sensors

MDPI

Article

Addressing Noise and Skewness in Interpretable Health-Condition Assessment by Learning Model Confidence †

Yuxi Zhou [1,2], Shenda Hong [3,4], Junyuan Shang [1,2], Meng Wu [1,2], Qingyun Wang [1,2], Hongyan Li [1,2,*] and Junqing Xie [5]

[1] School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China; joy_yuxi@pku.edu.cn (Y.Z.); sjy1203@pku.edu.cn (J.S.); 1501111363@pku.edu.cn (M.W.); wangqingyun@pku.edu.cn (Q.W.)

[2] Key Laboratory of Machine Perception, Ministry of Education, Peking University, Beijing 100871, China

[3] National Institute of Health Data Science at Peking University, Beijing 100191, China; hongshenda@pku.edu.cn

[4] Institute of Medical Technology, Health Science Center of Peking University, Beijing 100191, China

[5] Centre for Statistics in Medicine (CSM), Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (NDROMS), University of Oxford, Oxford OX3 7LD, UK; junqing.xie@ndorms.ox.ac.uk

* Correspondence: leehy@pku.edu.cn; Tel.: +86-010-6275-4911

† This paper is an extended version of our paper published in Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI), Macao, China, 10–16 August 2019; Yuxi Zhou, Shenda Hong, Junyuan Shang, et al. K-margin-based residual-convolution-recurrent neural network for atrial fibrillation detection.

Received: 20 November 2020; Accepted: 15 December 2020; Published: 19 December 2020

Abstract: Assessing the health condition has a wide range of applications in healthcare, military, aerospace, and industrial fields. Nevertheless, traditional feature-engineered techniques involve manual feature extraction, which are too cumbersome to adapt to the changes caused by the development of sensor network technology. Recently, deep-learning-based methods have achieved initial success in health-condition assessment research, but insufficient considerations for problems such as class skewness, noisy segments, and result interpretability make it difficult to apply them to real-world applications. In this paper, we propose a K-margin-based Interpretable Learning approach for health-condition assessment. In detail, a skewness-aware RCR-Net model is employed to handle problems of class skewness. Furthermore, we present a diagnosis model based on K-margin to automatically handle noisy segments by naturally exploiting expected consistency among the segments associated with each record. Additionally, a knowledge-directed interpretation method is presented to learn domain knowledge-level features automatically without the help of human experts which can be used as an interpretable decision-making basis. Finally, through experimental validation in the field of both medical and aerospace, the proposed method has a better generality and high efficiency with 0.7974 and 0.8005 F_1 scores, which outperform all state-of-the-art deep learning methods for health-condition assessment task by 3.30% and 2.99%, respectively.

Keywords: noise; class skewness; model interpretability; deep learning; health-condition assessment

1. Introduction

Assessing the health condition of complex systems has been an exciting research area for a long time. It has a wide range of applications in healthcare, military, aerospace, and industrial fields. For instance, in the field of aerospace, the structural safety state of an aircraft could be assessed by detecting whether there is any structural damage, and thus the maintenance of the aircraft could be arranged reasonably before a plane crash happened because of the structural damage which is invisible to the human eye.

Based on the time-series data recorded by various sensors in the course of the operation of complex systems, machine learning methods can be applied to learn effective characteristics which are associated with health-condition assessment; thereafter, they can be used for detecting and assessing the health condition of the complex systems.

Health-condition assessors can be divided into two categories including feature-engineered methods and deep-learning-based methods. Before the process of feature-engineered methods, a pre-processing phase is required to get the domain handcraft features from the health-condition monitoring data. Although extracting effective features is difficult and time-consuming, many excellent works have proposed various features and achieved good performance [1–8]. For instance, Bornn et al. [9] used a support vector machine (SVM) combined with an autoregressive (AR) model for structural damage detection, and used signal reconstruction and residual estimation to locate the unhealthy ones. Carrara et al. [10] use extracted features to classify the cardiac disease by logistic regression, k-Nearest Neighbors, and random forests. A Bayesian-based approach is employed for health-condition assessment of bridge expansion joints [11].

Nevertheless, these kinds of methods highly rely on the domain expertise of users, as they have to manually extract features, which are the key to making the model distinguish between different health conditions, from the dataset before using the model. Consequently, it is difficult to propose a general framework for the health-condition assessment using traditional feature-engineered methods. Moreover, with the development of sensor network technology, time-series records generated by various sensors of complex systems are bursting in data volume, expanding in data dimensions, and decreasing in data value density. However, traditional feature-engineered techniques, which involve time-consuming manual feature extraction, are too cumbersome to adapt to such changes.

Recently, deep-learning-based methods like Convolutional Neural Network (CNN) [12–16], Recurrent Neural Network (RNN) [17–20], Convolutional Recurrent Neural Network (CRNN) [21–23], and Residual Neural Networks (Res-Net) [24,25] have achieved initial success in health-condition assessment research by extracting features automatically. For example, CNN is employed for the automatic classification of electroencephalographic signals to diagnose the disease of Alzheimer's [26] and structural health monitoring of tall buildings subject to wind loads [27], while a CRNN-based approach is presented as a diagnostic tool for the classification of abnormal ECG signals [28].

Although the evaluation accuracy has been greatly improved compared with traditional methods [29–31], when facing problems in real-world applications, deep-learning-based health-condition assessment methods have insufficient considerations for problems such as class skewness, noisy segments, and result interpretability. These problems make it difficult to apply their success to real-world applications.

- **Class skewness:** As unhealthy conditions (e.g., diseases) occur rarely, the existence of class imbalance in health-condition monitoring records is inevitable. Classifying imbalanced data is a challenging problem [32]. Directly applying a machine learning algorithm may result in poor performance on the minority classes, as they are prone to ignore the subtle influence on the accuracy of an algorithm [33].
- **Noisy segments:** Most methods employ noise elimination approaches (e.g., grey spectral noise estimation [34], Dense Neural Network + spectrum-based noise elimination [35], CNN + Kalman filter [27], Encoder–Decoder [36], etc.) to denoise health-condition monitoring data, but they cannot deal with all of the noisy segments effectively, as the noisy segments refer to not only signal noise, but also the segments containing noisy labels. Taking a real-world electrocardiograph (ECG) record which is shown in Figure 1 as an example, its reference health-condition label given by cardiologists is 'Atrial Fibrillation'. However, it may still contain some other types of

segments (e.g., Normal sinus rhythm, etc.). Different from the irregular changes of the amplitude of segments with signal noises in the frequency domain or non-periodic variation in the time domain, only very subtle key differences in the segments that contain noisy labels can be found. For example, as shown in Figure 1, the morphological features of segments containing the noisy label of 'Normal sinus rhythm' are very similar to that containing only the ground-truth label of 'Atrial fibrillation', except that a 'P wave' only appears in the waveform of segments that contain the noisy label of 'Normal sinus rhythm'. However, the traditional noise elimination methods cannot effectively deal with these segments with noisy labels which are very similar to that containing only reference labels in morphological features. Furthermore, in the field of health-condition assessment, the accuracy of the model may be greatly reduced when a monitoring record contains too many noises, and thus it may lead to sudden death, air crash, and other tragic events due to the misdiagnosis. Therefore, in addition to accurate health-condition detection from the good-quality monitoring data, automatic recognition of low-quality noisy signals is required as well to remind domain experts to intervene in time. Nevertheless, the use of noise elimination methods may affect the accuracy of the detection of low-quality noisy signals, as it may cause changes or losses of features representing the quality of the monitoring signal itself.

- **Interpretability of prediction results:** Most existing deep-learning-based methods are often regarded as black-box models, as they can only output a specific diagnosis result without necessary interpretability. Nevertheless, in the real-life scenario of health-condition assessment, to realize "why" is, for the most part, more important. To explain the results of Deep Neural Network (DNN) models, most methods use the attention mechanism to highlight segments of an input record which are strongly associated with the model prediction and consider them as an explanation of prediction results [37,38], while others measure the role of each feature in the prediction process [39,40]. However, these kinds of highlighted segments and exploration of feature effects can only provide auxiliary information, which cannot be a decision-making basis actually, as they cannot provide domain technological proofs. For instance, to achieve a sound balance between the best commercial and safety decision-making carefully, airlines may require detailed technological information on why the flight plan of an aircraft has to be suspended immediately. Thus, interpretable explanations of the model must be technological proofs with admissive knowledge which can be widely accepted by domain experts.

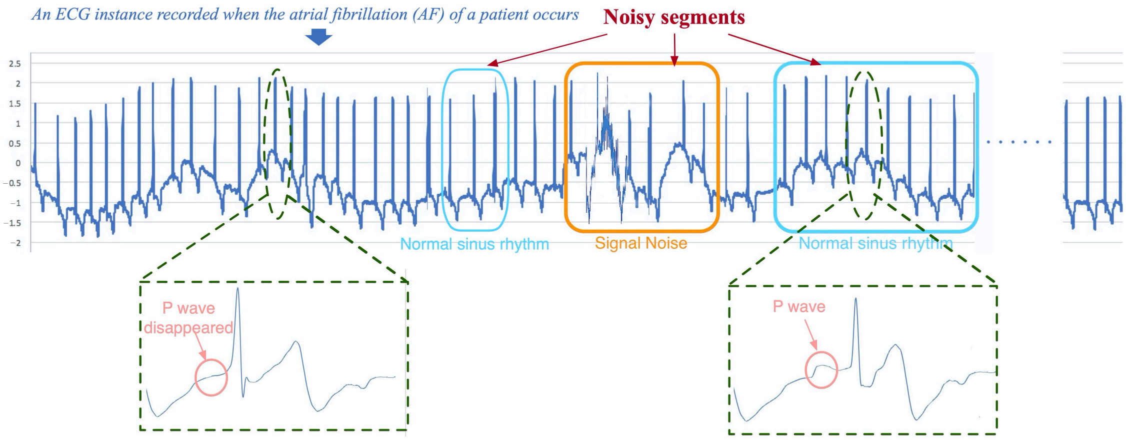


Figure 1. An example of a health-condition monitoring record with noisy segments whose label is Atrial Fibrillation.

Accordingly, there is a definite need for accurate, general, and interpretable health-condition monitoring solutions that can automatically deal with class skewness and noisy segments without changing or losing features representing the quality of the monitoring signal itself.

In this paper, we propose a K-margin-based intErpretable lEarNing (KEEN) for interpretable health-condition assessment.

- We first employ a skewness-aware RCR-Net model to establish a general deep learning model, which can automatically obtain the features and handle problems of class skewness in a wide range of health-condition monitoring data.
- A diagnosis model based on K-margin is presented to automatically focus on the most essential segments associated with a monitoring record, and tackle noisy segments by using the expected consistency among the segments which are related to the same record. The classification of each monitoring record can be identified by selecting the most likely label of the relevant top-K augmented segments.
- A knowledge-directed interpretation method is proposed to learn how to extract features that can be regarded as important domain knowledge without the help of human experts from the time-series data automatically. As a result, one can use these domain knowledge-level features as an interpretable decision-making basis.
- Thorough experiments are carried out on Atrial Fibrillation Monitoring dataset [41] and Structural Damage Monitoring Dataset [42]. The experimental results demonstrate that the proposed method with 0.7974 and 0.8005 F_1 scores, respectively, which outperform all state-of-the-art deep learning methods for health-condition assessment tasks by 3.30% and 5.71%.

2. Methods

2.1. Problem Definition and General Framework

Health-condition assessment is the task of automatically classifying a health-condition monitoring data into one of health-condition classes. Formally, we denote the training dataset as $\mathcal{D} = \{X, Y, Z\}$ where $X = \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(N)}\}$ are labelled health-condition monitoring sequence inputs, $Y = \{y^{(1)}, y^{(2)}, \ldots, y^{(N)}\}$ the corresponding label set, $Z = \{z^{(1)}, z^{(2)}, \ldots, z^{(N)}\}$ the set of corresponding domain knowledge attribute set, each domain knowledge attribute set $z^{(i)} = \{z^{(i)1}, z^{(i)2}, \ldots, z^{(i)F}\}$ has F domain knowledge attributes and N the total number of training data. Moreover, each health-condition monitoring record $\boldsymbol{x}^{(i)}$ has L monitoring channels $\boldsymbol{x} = \{\boldsymbol{x}^{(i)1}, \boldsymbol{x}^{(i)2}, \ldots, \boldsymbol{x}^{(i)L}\}$ and each channel is a time-series $\boldsymbol{x}^{(i)l} = \{\boldsymbol{x}_1^{(i)l}, \boldsymbol{x}_2^{(i)l}, \ldots\}$.

Given the unlabelled testing dataset $\mathcal{D}_{\sqcap} = \{X\}$ where $X = \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(N)}\}$ are unlabelled health-condition monitoring sequence inputs, the goal of the health-condition assessment is to learn a predictive model which takes unlabelled health-condition monitoring sequence $\boldsymbol{x}^{(i)}$ as input and outputs the prediction of its domain knowledge attributes $\hat{z}^{(i)}$ and its prediction of health-condition class $\hat{y}^{(i)} \in \mathcal{C}$ where $\mathcal{C} = \{c_1, c_2, \ldots, c_M\}$ is a set of M different health-condition classes.

As research results based on deep learning have achieved remarkable success in health-condition assessment tasks recently, which can naturally integrate and extract hierarchical features, the DNN model is used as the basic classifier. Moreover, to automatically extract features from CNN and capture long-term trends from RNN, we use a residual convolution recurrent neural network (RCR-Net), as shown in Figure 2.

Nevertheless, a RCR-Net model still cannot be directly applied for reliably detecting and assessing health conditions from monitoring data with noise and class imbalance. On the one hand, the class imbalance and semantical ambiguities caused by noisy segments may lead to poor and even unacceptable quality of DNN models. On the other hand, a proper explanation of the health-condition assessment result is required to support reliable decision-making.

A K-margin-based interpretable learning method is presented to solve the above difficulties. Specifically, a skewness-aware RCR-Net approach (Section 2.2) is presented to alleviate the problem of class imbalance. In addition, to tackle noise, a K-margin based diagnosis model (Section 2.3) is proposed, which can automatically focus on the most important segments and involve only part of the labeled segments in the skewness-aware RCR-Net learning process. Finally, a knowledge-directed diagnosis interpretation model is employed to extract features that can be regarded as important domain knowledge. The framework of our model is shown in Figure 2.

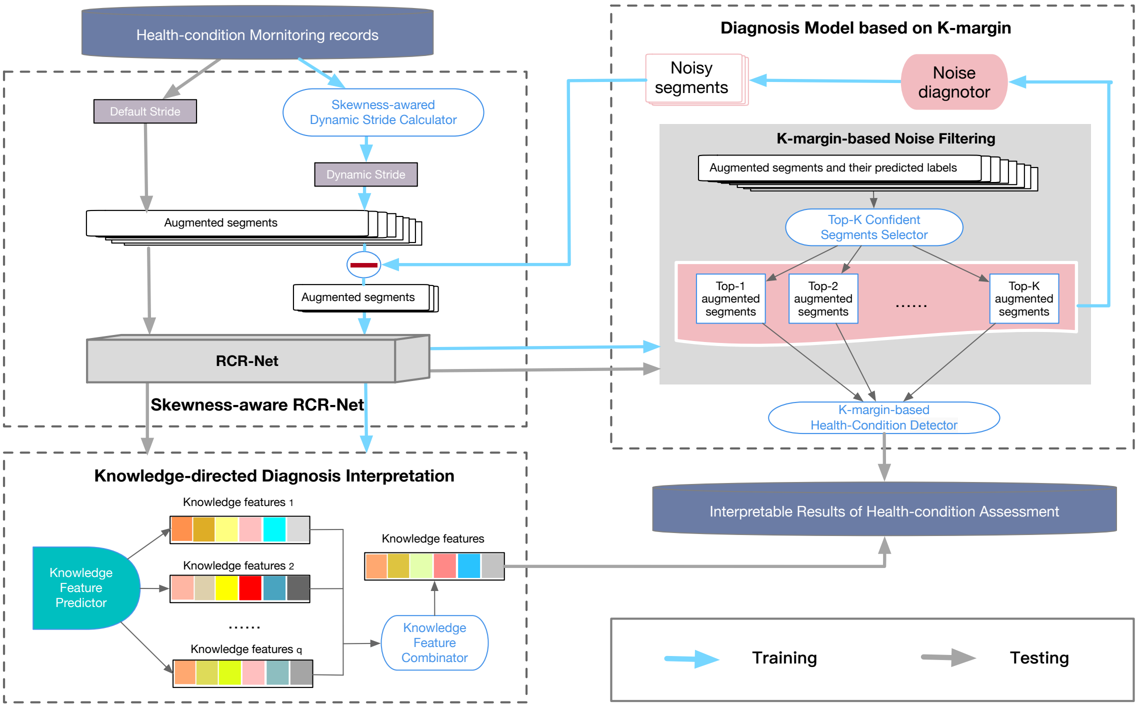

Health-condition Mornitoring records
Default Stride
Skewness-awared Dynamic Stride Calculator
Dynamic Stride
Augmented segments
Augmented segments
RCR-Net
Skewness-aware RCR-Net
Diagnosis Model based on K-margin
Noisy segments
Noise diagnotor
K-margin-based Noise Filtering
Augmented segments and their predicted labels
Top-K Confident Segments Selector
Top-1 augmented segments
Top-2 augmented segments
......
Top-K augmented segments
K-margin-based Health-Condition Detector
Interpretable Results of Health-condition Assessment
Knowledge-directed Diagnosis Interpretation
Knowledge Feature Predictor
Knowledge features 1
Knowledge features 2
......
Knowledge features q
Knowledge features
Knowledge Feature Combinator
Training
Testing

Figure 2. The framework of the k-margin-based interpretable model.

2.2. Skewness-Aware RCR-Net

To alleviate the problems of the lack of well-labelled data and class imbalance, a skewness-aware RCR-Net model is proposed to deal with these problems. First, a skewness-aware data augmentation is employed to generate more short-term segments from the long-term monitoring records. After that, a multi-view RCR-Net model is presented to establish a general deep learning model for capturing features automatically from the health-condition monitoring data.

Generally, two predefined parameters are required for the skewness-aware data augmentation—the length of the sliding windows $\mathscr{W}$) and the maximum stride threshold $\mathscr{M}$. For records with rare labels, the dynamic stride $\mathscr{S}$ becomes smaller, while, for records with common labels, it becomes larger. Formally, given the maximum stride threshold of $\mathscr{M}$ and the label set $C = \{c_1, c_2, \ldots, c_M\}$, the dynamic stride of records with labeled C_j is given by the following formula:

$$\mathscr{S}_{c_j} = \lceil \mathscr{M} \times \frac{|\text{records labelled } c_j\ |}{\max_{s=1}^{M}\ |\text{records labelled } c_s\ |} \rceil \tag{1}$$

Please note that, if the length of a monitoring record is less than the length of the sliding windows $\mathscr{W}$, a zero-padding approach would be employed to pad it with zeros at the end of it

Additionally, a multi-view RCR-Net is presented to establish a general deep learning model, which can automatically obtain features for a wide range of health conditions. Specifically, it consists of 33 layers of residual blocks [43], one layer of recurrent block, and one layer of a fully connected block. The purpose of using residual blocks is to construct a deeper model through the residual connection between blocks, and automatically extract more effective local morphological-level features. To capture potential trend-level features in monitoring data, a recurrent layer with Bi-directional Long-Short Term Memory (Bi-LSTM) cells is employed. Finally, the prediction is made by a fully connected layer and a softmax layer. The cross-entropy loss of the objective function is calculated to optimize the loss of the training neural network.

The high-level architecture of the multi-view RCR-Net is shown in Figure 3. It takes an augmented segment as input. After that, it splits the augmented segments into $\lceil \frac{\mathscr{W}}{\mathscr{N}} \rceil$ fragments with length of $\mathscr{N}$ and outputs the health-condition prediction of the augmented segment.

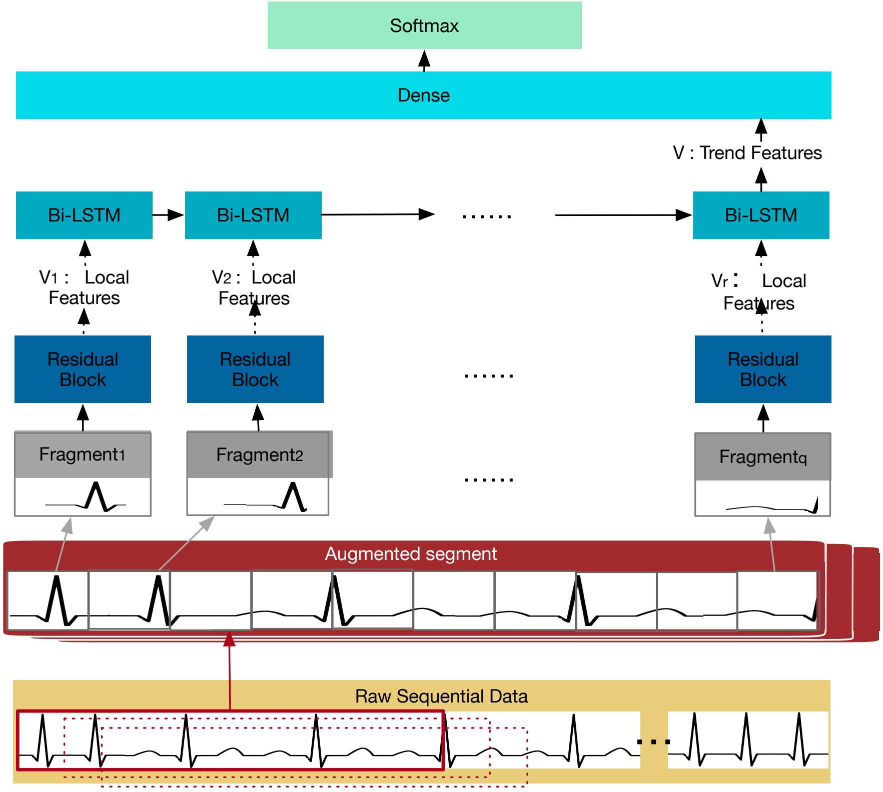


Figure 3. The high-level architecture of the multi-view RCR-Net model.

2.3. Diagnosis Model Based on K-Margin

Although the improvement of data inadequacy and class imbalance through data augmentation process is essential to enhancing the performance of the health-condition assessment model, it inevitably generates "hard" health-condition monitoring segments due to noisy segments. For example, as shown in Figure 4, the reference label of the 8th–14th augmented segments is 'Atrial Fibrillation' as the reference label of the ECG instance is 'Atrial Fibrillation'. Nevertheless, the main ECG signs are 'Too noisy to classify' for the 8th–11th augmented segments and 'Normal sinus rhythm' for the 12th–14th augmented segments. Therefore, to filter these noisy segments, we first enhance the robustness of our model by using a K-margin-based noise filtering approach to compute the cross-entropy objective function of only a portion of the selected segments. Moreover, a K-margin-based health-condition Detector is proposed to predict the health condition of a monitoring record according to the top-K confident segments.

2.3.1. K-Margin-Based Noise Filtering

A minimum uncertainty margin is proposed to select appropriate segments to calculate cross-entropy. As the prediction of all augmented segments of each health monitoring record can be obtained by a trained multi-view RCR-Net model, and thus we define the uncertainty margin of t-th segment $\boldsymbol{x}_t^{(i)}$ for a given monitoring record $\boldsymbol{x}^{(i)}$ as:

$$Margin(\boldsymbol{x}_t^{(i)}) = P(\ddot{y}_t^{(i)}|\boldsymbol{x}_t^{(i)}) \; - \; P(\dot{y}_t^{(i)}|\boldsymbol{x}_t^{(i)}) \tag{2}$$

where $\dot{y}_t^{(i)}$ and $\ddot{y}_t^{(i)}$ are the most probable and second-most probable prediction classes of the record $\boldsymbol{x}_t^{(i)}$.

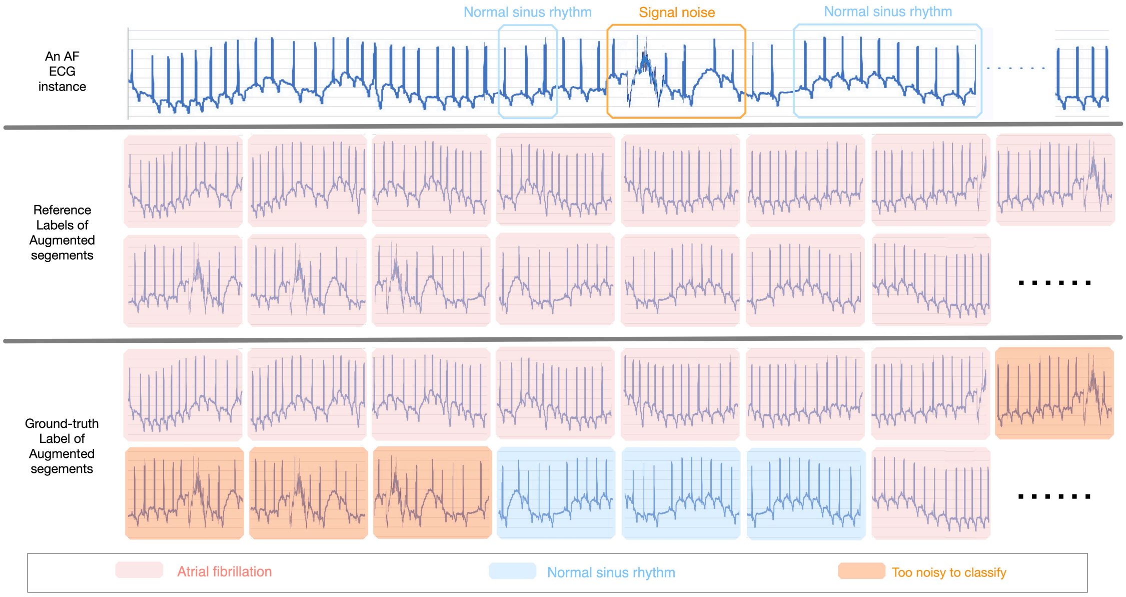


Figure 4. An example of "hard" segments with noisy labels caused by the data augmentation process.

Usually, the trained model has less doubt (that is, more confidence) in distinguishing the two most probable categories if the uncertainty margin is smaller for a given health monitoring record $\boldsymbol{x}_t^{(i)}$. Instead, segments with larger uncertainty margins are more ambiguous. Therefore, the most confident segment of i-th monitoring record $\boldsymbol{x}^{(i)}$ can be defined as:

$$\boldsymbol{x}_*^{(i)} = \underset{\boldsymbol{x}_t^{(i)}, t \in \{1,2,\ldots,T\}}{\arg\min} \quad (Margin(\boldsymbol{x}_t^{(i)})) \tag{3}$$

where $\boldsymbol{x}_t^{(i)}$ is the t-th segment of the record $\boldsymbol{x}_i$. A diagnosis based on minimum uncertainty margin is a strategy to find the predicted class of a monitoring record with the largest confidence. Similarly, the most confident label of a record $\boldsymbol{x}^{(i)}$ can be defined as follows:

$$\hat{y}_*^{(i)} = \underset{\hat{y}_t^{(i)}, t \in \{1,2,\ldots,T\}}{\arg\min} \quad (P(\ddot{y}_t^{(i)}|\boldsymbol{x}_t^{(i)}) - P(\hat{y}_t^{(i)}|\boldsymbol{x}_t^{(i)})) \tag{4}$$

A K-margin-based health-condition label prediction algorithm (K-margin) is presented (see Algorithm 1) to select top-K most confident segments for training the multi-view RCR-Net. Given a health-condition monitoring record $\boldsymbol{x}^{(i)}$ and its augmented segments $\boldsymbol{x}^{(i)} = \{\boldsymbol{x}_1^{(i)}, \boldsymbol{x}_2^{(i)}, \ldots, \boldsymbol{x}_T^{(i)}\}$, it requires K iterations to output the top-K most confident segments $\boldsymbol{X}_{1:K}^{(i)}$ and their label $\hat{Y}_{1:K}^{(i)}$ under trained multi-view RCR-Net model ϕ.

Intuitively, the top-K most confident fragments $\boldsymbol{X}_{1:K}^{(i)}$ can be used for the multi-view RCR-Net training process, instead of the entire segments array $\boldsymbol{x}^{(i)}$, to avoid learning features from noise segments and automatically focus on the most essential part that can represent the characteristics of a given record. Nevertheless, it is not reliable to use them for the training process when the model is not reliable yet, since they are selected by the KEEN model. Hence, we calculate the average probability of predictions for all augmented segments of a record $\boldsymbol{x}^{(i)}$:

$$\alpha_i = \frac{1}{T} \; (\sum_{t=1}^{T} P(\hat{y}_t^{(i)}|\boldsymbol{x}_t^{(i)})) \tag{5}$$

where $P(\hat{y}_t^{(i)}|\boldsymbol{x}_t^{(i)})$ is the prediction probability of t-th augmented candidate for a record $\boldsymbol{x}^{(i)}$, and T is the number of augmented candidates for the record $\boldsymbol{x}^{(i)}$. The top-K most confident candidates for the given record $\boldsymbol{x}_{1:K}^{(i)}$ would be used in the model fine-tuning process to achieve higher consistency among the augmented candidates and reduce the impact of noisy ones when $\alpha_i > 5$. Otherwise, all the candidates would be used. The reason is that the low average probability of predictions for all augmented candidates indicates a poor model performance. Therefore, the multi-view RCR-Net model is not accurate and reliable, a more "hard" sample is required for the training process. Thereafter, the selected $alpha$-segments $\boldsymbol{X}_{\alpha}^{(i)}$can be defined as follows:

$$\boldsymbol{X}_{\alpha}^{(i)} = \begin{cases} \boldsymbol{X}_{1:K}^{(i)}, & \text{if } \alpha_i > 0.5 \\ \boldsymbol{x}^{(i)}, & \text{otherwise} \end{cases} \tag{6}$$

Consequently, the task of health-condition assessment using our KEEN model can be expressed as optimizing the cross-entropy objective function:

$$L(X, \hat{y}) = \frac{1}{N \cdot |\boldsymbol{X}_{\alpha}^{(i)}|} \sum_{i=1}^{N} \sum_{t=1}^{|\boldsymbol{X}_{\alpha}^{(i)}|} \log P(Y = \hat{y}_t^{(i)} | X = \boldsymbol{X}_{\alpha t}^{(i)}) \tag{7}$$

Algorithm 1 K-margin($\boldsymbol{x}^{(i)}, K, \phi$)

Input: A T-segments array $\boldsymbol{x}^{(i)} = \{\boldsymbol{x}_1^{(i)}, \boldsymbol{x}_2^{(i)}, \ldots, \boldsymbol{x}_T^{(i)}\}$ of an input health-condition monitoring record $\boldsymbol{x}^{(i)}$, trained multi-view RCR-Net model ϕ

Parameter: An integer K

Output: Top-K most confident segments $\boldsymbol{X}_{1:K}^{(i)}$ and their label predictions $\hat{Y}_{1:K}^{(i)}$

```
$x'^{(i)} \leftarrow \boldsymbol{x}^{(i)}$
$\boldsymbol{X}_{1:K}^{(i)} \leftarrow \emptyset,\ \hat{Y}_{1:K}^{(i)} \leftarrow \emptyset,\ k \leftarrow K$
while $k > 0$ do
    $\boldsymbol{x}_*^{(i)} \leftarrow$ Most confident segment predicting by $\phi(x'^{(i)})$ using Equation (3)
    $\hat{y}_*^{(i)} \leftarrow$ Most confident label predicting by $\phi(x'^{(i)})$ using Equation (4)
    $x'^{(i)} \leftarrow x'^{(i)}) \setminus \boldsymbol{x}_*^{(i)}$ # Implementation: remove segment index of $\boldsymbol{x}_*^{(i)}$ from segment indexes list of $x'^{(i)}$
    $\boldsymbol{X}_{1:K}^{(i)} \leftarrow \boldsymbol{X}_{1:K}^{(i)} \cup \{\boldsymbol{x}_*^{(i)}\},\ \hat{Y}_{1:K}^{(i)} \leftarrow \hat{Y}_{1:K}^{(i)} \cup \{\hat{y}_*^{(i)}\}$
    $k \leftarrow k - 1$
end while
return $\boldsymbol{X}_{1:K}^{(i)}, \hat{Y}_{1:K}^{(i)}$
```

2.3.2. K-Margin-Based Health-Condition Detector

In order to diagnose the health-condition for the given health-condition monitoring record $\boldsymbol{x}^{(i)}$, given the training model Φ, we convert the label predictions $\hat{Y}_{1:K}^{(i)}$ of the top-K most confident segments into the matrix $\hat{\boldsymbol{Y}}_{1:K}^{(i)} \in \mathbb{R}^{K \times M}$ as follows:

$$\hat{\boldsymbol{Y}}_{k,j}^{(i)} = \begin{cases} 1, & \text{if } c_j = \arg\max_{1 \le j \le M} \hat{Y}_{k,j}^{(i)} \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

Furthermore, a K-majority weighted voting algorithm based on the minimum uncertainty margin is presented to output the most likely label of a given record $\boldsymbol{x}^{(i)}$. The k-majority weighted voting method is to vote on the classes that the record $\boldsymbol{x}^{(i)}$ may belong to, which can be defined as follows:

$$\hat{y}^{(i)} = \arg\max_{c_j \in \mathcal{C}} (\sum_{1}^{K} \hat{Y}_{kj}^{(i)}) \tag{9}$$

Thus far, by naturally exploiting expected consistency among the segments associated with each record, our KEEN model can make a diagnosis of a given record $\boldsymbol{x}^{(i)}$. It can automatically deal with noise as only a portion of segments that belong to the same record would be included in the learning process; thereafter, diagnosis would be made according to the predictions of the most essential segments as well.

2.4. Knowledge-Directed Interpretation

In the real-life scenario of health-condition assessment, the method of computer-aided diagnosis requires a high degree of interpretability so that humans can give a reliable conclusion based on the diagnosis basis.

Recently, some methods try to explain the DNN model by highlighting the most relevant segments of health-condition monitoring data [38] and exploring feature effects [40] in the prediction process. Nevertheless, this kind of method cannot provide detailed domain technological-level information on "why", as we still do not know the relationship between this kind of explanation and domain knowledge.

The main goal of knowledge-directed interpretation is to provide an indispensable domain technological-level information on "why", and to make the complex reliability of predicting results well explicable in order to make reliable decisions. Specifically, we integrate domain knowledge into the training process of the skewness-aware RCR-Net.

The function of knowledge-directed interpretation is similar to the observation process when domain experts try to classify health-condition monitoring data. When domain experts diagnose health-condition monitoring records, they first observe the characteristics of these records. Taking arrhythmia diagnosis as an example, cardiologists usually analyze the ECG record to see characteristics such as "P waves disappear", "RR-interval", etc. After that, using these characteristics to classify them. Inspired by this, a knowledge-directed interpretation method is proposed to convert the morphological-level local features, which are automatically extracted from the health-condition monitoring record, into the knowledge-level features.

The architecture of the knowledge-directed interpretation method is shown in Figure 5. The health-condition predictor is shown in the orange dashed box, which trains a multi-view RCR-Net for health-condition classification. The red one is a knowledge-level feature extraction step, which shares weights with the Residual layer of health-condition predictor, and then replaces the Bi-LSTM layer and Softmax layer with a Mean square error loss (MSE) layer to form a knowledge-level feature extractor.

It takes a fragment $\boldsymbol{x}_{tq}^{(i)}$ of a segment $\boldsymbol{x}_{t}^{(i)}$ which is augmented from a health-condition monitoring recode $\boldsymbol{x}^{(i)}$, and outputs the prediction of its knowledge-level features $\hat{z}_{tq}^{(i)}$. Subsequently, we concatenate all knowledge-level features of the fragments which belong to the same segment $\boldsymbol{x}_{t}^{(i)}$ to get the combined features.

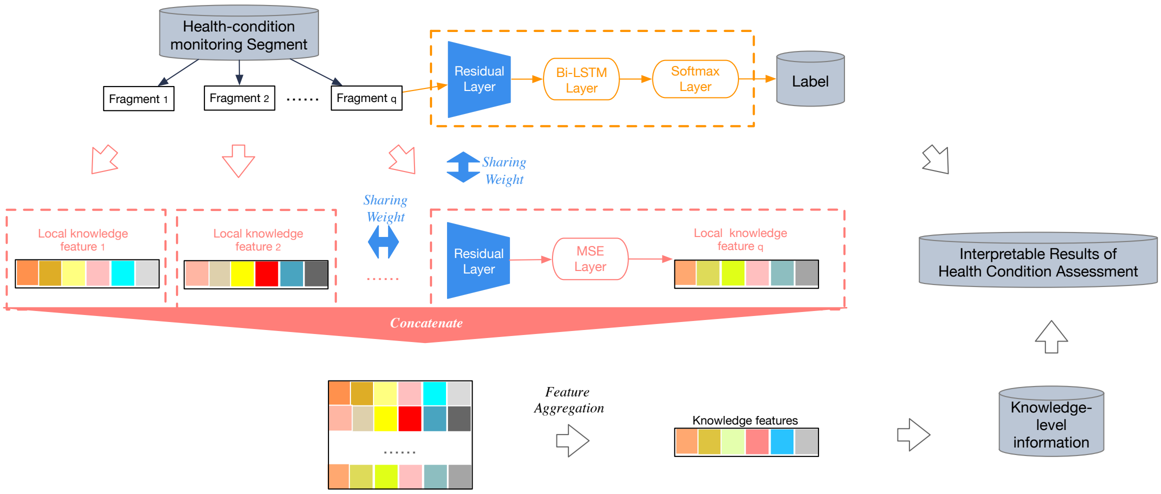

Health-condition monitoring Segment
Fragment 1
Fragment 2
……
Fragment q
Residual Layer
Bi-LSTM Layer
Softmax Layer
Label
Sharing Weight
Sharing Weight
Local knowledge feature 1
Local knowledge feature 2
……
Residual Layer
MSE Layer
Local knowledge feature q
Concatenate
Interpretable Results of Health Condition Assessment
……
Feature Aggregation
Knowledge features
Knowledge-level information

Figure 5. The architecture of a knowledge-directed interpretation method.

Formally, given extracted features $\hat{z}_t^{(i)} = \{\hat{z}_{t1}^{(i)}, \ldots, \hat{z}_{tq}^{(i)}\}$ corresponding to one segment $\boldsymbol{x}_t^{(i)}$, where each $\hat{z}_{tq}^{(i)} \in \mathbb{R}^F$, and $\hat{z}_t^{(i)} \in \mathbb{R}^{q \times F}$. The aggregated features are computed by column-wise aggregation operation of $\hat{z}_{tq}^{(i)}$. Then, the f-th knowledge-level feature of a segment $\boldsymbol{x}_t^{(i)}$ can be denoted as:

$$\hat{z}_t^{(i)f} = \psi_f(\{\hat{z}_{tq}^{(i)} \mid \boldsymbol{x}_{tq}^{(i)} \in \boldsymbol{x}_t^{(i)}\}) \tag{10}$$

where $\psi_f(\cdot)$ is an aggregation operation. Usually, pooling operation (e.g., sum, max, average, etc.) [44] can be used to aggregate these features. As each of the knowledge-level features varies very much in nature, it requires aggregating them into one unified feature vector according to their own properties. To name only a few, a Max Pooling operation would be employed as the aggregation operation for the knowledge-level attribute of "P waves disappear" to determine whether there is "P wave disappear" in some fragments of a segment. On the contrary, Mean Pooling operation would be better for the knowledge-level attribute of "ST slope". Sometimes, other aggregation operations may be used as well; for example, the calculation of the standard deviation coefficient may be employed for the knowledge-level attribute of "PR interval" to determine whether the interval is unequal on the entire record.

Similarly, the f-th knowledge-level feature of a record $\boldsymbol{x}^{(i)}$ can be denoted as:

$$\hat{z}^{(i)f} = \psi_f(\{\hat{z}_t^{(i)} \mid \boldsymbol{x}_t^{(i)} \in \boldsymbol{x}^{(i)}\}) \tag{11}$$

To avoid learning knowledge-level information from noisy segments, the aggregation operation would be only applied on the selected top-K most confident segments for the given health-condition monitoring record $\boldsymbol{x}^{(i)}$ under trained model ϕ. We chose the mean square error loss as the empirical loss; thereafter, we optimize the knowledge-directed interpretation model by gradient descent as follows:

$$L(\boldsymbol{X}, \hat{z}) = \frac{1}{N \cdot |\boldsymbol{X}_\alpha^{(i)}| \cdot F} \sum_{i=1}^{N} \sum_{t=1}^{|\boldsymbol{X}_\alpha^{(i)}|} \sum_{f=1}^{F} (z_t^{(i)f} - \hat{z}_t^{(i)f})^2 \tag{12}$$

The total loss could be defined as follows:

$$L(\boldsymbol{X}, \hat{y}, \hat{z}) = \frac{1}{N \cdot |\boldsymbol{X}_\alpha^{(i)}|} \sum_{i=1}^{N} \sum_{t=1}^{|\boldsymbol{X}_\alpha^{(i)}|} \log P(Y = \hat{y}_t^{(i)} | X = \boldsymbol{X}_{\alpha t}^{(i)}) + \frac{1}{N \cdot |\boldsymbol{X}_\alpha^{(i)}| \cdot F} \sum_{i=1}^{N} \sum_{t=1}^{|\boldsymbol{X}_\alpha^{(i)}|} \sum_{f=1}^{F} (z_t^{(i)f} - \hat{z}_t^{(i)f})^2 \tag{13}$$

In this way, domain technological-level features (e.g., P waves disappear, RR-interval, etc.), which correspond to domain knowledge that is convincing and understandable to a domain expert, can be extracted.

3. Experiments

3.1. Dataset

We carry out experiments on two datasets:

- Atrial Fibrillation Monitoring Dataset [41]: 8528 ECG monitoring records, which last from 9 s to slightly more than 60 s, is contained. All ECG records were sampled as 300 Hz and they have been band pass filtered by the AliveCor device. However, there are still many signal noises contained in this dataset. To filter the low-quality monitoring signals, all monitoring records which contain too many signal noises would be classified into a category called "Too noisy to classify". Moreover, all health conditions that are abnormal but not an atrial fibrillation (e.g., junctional arrhythmia, ventricular arrhythmia, etc.) are considered as a single category, that is, other rhythms. As a result, these records are divided into four categories of health conditions: (1) Normal sinus rhythm **N** (contains 5050 records, accounting for about 59.22%), (2) Atrial Fibrillation **A** (contains 738 records, accounting for about 8.65%), (3) Other rhythm **O** (contains 2456 records, accounting for about 28.80%), and (4) Too noisy to classify **P** (contains 284 records, accounting for about 3.33%).
- Aircraft Monitoring Dataset [42]: The aircraft structural damage monitoring dataset contains 808 labeled structural damage testing records lasting about 1 s sampling with 10 MHz. Each record has two channels—the baseline channel and monitoring channel. The data of the baseline channel were collected before the four-point bending test for composite laminates and impact test for stiffened plate were carried out. In addition, the data of the monitoring channel were collected during the four-point bending test and impact test. These records are classified as 2 health-condition class: (1) Normal **N** (contains 538 records, accounting for about 66.58%), (2) Structural damage **D** (contains 270 records, accounting for about 33.42%).

3.2. Performance Measurements

Several common metrics (Precision and Recall) are used to measure the model performance by evaluating the proximity of the predicted labels to the referenced labels given by domain experts:

- $Precision = \frac{1}{M}\sum_{c=1}^{M}\sum_{i\in\{i|y^{(i)}=c\}}\frac{\mathbb{1}(y^{(i)}=\hat{y}^{(i)})}{|\{i|y^{(i)}=c\}|}$.
- $Recall = \frac{1}{M}\sum_{c=1}^{M}\sum_{i\in\{i|y^{(i)}=c\}}\frac{\mathbb{1}(y^{(i)}=\hat{y}^{(i)})}{|\{i|\hat{y}^{(i)}=c\}|}$.

In addition, the following F_1 evaluation measurements are employed for the Atrial Fibrillation Monitoring dataset:

- *F_1 scores of each class*: Denoted as F_{1N} for normal sinus rhythm, F_{1A} for AF, F_{1O} for other rhythm, F_{1P} for noise. Detailed definitions can be found in [41].
- *Averages of F_1 scores*: $F_1 = \frac{F_{1N}+F_{1A}+F_{1O}+F_{1P}}{4}$.

Similarly, the following F_1 evaluation measurements are employed for the Aircraft Monitoring Dataset by using the counting rules shown in Table 1:

- *F_1 scores of Normal class*: $F_{1N} = \frac{2\cdot N_n}{\sum_N + \sum_n}$
- *F_1 scores of Structural Damage class*: $F_{1D} = \frac{2\cdot D_d}{\sum_D + \sum_d}$
- *Averages of F_1 scores*: $F_1 = \frac{F_{1N}+F_{1D}}{2}$

Table 1. Counting rules for the numbers of the variables used in the F_1 evaluation measurements for Structural Damage Monitoring Dataset.

		Predicted Class		
		Normal(N)	Structural Damage(D)	Total
Reference Class	Normal(N)	N_n	N_d	$\sum N$
	Structural Damage(D)	D_n	D_d	$\sum D$
	Total	$\sum n$	$\sum d$	
		……		

3.3. Implementation Details

The above performance measurements are used to evaluate the effectiveness of various state-of-the-art DNN methods and reported average results by running a 5-fold cross validation. Furthermore, the knowledge-level attributes (P wave disappear, P wave amplitude, QRS complex amplitude, QRS duration, T wave amplitude, and PR interval, and ST slope) are extracted from the Atrial Fibrillation Monitoring dataset using the ECGPUWAVE [45] on each fragment and the knowledge-level attributes of each segment and record are generated by aggregation operations to train the knowledge-directed interpretation model. Similarly, the knowledge-level attributes (Peak-to-peak amplitude [46], Pearson correlation coefficient [47], and Signal energy [48]) are extracted from the Aircraft Monitoring Dataset.

For the training dataset, the parameter value of $\mathscr{W}$ and max stride threshold $\mathscr{M}$ is set to be 6000 and 500 for skewness-aware data augmentation, respectively—while, for the testing dataset, as their labels are unknown, a sliding window with a window size $\mathscr{W} = 6000$ would still be used for generating the segments, but its stride would be set to a default value of 300. In addition, the parameter values of K and $\mathscr{N}$ are set to be 3 and 300, respectively, for training the KEEN (without knowledge-directed interpretation) model and KEEN+ (including knowledge-directed interpretation) model (See Section 2.3) which are implemented on Tensorflow r1.4 using Python 3.6.2.

3.4. Comparing with Other Methods

The following state-of-the-art deep neural network methods (their shape of input and output are similar to the Atrial Fibrillation Monitoring dataset and Aircraft Monitoring Dataset to avoid the changes of model architecture) are compared with the proposed methods:

- *CNN* uses a variation of multilayer perceptrons to automatically extract features. Recently, Sodmann et al. [49] constructed CNN architecture for health-condition assessment without manually extracted features.
- *RNN* can automatically extract the time-domain trend features by allowing it to exhibit temporal dynamic behavior. It is used as the major classifier for health-condition assessment in Faust et al. [20].
- *CRNN* takes advantage of convolutional neural networks (CNNs) for local feature extraction and recurrent neural networks (RNNs) for temporal summarization of the extracted features. The CRNN model is employed in Liu et al. [22] for detecting the unhealthy condition of myocardial infarction (MI).
- *ResNet* can build very deep networks through modules called residual layers, which can avoid higher errors rate due to naively adding CNN layers [43]. In Hannun et al. [24], it is employed for cardiac health-condition assessment, and the accuracy is comparable to or higher than that of cardiologists.

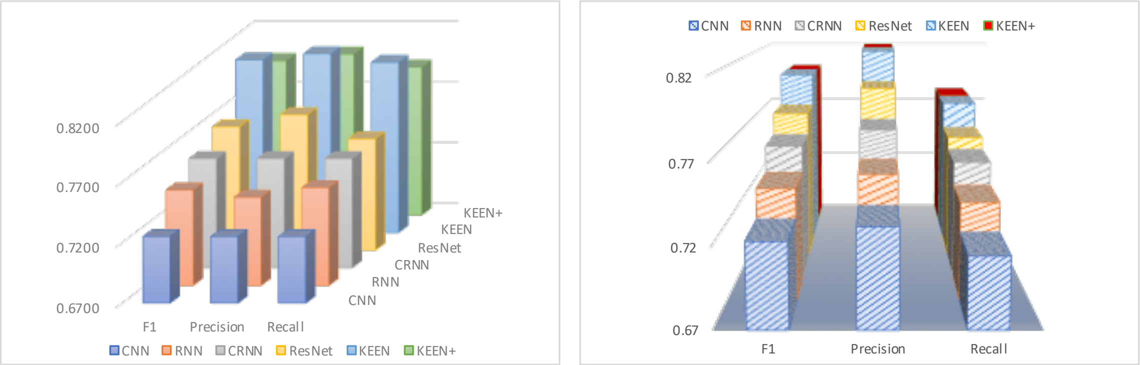
4. Results Discussion

4.1. Effect of Application in the Field of Medical Health

To validate the accuracy of the proposed methods in the medical field, the Atrial Fibrillation Monitoring dataset, which contains ECG records sampled from 8528 selected patients in real-world scenarios, is used for evaluating the effectiveness of our methods for the task of atrial fibrillation diagnosis.

4.1.1. Effectiveness of Atrial Fibrillation Diagnosis

It is clear from the diagrams in Figure 6a that the precision, recall, and F_1 scores of the KEEN and KEEN+ are better than all the state-of-the-art deep learning methods in the task of atrial fibrillation diagnosis. The methods of CNN, RNN, RCNN, and RESNET are not effective, mainly due to the high noise contamination in this data set, and such methods do not deal with the problem of noisy segments. The method of ResNet works better mainly because of deeper networks. However, it is limited by the ability of model expression of the first three methods to build very deep networks, which may lead to the information loss and feature loss.



(a) On the Atrial Fibrillation Monitoring dataset.

(b) On the Aircraft Monitoring dataset.

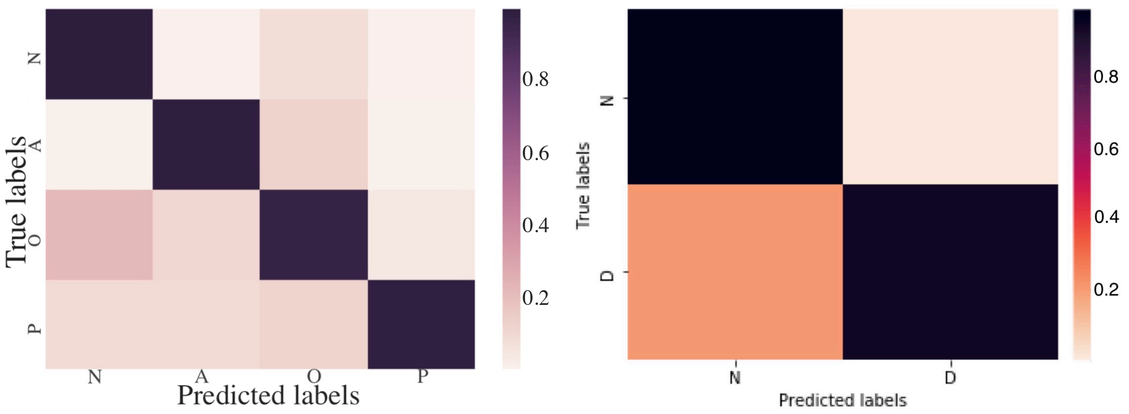
Figure 6. The result of Precision, Recall, F1 scores.

Additionally, the F_1 scores on the Atrial Fibrillation Monitoring dataset are demonstrated in Table 2. We can see that it is an effective method to improve prediction performance by automatically focusing on the most essential segments, as the F_1 scores of KEEN and KEEN+ are 0.8125 and 0.7974 respectively, which are 5.26–12.08% and 3.30–10.00% higher than other DNN methods. Remarkably, compared to other DNN methods, the prediction performance of our methods are improved significantly for detecting the low-quality records ("Too Noisy To Classify"). This is partly due to the improvement of class skewness (as the skewness-driven dynamic augmentation makes the proportion of records with different labels closer to each other, that is, the training records of "Too Noisy To Classify" increase greatly), and the improvement of noisy segments problem (as our method would be more attentive to the most important segments and avoid learning features from noise segments for all records, which could reduce, to some extent, the misclassification of "Too Noisy To Classify" records).

Moreover, the confusion matrix of the model predictions on the test dataset of Atrial Fibrillation Monitoring dataset is shown in Figure 7a. Mistakes made by models are almost understandable. To give just a few examples, many "Too Noisy To Classify" records are confused with "Other Rhythm" records which makes sense given that it can be subtle to detect "Other Rhythm" records, especially when their ECG morphologies are similar or when noise is present. In addition, it makes sense to confuse "Other Rhythm" records and "Normal Sinus Rhythm" ones, as it is difficult to distinguish among them sometimes, even for cardiologists.

Table 2. Comparing our method with other deep learning methods on the Atrial Fibrillation Monitoring dataset.

Method	F_{1N}	F_{1A}	F_{1O}	F_{1P}	F_1
CNN	0.9123	0.8140	0.7097	0.4636	0.7249
RNN	0.9187	0.8100	0.7834	0.4829	0.7488
RCNN	0.9190	0.8221	0.7319	0.5676	0.7602
ResNet	0.9056	0.8431	0.7415	0.5873	0.7719
KEEN	0.9061	0.8712	0.7166	0.7561	0.8125
KEEN+	0.9053	0.8704	0.7142	0.6995	0.7974



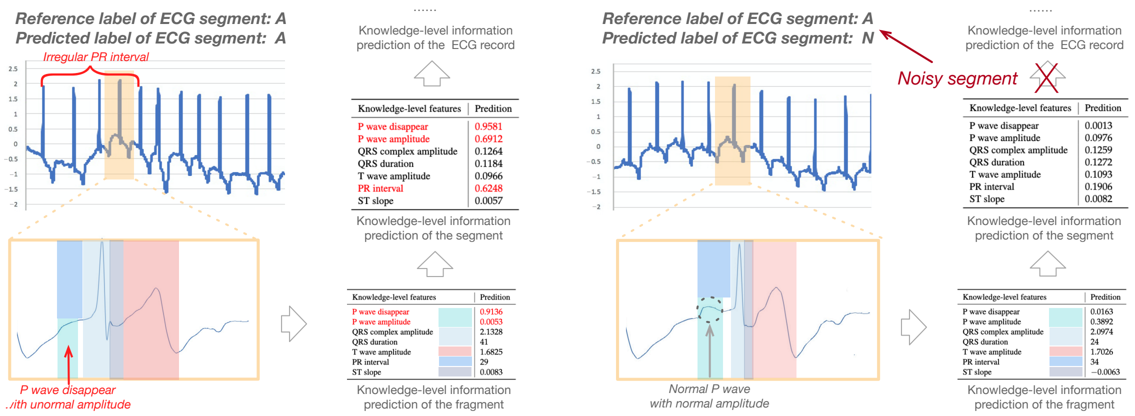
(a) On the Atrial Fibrillation Monitoring dataset.

(b) On the Aircraft Monitoring dataset.

Figure 7. Confusion matrix of KEEN.

4.1.2. Interpretation Results of Atrial Fibrillation Diagnosis

To show the interpretation results, we first evaluate the average error of all the knowledge-level attributes used in the experiments in Table 3. We can see that the average errors of all knowledge-level features are very small. Apart from quantitative analysis, we list two interpretation examples in Figure 8 to better understand how an interpretation model works. Each example contains one segment of the ECG record shown in Figure 1; the knowledge-level features generated by the KEEN+, thereafter, output them for the fragments, segments, and the ECG record (i.e., the final output of knowledge-level features are generated by aggregation operations to combine all the knowledge-level features of the segments which belong to the same ECG records into one unified feature vector).



Reference label of ECG segment: A
Predicted label of ECG segment: A
Irregular PR interval
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
P wave disappear
.ith unormal amplitude
......
Knowledge-level information prediction of the ECG record
Knowledge-level information prediction of the segment
Knowledge-level information prediction of the fragment
Reference label of ECG segment: A
Predicted label of ECG segment: N
Noisy segment
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
Normal P wave
with normal amplitude
......
Knowledge-level information prediction of the ECG record
Knowledge-level information prediction of the segment
Knowledge-level information prediction of the fragment

Knowledge-level features	Predition
P wave disappear	0.9581
P wave amplitude	0.6912
QRS complex amplitude	0.1264
QRS duration	0.1184
T wave amplitude	0.0966
PR interval	0.6248
ST slope	0.0057

Knowledge-level features		Predition
P wave disappear	🟩	0.9136
P wave amplitude	🟩	0.0053
QRS complex amplitude	⬜	2.1328
QRS duration	⬜	41
T wave amplitude	🟥	1.6825
PR interval	🟦	29
ST slope	▨	0.0083

Knowledge-level features	Predition
P wave disappear	0.0013
P wave amplitude	0.0976
QRS complex amplitude	0.1259
QRS duration	0.1272
T wave amplitude	0.1093
PR interval	0.1906
ST slope	0.0082

Knowledge-level features		Predition
P wave disappear	🟩	0.0163
P wave amplitude	🟩	0.3892
QRS complex amplitude	⬜	2.0974
QRS duration	⬜	24
T wave amplitude	🟥	1.7026
PR interval	🟦	34
ST slope	▨	−0.0063

(a) The first interpretation example.

(b) The second interpretation example.

Figure 8. Case studies of prediction interpretation on the Atrial Fibrillation Monitoring dataset.

Table 3. Error analysis of knowledge-level features on the Atrial Fibrillation Monitoring dataset.

Knowledge-Level Features	**Average Error**
P wave disappear	0.3918
P wave amplitude	0.0540
QRS complex amplitude	0.009
QRS duration	0.021
T wave amplitude	0.042
PR interval	0.196
ST slope	0.017

In the first example, the reference label and the predicted label of the first input ECG segment are both “A” as shown in Figure 8a. Moreover, the knowledge-level information predicted on both fragment and segment reveal characteristics of Atrial Fibrillation (e.g., high probability of P wave disappear, high value of standard deviation coefficient on P wave amplitude, and PR interval which reflect irregular changes of them between different fragments of an ECG segment, etc.). Consequently, the KEEN+ can automatically predict this important domain knowledge-level information, which could be provided as the domain technological decision-making basis, for the whole ECG record without the help of domain experts. Moreover, the KEEN+ could help cardiologists to accurately locate the ECG fragments when cardiac arrhythmia happens by analyzing the abnormal knowledge-level information (e.g., P wave disappear) of each fragment.

As another example, the reference label and the predicted label of the second input ECG segment are “A” and “N”, respectively. Some knowledge-level information (e.g., normal P wave with normal amplitude), which agree with the clinical symptoms of Normal Sinus Rhythm other than Atrial Fibrillation and indicate that the second input ECG segment is a noisy segment, can be obtained on the fragments and segment of the input ECG segment as shown in Figure 8b. As a matter of fact, these noisy segments should be avoided when using aggregation operations to generate the knowledge-level information for an ECG record, which is exactly the goal of this work. This problem is well addressed by the KEEN+ model, since aggregation operation would be only applied on the selected top-K most confident segments for the given ECG record, and therefore output a more consistent and accurate knowledge-level information.

4.2. Effect of Application in the Field of Aircraft Structural Health

We evaluate the effect of application in the field of aircraft structural health on the Aircraft Monitoring dataset, which is collected by the piezoelectric sensor networks on real aircraft parts.

4.2.1. Effectiveness of Aircraft Structural Damage Detection

The precision, recall, and F_1 results are shown in Figure 6b. We can see that the proposed KEEN and KEEN+ method with 0.8012 and 0.8005 F_1 scores outperform all of the most advanced deep learning methods for the task of aircraft structural damage detection by 2.99–10.71% and 2.99–10.62%, respectively.

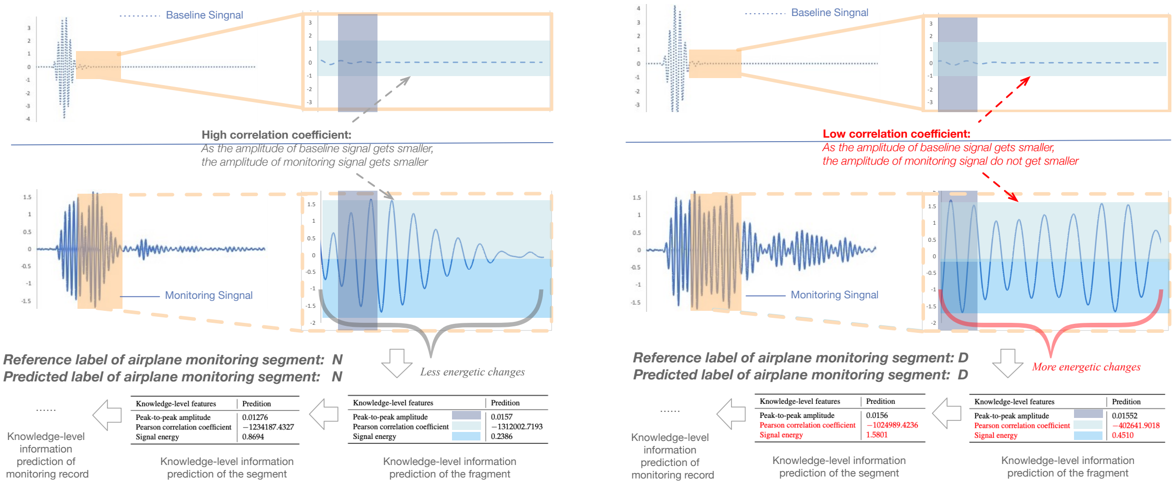
Figure 7b shows the confusion matrix of the model predictions on the test dataset of the Aircraft Monitoring dataset. Obviously, quite a few records of “Structural Damage” are confused with “Normal” ones, which is mainly because some invisible cracks to the human eye are really very difficult to distinguish from the normal ones in the aircraft structural health-condition monitoring records.

4.2.2. Interpretation Results of Aircraft Structural Damage Detection

The results of average error rates of the knowledge-level attributes used in the experiments are shown in Table 4, and we can see that the average errors of all knowledge-level features are very small as well. Similarly, we list two interpretation examples in Figure 9 to better understand how an interpretation model works in the task of aircraft structural damage detection. Each example contains one segment of an aircraft monitoring record, and each segment has two input channels—the baseline signal channel x_b and the monitoring signal channel x_m.

Table 4. Error analysis of knowledge-level features on the Aircraft Monitoring dataset.

Knowledge-Level Features	Average Error
Peak-to-peak amplitude	0.0124
Pearson correlation coefficient	0.1397
Signal energy	0.0024



Knowledge-level features	Predition
Peak-to-peak amplitude	0.01276
Pearson correlation coefficient	−1234187.4327
Signal energy	0.8694

Knowledge-level features		Predition
Peak-to-peak amplitude	▨	0.0157
Pearson correlation coefficient	▨	−1312002.7193
Signal energy	▨	0.2386

(a) The first interpretation example.

Knowledge-level features	Predition
Peak-to-peak amplitude	0.0156
Pearson correlation coefficient	−1024989.4236
Signal energy	1.5801

Knowledge-level features		Predition
Peak-to-peak amplitude	▨	0.01552
Pearson correlation coefficient	▨	−402641.9018
Signal energy	▨	0.4510

(b) The second interpretation example.

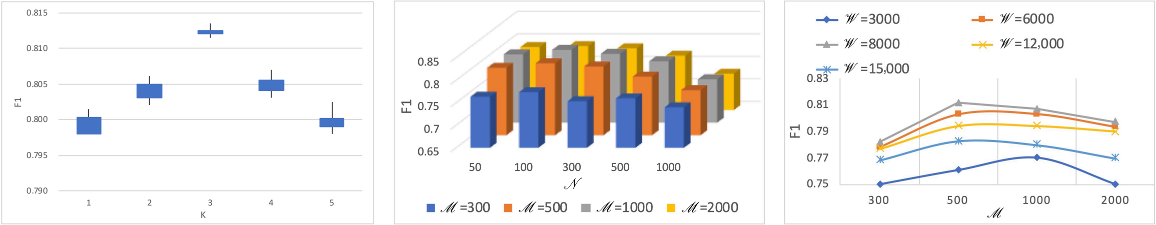
Figure 9. Case studies of prediction interpretation on the Aircraft Monitoring dataset.

The reference label and predicted label of the input aircraft monitoring segment shown in the first example are both "N", while the reference label and predicted label are both "D" in the second example. Compared with the second example as shown in Figure 9b, the amplitude of monitoring signal in the first example (as shown in Figure 9a) decreases significantly, which is much more similar to the baseline signal, while the second one shows an even stronger amplitude variation which reflects higher signal energetic changes. The results indicate that the second example has a much lower Pearson correlation coefficient (measuring the correlation of amplitude variations between baseline signal and the monitoring signal) and higher signal energy (measuring the changes of energy), which meets the diagnostic criteria in aircraft structural damage.

4.3. Analysis of the Influence of Hyper-Parameters

To evaluate the effect of different hyper-parameter settings on the performance of our method, four main hyper-parameters (threshold K, window size $\mathscr{W}$, maximum stride threshold $\mathscr{M}$, and length of each fragment $\mathscr{N}$) are further evaluated.

Figure 10a demonstrates the effects of threshold K. The F_1 score increase when the threshold K varies within the range of $[1, 3]$, yet they drop when the threshold K is larger than 3. The primary cause is that the accuracy of the proposed model depends on both the number of segments engaged in the algorithm and the confidence of their labels. The threshold K not only determines how many segments of each record participate in the algorithm but also the confidence of each segment involved in the training process. That is, the higher the reasonable K value, the lower the confidence of predictions (as we choose K segments from the most confident one to the less confident one). However, a much higher K value may lead to a decrease in the confidence of segments participating in the learning and voting phase, and degrade the performance of the proposed model as well.



(a) Varying threshold K.

(b) Varying maximum stride threshold $\mathscr{M}$ and length of each fragment $\mathscr{N}$.

(c) Varying windows size $\mathscr{W}$ and maximum stride threshold $\mathscr{M}$.

Figure 10. F_1 score results of KEEN for for different hyper-parameters.

Furthermore, the effects of two main hyper-parameters (maximum stride threshold $\mathscr{M}$ and length of each fragment $\mathscr{N}$) are shown in the Figure 10b. When parameters of $\mathscr{W}$ and $\mathscr{N}$ vary within the range of $[500, 1000]$and $[300, 2000]$ respectively, smaller fragments $\mathscr{N}$ may benefit the overall performance. This is probably because it could provide more fragments for a recurrent neural network layer to capture trend-level features. However, more critical local morphological-level features might be lost in a 'too small' fragment.

In addition, when the hyper-parameters of windows size $\mathscr{W}$ and maximum stride threshold $\mathscr{M}$ change within the range of $[3000, 8000]$and $[300, 2000]$, respectively, as shown on the Figure 10c, usually the smaller the parameters, the larger the augmented training data, and the better the performance of the proposed model. Nevertheless, if a 'too small' parameter of windows size $\mathscr{W}$ which could only provide a very narrow view for extracting local morphological-level features, the model performance might be degraded too.

4.4. Discussion

The K-margin-based intErpretable lEarNing (KEEN) is a general framework for health-condition assessment, which aims to provide interpretable health-condition detection results on a dataset with class skewness and noisy segments. The advantages of the proposed method are fourfold:

- It can automatically adapt to the dataset with class skewness (e.g., the Atrial Fibrillation Monitoring dataset) and improve the model performance on the dataset without class skewness as well by improving the inadequacy of well-labeled data (e.g., the Aircraft Monitoring Dataset) using the skewness-aware RCR-Net model.
- No extra process of signal noise elimination is needed to help the prediction of health conditions from monitoring data that contains noisy segments (both segments with too many signal noises and segments with noisy labels).
- Low-quality noisy records can be reliably detected, as our method would not cause the problem of changes or losses of features representing the quality of the monitoring signal itself.
- Important domain knowledge features can be extracted without the help of human experts from the time-series data automatically. As a result, KEEN+ can provide an interpretable decision-making basis or technological proof by using these domain knowledge-level features.

However, it has two disadvantages:

- It needs extra computation when filtering the noisy segments. Nevertheless, the computational cost, which is $\lceil \frac{\mathscr{W}}{\mathscr{N}} \rceil \times K$, is very low compared with the computational cost of DNN methods. For example, for the experimental settings mentioned in Section 3, the extra computational cost of each monitoring record is $\lceil \frac{6000}{300} \rceil \times 3$=60.
- The model performance could be influenced by the inaccuracy of the knowledge-level characteristics extracted from the dataset with high noise contamination. Comparing with the similar F_1 scores of KENN and KEEN+ on the Aircraft Monitoring Dataset, the accuracy of the KEEN+ method is slightly lower than that of the KEEN method on the Atrial Fibrillation

Monitoring Dataset. This may be mainly because the noise contamination of the Atrial Fibrillation Monitoring Dataset is much higher, and thus the accuracy of the extracted knowledge-level characteristics is lower. From a performance perspective, the knowledge-level interpretation method does lower the model performance to some extent on the dataset with high noise contamination, but, considering that it can make results more explicable, this performance degradation is acceptable.

5. Conclusions

In this paper, we propose a KEEN model for health-condition assessment which can automatically handle the problems of class skewness and noisy segments, and provide knowledge-level interpretations of the predicted results that can be used as a decision-making basis as well. Through experimental validation in the field of both medical and aerospace, the proposed method has a better generality and effectiveness comparing with the state-of-the-art deep learning methods for health-condition assessment tasks. Moreover, the following conclusions can be drawn from the test scene:

- The proposed methods KEEN and KEEN+ with 0.8125 and 0.7974 F_1 scores, which outperform all state-of-the-art DNN methods for health-condition assessment tasks by 3.30% and 5.71% on the Atrial Fibrillation Monitoring dataset with class skewness and noisy segments, indicate that the proposed methods can automatically adapt to the dataset with class skewness and noisy segments.
- The KEEN (with 0.8012 F_1 scores) and KEEN+ (with 0.8005 F_1 scores) are able to perform 2.99% better by improving the inadequacy of well-labeled data and filtering the noisy segments on the Aircraft Monitoring Dataset without class skewness.
- KEEN and KEEN+ show good performance with 0.7561 and 0.6995 F_1 scores for detecting the low-quality records (“Too Noisy To Classify”) which outperform other DNN methods by 28.74% and 19.10% on the Atrial Fibrillation Monitoring dataset.
- The model interpretation shows improvement in the results compared to other DNN methods, as KEEN+ can provide domain knowledge-level features as technological proofs for the results of health-condition assessment.

In future work, it will be important to investigate how the proposed approach could be performed online (e.g., taking the skewness drift problem into account, how to rapidly update the model with a new data stream, etc.). Moreover, another possible rewarding avenue of future research is to consider multi-modality data input and more fine-grained output categories to improve the model performance and apply it in a more practical situation.

Author Contributions: Conceptualization, Y.Z.; methodology, Y.Z.; software, Y.Z.; validation, Y.Z.; formal analysis, Y.Z.; investigation, Y.Z.; resources, Y.Z.; writing—original draft preparation, Y.Z.; visualization, Y.Z.; data curation, M.W. and Q.W.; writing—review and editing, S.H. and J.S.; supervision, H.L. and J.X.; project administration, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key Research and Development Program of China (No. 2020YFB2103402).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DNN	Deep Neural Network
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
CRNN	Convolutional Recurrent Neural Network

ResNet	Residual Neural Networks
Bi-LSTM	Bi-Directional Long-Short Term Memory
RCR-Net	Residual-Convolution-Recurrent Neural Network
K-margin	K-Margin-Based Health-Condition Label Prediction Algorithm
KEEN	K-Margin-Based Interpretable Learning
MSE	Mean Square Error Loss
ECG	Electrocardiograph
AF	Atrial Fibrillation

References

1. Tateno, K.; Glass, L. Automatic detection of atrial fibrillation using the coefficient of variation and density histograms of RR and ΔRR intervals. *Med. Biol. Eng. Comput.* **2001**, *39*, 664–671. [CrossRef] [PubMed]
2. Faust, O.; Acharya, U.R.; Min, L.C.; Sputh, B.H.C. Automatic identification of epileptic and background EEG signals using frequency domain parameters. *Int. J. Neural Syst.* **2010**, *20*, 159–176. [CrossRef] [PubMed]
3. Faust, O.; Acharya, U.R.; Molinari, F.; Chattopadhyay, S.; Tamura, T. Linear and nonlinear analysis of cardiac health in diabetic subjects. *Biomed. Signal Process. Control* **2012**, *7*, 295–302. [CrossRef]
4. Linker, D. Accurate, automated detection of atrial fibrillation in ambulatory recordings. *Cardiovasc. Eng. Technol.* **2016**, *7*, 182–189. [CrossRef]
5. Hong, S.; Zhou1, Y.; Wu, M.; Shang, J.; Wang, Q.; Li1, H.; Xie, J. Combining deep neural networks and engineered features for cardiac arrhythmia detection from ECG recordings. *Physiol. Meas.* **2019**, *40*, 054009. [CrossRef]
6. Ichikawa, K.; Tamano, H. Unsupervised Qualitative Scoring for Binary Item Features. *Data Sci. Eng.* **2020**, *5*, 317–330. [CrossRef]
7. Kumar, R.; Bishnu, P.S. Identification of k-Most Promising Features to Set Blue Ocean Strategy in Decision Making. *Data Sci. Eng.* **2019**, *4*, 367–384. [CrossRef]
8. Ieracitano, C.; Mammone, N.; Hussain, A.; Morabito, F.C. A Convolutional Neural Network based self-learning approach for classifying neurodegenerative states from EEG signals in dementia. In Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN)), Shenzheng, China, 18–22 July 2020; pp. 1–8.
9. Bornn, L.; Farrar, C.R.; Park, G.; Farinholt, K. Structural health monitoring with autoregressive support vector machines. *J. Vib. Acoust.* **2009**, *131*, 021004. [CrossRef]
10. Carrara, M.; Carozzi, L.; Moss, T.J.; de Pasquale, M.; Cerutti, S.; Ferrario, M.; Lake, D.E.; Moorman, J.R. Heart rate dynamics distinguish among atrial fibrillation, normal sinus rhythm and sinus rhythm with frequent ectopy. *Physiol. Meas.* **2015**, *36*, 1873. [CrossRef] [PubMed]
11. Ni, Y.Q.; Wang, Y.W.; Zhang, C. A Bayesian approach for condition assessment and damage alarm of bridge expansion joints using long-term structural health monitoring data. *Eng. Struct.* **2020**, *212*, 110520. [CrossRef]
12. Kiranyaz, S.; Ince, T.; Gabbouj, M. Real-time patient-specific ECG classification by 1D convolutional neural networks. *IEEE Trans. Biomed. Eng.* **2015**, *63*, 664–675. [CrossRef] [PubMed]
13. Kamaleswaran, R.; Mahajan, R.; Akbilgic, O. A robust deep convolutional neural network for the classification of abnormal cardiac rhythm using single lead electrocardiograms of variable length. *Physiol. Meas.* **2018**, *39*, 035006. [CrossRef] [PubMed]
14. Hu, B.; Hu, Z.; Ran, L.; Mawby, P.; Jia, C.; Ng, C.; McKeever, P. Deep Learning Neural Networks for Heat-Flux Health Condition Monitoring Method of Multi-Device Power Electronics System. In Proceedings of the IEEE Energy Conversion Congress and Exposition (ECCE), Baltimore, MD, USA, 29 September–3 October 2019; pp. 3769–3774.
15. San, K.T.; Sohn, S.Y. Multitask learning for health condition identification and remaining useful life prediction: Deep convolutional neural network approach. *J. Intell. Manuf.* **2020**, 1–11. [CrossRef]
16. Hong, S.; Xu, Y.; Khare, A.; Priambada, S.; Maher, K.O.; Aljiffry, A.; Sun, J.; Tumanov, A. HOLMES: Health Online Model Ensemble Serving for Deep Learning Models in Intensive Care Units. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual Event, Los Angeles, CA, USA, 23–27 August 2020; pp. 1614–1624.

17. Schwab, P.; Scebba, G.C.; Zhang, J.; Delai, M.; Karlen, W. Beat by beat: Classifying cardiac arrhythmias with recurrent neural networks. In Proceedings of the IEEE Computing in Cardiology (CinC), Nice, France, 24–27 September 2017; pp. 1–4.
18. Faust, O.; Shenfield, A.; Kareem, M.; San, T.R.; Fujita, H.; Acharya, U.R. Automated detection of atrial fibrillation using long short-term memory network with RR interval signals. *Comput. Biol. Med.* **2018**, *102*, 327–335. [CrossRef]
19. Warrick, P.A.; Homsi, M.N. Ensembling convolutional and long short-term memory networks for electrocardiogram arrhythmia detection. *Physiol. Meas.* **2018**, *39*, 114002. [CrossRef]
20. Faust, O.; Kareem, M.; Shenfield, A.; Ali, A.; Acharya, U.R. Validating the robustness of an internet of things based atrial fibrillation detection system. *Pattern Recognit. Lett.* **2020**, *133*, 55–61. [CrossRef]
21. Liang, K.; Qin, N.; Huang, D.; Fu, Y. Convolutional recurrent neural network for fault diagnosis of high-speed train bogie. *Complexity* **2018**, *2018*, 4501952. [CrossRef]
22. Liu, W.; Wang, F.; Huang, Q.; Chang, S.; Wang, H.; He, J. MFB-CBRNN: A hybrid network for MI detection using 12-lead ECGs. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 503–514. [CrossRef]
23. Hong, S.; Wang, C.; Fu, Z. Gated Temporal Convolutional Neural Network and Expert Features for Diagnosing and Explaining Physiological Time Series: A Case Study on Heart Rates. *Comput. Methods Programs Biomed.* **2020**, 105847. [CrossRef]
24. Hannun, A.Y.; Rajpurkar, P.; Haghpanahi, M.; Tison, G.H.; Bourn, C.; Turakhia, M.P.; Ng, A.Y. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. Med.* **2019**, *25*, 65. [CrossRef]
25. Shen, Y.; Voisin, M.; Aliamiri, A.; Avati, A.; Hannun, A.; Ng, A. Ambulatory atrial fibrillation monitoring using wearable photoplethysmography with deep learning. In Proceedings of the 25th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Anchorage, AK, USA, 4–8 August 2019; pp. 1909–1916.
26. Ieracitano, C.; Mammone, N.; Hussain, A.; Morabito, F.C. A novel multi-modal machine learning based approach for automatic classification of EEG recordings in dementia. *Neural Netw.* **2020**, *123*, 176–190. [CrossRef] [PubMed]
27. Oh, B.K.; Glisic, B.; Kim, Y.; Park, H.S. Convolutional neural network-based wind-induced response estimation model for tall buildings. *Comput. Aided Civ. Infrastruct. Eng.* **2019**, *34*, 843–858. [CrossRef]
28. Lih, O.S.; Jahmunah, V.; San, T.R.; Ciaccio, E.J.; Yamakawa, T.; Tanabe, M.; Kobayashi, M.; Faust, O.; Acharya, U.R. Comprehensive electrocardiographic diagnosis based on deep learning. *Artif. Intell. Med.* **2020**, *103*, 101789. [CrossRef] [PubMed]
29. Zhou, Y.; Hong, S.; Shang, J.; Wu, M.; Wang, Q.; Li, H.; Xie, J. K-margin-based residual-convolution-recurrent neural network for atrial fibrillation detection. In Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI), Macao, China, 10–16 August 2019; pp. 6057–6063.
30. Sun, C.; Hong, S.; Song, M.; Li, H. A Review of Deep Learning Methods for Irregularly Sampled Medical Time Series Data. *arXiv* **2020**, arXiv:2010.12493.
31. Hong, S.; Zhou, Y.; Shang, J.; Xiao, C.; Sun, J. Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review. *Comput. Biol. Med.* **2020**, *122*, 103801. [CrossRef] [PubMed]
32. Branco, P.; Torgo, L.; Ribeiro, R.P. A survey of predictive modeling on imbalanced domains. *ACM Comput. Surv. (CSUR)* **2016**, *49*, 1–50. [CrossRef]
33. Lu, Y.; Cheung, Y.; Tang, Y.Y. Dynamic Weighted Majority for Incremental Learning of Imbalanced Data Streams with Concept Drift. In Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI), Melbourne, Australia, 19–25 August 2017; pp. 2393–2399.
34. Liu, S.-H.; Hsieh, C.-H.; Chen, W.; Tan, T.-H. ECG noise cancellation based on grey spectral noise estimation. *Sensors* **2019**, *19*, 798. [CrossRef]
35. Zhao, Z.; Särkkä, S.; Rad, A.B. Kalman-based Spectro-Temporal ECG Analysis using Deep Convolutional Networks for Atrial Fibrillation Detection. *J. Signal Process. Syst.* **2020**, *92*, 621–636. [CrossRef]
36. Reljin, N.; Lazaro, J.; Hossain, M.B.; Noh, Y.S.; Cho, C.H.; Chon, K.H. Using the Redundant Convolutional Encoder-Decoder to Denoise QRS Complexes in ECG Signals Recorded with an Armband Wearable Device. *Sensors* **2020**, *20*, 4611. [CrossRef]

37. Xu, Y.; Biswal, S.; Deshpande, S.R.; Maher, K.O.; Sun, J. Raim: Recurrent attentive and intensive model of multimodal patient monitoring data. In Proceedings of the 24th ACM International Conference on Knowledge Discovery & Data Mining (SIGKDD), London, UK, 19–23 August 2018; pp. 2565–2573.
38. Hong, S.; Xiao, C.; Ma, T.; Li, H.; Sun, J. Mina: Multilevel knowledge-guided attention for modeling electrocardiography signals. In Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI), Macao, China, 10–16 August 2019; pp. 5888–5894.
39. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
40. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 4765–4774.
41. Clifford, G.D.; Liu, C.; Moody, B.; Lehman, L.H.; Silva, I.; Li, Q.; Johnson, A.E.; Mark, R.G. AF Classification from a short single lead ECG recording: The PhysioNet/Computing in Cardiology Challenge 2017. In Proceedings of the IEEE Computing in Cardiology (CinC), Nice, France, 24–27 September 2017; pp. 1–4.
42. Yang, Y.; Zhou, Y.; Wang, L. Integrated Method of Multiple Machine-Learning Models for Damage Recognition. *J. Data Acquis. Process.* **2020**, *35*, 278–287.
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
44. Saeedan, F.; Weber, N.; Goesele, M.; Roth, S. Detail-preserving pooling in deep networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 9108–9116.
45. Goldberger, A.L.; Amaral, L.A.N.; Glass, L.; Hausdorff, J.M.; Ivanov, P.C.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.-K.; Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **2000**, *101*, e215–e220. [CrossRef] [PubMed]
46. Betz1, D.C.; Staszewski1, W.J.; Thursby, G.; Culshaw, B. Structural damage identification using multifunctional Bragg grating sensors: II. Damage detection results and analysis. *Smart Mater. Struct.* **2006**, *15*, 1313. [CrossRef]
47. Michaels, J.E.; Michaels, T.E. An integrated strategy for detection and imaging of damage using a spatially distributed array of piezoelectric sensors. In Proceedings of the Health Monitoring of Structural and Biological Systems, San Diego, CA, USA, 18–22 March 2007.
48. Gao, H.; Shi, Y.; Ros, J.L. Guided wave tomography on an aircraft wing with leave in place sensors. In Proceedings of the AIP Conference, Paris, France, 18–22 July 2005; Volume 760, pp. 1788–1794.
49. Sodmann, P.; Vollmer, M.; Nath, N.; Kaderali, L. A convolutional neural network for ECG annotation as the basis for classification of cardiac rhythms. *Physiol. Meas.* **2018**, *39*, 104005. [CrossRef] [PubMed]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.