interpret the collected data.

To address these issues, we propose a machine learning framework for automated diagnosis that integrates HSNs and IoT. Specifically, we present Keyword-based Integrated HSN of Things (KIHoT), an approach that utilizes electrocardiogram (ECG) signals to predict disease-related keywords and make it easier for researchers and clinicians to access valuable medical knowledge. By automating data collection, data labeling, and model training processes, the proposed system expands accessibility to healthcare information and helps users to retrieve relevant information from HSNs based on their biosignals objectively. The proposed system aims to provide an end-to-end HSN service with no expert knowledge required from users, significantly expanding accessibility to healthcare information.

The proposed method was evaluated using 11,936 ECG signals from patients with heart disease and achieved an averaged accuracy of 98% for disease prediction. The rate of valid keywords, namely sensitivity of keyword extraction, was over 90% for all instances and over 95% for 80% of instances. Our results suggest that KIHoT can effectively extract relevant information from HSN portals, making it easier for researchers and clinicians to access valuable medical knowledge.

The major contributions of this paper are as follows. First, the study presents a feasible solution to take advantage of HSNs via novel IoT approaches, converting biosignals to human readable keywords using word embedding vector features and CNN predictors. Second, the labels for training this model have the potential to be collected from the internet without any expert knowledge. Third, in this approach, the large amounts of data from IoT and HSNs are integrated to provide a cost-efficient method for health monitoring.

The rest of paper is organized as follows. We summarize the previous approach and analyze the requirements of this approach, as shown in Sect. "Problem Statement and Objective of the Study". Then, we propose a keyword prediction framework based on word embedding, as shown in Sect. "Literature Review". Next, we compare the proposed approach with conventional approaches, as shown in Sect. "Research Design/Methodology". Finally, we conclude by discussing the impact of the proposed approach in Sect. "Discussion/Analysis of Research Findings".

Problem Statement and Objective of the Study

The largest barrier is to search for information from HSN. Data from HSN is huge and this is a challenge for users to find related information. HSNs often rely on user input keywords to search for information, which can lead to incomplete and inaccurate descriptions due to users' lack of professional knowledge and experience with internet searches.

Finding right information requires in-depth knowledge about diseases and experience in Internet search, but current HSNs rely on users to describe their conditions based on subjective feelings. This can result in inaccurate analyses and a time-consuming process of refining keywords to find the correct information. Furthermore, information in HSNs also contains mistakes and informal terms [8, 9]. The resulting incomplete descriptions usually lead to inaccurate analysis (Fig. 1). Thus, users end up needing to refine their keywords by finding how others describe their conditions. This can be time consuming and ineffective since other users of the internet may not have similar conditions and may be equally lacking in knowledge. Also, it does not search automatically requiring extensive user involvement and time to find the right information. The difficulty lies in users' lack of accuracy in choosing suitable keywords for searching related information. As a result, HSN is time consuming and challenging to search the right relevant information.

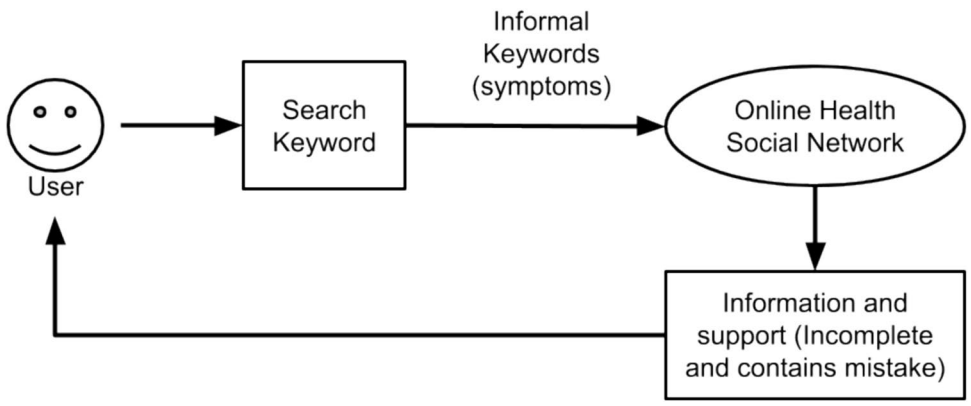

Fig. 1 Common usage of Health Social Network (HSN)

Another problem is that the existing automated diagnosis tool based on machine learning does not solve the problem either. We still lack tools that can provide an accurate diagnosis for people to find the exact information using the diagnosis term. The current automated diagnosis model is trained with labeled data, which is expensive and slow for collection. The number of labeled data as well as certain types of disease labels are limited. As a result, only a limited number of diseases can be treated (Fig. 2). This limits the ability of automated diagnosis to provide information and keywords to users for searching more information.

To solve the above-mentioned problems of HSN, a machine learning framework for automated diagnosis can be proposed for fully automating data collection, data labeling, and model training processes based on the exploding amount of data from both HSNs and IoT. From the perspective of HSNs, our approach helps users refine their keywords based on their biosignal objectively.

The proposed system aims to provide an end-to-end HSN service, which requires no expert knowledge from users (Fig. 3). This study develops Keyword based Integrated HSN of Things (KIHoT) for integrating HSNs and IoT. This approach provides related keywords to users via Electrocardiogram (ECG). The keyword prediction model can be trained with association between keywords and ECG. Hence, they can be used to search for related information from existing HSN portals. With no human intervention, the proposed integrated HSNs significantly expand accessibility to HSN.

Literature Review

Collection of Condition-Related Expression from HSNs

Social media data sharing has caused a data explosion, which facilitates data mining and AI. Compared to traditional data gathering approaches, data mining in social networks is fast and low-cost. Many studies have collected data on mental health issues [10, 11], influenza epidemics [12, 13] and Adverse Drug Reactions (ADR) [8, 9, 14–16] from social media such as Twitter.

Data collection from HSN is faster and cheaper than traditional data collection methods and provides large amounts of data on mental health issues [10, 11], influenza [12, 13] and Side Effects (ADR) [8, 9, 14–16]. For example, manually collecting data from a doctor to monitor the flu results in a delay of one to two weeks

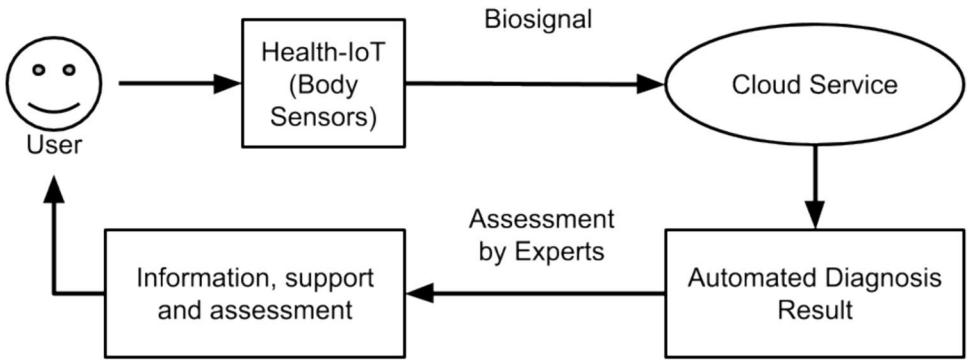

Fig. 2 Common usage of Internet of Things (IoT)

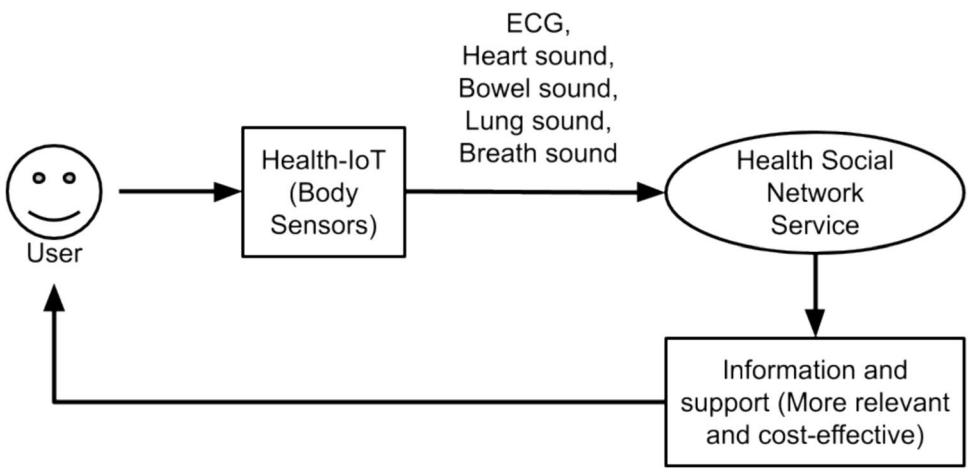

Fig. 3 Common usage of our proposed Integrated Health Social Network

[12]. However, the data from HSN is too much to process manually. To process large amounts of data, natural language process (NLP) approaches are applied[8, 9].

The approaches for detecting mental disorders are Linguistic Inquiry, Word Count [10, 17], quantifying mental health signals in Twitter posts with NLP features [10], and indexing anonymous keywords for autism and ADHD [3]. Discovering ADR is an important part in postmarketing surveillance [11]. The common approaches for extracting an ADR-related phrase in HSNs are conventional machine learning (SVM) [8, 14], the linguistic approach [15], deep learning (RNN) [11] [16] and Conditional Random Fields (CRFs) [9, 11]. Using social media to detect the spread of influenza focus on filtering internet content to extract disease-related expressions, mostly conditions, from HSNs [12, 13].

Those condition related expressions can be used to label ECG or other biosignals. By collecting the disease-related expressions in HSNs together with the associated ECG from same users, a keyword recommendation model can be trained without any manual labelling. The labelling is done based on existing information in HSNs. Our previous approach [18] also labels the ECG with emulated keywords. The information extracted from HSNs can be used to replace the emulated keywords and give more realistic results.

Word Embedding

In this study, diseases were associated as sets of keywords. Word embedding is a widely used vector space word projection method of natural language processes. Unlike one hot annotation, word embedding represents important semantic features. Similar words, such as cough and breathless, have word embedding with higher cosine similarity. In contrast, no such similarity exists in their one-hot representation. The most important feature of word embedding is that the vectors of words with similar semantics have higher cosine similarity as (1).

$$C(a,b) = \frac{a \bullet b}{\|a\|\|b\|} \tag{1}$$

Word embedding are learnt by neural network that predicts words given the context [19]. Vector representations can be stored and used like a dictionary. Table 1 shows the performance (by default is F-measure; using accuracy when F-measure is not available) of previous approaches using Word Embedding. Some studies show a promising result for using word embedding as a feature.

Table 1 Performance of HSN Approaches Using Word Embedding

Reference	Performance (F-measure)	Sample Number
[8]	80.3%	6320
[16]	81.41% (accuracy)	1250
[20]	85.2% (accuracy)	1824
[21]	65%	33,332
[22]	50%	33,332
[9]	79.4%	1250
Weighted Average	60.8%	77,308

Table 2 Performance of Previous CNN Approaches

Feature	Performance	Dataset size	Reference
MFCC	84.0%	3240	[23]
DWT	82%	3240	[24]
MFCC	81.3%	3240	[25]
MFCC	81.4%	3240	[26]
FT	95.2%	3240	[27]
RGB image	94.2%	3240	[28]
Sonogram	94.2%	3126	[29]
Spectrogram	89.8%	50	[30]
Spectrogram	86%	1630	[31]
Spectrogram	98.5%	50	[32]
STFT Spectrogram	98.2%	817	[33]
Spectrogram	74.0%	176	[34]
ECG Time Domain	86.4%	12,186	[6]
	87.0%		

Convolutional Neural Network

The conventional approaches to biosignal-based diagnosis are pre-processing, segmentation, feature extraction, and classification. A segmentation algorithm is usually required to select the parts of the whole biosignal which contain more information, such as a heartbeat or a bowel activity. Many hand-crafted feature extraction methods are also needed for researchers to summarize the features of the signal, such as frequency information (methods based on Fourier transformation, such as FFT, STFT, MFCC), wavelets, and the derivatives of those features. The extraction, configuration, and selection need a high degree of expert knowledge and manual work. Fortunately, deep learning methods provide end-to-end approaches, which accept raw signals of ECGs [6] or spectrograms of audio as input and generate features based on tasks. Some approaches also do not require segmentation, as deep learning approaches are able to locate the information from raw signals automatically and discard unrelated information [6]. Table 2 shows the performance of

previous CNN approaches with biosignal. The good performance shows that CNN is reliable for fitting biosignal models.

The convolutional layer consists of a set of convolution kernels. A convolution kernel is a set of trainable shared weights which detect elementary features from the previous layer with a sliding window, forming a feature map. The exact position in a feature map is less critical and can be harmful when there is a shift in the input. Afterwards, a pooling layer down samples the feature map with the maximum value and makes CNN less sensitive to the exact position.

In previous CNN approaches, the Fourier Transform-based spectrogram is the most common feature used in CNN for sound diagnosis [27] [30] [31] [32] [29] [34]. The STFT (Short Time Fourier Transform) spectrogram is an example of an implementing spectrogram [33]. Likewise, DWT is used for CNN Classification [24]. The MFCC spectrogram feature is also commonly used [23] [25] [26]. On the other hand, Deperlioglu [28] converts heart sound data into RGB images to reduce computation requirements. Xiong et al. [6] also proposed an ECG classification for cardiac arrhythmias' detection with 1D CNN, achieving 86.4% F1 accuracy.

The current approaches to diagnosis of heart disease are based on heart sounds or ECG; it is used for automatic diagnosis of heart disease [5]. ECG provides information about heart function, such as heart rhythm, and there are approaches that use ECG readings for classification of heart rhythms [6]. In addition, some approaches apply IoT to heart diagnosis using ECG as well [7]. The state-of-the-art of ECG diagnosis are atrial fibrillation detection with CRNN [35], heart disease classification using DBLTSM [5], arrhythmia classification using CNN [36] and heart disease classification Self-supervised representation learning with LSTM + MLP [37].

Research Design/Methodology

The proposed method aims to automatically convert ECG signals into meaningful and readable keywords by contrast learning. Thus, HSN content can be searched using automatically generated keywords from ECGs instead of users' keying in keywords. As a result, the autonomous nature of HSNs is improved by integrating with biosignals from IoT.

Figure 4 illustrates the overall process involved in this approach. First, the ECG signals and labels are collected from PTBDB dataset. Each record in PTBDB dataset contains a diagnosis report and a ECG recording with varying length. The ECG signals are then segmented by beats and normalized as pre-process for model training and disease prediction. The sampling rate of ECG in PTBDB dataset is 1000 per seconds. We select each heart beats based on R peaks. Each beat includes ECG 251 ms before R peak and 400 ms after R peaks.

We then label the ECG with keywords that describes the symptom. In a previous study [38], the researchers extracted keywords based on similarity, and the keywords were manually chosen. In this study, the keywords were automatically chosen based on the description on the internet, and the keywords were predicted with a greedy search method instead. Symptoms in the disease expression are selected as keywords.

The keywords are then converted into word embedding vectors, and the sums of the word embedding vectors (SOWE) are calculated. The normalized ECG and SOWE are used to train the CNN predictor. After that, CNN predicts the SOWE given ECG. To verify the performance of CNN, we also compare CNN with linear regression to predict the SOWE. The linear regression approach uses a single linear layer instead, takes the raw signal as input and output SOWE.

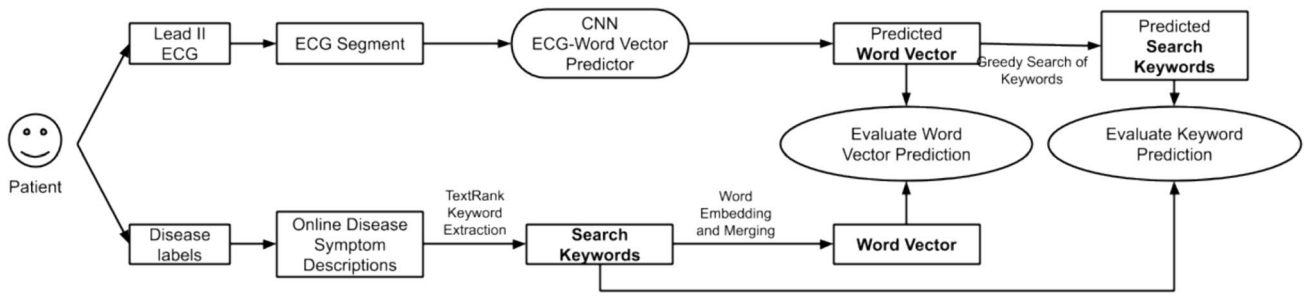

Fig. 4 The labels in PTB database were converted into keyword sets. The keyword sets were then converted into SOWEs. The CNN encoder was trained with ECG as input and the SOWE as output. After that, the CNN encoder was able to predict the most likely SOWE given ECG. The predicted SOWEs were then converted into keywords, to compare with the original keyword sets. Also, the SOWEs were also used to predict diseases by the cosine similarities between the original SOWE

The predicted keywords are then extracted from SOWE using a greedy algorithm. Predicted SOWE are also used to predict disease with cosine similarity.

Dataset

Since there was no ECG data associated with the symptom expression in HSNs, this study used made-up data for testing the approach. A new evaluation model needed to be used for this study. Data were collected from a Physionet dataset called PTBDB. This database contains 70,207 heart beats from 268 subjects with diagnostic results. Diagnostic labels include cardiomyopathy, bundle branch block, hypertrophy, myocarditis, myocardial infarction, valvular heart disease, dysrhythmia, and healthy controls. Each recording in PTBDB contains 15 signals: 12 conventional leads with 3 Frank lead ECG. The sampling rate of the signal is 1000.

In the HSN, patients are not expected to know their diseases. They are only expected to know their symptoms and write related posts. Since there is still no such dataset with associated condition related posts and actual diagnosis result, a dataset is mocked up based on the PTBDB. For each disease, a description of its symptoms is chosen from the internet to use as the posts for each disease. The symptoms for patients with the same disease were assumed to be the same. We also assumed the patients' description of their conditions was accurate. Thus, we extracted keywords from the symptom description of each condition on the Internet, as symptom description is more common in HSN. The description was from the top search result of the disease. The details of the description are in the Appendix.

Table 3 shows detailed statistics of samples and the distribution of study subjects for each class. Physionet toolkit is used to select QRS peak based on lead I. We selected ECG beats from lead II since lead II is widely used in other approaches. To balance the dataset, under sampling is used. Due to some classes have too few samples, they are not included in this study. Only 5% of MI and 20% of healthy ECG beats were evenly selected from each subject. The classes with too few samples were discarded.

Table 3 Number of Beats in the Dataset

Disease	Original Dataset	Selected Dataset
myocardial infarction	52,326	2496 (5%)
healthy control	10,551	2110 (20%)
valvular heart disease	499	499 (100%)
Dysrhythmia	1290	1290 (100%)
Cardiomyopathy	2227	2227 (100%)
Hypertrophy	991	991 (100%)
bundle branch block	2323	2323 (100%)

Sum of Word Embedding (SOWE)

The main contribution of this study is to generate keywords given ECG. The model predicts the keywords in the form of SOWE instead of one-hot label. Thus, the SOWE represents the disease as well as the bag of keywords. Compared to one-hot annotation, word embedding has two advantages: it supports a huge amount of candidate keywords, and it maintains the semantic meaning of the bag of keywords to make disease prediction easier for a given vector.

It is very difficult to determine the number of possible keywords to describe a condition, while word embedding can represent a huge number of keywords with a fixed number of attributes. When the number of keywords grows in the real HSN application, the word embedding approach will have better efficiency.

An existing word embedding dictionary [9] was used to represent the keywords. This dictionary is based on CBOW with 200 attributes. This dictionary was trained with 2.5 million unlabeled comments from online social networks and scientific lectures. This word embedding model projects each word into a vector with 200 dimensions. The model was trained with Gensim library.

Given the emulated post dataset as label, keywords are needed to be selected to remove unnecessary stop words. Key parses were extracted by TextRank algorithm [39]. The top five keywords for each condition were selected via this approach. For the health subjects, keyword "health" was chosen instead. For each keyword, the word vector is selected from the existing dictionary mentioned above. Since there is more than one keyword for each disease, the keyword vector is summed up.

The SOWE of all the diseases also forms a matrix for disease diagnosis. The reason for labelling ECG with SOWE directly is that it is easy to add new keywords, if they appear in the word embedding dictionary. This procedure is also equivalent to a classification model when SOWE are compared with other conditions using cosine similarity measurement as in Eqs. (2), (3), and (4):

$$Output(X) = W_0 X + b_0 \tag{2}$$

$$Classification(X) = W_1 Output(X) \tag{3}$$

$$Classification(X) = W_1 W_0 X + W_1 \cdot b_0 \tag{4}$$

where W_0, W_1 are learnable weight, b_0 is learnable bias and X is input.

In addition, each hidden node represented an interpretable meaning by extracting the related bag of keywords. The activation function of classification was SoftMax, and the loss function for classification was cross entropy.

Keyword Extraction from SOWE

For extract the keywords from word embedding, a greedy search algorithm [40] was used in this approach, which also was expected to extract most important keywords given a sum of word embedding vector. The keyword extraction method has two steps: greedy searching step and refining step.

For the greedy searching step, the Bag of Words (BOW) is initialed as an empty set. In each step, the candidate words that can minimize the Euclidian distance between the SOWE of selected keyword set and given keyword vectors are selected and added into the keyword set. This step repeats until no further keyword could minimize the difference between the SOWE of keyword set and given word vector.

The refining step attempts to replace each chosen keyword with any other candidate words. For each keyword, they are firstly removed from the BOW. Then each candidate word is added into the BOW to calculated if the difference could be further minimized. If a candidate word is found to minimize the difference, this candidate word is then retained. Otherwise, this refining step will be reversed. This step will keep repeating until no further improvement to the Euclidian similarity or reach maximum repeating limit.

CNN ECG-Word Vector Predictor

A CNN from a previous study is trained to convert an ECG signal segment into a word embedding vector with 200 attributes. Figure 5 shows the architecture of CNN in our study. This CNN had three convolution blocks, which were feature extracting layers for extracting abstract features from raw signals. The feature extractors were used to extract the fundamental features from previous layers. The ECG signals are inputted into the first convolutional block directly. The input size of ECG raw signal is 651. In the end, a dense layer performed the regression of 200 attributes of word embedding vectors with linear kernels. The loss function for regression was mean square error. As a baseline, an identical CNN is trained to predict the keyword with one-hot label as a disease classifier. The only difference between the two CNN is the output layer.

Percentage split was the evaluation method in our approaches. We used 70% of the samples for training and 30% of the samples for evaluation. We trained two different CNNs with 50 epochs and batch sizes of 64. The first CNN was the proposed approach, which predicted the keyword vector of a condition given ECG. The other CNN has output layer as a conventional multiclass classifier, serving as a benchmark.

Evaluation Matrices

We evaluated the proposed method in two ways. The first form of evaluation is to measure the ability of the model to predict vectors. The second form of evaluation measures the accuracy of retrieving keyword back from predicted vector. The third form of evaluation involved measuring the difference between the predicted vectors and target vectors. The measurement of difference was done by a classification task of the diseases. For each predicted vector, the most similar vector from the conditions was selected. The condition was then compared with the actual condition for evaluation of the prediction. This task evaluated the likelihood of confusion among the conditions.

To evaluate the advantage of SOWE over one-hot label, leave-one-class-out testing was done. The procedure of the leave-one-class-out is to remove one class completely from the training set but still attempt to predict the class in the testing set. This evaluates the ability of the proposed model to predict unknown diseases which do not exist in the training data.

Discussion/Analysis of Research Findings

Keyword Vector Prediction Result

Table 4 shows the training MSE, testing MSE and keyword extraction accuracy of CNN and linear regression. CNN have significantly higher performance to linear regression.

To measure the usefulness of the approach, an evaluation was performed by measuring the number of keywords correctly extracted with the predicted vector. The percentage of corrected extracted keywords was compared with original

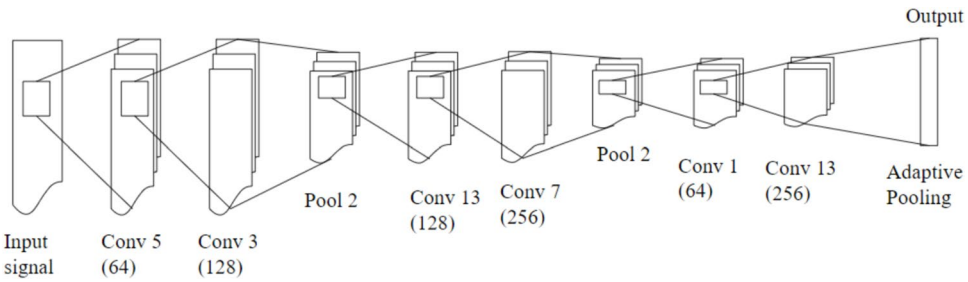

Fig. 5 Structure and Detailed Arguments of The Purposed CNN

Table 4 Training MSE, Testing MSE and Keyword Extraction Accuracy of CNN and Linear Regression

	CNN	Linear Regression
Training MSE loss	0.276	10.154
Testing MSE loss	0.468	12.903
Keyword Extraction Accuracy	98.604%	57.498%

keywords. The predicted keywords are extracted with the above-mentioned greedy search algorithm. For example, an ECG with MI will output a vector with 200 dimensions.

After that, keyword is selected within a candidate keyword dictionary. The candidate keywords are symptoms for all known disease and prepared for keyword selection. The large difference in MSE results in different keyword extraction accuracy, which CNN extracted 98.6% of keywords correctly while linear regression only has 57.5% of accuracy.

The sensitivity is calculated as (5):

$$Sensitivity = \frac{number\ of\ correct\ predicted\ keywords}{number\ of\ ground\ truth\ keywords} \quad (5)$$

while the precision is calculated as (6):

$$Precision = \frac{number\ of\ correct\ predicted\ keywords}{number\ of\ predicted\ keywords} \quad (6)$$

The keywords, such as pain, dizziness, rhythms, and vomiting, are then selected with a greedy search of candidate keywords from the vector. Since the ground truth keywords for MI are Pain, dizziness, weakness, heaviness, and vomiting, this example gets three correct prediction out of five ground truth keywords, the sensitivity of this prediction is 60%. On the other hand, there are four keywords selected but only three is correct, so the precision is 75%.

Figure 6 (scale from 70 to 100%) shows the number of instances with different sensitivity of keywords correctly predicted, while Fig. 7 (scale from 0 to 100%) shows the result of the baseline approach. For all conditions, the model managed to extract all possible correct keywords for over 70% of instances, and it also managed to extract over 80% correct keywords for over 95% of instances. Most of conditions, except hypertrophy and health control, extracted over 95% of all keywords correctly. As a result, the model provided sufficient information for most instances. Figure 7 shows the result of a conventional approach using linear regression. The conventional approach provided sufficient information for only very few instances.

Figure 8 (scale from 70 to 100%) shows the population of instances that met various performance standards to extract valid keywords, while Fig. 9 (scale from 0 to 100%) shows the results of the baseline approach. The rate of valid keywords, namely sensitivity of keyword extraction, was over 90% for all instances and over 95% for 80% of instances. In contrast, the baseline method was only able to provide useful keywords for a few conditions with a much lower sensitivity.

Tables 5 and 6 show the average sensitivity and specificity of keyword extraction of the CNN approach and the linear regression approach. Compared to linear regression, CNN

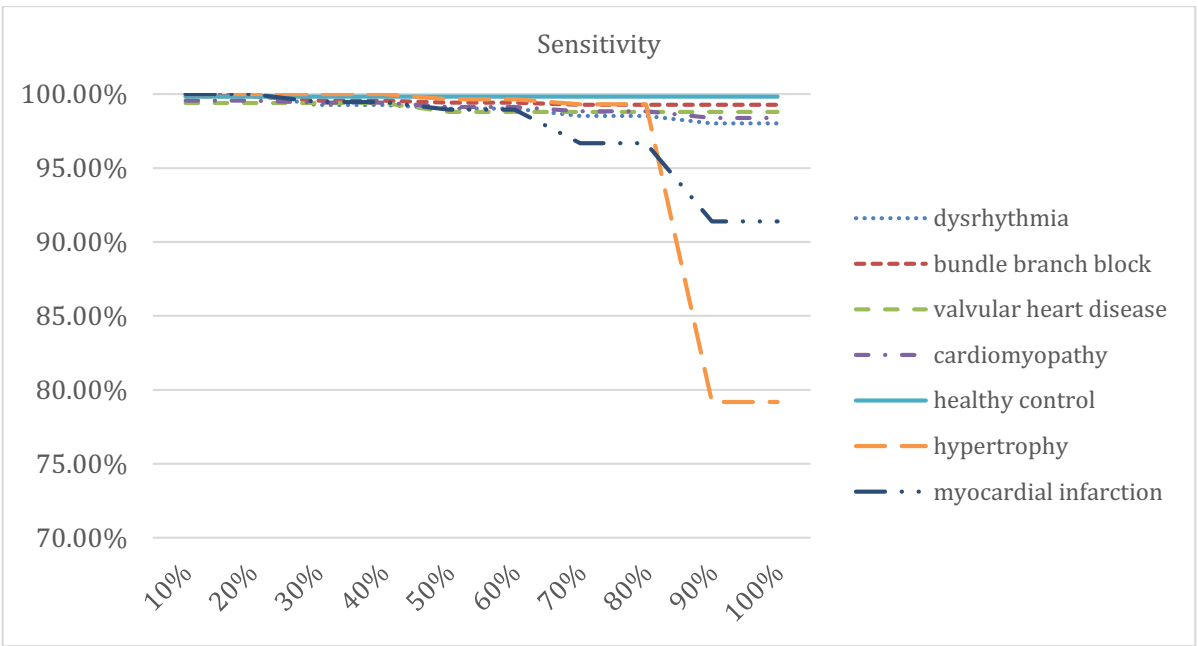

Fig. 6 Sensitivity of keyword extraction of CNN

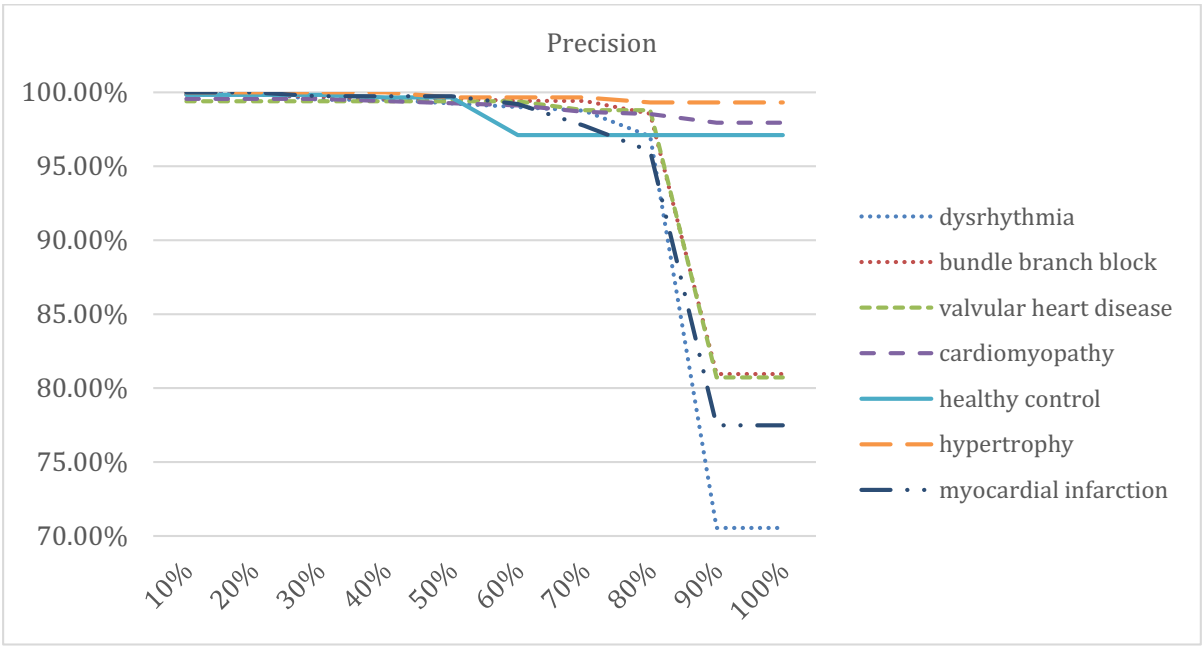

Fig. 7 Precision of keyword extraction of CNN

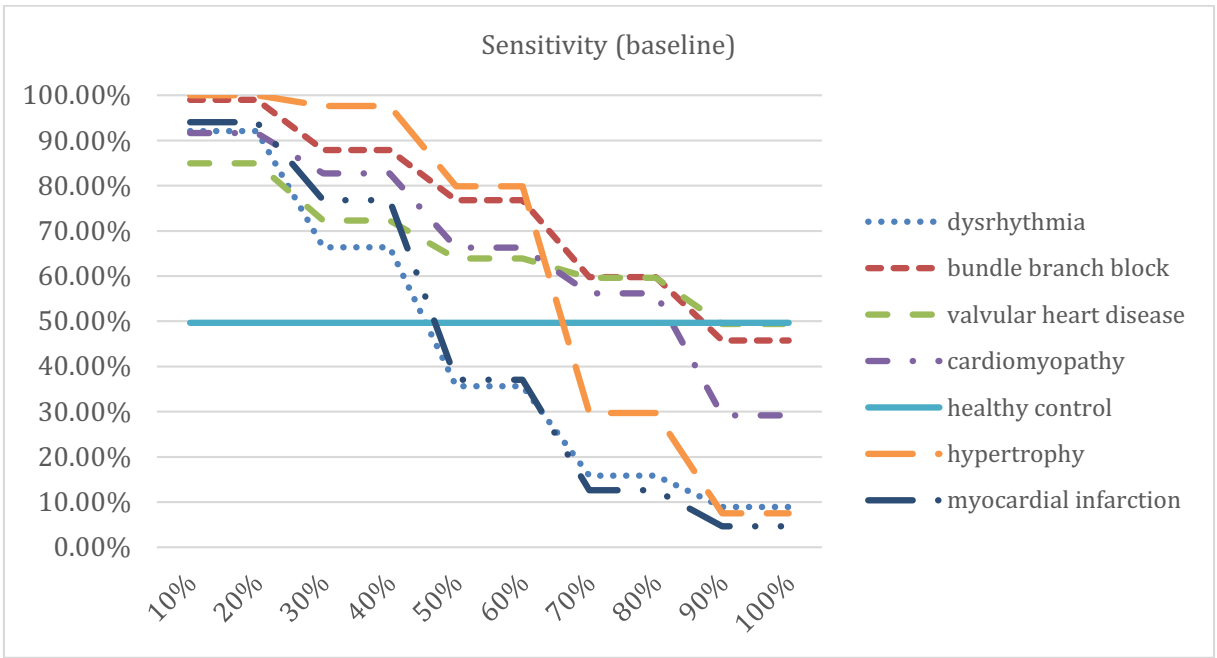

Fig. 8 Sensitivity of keyword extraction of linear regression

achieved good performance for all conditions in both sensitivity and specificity.

Compared to conventional machine learning approaches, our deep learning approach achieved more stable results and provided more useful information. The CNN approach had a larger population of instances which had high sensitivity and specificity compared to the conventional machine learning approach.

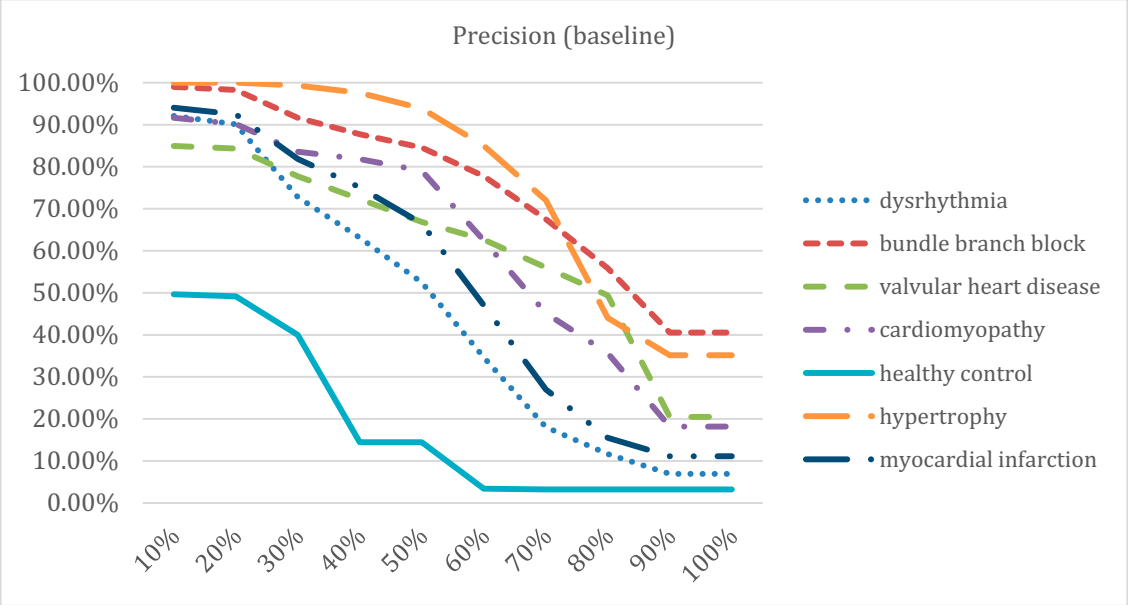

Fig. 9 Precision of keyword extraction of linear regression

Table 5 Average Performance of Keyword Extraction by word embedding from CNN

	Sensitivity	Specificity
dysrhythmia	98.960%	94.345%
bundle branch block	99.481%	96.403%
valvular heart disease	99.036%	96.185%
cardiomyopathy	99.062%	99.005%
healthy control	99.830%	98.441%
hypertrophy	95.631%	99.710%
myocardial infarction	97.298%	95.327%

Table 6 Average Performance of Keyword Extraction word embedding from linear regression

	Sensitivity	Specificity
dysrhythmia	43.762%	46.863%
bundle branch block	73.824%	76.047%
valvular heart disease	66.024%	61.228%
cardiomyopathy	65.191%	62.467%
healthy control	49.660%	19.694%
hypertrophy	62.935%	78.188%
myocardial infarction	45.033%	54.503%

Table 7 Classification Result by CNN Classifier with Softmax

	Accuracy	Sensitivity	Specificity
dysrhythmia	90.347%	96.053%	99.528%
bundle branch block	98.268%	94.452%	98.615%
valvular heart disease	84.337%	97.902%	99.912%
cardiomyopathy	98.974%	89.286%	97.206%
healthy control	90.476%	93.662%	98.797%
hypertrophy	87.372%	96.241%	99.696%
myocardial infarction	89.139%	90.094%	97.381%
Meta-Average	91.273%	93.956%	98.734%

Table 8 Searching Performance of word embedding predicted by CNN

	Nearest	2nd	3rd
dysrhythmia	99.505%	99.505%	100.000%
bundle branch block	99.711%	100.000%	100.000%
valvular heart disease	98.795%	99.398%	99.398%
cardiomyopathy	99.413%	99.413%	99.707%
healthy control	99.490%	99.490%	99.830%
hypertrophy	99.317%	99.317%	99.659%
myocardial infarction	99.603%	99.735%	99.735%
Meta-Average	99.405%	99.551%	99.761%

Disease Prediction Result

The classification performance of word embedding with linear kernel was compared with that of normal classification. To perform classification tasks with word embedding, the cosine similarities between the predicted vector and the sum of keyword vectors of each disease were calculated. Table 7 shows the classification result of CNN classifier, while Tables 8 and 9 show the classification results of searching

Table 9 Searching Performance predicted by linear regression

	Nearest	2nd	3rd
dysrhythmia	57.426%	76.238%	91.337%
bundle branch block	92.496%	98.990%	99.567%
valvular heart disease	63.253%	68.675%	78.313%
cardiomyopathy	78.299%	81.671%	86.217%
healthy control	13.095%	18.027%	25.510%
hypertrophy	78.157%	93.857%	98.294%
myocardial infarction	62.384%	81.192%	87.550%
Meta-Average	63.587%	74.093%	80.970%

Table 10 Searching Performance (Leave One Class Out)

	Nearest	2nd	3rd
dysrhythmia	0.000%	27.228%	81.931%
bundle branch block	0.000%	13.131%	71.573%
myocardial infarction	0.000%	9.536%	48.079%

based on cosine similarity of word vectors. Tables 8 and 9 evaluate the performance of classifiers in finding the right condition in top N number of nearest conditions. Our cosine similarity-based classification achieved slightly better performance compared to the CNN classifier. More importantly, for the linear regression approach, most of conditions still achieved acceptable performance given top 2 or top 3 similar conditions. This result shows that the sum of the symptom keyword vector had a positive effect in classification and diagnosis of diseases.

Table 10 shows that using SOWE as labels has the potential to predict the diseases which is not exist in training set. Unlike existing disease predictors, which can only output a finite number of labels, the proposed approach predicts a word embedding vector, which is related to symptoms. The proposed approach can theoretically predict disease-related symptoms that do not appear in training data, since such new disease and the related symptom is given. To test the effect of predicting the symptoms of unseen disease, leave-one-class-out validation is done. Each disease is hold in training to produce a model and that model is used to predict the hold out disease. Despite the keywords set between different diseases can be very different, SOWE extract the semantics of the keywords set. Word embedding improve the performance of CNN and have acceptable result with linear regression. This improvement shows that the embedding dictionary contains the information of the similarities of the diseases for models to build relations between signal and embedding. Our method could be a zero-shot learning approach because it learns semantic embeddings of ECG rather a particular class. Therefore, the semantic embeddings have the potential to predict unknown disease given the disease description.

Table 11 compares our approach with the recent state-of-the-art ECG deep learning approaches. Our approach shows outstanding performance compared to both supervised classification and self-supervised learning.

Discussion

The result shows that when predicting diseases given the sum of word embedding instead of one-hot annotation of Bag-of-Words, the distributed representation is also more robust. The sum of the keyword vector provides an easy method for estimating disease given keywords. This model has potential to predict unknown diseases which do not exist in the training set. This provides a scalable diagnostic framework using information in HSNs. Similar symptoms have similar values in word embedding; so, the sum of word embeddings is a robust method for differentiating diseases based on their symptoms.

Recommendation

With an average accuracy of 98% for disease prediction, KIHoT has the potential to become a valuable tool for researchers, clinicians, and patients alike. The integration of HSNs and IoT provides an innovative method for collecting data and labels without requiring expert knowledge. The proposed system expands accessibility to healthcare information while simplifying the data collection and analysis processes. KIHoT's success in predicting relevant keywords from ECG signals makes it a promising solution for monitoring health

Table 11 Comparison with state-of-the-art ECG deep learning approaches

Approach	Task	Performance
CRNN [35]	Atrial Fibrillation Detection	90.6% Score
DBLTSM [5]	ECG Classification	100.0% Macro F1 score
CNN [36]	Arrhythmia classification	98.96%
Self-supervised representation learning with LSTM + MLP [37]	Disease classification	94.18%
SOWE regression with CNN (Ours)	Disease classification	99.40%

conditions and providing timely interventions. Overall, we believe that KIHoT has the potential to significantly improve healthcare services by making reliable health information more easily accessible to all users.

KIHoT can provide valuable insights into patient health status and diagnose diseases more accurately, leading to better-informed medical decisions and improved patient outcomes. By automating data collection, data labeling, and model training processes, KIHoT expands accessibility to healthcare information and helps users refine their keywords based on their biosignals objectively. This can lead to cost-efficient healthcare services while maintaining the quality of care.

Future Research Focus

While this study offers a feasible solution for integrating HSNs and IoT for cost-efficient healthcare services, there are limitations to our approach that require further investigation, such as the accuracy of the disease-related keyword prediction model and the quality of extracted keywords and the word embedding dictionary. The limitations of the present study were that the data were emulated, which may render the proposed approach unsuitable in real social network environments. The prediction of keywords and disease also depends on the quality of extracted keywords and the word embedding dictionary. This may be the reason for the good result on certain diseases while have a lower accuracy for the other diseases. For further study, the performance can be further improved by using better keyword extraction methods. Thus, a new, more robust model may still be required. Future research also should focus on expanding KIHoT's capabilities to cover a wider range of diseases and investigate its scalability with larger amounts of data from multiple sources.

Conclusion

In conclusion, this study proposes an automated diagnosis framework that integrates HSNs and IoT to address the challenges of rising healthcare costs, limited accessibility to healthcare information, and the difficulty of finding relevant medical knowledge in HSNs. The proposed approach, KIHoT, utilizes electrocardiogram (ECG) signals to predict disease-related keywords and expand accessibility to healthcare information. The study demonstrates that KIHoT can effectively extract relevant information from HSN portals, achieving an averaged accuracy of 98% for disease prediction and a high rate of valid keywords.

The proposed KIHoT provides a cost-efficient method for health monitoring, automating data collection, data labeling, and model training processes, and requires no expert knowledge from users, significantly expanding accessibility to healthcare information. By integrating pervasive and autonomous IoT data feeds with HSN information and a community network, (a) usability of HSNs will be significantly improved; (b) cost of medical care will be significantly reduced; and (c) efficiency and effectiveness of medical services will be significantly improved. This approach focuses on helping users to search about their health condition on the Internet. Based on the highly active users in HSN, the integrated HSN could handle more diseases compared to normal automated diagnosis approaches.

Appendices

The word embedding model used in our study:

https://github.com/dartrevan/ChemTextMining/blob/master/word2vec/Health_2.5mreviews.s200.w10.n5.v15.cbow.bin

Extracted Keywords for Each Disease:

Dysrhythmia

Source: https://www.webmd.com/heart-disease/atrial-fibrillation/heart-disease-abnormal-heart-rhythm#3

Keyword: Beats, chest, heart headedness, breath

Bundle Branch Block

Source: https://www.webmd.com/heart-disease/what-is-heart-block#1

Keywords: Pain, breath, feeling, heart, beat

Valvular Heart Disease

Source: https://www.webmd.com/heart-disease/guide/heart-valve-disease#1

Keywords: Chest, weight, flip, beats, breath

Cardiomyopathy

Source: https://www.mayoclinic.org/diseases-conditions/cardiomyopathy/symptoms-causes/syc-20370709

Keywords: Abdomen, buildup, exertion, rest, legs

Healthy Control

Keywords: Health

Hypertrophy

Source: https://www.webmd.com/heart-disease/guide/hypertrophic-cardiomyopathy#1

Keywords: Exercise, rhythms, heart, breath, dyspnea

Myocardial Infarction

Source: https://www.webmd.com/heart-disease/guide/heart-disease-heart-attacks#1

Keywords: Pain, dizziness, weakness, heaviness, vomiting

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Data availability The data used in this study is publicly available and can be accessed through the PTB Diagnostic ECG Database (PTBDB) at the following URL: https://physionet.org/content/ptbdb/1.0.0/. All the documents related to this database are publicly available through the provided link.

Declarations

Conflict of interest The authors declare no conflicts of interest.

References

1. Mozaffarian D, Benjamin EJ, Go AS, Arnett DK, Blaha MJ, Cushman M, et al. Heart disease and stroke statistics–2015 update: A report from the American Heart Association. Circulation. 2015;131(4): e29.
2. Mozaffarian D, Benjamin E, Go A, Arnett D, Blaha M, Cushman M, et al. Heart disease and stroke statistics-2016 update: a report from the American Heart Association. Circulation. 2016;133(4): e38.
3. Song I, Marsh NV. Anonymous indexing of health conditions for a similarity measure. Inform Technol Biomed IEEE Trans. 2012;16(4):737–44.
4. Vong J, Song I. Automated Health Care Services. Emerging Technologies for Emerging Markets. Springer; 2015. p. 89–102.
5. Yildirim O. A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification. Comput Biol Med. 2018;96:189–202. https://doi.org/10.1016/j.compbiomed.2018.03.016.
6. Xiong ZH, Nash MP, Cheng E, Fedorov VV, Stiles MK, Zhao JC. ECG signal classification for the detection of cardiac arrhythmias using a convolutional recurrent neural network. Physiol Meas. 2018;39(9):10. https://doi.org/10.1088/1361-6579/aad9ed.
7. Azariadi D, Tsoutsouras V, Xydis S, Soudris D. ECG signal analysis and arrhythmia detection on IoT wearable medical devices. 2016 5th International Conference on Modern Circuits and Systems Technologies (MOCAST)2016. p. 1–4.
8. Alimova I, Tutubalina E. Automated detection of adverse drug reactions from social media posts with machine learning. International Conference on Analysis of Images, Social Networks and Texts: Springer; 2017. p. 3–15.
9. Miftahutdinov Z, Tropsha A, Tutubalina E. Identifying disease-related expressions in reviews using conditional random fields. 2017.
10. Coppersmith G, Dredze M, Harman C. Quantifying mental health signals in Twitter. Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality2014. p. 51–60.
11. Tutubalina E, Nikolenko S. Combination of deep recurrent neural networks and conditional random fields for extracting adverse drug reactions from user reviews. Journal of Healthcare Engineering. 2017;2017.
12. Achrekar H, Gandhe A, Lazarus R, Yu S-H, Liu B. Online Social Networks Flu Trend Tracker: A Novel Sensory Approach to Predict Flu Trends. In: Gabriel J, Schier J, Van Huffel S, Conchon E, Correia C, Fred A, et al., editors. Biomedical Engineering Systems and Technologies. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013. p. 353–68.
13. Sun X, Ye J, Ren F. Hybrid Model Based Influenza Detection with Sentiment Analysis from Social Networks. Chinese National Conference on Social Media Processing: Springer; 2015. p. 51–62.
14. Sharif H, Zaffar F, Abbasi A, Zimbra D. Detecting adverse drug reactions using a sentiment classification framework. 2014.
15. Na J-C, Kyaing WYM, Khoo CS, Foo S, Chang Y-K, Theng Y-L. Sentiment classification of drug reviews using a rule-based linguistic approach. International conference on asian digital libraries: Springer; 2012. p. 189–98.
16. Tutubalina E, Miftahutdinov Z, Nikolenko S, Malykh V. Medical concept normalization in social media posts with recurrent neural networks. J Biomed Inform. 2018;84:93–102. https://doi.org/10.1016/j.jbi.2018.06.006.
17. Cheng QJ, Li TMH, Kwok CL, Zhu TS, Yip PSF. Assessing suicide risk and emotional distress in chinese social media: a text mining and machine learning study. J Med Internet Res. 2017;19(7):10. https://doi.org/10.2196/jmir.7276.
18. Huang Y, Song I. Indexing Biosignal for Integrated Health Social Networks. ICBBE 2019. China, Shanghai: ACM; 2019
19. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:13013781. 2013.
20. Zhang L, Hall M, Bastola D. Utilizing Twitter data for analysis of chemotherapy. Int J Med Informatics. 2018;120:92–100. https://doi.org/10.1016/j.ijmedinf.2018.10.002.
21. Zhang S, Grave E, Sklar E, Elhadad N. Longitudinal analysis of discussion topics in an online breast cancer community using convolutional neural networks. J Biomed Inform. 2017;69:1–9. https://doi.org/10.1016/j.jbi.2017.03.012.
22. Tutubalina E, Nikolenko S. Exploring convolutional neural networks and topic models for user profiling from drug reviews. Multimedia Tools and Applications. 2018;77(4):4791–809. https://doi.org/10.1007/s11042-017-5336-z.
23. Rubin J, Abreu R, Ganguli A, Nelaturi S, Matei I, Sricharan K. Classifying heart sound recordings using deep convolutional neural networks and mel-frequency cepstral coefficients. Computing in Cardiology Conference (CinC), 2016: IEEE; 2016. p. 813–6.
24. Potes C, Parvaneh S, Rahman A, Conroy B. Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds. Computing in Cardiology Conference (CinC), 2016: IEEE; 2016. p. 621–4.
25. Nilanon T, Yao J, Hao J, Purushotham S, Liu Y. Normal/abnormal heart sound recordings classification using convolutional neural network. Computing in Cardiology Conference (CinC), 2016: IEEE; 2016. p. 585–8.
26. Bozkurt B, Germanakis I, Stylianou Y. A study of time-frequency features for CNN-based automatic heart sound classification for pathology detection. Comput Biol Med. 2018;100:132–43. https://doi.org/10.1016/j.compbiomed.2018.06.026.
27. Kucharski D, Grochala D, Kajor M, Kantoch E. A Deep Learning Approach for Valve Defect Recognition in Heart Acoustic Signal. In: Borzemski L, Swiatek J, Wilimowska Z, editors. Information Systems Architecture and Technology, Pt I. Advances in Intelligent Systems and Computing. Cham: Springer International Publishing Ag; 2018. p. 3–14.
28. Deperlioglu O. Classification of phonocardiograms with convolutional neural networks. Brain-Broad Res Artificial Intell Neurosci. 2018;9(2):22–33.

29. Dominguez-Morales JP, Jimenez-Fernandez AF, Dominguez-Morales MJ, Jimenez-Moreno G. Deep neural networks for the recognition and classification of heart murmurs using neuromorphic auditory sensors. IEEE Trans Biomed Circuits Syst. 2018;12(1):24–34. https://doi.org/10.1109/tbcas.2017.2751545.
30. Kang SH, Joe B, Yoon Y, Cho GY, Shin I, Suh JW. Cardiac auscultation using smartphones: pilot study. JMIR Mhealth Uhealth. 2018;6(2):11. https://doi.org/10.2196/mhealth.8946.
31. Aykanat M, Kilic O, Kurt B, Saryal S. Classification of lung sounds using convolutional neural networks. Eurasip Journal on Image and Video Processing. 2017:9. doi: https://doi.org/10.1186/s13640-017-0213-2.
32. Bardou D, Zhang K, Ahmad SM. Lung sounds classification using convolutional neural networks. Artif Intell Med. 2018;88:58–69. https://doi.org/10.1016/j.artmed.2018.04.008.
33. Kochetov K, Putin E, Azizov S, Skorobogatov I, Filchenkov A. Wheeze detection using convolutional neural networks. In: Gama J, Vale Z, Cardoso HL, editors. Oliveira E. Progress in Artificial Intelligence. Lecture Notes in Artificial Intelligence. Cham: Springer International Publishing Ag; 2017. p. 162–73.
34. Zhang WJ, Han JQ, Ieee. Towards Heart Sound Classification Without Segmentation Using Convolutional Neural Network. 2017 Computing in Cardiology. Computing in Cardiology Series. Los Alamitos: Ieee Computer Soc; 2017.
35. Limam M, Precioso F, Ieee. Atrial Fibrillation Detection and ECG Classification based on Convolutional Recurrent Neural Network. 2017 Computing in Cardiology. Computing in Cardiology Series. Los Alamitos: Ieee Computer Soc; 2017.
36. Liu Z, Zhang X. ECG-based heart arrhythmia diagnosis through attentional convolutional neural networks. IEEE Int Conf Internet Things Intell Syst (IoTaIS). 2021;2021:156–62.
37. Mehari T, Strodthoff N. Self-supervised representation learning from 12-lead ECG data. Comput Biol Med. 2021;141: 105114.
38. Huang Y, Song I, Rana P, Koh G. Fast diagnosis of bowel activities. 2017 International Joint Conference on Neural Networks (IJCNN)2017. p. 3042–9.
39. Mihalcea R, Tarau P. Textrank: Bringing order into text. Proceedings of the 2004 conference on empirical methods in natural language processing2004. p. 404–11.
40. White L, Togneri R, Liu W, Bennamoun M. Generating bags of words from the sums of their word embeddings. International Conference on Intelligent Text Processing and Computational Linguistics: Springer; 2016. p. 91–102.