Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 207 (2022) 3032–3040

Procedia Computer Science

www.elsevier.com/locate/procedia

26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2022)

Main requirements of end-to-end deep learning models for biomedical time series classification in healthcare environments

Ángel Serrano Alarcón*[a], Natividad Martínez Madrid[a], Ralf Seepold[b], Juan Antonio Ortega Ramirez[c]

[a]*Reutlingen University, Alteburgstr. 150, 72762 Reutlingen, Germany*
[b]*HTWG Konstanz, Alfred-Wachtel-Str. 8, 78462 Konstanz, Germany*
[c]*University of Seville, San Fernando 4, 41004 Sevilla, Spain*

Abstract

The use of deep learning models with medical data is becoming more widespread. However, although numerous models have shown high accuracy in medical-related tasks, such as medical image recognition (e.g. radiographs), there are still many problems with seeing these models operating in a real healthcare environment. This article presents a series of basic requirements that must be taken into account when developing deep learning models for biomedical time series classification tasks, with the aim of facilitating the subsequent production of the models in healthcare. These requirements range from the correct collection of data, to the existing techniques for a correct explanation of the results obtained by the models. This is due to the fact that one of the main reasons why the use of deep learning models is not more widespread in healthcare settings is their lack of clarity when it comes to explaining decision making.

Keywords: Deep learning; biomedical time series; healthcare.

* Corresponding authors. Tel.: +49 7121 271 4094
E-mail address: angel.serrano_alarcon@reutlingen-university.de

1877-0509
10.1016/j.procs.2022.09.532

1. Introduction

The impact of artificial intelligence algorithms in the industry continues to grow, and their use is becoming more and more noticeable in people's daily lives. We can find artificial intelligence algorithms in routine tasks such as object detection, weather forecasting, or disease diagnosis with increasing regularity [1]. The most promising results derived from the use of artificial intelligence are expected in medicine, where the huge growth of biomedical databases that are now more easily accessible allows for the further development of artificial intelligence models [1-3]. Despite its growing popularity, today it is still difficult to see artificial intelligence models on medical devices and there is very little evidence of clinical or economic impact [4,5]. So far, different types of algorithms can be found within the field of artificial intelligence, ranging from using classical algorithms (Decision Trees, Random Forest, k-nearest neighbours, and others) that fall within the field of machine learning to using deep learning algorithms [1,3,4]. In recent years, deep learning algorithms have often prevailed over the use of machine learning algorithms for various reasons, such as outperforming existing algorithms in terms of accuracy and reducing the data processing prior to model training [7]. This fact is very relevant when it comes to developing algorithms for the diagnosis and treatment of diseases in which the degree of specialization is very high [8]. In principle, when using deep learning models, feature extraction, which is crucial in the use of machine learning algorithms, is practically avoided. This process is crucial in any discipline in which machine learning algorithms are used. However, it is especially relevant in medicine, where it could be considered critical to succeed in classifying biomedical signals to detect a specific disease or physiological event in patients by using global or local time series features [3]. Consequently, the use of deep learning algorithms is considered an important key to what is known as personalized medicine [9,10].

Despite the advantages of using deep learning over machine learning, there are also disadvantages. The most notable is the lack of explainability of the models [11]. A deep learning model in which the reason for a decision cannot be explained in detail can never be used in healthcare. These models are intended to assist the physician who makes the decision, and therefore the physician must understand the reason for the decision made by the algorithm [2]. In addition to the above, an appropriate selection of the model is always accompanied by adequate collection and processing of the data. In deep learning, the quantity of data and its quality are essential, sometimes much more important than the correct selection of the model architecture. There is a wide variety of medical data, which is very heterogeneous due to the diversity that exists when collecting it in healthcare settings such as hospitals [12]. Most of the data collected from patients are time series, where expert analysis is essential to draw conclusions from them [13]. With the constant development of deep learning models over time, better and better results are being obtained from the raw time series that feed the models.

One of the tasks where time series are most commonly used is in classification, where the main goal is often to use labelled data to predict whether a patient has a certain disease or not. Currently, there are a large number of papers on deep learning models that have achieved good results in this task [1,7,14] . However, there are different aspects to take into account when using biomedical time series with deep learning models for classification tasks. In general, it is not easy to find databases that contain a large amount of data and are of high quality. In addition to this, these data must contain representative information within our case study to allow the models to generalize well [15]. With the use of time series, this fact can be crucial, as sometimes the achievement or not of satisfactory outcomes may depend on the value given at a certain timestamp. Without a certain level of data quality, there will be no success in training models that work with time series. As mentioned above, another aspect to take into account when working with time series is the feature extraction process, which has been essential in time series processing in the past in machine learning (distance-based, shapelets, etc.) [16,17]. This task requires highly specialised technicians in addition to being time-consuming [6,15,16]. With the use of deep learning models, this task is made easier, although, as shown later, this is not simple. Finally, and probably the most important aspect when working with time series to achieve accurate results in the field of deep learning is the fact of explainability [14]. As mentioned above, this matter is essential for models to be used effectively in medicine. However, many of the models trained in biomedical time series and that perform well, suffer from a lack of explainability and give rise to what are known as black boxes, as a consequence, this fact prevents these models from being used in real healthcare settings [19].

Taking into account all of the above, the aim of this paper, after analysing existing studies on the subject, is to study in-depth the use of deep learning models with biomedical time series for classification tasks, with an emphasis on the data processing and explainability of the models [1,3,7,13,16,19]. Although the regulation of artificial intelligence for

clinical use is in continuous development and addresses numerous points of interest such as challenges in the generalization of models, algorithm bias, susceptibility to attacks, implementation difficulties, among others [21]. In this work, a review of the main points that must be taken into account when developing algorithms is made, without covering other aspects that must be fulfilled for a correct delivery of an artificial intelligence system in healthcare. Other works also analyze the use of machine learning models in healthcare, although they do not focus on the use of end-to-end deep learning models [22]. In this way, a series of requirements that are exclusive for the development of deep learning models will be presented.

2. Methods

This section addresses the main requirements for the development of deep learning models with biomedical time series in classification tasks for their subsequent implementation in real healthcare settings. A priori, these requirements could be adapted to any model that does not use time series; however, there are some points that are very unique to time series that are taken into account in this work.

2.2 Main requirements

2.2.1 Data acquisition and data preprocessing

As mentioned in the Introduction, the first step in working with deep learning models is to pay attention to the quality and extent of the available data. The problem of data extent is not so significant in machine learning models. However, this aspect is decisive in deep learning models, which will be fed with time series with very little processing and usually no feature extraction. The use of medical data is not as simple as the use of any type of data from other disciplines for several reasons, one of the main ones being patient privacy. When implementing deep learning models with time series, it is essential to define the main task to be performed by the algorithm beforehand. Typically, when working with biomedical time series, the objective is to detect certain physiological events or determine a particular disease. Therefore, depending on the task to be performed, this should be clarified before searching for data in existing biomedical databases. This fact is vital in the first place because in most cases the data to be used by the algorithm in its production stage are insufficient for the training period. Therefore, external databases should be used for the training stage, for example, in the development of portable devices to detect certain sleep diseases [23]. For the development of portable devices, a biomedical database must be found that contains data that are as close as possible to the data with which the algorithm will work in the real healthcare environment. There are numerous biomedical databases easily accessible through sites such as PhysioNet or The National Sleep Research Resource (NSRR) [25,26]. In this step, medical experts play a crucial role [5,15]. They not only help to define the basic features of the time series to be used in training, but they also have to verify that the databases to be used have the necessary quality for the development of the models. Once databases containing data with information relevant to our case study have been found, it is time to define the strategy that sets out how to feed the deep learning algorithm. This varies whether working with tabulated data, images, or time series. However, this work focuses on biomedical time series, so the most common or effective way is usually to classify signals by using two approaches, according to [3,16,23,24]:

- Subject-based
- Event-based

Typically, for the use of patient-based biomedical time series, the entire time series obtained from the patient, such as an entire electrocardiogram, is used. This approach is generally effective for the diagnosis of a given disease, but not for the finding of a given physiological event, such as an arrhythmia [27]. The other approach addresses this fact, and focuses on the search for specific events. A time window should be established in which the time series is divided, and this must be agreed with the expert, depending on the duration of the event to be searched for [28]. This fact implies that the databases used should contain annotations together with the time series. Otherwise, it would have to be done with the help of an expert, although this would be very costly both economically and temporally. Although

not commonly deployed, there are new approaches where events can be classified without applying windowing, this technique is known as segmentation, which seems very promising, but will not be the subject of study in this manuscript [29]. Before feeding the time series to the deep learning models, signal processing occurs. One of the great strengths of deep learning is the use of raw time series with little or no processing, which is the big difference from machine learning [6,26]. Despite this, artefact removal tasks and standardisation of biomedical time signals usually occur, as signals collected from patients can contain values in very different ranges. The standardisation should be done based on signals from the same patient or when the signals have already been divided into windows. That is, not using the entire dataset as a whole, thus maintaining variability between patients.

2.2.2 Model selection and training

The choice of the model architecture along with the selection of data is the other crucial point in the use of biomedical time series for classification purposes. There are extensive reviews on which models work best with time series of any kind and biomedical time series in particular, so therefore many deep learning architectures have been proposed. The publications reviewed to extract information on commonly used architectures can be found in table 1.

Table 1. Papers on deep learning models for biomedical time series classification

Publication	Objective of the research article
Zemouri et al., 2019 [1]	This article reviews the main deep learning concepts relevant to biomedical applications related to the Omics, bioimaging, medical imaging, BBMI (study of the brain and body machine interface) and public and medical health management (PmHM). Concise summaries of omics and BBMI applications are provided.
Bock et al, 2021 [3]	This article aims to provide an introduction to the classification of time series. The manuscript first addresses the characteristics that biomedical time series can have and then continue with an overview of common machine learning algorithms for time series classification. Real use cases are used for this purpose.
Fawaz et al, 2019 [7]	This paper addresses the current performance of deep learning algorithms for time series classification (TSC) by presenting an empirical study of the most recent DNN architectures for TSC. An overview of the most successful deep learning applications in various time series domains is provided under a unified taxonomy of DNNs for TSC. An open source deep learning framework is also provided to the TSC community with implementation of each of the compared approaches and evaluation on a univariate TSC benchmark (the UCR/UEA archive) and 12 multivariate time series datasets.
Xiao et al, 2018 [12]	This article is a systematic review of deep learning models for electronic health record (EHR) data. In addition the article contains several deep learning architectures with analysis of different data sources and their target applications.
Fauvel et al, 2021 [14]	This paper presents XCM, a convolutional neural network for multivariate time series (MTS) classification. XCM improves explainability by providing faithful and more informative explanations of the decisions made by the model.
Wang et al, 2017 [16]	A simple but strong baseline is proposed for time series classification from scratch with deep neural networks. The proposed baseline models are pure end-to-end models from start to finish, without any heavy pre-processing of the raw data or feature extraction. A general analysis is provided to discuss the generalizability of the models, learned features, network structures and classification semantics.
Ruiz et al, 2021 [17]	This paper reviews recently proposed algorithms for MTS classification, based on deep learning, shapelets and bag of words approaches.
Mostafa et al, 2019 [18]	The aim of this paper is to analyse scientific research papers published in the last decade, providing an answer to research questions such as how to implement different deep networks, what kind of pre-processing or feature extraction is necessary, and the advantages and disadvantages of different types of networks.
Biswal et al, 2018 [30]	This work focuses on the diagnosis of sleep disorders through the use of neural networks. Analyzing the characteristics of the data as well as the architectures of the deep learning models.

However, after analyzing these reviews, three architectures yield the best results in most disciplines, not just medicine, and can be considered state-of-the-art models for univariate and multivariate time series classification. These are:

- Recurrent neural network (RNN)
    - Long short-term memory (LSTM)
- Residual Networks (ResNets)
- Convolutional Neural Network (CNN)
    - Convolutional Neural Network 1-D
    - Convolutional Neural Network 2-D

This fact does not mean that deep learning models should not be developed with other architectures [14]. However, these architectures have proven to be very effective, as they are easily programmable, execution times are not very long (especially CNN1), and they achieve good results with not necessarily a large number of data. Despite the long tradition of using RNNs, ResNets and CNNs have made a strong entry into deep learning with temporal biomedical signals. In particular, CNN, which was initially used for image detection. These architectures enable something that is crucial to explain the decision made during training [18]. In addition to this, CNN has shown that they require less computational resources than other models (which means much shorter development cycles) plus less training data is needed to achieve maximum performance [26]. This fact is especially relevant in this field, where the number of data referring to biomedical time series is not very extensive since the labelling of signals is time-consuming and expensive.

Therefore, these networks allow to achieve a proper balance between achieving good model performance and an adequate explanation for the end user in this case, the clinicians such as doctor, nurses, pharmacist, etc. It should be noted that there are numerous studies that cover the best training techniques for models, the best hyperparameter configuration, etc [7]. However, these aspects are not covered in this article since it focuses on the analysis of basic requirements for the production of deep learning models in healthcare.

2.2.3 Interpretability of the model

As already mentioned throughout this article, model interpretability is a crucial aspect of the use of artificial intelligence models in healthcare settings. Therefore, when choosing the architecture for the deep learning model, it must meet the basic requirement of allowing a detailed explanation of the factors that influence the decision making when generating the outputs. This fact is not that simple and is one of the significant problems deep learning has always faced. It is also relevant that not just any explanation is accepted, since the end-user will have a high level of knowledge in the discipline of study. Therefore, an explanation should be provided that allows clinicians to understand and study the result in a quick, simple and concise way.

There are different techniques that have been developed to facilitate the explanation of deep learning models when working with time series [34-36]. However, not all of these techniques will be covered. Due to our experience, we will focus on two of them due to their ease of implementation and their widespread use among machine learning engineers [34-36]. These techniques are known as Class Activation Map (CAM) and Gradient-weighted Class Activation Mapping [37,39]. CAM explains the classification made by a certain deep learning model by highlighting the subsequences of the signals that contributed the most to a certain output and also allow to find out the CNNs behaviour during the prediction on new biomedical time series [7,13]. It has been mentioned before, that there are architectures that facilitate the explainability of the models. Those architectures include a Global Average Pooling (GAP) layer that precedes a SoftMax layer that generates the final labelled output [16]. Therefore, CNNs and ResNets are the architectures that have the best balance in terms of accuracy-interpretability by allowing the use of these techniques since they include the mentioned layers.

Although CAM allows visualization of the regions of time series that have contributed the most to generate the output in the classification, this technique does not work for all deep learning architectures such as networks containing

fully connected layers [35]. Therefore, the use of a generalization of CAM is recommended. This generalization derives in Grad-CAM. Grad-CAM provides visual explanation for any type of CNN-based, regardless of its architecture [36]. Grad-CAM consists of taking the output feature map of a convolution layer, given an input signal(s), and weighing each channel in that feature map by the gradient of the class with respect to the channel [37].

Through the use of these techniques, it will also be possible to monitor whether the neural network is learning to detect the regions of the signal of interest based on the guidelines given by the clinicians. In the end, these two techniques allow the clinician to visualise whether the model has made the right decision and, finally, whether that prediction makes sense or not. The use of CAM or the Grad-CAM will depend on the type of architecture to be used, although the use of GRAD-CAM is usually chosen.

3. Results

As shown in Figure 1, the development of deep learning models with biomedical time series for production in healthcare settings, has several essential requirements that must be taken into account to be successful. It is vital to recognise the role of the expert within the medical field. The advice of a medical expert should be sought in both the early and late stages of development. First of all, for a correct selection of data, it is essential to know which set of signals best represents the relevant information for the classification to be carried out. In addition to which features and intervals of the signals represent those events of interest that are being sought. It is essential to know these basics before moving on to the signal collection stage.

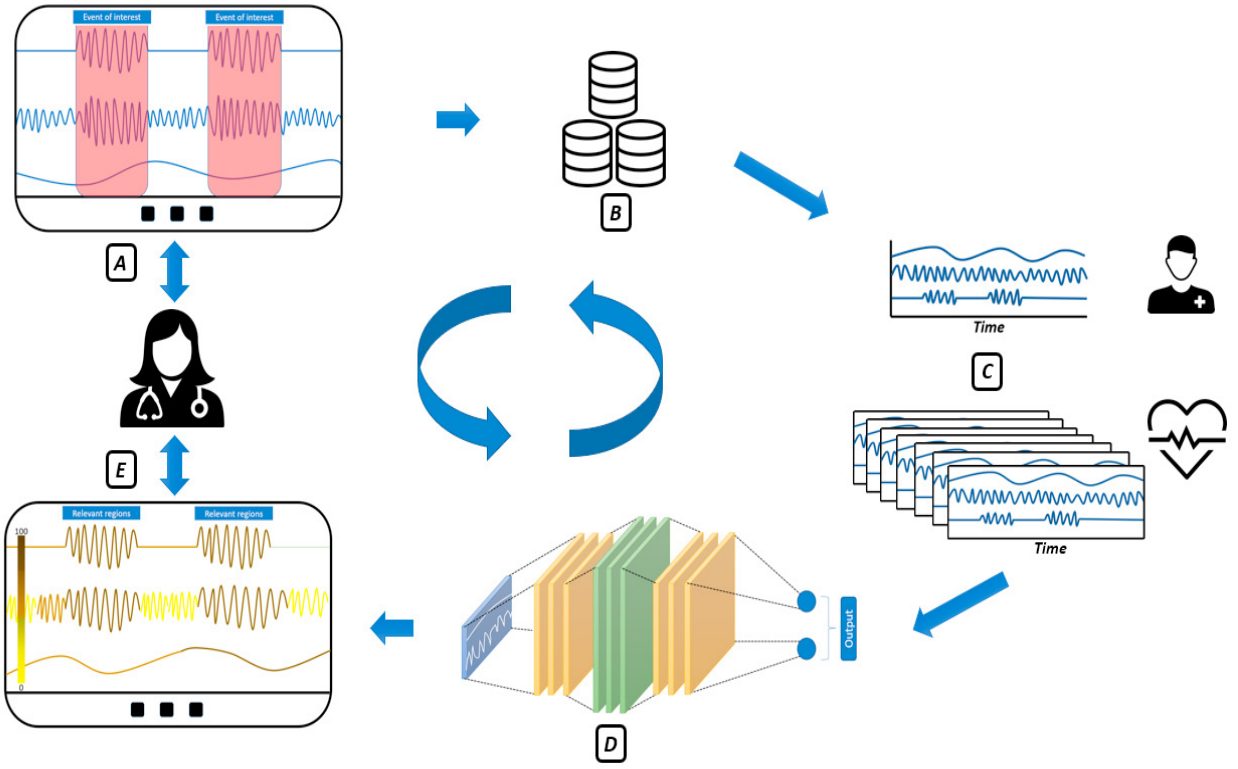

Figure 1. Development cycle of deep learning models for the classification of biomedical time series in health environments. (A) Study of the problem and requirements of the databases made by the clinician. (B) Data Collection and storage (C) Data preprocessing (D) Model selection and training (E) Model interpretability.

The second part of the cycle refers to signal collection. As mentioned above, deep learning algorithms will not always work in production with the signals used during training. This, together with the lack of an extensive number of signals, makes it relevant to search for databases that provide biomedical signals that are as relevant as possible to our field of study.

Generally, the use of the signals must be justified and the privacy agreement must be accepted, as they are signals from real patients, where privacy prevails. Once the database to be used has been decided, data storage and processing should be performed. In general, the development of deep learning models aims to achieve good performance with raw temporal biomedical signals, that is, with minimal preprocessing. However, the signals are usually standardised per patient, or, once the signals have been divided into windows. Despite this, there are occasions where signal processing is essential, either to remove noise or to improve signal quality [28]. However, these techniques are not addressed in this manuscript. After data preprocessing, the input to be fed to the deep learning models must be determined. A clear example would be to detect a patient suffering from sleep apnea, to detect whether the patient has apnea or not, the whole set of measured signals from the patient could be used [26]. However, for the detection of apnea events or the number of appearances, the way to operate is by windowing the signals in time periods (seconds). These windows are usually 30 to 60 seconds when working with biomedical time series [28,30]. It should be noted that information is usually lost when dividing the signal into windows since, depending on the duration of the event of interest, it could be framed within a window or not be represented at all. Although this drawback can be alleviated by using sliders of short duration when dividing or using windows of short duration, there are new techniques that try to adjust as much as possible to the annotations contained in the time series, as in the case of segmentation. Once the data has been processed and decided on which is the input for the models, the architecture should be designed with a focus on the subsequent explanation of the decision-making process. After the analysis of different publications, convolutional networks allow the use of a layer that allows the network to be traversed backwards and to know which parts of the temporal signal have been most important when deciding about the output (CAM and Grad-CAM). The beauty of this technique is that it allows the clinician to visualise which regions of the signal have contributed the most to the decision. Thus the clinician can see at a glance why this has occurred concisely and briefly. In addition, this technique can help improve the training of deep learning models by discovering which regions of the signal contribute the most to generating a specific output, as a result it is possible to study whether the model detects the event of interest and thus readjusts the model, as can be seen in Figure 1. A clear example of this is the growing number of developed portable devices, which often fail to be considered real medical devices [38]. They do not provide a clear and detailed explanation of the decision-making process.

During the training of the model, drawbacks such as unbalanced data, generalisation problems, hyperparameters optimization, overfitting may also arise. However, all these aspects are not addressed in this manuscript, as the main mission of this paper is to provide a broad overview of the requirements needed to develop a deep learning model for use in a real healthcare environment. Another important aspect to take into account when working with time series is the software used. Currently there are different solutions depending on the programming language. The Python programming language is one of the most used since it allows the use of libraries such as Keras (neural network library), sktime (machine learning for time series) or Tensorflow (deep learning framework) among others [3,35]. In addition to all of the above, new regulations are continually being developed, which on the one hand allow standardizing the development of deep learning models for medical devices and also facilitate the task of producing these medical devices so that they meet safety, efficacy and privacy standards [40].

4. Conclusion and outlook

The role that artificial intelligence can play in medicine is promising. Nowadays, there are already artificial intelligence models with an accuracy in recognition tasks that surpasses human capabilities so that these tools can provide the clinician with immense support in decision-making. Much of the data generated in healthcare centres tends to be time-series. The processing of time series is not actual and has been done for a long time. This field has always required highly specialized technicians. The use of deep learning models facilitates the tedious task of feature extraction and improves the results of classical models. However, it also generates much uncertainty in its decision making and gives rise to the well-known problem of black boxes. There are now models that allow the use of techniques that provide the opportunity to study which parts of the biomedical signals have been most influential in

decision making. Models will continue to be developed to improve outcomes, but the explainability of models must be intrinsic if they are to be used in a real healthcare environment. For future work, the requirements outlined in this manuscript should be developed.

Acknowledgements

This research was partially funded by the German Federal Ministry for Economic Affairs and Energy, ZiM project "Sleep Lab at Home" (SLaH) grant: ZF4825301AW9.

References

[1] R. Zemouri, N. Zerhouni, and D. Racoceanu, Deep Learning in the Biomedical Applications: Recent and Future Status, *Appl. Sci. 2019, Vol. 9, Page 1526*. **9** (2019) 1526. doi:10.3390/APP9081526.

[2] S. Gaube, H. Suresh, M. Raue, A. Merritt, S.J. Berkowitz, E. Lermer, J.F. Coughlin, and J. V Guttag, ARTICLE Do as AI say: susceptibility in deployment of clinical decision-aids, (n.d.). doi:10.1038/s41746-021-00385-9.

[3] C. Bock, M. Moor, C.R. Jutzeler, and K. Borgwardt, Machine Learning for Biomedical Time Series Classification: From Shapelets to Deep Learning, *Methods Mol. Biol.* **2190** (2021) 33–71. doi:10.1007/978-1-0716-0826-5_2.

[4] M.P. Sendak, J.D. Arcy, S. Kashyap, M. Gao, M. Nichols, K. Corey, W. Ratliff, and S. Balu, A Path for Translation of Machine Learning Products into Healthcare Delivery, *EMJ Innov.* (2020). doi:10.33590/emjinnov/19-00172.

[5] M.E. Matheny, D. Whicher, and S. Thadaney Israni, Artificial Intelligence in Health Care: A Report from the National Academy of Medicine, *JAMA - J. Am. Med. Assoc.* **323** (2020) 509–510. doi:10.1001/jama.2019.21579.

[6] V. Thorey, A.B. Hernandez, P.J. Arnal, and E.H. During, AI vs Humans for the diagnosis of sleep apnea, *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*. (2019) 1596–1600. doi:10.1109/EMBC.2019.8856877.

[7] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.A. Muller, Deep learning for time series classification: a review, *Data Min. Knowl. Discov.* **33** (2019) 917–963. doi:10.1007/s10618-019-00619-1.

[8] I. Perez-Pozuelo, B. Zhai, J. Palotti, R. Mall, M. Aupetit, J.M. Garcia-Gomez, S. Taheri, Y. Guan, and L. Fernandez-Luque, The future of sleep health: a data-driven revolution in sleep science and medicine, *Npj Digit. Med.* **3** (2020) 1–15. doi:10.1038/s41746-020-0244-4.

[9] M.N. Kamel Boulos, and P. Zhang, Digital twins: From personalised medicine to precision public health, *J. Pers. Med.* **11** (2021). doi:10.3390/jpm11080745.

[10] B. Björnsson, C. Borrebaeck, N. Elander, T. Gasslander, D.R. Gawel, M. Gustafsson, R. Jörnsten, E.J. Lee, X. Li, S. Lilja, D. Martínez-Enguita, A. Matussek, P. Sandström, S. Schäfer, M. Stenmarker, X.F. Sun, O. Sysoev, H. Zhang, and M. Benson, Digital twins to personalize medicine, *Genome Med.* **12** (2019) 10–13. doi:10.1186/s13073-019-0701-3.

[11] F.J. Baldán, and J.M. Benítez, Multivariate times series classification through an interpretable representation, *Inf. Sci. (Ny).* **569** (2021) 596–614. doi:10.1016/j.ins.2021.05.024.

[12] C. Xiao, E. Choi, and J. Sun, Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review, *J. Am. Med. Inform. Assoc.* **25** (2018) 1419–1428. doi:10.1093/JAMIA/OCY068.

[13] A. Subasi, Practical guide for biomedical signals analysis using machine learning techniques : a MATLAB based approach, (n.d.).

[14] K. Fauvel, T. Lin, V. Masson, É. Fromont, and A. Termier, XCM: An Explainable Convolutional Neural Network for Multivariate Time Series Classification, *Mathematics*. **9** (2021) 3137. doi:10.3390/math9233137.

[15] C.A. Goldstein, R.B. Berry, D.T. Kent, D.A. Kristo, A.A. Seixas, S. Redline, and M. Brandon Westover, Artificial intelligence in sleep medicine: background and implications for clinicians, *J. Clin. Sleep Med.* **16** (2020) 609. doi:10.5664/JCSM.8388.

[16] Z. Wang, W. Yan, and T. Oates, Time series classification from scratch with deep neural networks: A strong baseline, *Proc. Int. Jt. Conf. Neural Networks*. **2017-May** (2017) 1578–1585. doi:10.1109/IJCNN.2017.7966039.

[17] A.P. Ruiz, M. Flynn, J. Large, M. Middlehurst, and A. Bagnall, The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances, Springer US, 2021. doi:10.1007/s10618-020-00727-3.

[18] S.S. Mostafa, F. Mendonça, A.G. Ravelo-García, and F. Morgado-Dias, A systematic review of detecting sleep apnea using deep learning, *Sensors (Switzerland)*. **19** (2019) 1–26. doi:10.3390/s19224934.

[19] F. Wang, R. Kaushal, and D. Khullar, Should health care demand interpretable artificial intelligence or accept "black Box" Medicine?, *Ann. Intern. Med.* **172** (2020) 59–61. doi:10.7326/M19-2548.

[20] P. Ivaturi, M. Gadaleta, A.C. Pandey, M. Pazzani, S.R. Steinhubl, and G. Quer, A Comprehensive Explanation Framework for Biomedical Time Series Classification, *IEEE J. Biomed. Heal. Informatics*. **25** (2021) 2398–2408. doi:10.1109/JBHI.2021.3060997.

[21] C.J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, Key challenges for delivering clinical impact with artificial intelligence, *BMC Med.* **17** (2019) 1–9. doi:10.1186/s12916-019-1426-2.

[22] J. Wiens, S. Saria, M. Sendak, M. Ghassemi, V.X. Liu, F. Doshi-Velez, K. Jung, K. Heller, D. Kale, M. Saeed, P.N. Ossorio, S. Thadaney-Israni, and A. Goldenberg, Do no harm: a roadmap for responsible machine learning for health care, *Nat. Med.* **25** (2019).

doi:10.1038/s41591-019-0548-6.

[23] A. Manoni, F. Loreti, V. Radicioni, D. Pellegrino, L. Della Torre, A. Gumiero, D. Halicki, P. Palange, and F. Irrera, A new wearable system for home sleep apnea testing, screening, and classification, *Sensors (Switzerland)*. **20** (2020) 1–26. doi:10.3390/s20247014.

[24] Sleep Data - National Sleep Research Resource - NSRR, (n.d.). https://sleepdata.org/ (accessed April 27, 2022).

[25] PhysioNet, (n.d.). https://physionet.org/ (accessed April 27, 2022).

[26] S. Kristiansen, K. Nikolaidis, T. Plagemann, V. Goebel, G.M. Traaen, B. Øverland, L. Aakerøy, T.-E.E. Hunt, J.P. Loennechen, S.L. Steinshamn, C.H. Bendz, O.-G.G. Anfinsen, L. Gullestad, and H. Akre, Machine Learning for Sleep Apnea Detection with Unattended Sleep Monitoring at Home, *ACM Trans. Comput. Healthc.* **2** (2021) 1–25. doi:10.1145/3433987.

[27] S. Vijayarangan, B. Murugesan, R. Vignesh, S.P. Preejith, J. Joseph, and M. Sivaprakasam, Interpreting Deep Neural Networks for Single-Lead ECG Arrhythmia Classification, *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*. **2020-July** (2020) 300–303. doi:10.48550/arxiv.2004.05399.

[28] E.S. JeyaJothi, J. Anitha, S. Rani, and B. Tiwari, A Comprehensive Review: Computational Models for Obstructive Sleep Apnea Detection in Biomedical Applications, *Biomed Res. Int.* **2022** (2022) 1–21. doi:10.1155/2022/7242667.

[29] M. Perslev, M.H. Jensen, S. Darkner, P.J. Jennum, and C. Igel, U-Time: A fully convolutional network for time series segmentation applied to sleep staging, *Adv. Neural Inf. Process. Syst.* **32** (2019) 1–12.

[30] S. Biswal, H. Sun, B. Goparaju, M. Brandon Westover, J. Sun, and M.T. Bianchi, Expert-level sleep scoring with deep neural networks, *J. Am. Med. Informatics Assoc.* **25** (2018) 1643–1650. doi:10.1093/jamia/ocy131.

[31] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, Understanding Neural Networks Through Deep Visualization, (n.d.).

[32] T.Y. Hsieh, S. Wang, Y. Sun, and V. Honavar, Explainable Multivariate Time Series Classification: A Deep Neural Network Which Learns to Attend to Important Variables As Well As Time Intervals, Association for Computing Machinery, 2021. doi:10.1145/3437963.3441815.

[33] M. Han, and X. Liu, Feature selection techniques with class separability for multivariate time series, *Neurocomputing*. **110** (2013) 29–34. doi:10.1016/J.NEUCOM.2012.12.006.

[34] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, Learning Deep Features for Discriminative Localization, *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* **2016-Decem** (2016) 2921–2929. doi:10.1109/CVPR.2016.319.

[35] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, *Int. J. Comput. Vis.* **128** (2020) 336–359. doi:10.1007/s11263-019-01228-7.

[36] R.R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, D. Batra, and V. Tech, Grad-CAM: Why did you say that?, (2016). doi:10.48550/arxiv.1611.07450.

[37] 9 Advanced deep learning for computer vision - Deep Learning with Python, Second Edition, (n.d.). https://livebook.manning.com/book/deep-learning-with-python-second-edition/chapter-9/217 (accessed April 27, 2022).

[38] M.T. Bianchi, Sleep devices: wearables and nearables, informational and interventional, consumer and clinical, *Metabolism.* **84** (2018) 99–108. doi:10.1016/j.metabol.2017.10.008.

[39] J. Siebert, J. Groß, and C. Schroth, A Systematic Review of Packages for Time Series Analysis, (2021) 22. doi:10.3390/engproc2021005022.

[40] US FDA, Proposed Regulatory Framework for Modifications to Artificial Intelligence / Machine Learning (AI / ML) -Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback, *U.S Food Drug Adm.* (2019) 1–20.