Validating the robustness of an internet of things based atrial fibrillation detection system

FAUST, Oliver, KAREEM, Murtadha, SHENFIELD, Alex, ALI, Ali and RAJENDRA ACHARYA, U

Available from Sheffield Hallam University Research Archive (SHURA) at:
http://shura.shu.ac.uk/25812/

This document is the author deposited version. You are advised to consult the publisher's version if you wish to cite from it.

Published version


Copyright and re-use policy

See http://shura.shu.ac.uk/information.html
Highlights

- We validated a mature LSTM deep learning model with more and more varied data to mimic the medical diagnosis process.

- The validation data was distinct from the training data: it was measured from different patients.

- The model was created with the data from only 20 patients and we have validated it with data from 82 patients.

- Under these difficult circumstances, our LSTM based deep learning model achieved an accuracy of 94%.

- The classification performance indicates that the mature LSTM based deep learning model is fit for practical service.
Validating the robustness of an internet of things based atrial fibrillation detection system

Oliver Faust\textsuperscript{a}, Murtadha Kareem\textsuperscript{a}, Alex Shenfield\textsuperscript{a}, Ali Ali\textsuperscript{b}, U Rajendra Acharya\textsuperscript{c,d}

\textsuperscript{a}Department of Engineering and Mathematics, Sheffield Hallam University, United Kingdom
\textsuperscript{b}Sheffield Teaching Hospitals NHS Foundation Trust, University of Sheffield, United Kingdom
\textsuperscript{c}Ngee Ann Polytechnic, Singapore
\textsuperscript{d}Department of Bioinformatics and Medical Engineering, Asia University, Taiwan

ABSTRACT

This paper describes the validation of a deep learning model for Internet of Things (IoT) based health care applications. As such, the deep learning model was created to detect episodes of Atrial Fibrillation (AF) using Heart Rate (HR) signals. The initial Long Short-Term Memory (LSTM) model was developed using 20 data sets, from distinct subjects, obtained from the AFDB database on PhysioNet. This model achieved an AF detection accuracy of 98.51% with ten fold cross validation. In this study, we validated the initial results by testing the developed deep learning model with unknown data. To be specific, we fed the data from 82 subjects to the deep learning system and compared the classification results with the diagnosis results indicated by human practitioners. The validation results show 94% accuracy with an area under the Receiver Operating Characteristic (ROC) curve of 96.58. These results indicate that the LSTM model is able to extract the feature maps from the unknown data and hence detect the AF periods accurately. With this blindfold validation testing we violated a well known design rule for learning systems which states that more data should be used for training than for testing. By doing so, we have established that our deep learning system is fit for practical deployment, because in a practical situation the diagnosis support system must apply the knowledge, extracted from a limited training data set, to a HR trace from a patient.

© 2020 Elsevier Ltd. All rights reserved.

1. Introduction

AF is the most common serious irregular heart rhythm associated with rapid heart rate in adults (Benjamin et al., 1998). Atrial means to the atria (plural of atrium), which describes the locations at the top two chambers of the heart (Kenny, 2018; Swapna et al., 2018). Fibrillation refers to irregular, rapid and unsynchronised contraction of muscle fibers. Sinus rhythm represents the normal beat of the heart, which is managed by a sophisticated electrical control system. This system controls the timing of the heart pump. When the electrical system is functioning properly, it maintains a normal heart rate and rhythm (Martis et al., 2013b). Problems with this electrical system can cause two types of arrhythmia; tachycardia and bradycardia when the heart beats too fast or too slow respectively (Faust et al., 2013). The sinus node consists of a cluster of special cells, which act as the heart’s natural pacemaker. In AF, disordered electrical activity progresses in the walls of the atria, exceeding the sinus node. As a result, the atria begins to fibrillate and the rhythm will change from normal to abnormal. That leads to a rapid rhythm as their muscular walls fail to contract with coordination and regularity. To be accurate, 0.4% of adults are affected by this disease, and that prevalence increases with age (Ali et al., 2012). Less than 1% of people, aged of 60 or younger, are affected by AF and in surplus of 6% for those aged 80 and older (Adamson et al., 2004). It is anticipated that the occurrence of AF increases in accordance with the aging population. In addition, AF incidence is related to a significant increase in stroke, heart failure, poor mental health, diminished life quality, and as such it is a leading cause of death (Stewart et al., 2002). This disease is associated with various types of symptoms, such as chest pain, shortness of breath, fainting,
fatigue, palpitation and light-headedness (Thrall et al., 2007). The fact that AF is an independent risk factor for stroke has significant clinical and economic implications (Ali et al., 2015). Overall, treatment of AF often includes cardioversion (restoring sinus rhythm) which may involve using medication or electrical cardioversion.

Before we can treat this serious disease, it is necessary to diagnose it first. One way of diagnosing AF is to record HR signals and subsequently look for signs of AF in these signals. The most promising approaches, documented in the scientific literature, use computer support for that detection task (Hagibara et al., 2018; Song et al., 2012). The Computer-Aided Diagnosis (CAD) systems are trained and tested on labeled data, i.e. data that was analyzed by human cardiologists. Looking at the scientific literature in more detail revealed that all studies stopped at a one-time evaluation of the learning model (Chen et al., 2017; Gotlibovych et al., 2018; Xia et al., 2018; Yildirim et al., 2018; Yuan et al., 2016). That means these studies did not test the efficacy of their model with more and more varied HR data. Unfortunately, using the model with more and more varied data is the use case scenario which arises in clinical practice. To be specific, in a practical setting the patient data represents a sample from the continuum of all possible data. That continuum is better represented with more and more varied data. Hence, testing a model with more and more varied data instills trust that the model will work in a practical scenario. The absence of such blindfold validation is a research gap, because the lack of follow up verification results in a diminished trust in the CAD systems. This lack of trust is one of the reasons why we do not see the wide spread application of CAD systems in clinical practice.

The current study tried to overcome the above-mentioned problem by evaluating the performance of a mature deep learning model with unseen data. We have asked the model to detect AF in signals from 82 long-term HR recordings of subjects with paroxysmal or sustained AF. The system achieved a classification accuracy of 94%, a precision of 95% and it had 95% recall. This result is significant as it is obtained using totally unknown data. As such, the model was established with a training data set that came from 20 patients and now we have validated it with data from 82 patients. Doing so goes against the common concept of using a smaller portion of the data for testing. Hence, it is more difficult for the deep learning model to establish the correct result. Furthermore, the training data was measured with a different setup from the validation data, i.e. sampling frequency, voltage range, and resolution were different. The fact that the results were achieved in that different environment indicates that the deep learning system knows what an AF affected HR signal looks like. Hence, these results establish that the deep learning model is reliable to detect AF in clinical settings. Having that impetus is a prerequisite for Intelligent Internet of Things (IoIoT) based diagnosis support tools.

To communicate our findings, we have structured the remainder of the paper as follows. The next section provides background on IoIoT and the deep learning model. The subsequent section introduces the methods used for validating that model. Section 4 states the validation results in the form of confusion matrix and ROC curve. The subsequent discussion takes center stage in this paper, because in this section we outline why a robust decision-making process is a necessary prerequisite for IoIoT based AF monitoring. The Discussion section also provides limitations and future work, before we wrap up the paper with the conclusions.

2. Background on intelligent Internet of medical things

IoIoTs deliver measurement data to a central storage location for centralized decision-making (Gubbi et al., 2013). In the medical domain, such measurement data are usually physiological signals, such as HR signals (Dimitrov, 2016). Figure 1 depicts a use case scenario for a medical IoIoTs. Patients wear a HR sensor, which is depicted in the figure as a black oval with a red dot. That sensor communicates the HR signal to a mobile device which relays the data to the cloud server via an IoT protocol. The LSTM based deep learning model, which was validated in this study, analyzes the incoming HR signals in real-time. If AF is detected, a cardiologist is informed. The cardiologist can review the suspicious HR trace and reach a diagnosis. That diagnosis can be communicated to the patients in the form of a simple traffic light scheme, as indicated in the figure.

2.1. Long short term memory based deep learning

The use case scenario features LSTM based deep learning. As such, LSTM is an extension of Recurrent Neural Network (RNN) (LeCun et al., 2015). That algorithm was chosen because LSTM can process data sequences (Sak et al., 2014). This functionality is distinct from other deep learning approaches which consider an input vector as data point in a multidimensional space (Faust et al., 2018). LSTM includes the idea of time by incorporating memory cells and a sophisticated updating algorithm (Oh et al., 2018). The updating algorithm uses three regulators or gates which control the flow of information within the LSTM. The input gate determines how much new data flows into the cell. The forget gate controls how much information is retained and the output gate controls the outflow of information. That flexible use of memory enables LSTM to approach the utility of a "general purpose computer" (Siegelmann and Sontag, 1995).

Figure 2 shows a functional diagram of the LSTM algorithm. The upper part of the diagram indicates the RNN loop unrolling, which results in individual LSTM cells. The hidden state vector \( h_t \in \mathbb{R}^h \) and the cell state vector \( c_t \in \mathbb{R}^h \) are passed from one cell to the next. The cells consume the input vector \( \tilde{x}_t \) at different time instances \( t \). Each cell A has LSTM functionality, as indicated in the lower part of the figure.

Each cell incorporates the three gates to establish the LSTM functionality (Gers et al., 2002). The forget gate regulates the information content stored within the cell and thereby it plays a vital role in modelling the way humans remember and indeed forget (Gers et al., 2000). It is implemented as the first multiplier from the left, highlighted in orange. The vector \( \tilde{f}_t \in \mathbb{R}^h \) controls the forget gate. Mathematically, that vector is established with:

\[
\tilde{f}_t = \sigma(\mathbf{W}_f \tilde{x}_t + \mathbf{U}_f h_{t-1} + \mathbf{b}_f)
\]
where $σ(...)$ is the sigmoid activation function. $\mathbf{W}_i \in \mathbb{R}^{h \times d}$ and $\mathbf{U}_f \in \mathbb{R}^{h \times d}$ are weight matrices and $\mathbf{b}_f \in \mathbb{R}^h$ is a forget specific bias vector. The parameters $h$ and $d$ indicate the number of hidden units and the dimensionality of the input vector respectively.

The input gate is implemented as the second multiplier from the left, highlighted in blue. The vector $\mathbf{i}_t \in \mathbb{R}^h$ controls the output gate. The mathematical definition of $\mathbf{i}_t$ is similar to $f_t$:

$$\mathbf{i}_t = σ(\mathbf{W}_i \mathbf{x}_t + \mathbf{U}_i \mathbf{h}_{t-1} + \mathbf{b}_i)$$

(2)

the weight matrices $\mathbf{W}_i \in \mathbb{R}^{h \times d}$ and $\mathbf{U}_i \in \mathbb{R}^{h \times d}$ as well as the bias vector $\mathbf{b}_i \in \mathbb{R}^h$ are different.

The output gate is implemented as the third multiplier from the left, highlighted in green. The vector $\mathbf{f}_t \in \mathbb{R}^h$ controls the input gate. The mathematical definition of $\mathbf{f}_t$ is similar to the outer two gates:

$$\mathbf{f}_t = σ(\mathbf{W}_f \mathbf{x}_t + \mathbf{U}_f \mathbf{h}_{t-1} + \mathbf{b}_f)$$

(3)

the weight matrices $\mathbf{W}_f \in \mathbb{R}^{h \times d}$ and $\mathbf{U}_f \in \mathbb{R}^{h \times d}$ as well as the bias vector $\mathbf{b}_f \in \mathbb{R}^h$ are different.

Having established the control vectors for the gates, we can progress to formulate the equations which define the cell output. The cell state vector for the current time is established with the following equation:

$$\mathbf{c}_t = f_t \circ \mathbf{c}_{t-1} + i_t \circ \text{Tanh}(\mathbf{W}_c \mathbf{x}_t + \mathbf{U}_c \mathbf{h}_{t-1} + \mathbf{b}_c)$$

(4)

where $\circ$ indicates element-wise multiplication and Tanh(...) is the hyperbolic tangent function. Finally, the current hidden state vector is established with:

$$\mathbf{h}_t = f_t \circ \text{Tanh}(\mathbf{c}_t)$$

(5)

The weights and biases are established during the training phase and they constitute the LSTM model. During the testing phase, the model is used to classify an input sequence $\mathbf{x}_t$. In our case, the model establishes if there are signs of AF in a HR sequence. The methods used for testing the LSTM model are introduced in the next section.

3. Methods

Recently, we developed an LSTM based deep learning model to detect the AF episodes using HR signals (Faust et al., 2018b). That model can provide the intelligence needed for state of the art IoT based diagnosis support systems. The current study setup tests the same model with more and more varied data obtained from a different source.

Figure 3 provides an overview of the study setup. The upper half of the figure shows the initial study setup. The LSTM based deep learning system was trained and tested with labeled HR signal data from 20 subjects sourced from PhysioNet’s Atrial Fibrillation Database (AFDB). Training the deep learning network means establishing the weight values of the individual neurons. The weight vectors from the initial study were used in the validation setup. However, more and more varied data were used for the blindfold validation testing. The following sections detail the validation setup by introducing the data sets and the processing steps in more detail.

3.1. Data used

The initial study was based on data collected from the MIT-BIH AFDB which is available on PhysioNet (Moody, 1983; Goldberger et al., 2000). This database includes twenty three 10 hour Electrocardiogram (ECG) recordings. These recordings were labelled with so called rhythm annotation files, the content of which was prepared manually. The specific types of rhythm annotations were (AFIB (atrial fibrillation), (AFL (atrial flutter), (J (AV junctional rhythm), and (N (used to indicate all...
other rhythms). Furthermore, the database contains beat annotation files that were used to extract the HR signals.

For the blindfold validation, we used the data from 82 subjects, which were sourced from the Long-Term AF Database (LTAFDB) (Petrutiu et al., 2007). All patients in that database were diagnosed with paroxysmal or sustained AF. Each dataset is composed from two ECG signals each of which has a duration of 24 to 25 hours and was sampled at 128 Hz. The datasets incorporate also rhythm and beat annotations, which were established manually by experienced cardiologists. Moreover, the R peaks are labelled, and the RR interval sequence was extracted based on these labels. Rhythm annotations were used to label the RR interval sequence as either AF or non-AF. Table 1 provides the measurement parameters that were used to acquire the data for the two individual databases. The fact that there is a significant difference in the sampling frequency indicates the measurement setups were quite different.

The pre-processing of the data in the validation setup follows closely the initial study. A sliding window was used to partition the HR signals into overlapping sequences of 100 beats. Each sequence of 100 beats was labeled as AF if one of the beats, within the sequence, was labeled AF, all other sequences were labeled non-AF, indicating normal or other cardiac disease. Data from 82 subjects were used with the blindfold validation strategy to evaluate the performance of the robust deep learning system. This means the proposed methods can be used to generalize not only unknown data, but also the unknown patients as well.

3.2. Performance measures

The confusion or error matrix summarizes the prediction results of a classification problem (Hay, 1988). The layout allows visualization of the performance of a decision-making algorithm. To be specific, the matrix has two rows and two columns that represent the classes of actual and predictive analysis. Therefore, these performance measures report the number of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Most medical test results refer to a positive case (classifying the subject having the disease) and a negative case (classifying the patient not having the disease).

The ROC curve is a method used to evaluate the diagnostic accuracy of tests in modern medicine (Mas, 2018). It is widely used to illustrate how well a diagnostic model can distinguish between the presence and absence of disease and works equally well with data sets that exhibit class imbalance. The ROC represents the TP rate (sensitivity) plotted against the FP rate (1-specificity) for various cut-off points (Powers, 2011). Each point on the ROC graph indicates the sensitivity/specificity corresponding to a specific decision threshold. The Area Under Curve (AUC) is a summary metric indicating the discriminatory power of a classifier.

3.3. Initial study setup and performance

For the initial study, we designed a deep RNN (Pascanu et al., 2013; Pearlmutter, 1989) with LSTM (Hochreiter and Schmidhuber, 1997; Bengio et al., 1994) model to detect AF based on HR traces. The pre-processed data blocks were fed directly into the model without performing any feature engineering (Faust, 2018). Table 2 provides the ten-fold cross validation results obtained for the initial model based on the data from 20 AFDB subjects.

4. Validation results

This section documents the blindfold validation results of the LSTM based deep learning model obtained in (Faust et al., 2018b). The results are based on HR signals from the 82 LTAFDB subjects. The radar plot, shown in Figure 4, provides a graphical representation of the model performance in terms of accuracy, precision, recall and F1 score. The labels, around the radar plot, correspond to the patient ID as mentioned in the
Table 1: Comparison between AFDB and LTAFDB.

<table>
<thead>
<tr>
<th></th>
<th>AFDB</th>
<th>LTAFDB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>10 h</td>
<td>Between 24 h and 25 h</td>
</tr>
<tr>
<td>ADC resolution</td>
<td>12-bit</td>
<td>12-bit</td>
</tr>
<tr>
<td>Voltage range</td>
<td>+10 mV</td>
<td>20 mV</td>
</tr>
<tr>
<td>Sampling frequency</td>
<td>250 Hz</td>
<td>128 Hz</td>
</tr>
<tr>
<td>Measurement location</td>
<td>Boston’s Beth Israel Hospital</td>
<td>Not reported</td>
</tr>
</tbody>
</table>

LTAFDB. The IDs were ordered in terms of accuracy. The signal with label 10, shown at 12 o’clock, has the highest accuracy. The accuracy performance of the model decreases clockwise, i.e. the accuracy was lower for signal 12 than for signal 10 etc. The model achieved the lowest accuracy of just under 70% for the HR signal from patient 22.

We have evaluated the model performance with a confusion matrix in order to validate AF and normal HR sequences obtained from 82 subjects. Figure 5 shows the results obtained using the new test data: (a) confusion matrix by blindfold validation, (b) ROC of the model. It can be noticed from the results in Table 3, that the LSTM deep learning model achieved a classification accuracy of 94% on the LTAFDB blindfold validation data set - classifying 95% of normal and 95% of HR sequences showing signs of AF correctly. The deep learning classifier achieves an AUC of 0.9658 which is close to a perfect score of 1, hence the classifier discriminates well between disease and no disease. Most classifier values varied between 0.0 (definitely negative) and 1.0 (definitely positive). The closer the ROC plot to the TP rate, the higher overall accuracy of the test performed. Figure 5b shows the ROC curve which indicates the AF diagnostic performance.

Table 2: Ten fold cross validation result established during the initial training and testing of the LSTM deep learning model.

<table>
<thead>
<tr>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>Accuracy</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>430,615</td>
<td>523,241</td>
<td>7,040</td>
<td>7,407</td>
<td>98.51%</td>
<td>0.9986</td>
</tr>
</tbody>
</table>

Blindfold validation should become a standard method to evaluate deep learning and other decision support algorithms.

The fact that the deep learning system was trained with HR signals is of particular importance for home health care. HR signals capture the beat-to-beat interval of the human heart (Acharya et al., 2013). The beat impulse is the most prominent signal feature when the electrical activity of the human heart is recorded. That prominence is reflected in the high signal-to-noise ratio. Hence, measuring the beat-to-beat interval is robust to noise. As a consequence, the measurement setup is simple, when compared to other physiological signals, such as ECG (Martis et al., 2013a). Another advantage of HR is that the time from one beat to the next can be encoded with a 2-byte integer. Hence, a digital HR signal consists of approximately 2 bytes a second. In contrast, to sample the complete electrical activity of the human heart requires around 256 samples a second, each of which is encoded with 2 bytes. Therefore, ECG signals have a 256 times higher data rate. The lower data rate has practical benefits, in terms of data communication, storage and processing. For example, HR signals can be communicated via Bluetooth Low Energy (BLE) whereas ECG signals require a broader wireless channel, such as WiFi. Using BLE instead of WiFi has beneficial implications for battery powered devices.

The low data rate and patient led signal acquisition makes HR signals a good choice for IoT based health care applications. For such applications the HR data travels from the point of measurement to a central cloud service over communication infrastructure. Having the data at a central location has a number of advantages. Deep learning can be used to detect AF in real time. That is a significant advantage over ECG measurements with Holter monitors, because Holter data can only be analyzed after the measurement period is complete. Validating the deep learning model for AF detection has paved the way for an IoT based AF diagnosis support tool. The knowledge extracted from a limited amount of labeled data can be used to provide real-time decision support. In order to reach the diagnosis, the deep learning result should be validated by a cardiologist. The basis for this validation can be the HR data which has been flagged as showing the subtle waveform alterations caused by AF.

A limitation of our study is that we did not address the important concept of training on the job. To be specific, a cardiologist learns while doing the job. The proposed deep learning model is static, i.e. it did not learn during the validation. At one point the knowledge, extracted from 20 patients, will be insufficient to cope with practical scenarios. In the future, we have to find a way to model that continuous training in order to improve the diagnostic quality of the proposed AF detection system. One
way of providing this continuous learning is to retrain the deep learning model with measurement data. A prerequisite for such a methodology is to have the HR data stored in a central location. Hence, streaming the HR data to a central cloud server might prove to be an advantage when the continuous learning problem is tackled.

6. Conclusion

The current study indicates that deep learning can acquire the knowledge to understand specific aspects of HR signals. In this case, our deep learning model understands HR signals, such that it can differentiate AF from non-AF affected signals. This is different from the traditional machine learning approach which is based on feature engineering. To be specific, the traditional approach differentiates AF from non-AF signals based on parameters. It is likely that these parameters change when more and more varied data are analyzed. Hence, it is common practice for studies based on traditional methods to state this as a limitation, i.e. more and more varied data is needed as well as retraining the machine learning model to improve the diagnostic quality.
In the current study, we have used more and more varied data without retraining the deep learning model. Data from 82 patients were used for this blindfold validation. With this data, our model achieved a classification accuracy of 94%, a precision of 95% and it had 95% recall. These results indicate that our deep learning model has fewer limitations when compared to traditional machine learning approaches. To be specific, we showed that knowledge extracted from a small training data set can be applied to a larger and more varied validation data set. This concept is significant for practical diagnostic support, because applying knowledge acquired during the limited training period is exactly what a cardiologist does. Hence, a practical IoT based medical decision support system must apply knowledge, extracted during training, to samples, i.e. patient data, from a very large data set.

**References**


Gubbi, J., Buyya, R., Marusic, S., Palaniswami, M., 2013. Internet of things (iot): A vision, architectural elements, and future directions. Future genera-

**Table 3: Blindfold validation obtained from the LSTM deep learning model for both 3 subjects from the AFDB and 82 subjects from the LTAFDB**

<table>
<thead>
<tr>
<th>Database</th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>No beats</th>
<th>Accuracy</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFDB</td>
<td>91,888</td>
<td>65,699</td>
<td>255</td>
<td>116</td>
<td>92,325</td>
<td>99.77%</td>
<td>1</td>
</tr>
<tr>
<td>LTAFDB</td>
<td>5,009,401</td>
<td>3,238,048</td>
<td>241,416</td>
<td>316,318</td>
<td>8,805,183</td>
<td>94%</td>
<td>96.58</td>
</tr>
</tbody>
</table>

![Fig. 5: Results obtained using the new test data: (a) confusion matrix by blindfold validation, (b) ROC of the model.](image)
tion computer systems 29, 1645–1660.
Conflict of interest

I would like to declare that there is no conflict of interest.

Oliver Faust
Corresponding author