Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 207 (2022) 2924–2933

Procedia Computer Science

www.elsevier.com/locate/procedia

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

Explainable machine learning for sleep apnea prediction

A. R. Troncoso-García[a,*], M. Martínez-Ballesteros[b], F. Martínez-Álvarez[a], A. Troncoso[a]

[a]*Data Science and Big Data Lab, Pablo de Olavide University, 41013 ES-Seville, Spain*
[b]*Department of Computer Science, University of Seville, 41012 ES-Seville, Spain*

Abstract

Machine and deep learning has become one of the most useful tools in the last years as a diagnosis-decision-support tool in the health area. However, it is widely known that artificial intelligence models are considered a black box and most experts experience difficulties explaining and interpreting the models and their results. In this context, explainable artificial intelligence is emerging with the aim of providing black-box models with sufficient interpretability so that models can be easily understood and further applied. Obstructive sleep apnea is a common chronic respiratory disease related to sleep. Its diagnosis nowadays is done by processing different data signals, such as electrocardiogram or respiratory rate. The waveform of the respiratory signal is of importance too. Machine learning models could be applied to the signal's analysis. Data from a polysomnography study for automatic sleep apnea detection have been used to evaluate the use of the Local Interpretable Model-Agnostic (LIME) library for explaining the health data models. Results obtained help to understand how several features have been used in the model and their influence in the quality of sleep.

Keywords: Explainable artificial intelligence; health data; polysomnography; LIME

1. Introduction

Deep learning (DL) and machine learning (ML) algorithms are tools used to make predictions and classify large and heterogeneous data in different fields [25]. They are the technology behind artificial intelligence applications in industries like object recognition, natural language treatment, or self-driving cars. Specifically, they are widely used in the health scope, for example, in the processing of health data and health images or as diagnostic decision support tools. This has led to tremendous progress in analysing and processing huge amounts of data. The objective of these technical solutions is to help health professionals in their daily work. However, physicians could not feel completely confident in the tools mentioned above, due to the fact that it is not known how predictions are made. That means that one of their most serious disadvantages is that they are considered black-box models. It is impossible to figure out

* Corresponding author. Tel.: +34-954-349-230 ; fax: +0-000-000-0000.
E-mail address: artrogar@upo.es

1877-0509
10.1016/j.procs.2022.09.351

how the model obtains the outputs by applying inner non-linear operations to the inputs. This is concretely a mayor problem when concerning the health area. Professionals need to know *how* and *why* algorithms are obtaining results in order to feel trustable about it. Because of that, concepts such as explainability and interpretability in artificial intelligence have been introduced. This is called eXplainable Artifical Intelligence (XAI). XAI is a subfield of AI that aims to provide explanations to the final user of the AI model. Therefore, explainability can be defined as the idea that a machine learning model and its output can be explained in a way that humans can understand it. Certain classes of algorithms, including more traditional machine learning algorithms, for instance Decision Trees, tend to be more readily explainable. Others, such as deep learning systems, while obtaining better performance, remain much harder to explain [2]. XAI algorithms are classified according to the phase of experimentation and prediction in which they are used. Pre-model interpretability is related to data interpretability and exploratory and data visualization. In-model interpretability is focused on creating models that could explain themselves, for example, decision tree-based algorithms. Then, post-model interpretability is about the interpretation of the outputs according to the inputs [3]. These techniques are independent of the model and are applied to increase the explainability of different algorithms. In that way, Local Interpretable Model-Agnostic (LIME) is one popular technique. It is widely used and is also applied to this study due to its simplicity [18]. It can be used as a Python library, meaning it has an easy usage. Both creating XAI models and adding explainability to pre-existing algorithms are ample research fields nowadays. These concepts became more crucial as AI models are used today to make high-stakes decisions in essential areas such as health care [20]. The need for XAI systems in this field is also related to ethical and fair decision making [1].

Sleep apnea is a chronic respiratory disorder that occurs while sleeping. It is defined as repeated episodes of total or partial obstruction during sleep. Normal breathing is interrupted during an episode of apnea, with pauses from seconds to minutes, causing worse sleep quality. It is associated with obesity and altered cardiopulmonary function [6]. As a result, patients usually have symptoms such as dizziness, daytime sleepiness, irritability, or loss of memory. This causes problems for daily activities including working, studying, or driving. Most cases are undiagnosed and untreated [26]. The Apnea Hypopnea Index (AHI) is considered a frequently used metric to diagnose and estimate the severity of the disorder, indicating the number of apnea events per hour of sleep. Between 4% and 20% of the adult world population has been estimated to be affected by sleep apnea, and there are pediatric patients as well. Its prevalence is increasing due to the increased prevalence of obesity and sedentary habits [8]. The most common way to detect and diagnose that disease is polysomnography (PSG). Patients must sleep on a hospital scene, connected to different sensors that control some body measures like heart rate, cerebral signals, breathing rithm or nasal airflow. Information is received through a number of separate channels [17]. During a PSG analysis, signals are collected by different sensors, usually between 3 and 8 hours. The physicians supervise the process. Then they observe the data and discuss the diagnosis. Machine Learning models are really helpful for simplifying and automatising this health exam.

In this paper, the addition of explainability to ML models applied to health data is studied. PSG health data is used as an input for train and test several classification models. The existence of an apnea period is the target class. Then, the best one in terms of certain quality measures is selected. After that, a model-agnostic technique is applied to obtain explanation of which attributes of the input data are used to make predictions.

The remainder of the article is structured as follows. In Section 2, recent advances in deep learning interpretability related to apnea detection are reviewed. Then Section 3 describes and details the experiments carried out, and Section 4 presents the results that have been obtained from polysomnography data. Finally, Section 5 concludes the paper.

2. Related work

Deep learning methods are widely applied in the detection of sleep apnea disorders. Different approaches have been used for its automatic or pseudoautomatic diagnosis. As it is related to respiratory and sleeping problems, several signals and measures should be recorded for this. The results show the incredible potential of these solutions, even with complex classification patterns and heterogeneous data [9].

The literature shows that machine learning and deep learning methods are applied to the detection and diagnosis of sleep apnea. Although they can achieve the correct results, human experts are required to validate the process. A key part of the procedure is selecting the input data for the model, including not only individual data but also patients'

habits and health and body characteristics. The preprocessing of the data is also important. Models that obtain better results are recurrent neural networks (RNN) and convolutional neural networks (CNN) [14].

The systematic review of the literature carried out in [10], shows the reliability of ML models in the detection of sleep apnea in pediatric patients. Their diagnosis is more difficult than that of adults due to the characteristics of their bodies. The signals most used for that are related to respiration, such as oxygen in blood and airflow. However, it is said that explainability is not reached in this scope. In addition, in [24] a principal component analysis (PCA) method is applied to processing respiratory data overnight, showing a high-speed computation. The data used are from the MIT-PSG database, similar to the data used here. The detection of obstructive respiratory events is another approach in which deep learning algorithms are also used [12].

The combination of traditional science and data science has led to tremendous progress in healthcare. Nevertheless, in most cases, they give intelligible explanations to the users. Health professionals would be confident applying models with transparency and human understanding. At that point, it has been demonstrated that achieving explainability in machine learning models applied to health data is a challenging open issue [20]. The difficulties of applying XAI models to health data have been widely discussed. For example, in [23] there is a focus on mental health detection. This is one of the most tough diagnosis in the medical scope due to the complexity of the brain. XAI techniques are promising for a stunning future.

In general, XAI methods can be generated in two ways: by creating interpretable models themselves or adding explainability to the existing ones. One example of the first case, is the work of Panigutti et al. [19]. A model-agnostic explainability technique called *Doctor XAI* is created. It is specific to health data and is based on ontology information. It achieves adequate results that explain the predictions of an RNN.

In other lines, there are several algorithms that can add explanation to DL and ML models. LIME is one popular technique for that. It is a library for Python widely used due to its simplicity: LIME explains the behavior of black-box models based on linear models around one instance of interest. Input data are perturbed and new predictions are made. Thus, the critical values to make predictions are detected [18]. On the basis of LIME, the authors have developed a new approach called Anchors. The idea works similarly to LIME but IF-THEN rules are generated [22]. Another interesting approach is SHapely Additive exPlanation (SHAP) [16]. This method calculates the contribution of each aspect for a concrete prediction based on game theory.

A comparison of the use of this kind of technique applied to health data is shown in [5]. They use data on patients with lung cancer and try to predict mortality. Then explainability is added using the mentioned methods. The results have shown that all of them provide more than just a prediction. They generate a clear importance of the feature. All of this is better for the reasoning of clinicians and is helpful for the support of clinical decision. Anyway, models are not able to substitute human expert knowledge.

Finally, this section looks at information on the scope of XAI and sleep apnea, which is a hot topic as XAI for health area in general. An example is found, in [15], where a random forest model is applied to detect patterns in the datasets of both apnea patients and healthy subjects. As decision trees are highly interpretable, the random forest as a set of trees allows one to extract information about how predictions are done. The results show that high frequencies in respiratory sounds tend to be employed for classification.

3. Methodology

In this paper, the LIME library for Python is used to explain predictions made using PSG data. LIME has been chosen due to its simplicity, being used as a Python library and its demonstrated interpretable capability [4]. It is an agnostic model-independent method for XAI that shows the explanation for a particular instance. It needs both the trained model and the instance as input data. Then, LIME makes perturbations around this local point and calculates the predictions that the trained model would make. This allows the library to determine which attributes and which of their values are crucial to determine the classification class. During the experimentation process, first several Python models from *scikit-learn* library are applied to the data. Then, the library SMOTE is used to balance the data in order to try to improve the models' results. Finally, LIME is applied to the predictions that have been obtained. These are the output data of the model showing the best results, that is, the Random Forest classifier using $n=100$ trees to obtain the classification. The explainable output is evaluated in Section 4.

3.1. Data

The data used in this experiment are from the MIT-BIH Polysomnographic Database [7]. Data have been recorded at the Beth Israel Hospital Sleep Laboratory of Boston. There are 18 records that include different files with the signals and their annotations. Data are recorded from 16 different patients. All records include an ECG signal, an invasive blood pressure signal (measured by using a catheter in the radial artery), an EEG signal, and a respiration signal from a nasal thermistor. Annotations are made by health professionals after carefully studying the signals. They label the data according to the existence of an episode of apnea and the stage of sleep in which the patient is. These annotations affect to the 30 seconds of the record that follow the annotation [11]. The PSG data are processed to create a dataset for classification tasks. The annotations given are used to establish the class: apnea event (1) or not (0). Each record is processed as a time series dataset. They are divided using a time window of 30 seconds (as annotations are labeling). The sampling rate at which signals were collected is 250 Hz, meaning that the final dataset has 7500 attributes per instance. Each instance contains the measure of airflow during this 30 seconds. Figure 1 illustrates an example of a complete PSG record with blood pressure, airflow, ECG, and EEG signals.

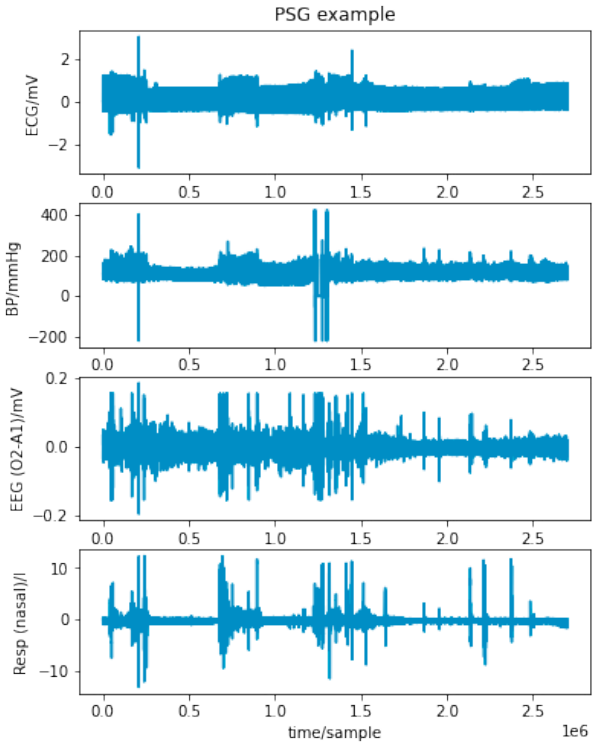

Fig. 1. PSG record with ECG, BP, EEG and nasal airflow signals

3.2. Predictive models

MIT-BIH-PSG data are used for a binary classification problem in which labels are related to sleep apnea. Here only machine learning models have been used due to their lower computation time. Labels used are the existence (1) or not (0) of an episode of apnea during each slice of the record. Data have been randomly divided into train (70%) and test (30%) dataset. A seed value has been establish to ensure that the experiment is reproducible. These dataset have been used during training and testing phases of the machine learning models. When it is not specified, the default options are used. ML models from the *scikit-learn* library in Python [21] are used. In particular, the models used are as follows.

- Logistic Regression (LR). This is one of the simplest models used for binary classification. The targets are predicted by a linear approximation.
- K Nearest Neighbors (KNN) classifier. The well-known classifier implementing the k-nearest neighbors vote. Here, *k=5*, meaning 5 neighbors, is set as a parameter of the algorithm.
- Decision Tree (DT) classifier. They are a nonparametric supervised learning method where the goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.
- Random Forest (RF) classifier. It is a meta-estimator that fits a number of decision tree classifiers on various subsamples of the dataset and uses averaging to improve the predictive accuracy and control overfitting. Here, it is applied three times for different numbers of trees, namely, *n=10, n=100* and *n=1000*.
- Gradient Boosting Classifier (GBC). This method builds an additive model in a forward stage-wise fashion. Allows for the optimization of arbitrary differentiating loss functions. In each stage, the regression trees are fitted on the negative gradient of the binomial or multinomial deviance loss function. For binary classification, only a single regression tree is induced.

3.3. Quality parameters

The models are trained and tested independently. It is done using the same training and test datasets. The results are evaluated as usual in binary health classification. We compute true positive (TP), which means that an apnea event was correctly detected, true negative (TN) as a well-predicted nonapnea event, false positive (FP), when we have predicted an apnea event that has no occur, and false negative (FN) in the opposite case. The metrics used to compare their performance are detailed as follows:

- $$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$
- *ROC-AUC*. The well-known Area Under the ROC Curve. This means that AUC measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1).
- $$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \quad \text{where:} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

3.4. Data balance

Predictive models are used to classify the instances into two categories: apnea event (1) or not (0). The classes in the dataset are imbalanced. This means that there are many more cases with no apnea (0) than with apnea (1), as shown in Figure 2.

For this reason, more instances are generated synthetically using SMOTE [13]. Usually predictive models obtain better performance when the classes are balanced. This tool makes possible the balance of the classes by creating fake data of the minority class. The performance of the selected algorithms in Section 3.2 is shown in Tables 1 and 2.

Here, results are presented using both the initial dataset and the balanced one. In general, they are quite similar and one could no appreciate a significant improvement. As anyone could think, the balanced of the classification categories is causing a better performance of the algorithms, if the focus is on *F1* metric and in simpler methods such as Decision Tree and K-Neighbors Classifier. However, for the ones with better global results such as Random Forest there is not any upgrade after the balanced of the class.

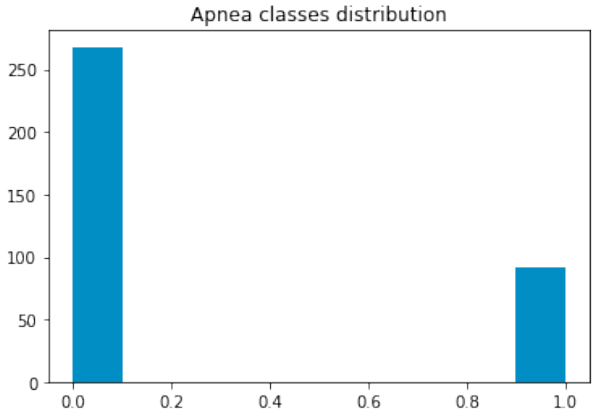

Fig. 2. Distribution of the classification categories in the initial dataset

Table 1. Performance of the ML models with unbalanced data.

Model	Accuracy	ROC-AUC	F1
LR	0.698	0.514	0.363
KNN (*k=5*)	0.750	0.711	0.408
DT	0.750	0.567	0.292
RF (*n=10*)	0.776	0.744	0.381
RF (*n=100*)	0.802	0.799	0.566
RF (*n=1000*)	0.784	0.802	0.528
GBC	0.758	0.686	0.333

Table 2. Performance of the ML models with balanced data.

Model	Accuracy	ROC-AUC	F1
LR	0.689	0.524	0.400
KNN (*k=5*)	0.629	0.728	0.463
Decision Tree	0.707	0.600	0.393
RF (*n=10*)	0.767	0.720	0.491
RF (*n=100*)	0.707	0.728	0.393
RF (*n=1000*)	0.741	0.751	0.483
GBC	0.752	0.656	0.400

Because of all of this, we conclude that the best in terms of metrics that have been used is Random Forest (RF) with *n=100*, trained with unbalanced data. LIME is used to add the explainability to this method, and results are presented in the following Section 4.

4. Results

The LIME library is used to give human-understandable explanations to the predictions made by the Random Forest model using 100 trees. LIME has a local behaviour, meaning that explanations are calculated for a single instance of the dataset. They are given for each particular instance instead of the complete dataset. Experiments have been carried out to evaluate the explanations of both classes, apnea (1) and nonapnea (0). Two examples of the LIME output for that are shown in Figures 3 and 4.

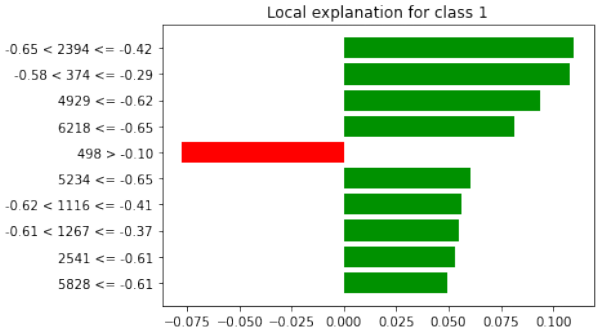

Fig. 3. LIME output for an instance in which the prediction is class 1

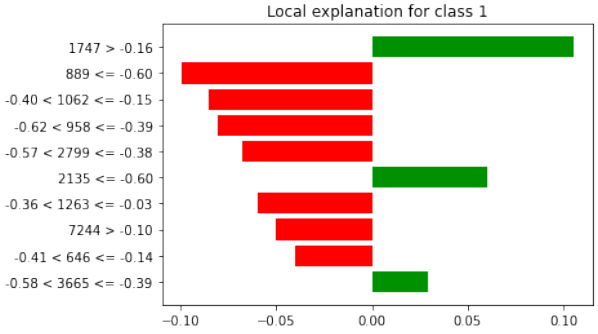

Fig. 4. LIME output for an instance in which the prediction is class 0

After applying LIME, a graph representation is created as an output. It displays a set of the most important features and their values, which are used to predict the class, on the left of the diagram. The colored bars are showing which of these sentences are more related to which of the two classes. Down, in X ex, there is the measure of the participation of each feature to predict class 1 (positive numbers) or class 0 (negative numbers). The size of the colored bars represents the level of importance of each attribute for the concrete predicted instance.

Concerning the LIME output diagrams, in Figure 3, the predicted class is 1 (apnea), whereas in Figure 4, it is 0 (non-apnea). The model has predicted class 1 because most of the features are related with class 1 (green, to the right) in Figure 3, whilst in Figure 4 they support class 0 (red, to the left).

In the mentioned Figures one could observe 10 features. The number of features is a parameter that the user can select when using LIME. In our case, these features are the time points for each record, labeled between 0 and 7499. Then, LIME provides information about the concrete value of these attributes. For each of them, a numerical interval is given. The model decides the classification category depending on the range in which a certain characteristic is found. The code of color informs about the connection between each of the features and the predicted class. The green features are the those that could make the model predict class 1, whereas the red ones are important for class 0 detection. For example, regarding Figure 4, feature 889 is less than or equal to -0.6, feature 1062 is between -0.4 and -0.15 and so on. The features claiming for class 0 are more and they are more relevant than the ones for class 1. Because of that, the predicted class for this specific instance of the test data set is 0 (non-apnea).

In view of the fact that the input data used for the detection of sleep apnea are usually signals, the conventional LIME output diagram is not completely clarified. Due to this, the results are again presented in Figures 5 and 6 where the airflow signal is drawn in blue, and red points are the important features that LIME has pointed out. This graphical representation is showing the same info as Figures 3 and 4, that is the essential points discovered by LIME. However, representing them together with the nasal airflow signal provides more understandable information for physicians

which are used to evaluate this kind of signals. It can be done because LIME provides explanation of how predictions are made for each particular instance of the dataset.

To sum up, it can be observed that minimum and maximum points are the key for the detection of apnea in Figure 5 (class 1). This is because the apnea event is an abrupt interruption of the respiratory process, leading to an abnormal pattern in nasal airflow. On the other hand, a nonapnea period could be observed in Figure 6 (class 0). This is a periodic signal that shows a normal breathing pattern while sleeping. The important features that are highlighted in red are showing that.

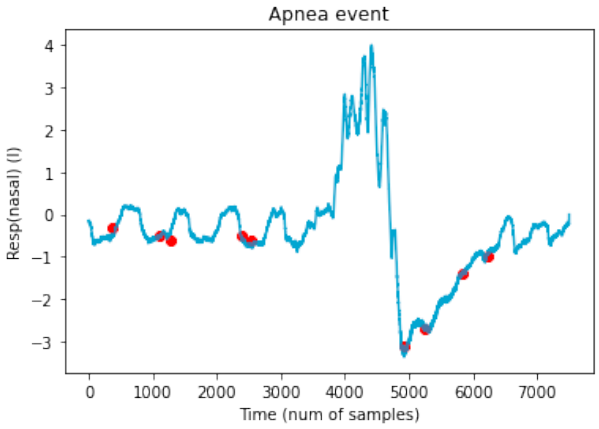

Fig. 5. PSG record showing an apnea event (class 1)

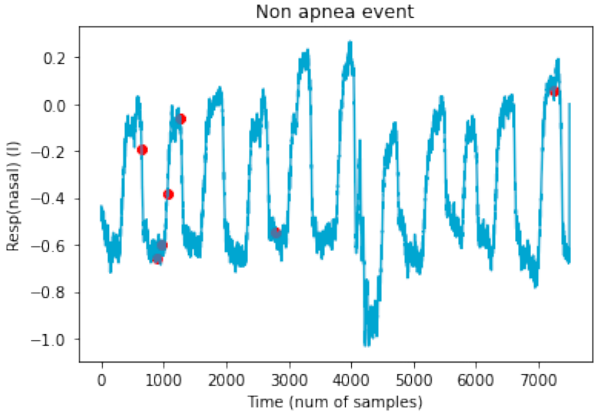

Fig. 6. PSG record showing a non apnea event (class 0)

5. Conclusions

Machine learning algorithms have a promising future in the realm of health. They support the detection and diagnosis of a wide range of diseases. The case of sleep apnea is particularly complicated due to the heterogeneous information used as input. Body measures such as nasal airflow are usually used. Automated collection and processing of this kind of data is still an open research issue. Current studies on sleep apnea are conducted to obtain efficient and accurate diagnosis methods. The fact that physicians needs to trust in these models creates the need to add explainability to them.

In this work, the application of the popular LIME method to health data is studied. PSG signals are used as input data to a set of machine learning predictive models. The most accurate one, Random Forest, is selected for the rest of the experimentation part. Then, explanations to the predictions are generated using LIME. The high potential of this tool is demonstrated, as important features and their value range, depending on the predicted class (apnea 1 or non-apnea 0) are highlighted. Knowing which points are used for making predictions makes easier for physicians trust on the model.

However, since the input data are signals, each feature is a time point. This leads to a non-completely understandable diagram. Then, a graphical representation of the results is created. The diagram shows the specific airflow signal and also the important features points. This is a visual explanation of the predictions made by Random Forest. Health professionals could easily see the specific points that the ML model has used to make the predictions. After that, they can find the model more trustable, which is an essential condition for health decision support tools.

To sum up, LIME has been shown as a powerful tool to make ML models explainable. Future work could lead to its application to further methods in the field of deep learning.

Acknowledgements

The authors would like to thank the Spanish Ministry of Science and Innovation for the support under the project PID2020-117954RB-C21 and the European Regional Development Fund and Junta de Andalucía for projects PY20-00870 and UPO-138516.

References

[1] Antoniadi, A.M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B.A., Mooney, C., 2021. Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: a systematic review. Applied Sciences 11, 5088.

[2] Burkart, N., Huber, M.F., 2021. A survey on the explainability of supervised machine learning. Journal of Artificial Intelligence Research 70, 245–317.

[3] Carvalho, D.V., Pereira, E.M., Cardoso, J.S., 2019. Machine learning interpretability: A survey on methods and metrics. Electronics 8, 832.

[4] Dieber, J., Kirrane, S., 2020. Why model why? assessing the strengths and limitations of lime. arXiv preprint arXiv:2012.00093 .

[5] Duell, J., Fan, X., Burnett, B., Aarts, G., Zhou, S.M., 2021. A comparison of explanations given by explainable artificial intelligence methods on analysing electronic health records, in: 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), IEEE. pp. 1–4.

[6] Gami, A.S., Hodge, D.O., Herges, R.M., Olson, E.J., Nykodym, J., Kara, T., Somers, V.K., 2007. Obstructive sleep apnea, obesity, and the risk of incident atrial fibrillation. Journal of the American College of Cardiology 49, 565–571.

[7] Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E., 2000. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. circulation 101, e215–e220.

[8] Gottlieb, D.J., Punjabi, N.M., 2020. Diagnosis and management of obstructive sleep apnea: a review. Jama 323, 1389–1400.

[9] Gutiérrez-Tobal, G.C., Álvarez, D., Crespo, A., Del Campo, F., Hornero, R., 2018. Evaluation of machine-learning approaches to estimate sleep apnea severity from at-home oximetry recordings. IEEE journal of biomedical and health informatics 23, 882–892.

[10] Gutiérrez-Tobal, G.C., Álvarez, D., Kheirandish-Gozal, L., Del Campo, F., Gozal, D., Hornero, R., 2021. Reliability of machine learning to diagnose pediatric obstructive sleep apnea: Systematic review and meta-analysis. Pediatric Pulmonology .

[11] Ichimaru, Y., Moody, G., 1999. Development of the polysomnographic database on cd-rom. Psychiatry and clinical neurosciences 53, 175–177.

[12] Korkalainen, H., Aakko, J., Nikkonen, S., Kainulainen, S., Leino, A., Duce, B., Afara, I.O., Myllymaa, S., Töyräs, J., Leppänen, T., 2019. Accurate deep learning-based sleep staging in a clinical population with suspected obstructive sleep apnea. IEEE journal of biomedical and health informatics 24, 2073–2081.

[13] Lemaître, G., Nogueira, F., Aridas, C.K., 2017. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research 18, 1–5. URL: http://jmlr.org/papers/v18/16-365.html.

[14] Mostafa, S.S., Mendonça, F., G Ravelo-García, A., Morgado-Dias, F., 2019. A systematic review of detecting sleep apnea using deep learning. Sensors 19, 4934.

[15] Nakari, I., Kitajima, E., Tajima, Y., Takadama, K., 2020. Non-contact sleep apnea syndrome detection based on what random forests learned, in: 2020 IEEE 2nd Global Conference on Life Sciences and Technologies (LifeTech), IEEE. pp. 240–244.

[16] Nohara, Y., Matsumoto, K., Soejima, H., Nakashima, N., 2019. Explanation of machine learning models using improved shapley additive explanation, in: Proceedings of the ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, pp. 546–546.

[17] Osman, A.M., Carter, S.G., Carberry, J.C., Eckert, D.J., 2018. Obstructive sleep apnea: current perspectives. Nature and science of sleep 10, 21.

[18] Palatnik de Sousa, I., Rebuzzi Vellasco, M., Costa da Silva, E., 2019. Local interpretable model-agnostic explanations for classification of lymph node metastases. Sensors 19, 2969.

[19] Panigutti, C., Perotti, A., Pedreschi, D., 2020. Doctor xai: an ontology-based approach to black-box sequential data classification explanations, in: Proceedings of the 2020 conference on fairness, accountability, and transparency, pp. 629–639.

[20] Payrovnaziri, S.N., Chen, Z., Rengifo-Moreno, P., Miller, T., Bian, J., Chen, J.H., Liu, X., He, Z., 2020. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. Journal of the American Medical Informatics Association 27, 1173–1185.

[21] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

[22] Ribeiro, M.T., Singh, S., Guestrin, C., 2018. Anchors: High-precision model-agnostic explanations, in: Proceedings of the AAAI conference on artificial intelligence.

[23] Roessner, V., Rothe, J., Kohls, G., Schomerus, G., Ehrlich, S., Beste, C., 2021. Taming the chaos?! using explainable artificial intelligence (xai) to tackle the complexity in mental health research.

[24] Sadr, N., de Chazal, P., 2018. A fast principal component analysis method for calculating the ecg derived respiration, in: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE. pp. 5294–5297.

[25] Torres, J.F., Hadjout, D., Sebaa, A., Martínez-Álvarez, F., Troncoso, A., 2021. Deep learning for time series forecasting: a survey. Big Data 9, 3–21.

[26] White, D.P., 2006. Sleep apnea. Proceedings of the American Thoracic Society 3, 124–128.