Making AI trustworthy in multimodal and healthcare scenarios

Ermanno Cordelli[1], Valerio Guarrasi[1], Giulio Iannello[1], Filippo Ruffini[1], Rosa Sicilia[1,*], Paolo Soda[1] and Lorenzo Tronchin[1]

[1]*Unit of Computer Systems and Bioinformatics, Department of Engineering*
University Campus Bio-Medico of Rome

Abstract

The pervasiveness of artificial intelligence in our daily lives has raised the need to understand and trust the outputs of the learning models, especially when involved in decision processes. As a result, eXplainable Artificial Intelligence has captured more and more interest in the scientific community, providing insights into the behaviour of these systems, ensuring algorithms fairness, transparency and trustworthiness. In this contribution we overview our work on the explainability of deep learning models applied to time series, multimodal data and towards extracting meaningful medical concepts.

Keywords

eXplainable Artificial Intelligence, Multimodal Learning explanations, medical concepts, multivariate time series

1. Introduction

Artificial Intelligence (AI) has proven to effectively support the decision process [1] and in particular deep learning techniques have achieved state-of-the-art performance [2]. Despite the impressive prediction accuracy attained in several applications, there is still the need to *explain* the decisions of the learning models proposed. As a result, eXplainable Artificial Intelligence (XAI) has captured more and more interest in the scientific community [3, 4, 5] since the complex nature of the models, such as deep neural networks, makes it impossible for the user to understand and validate the decision process. XAI aims to provide an insight into the behaviour and processes of these systems, ensuring algorithms fairness, identifying any potential bias in the training data and allowing complex AI models to be more transparent and understandable to humans [5].

Among a growing body of literature about XAI, in our laboratory we are directing our efforts toward three issues. The first concerns the explainability of Deep Learning (DL) models working on Time Series (TS) data. Indeed, the rising capabilities of storing and registering data have increased the number of temporal datasets, boosting the attention on TS classification models and raising the need to explain their decision. In this context, we present the application and evaluation of three XAI methods in a real-world multimodal task of anomaly detection on telematics data. We dealt with the challenge of explaining multivariate TS (MTS) and showing how to adapt different methodologies, originally designed for images, to this domain.

The second concerns the explainability of Multimodal Deep Learning (MDL) models, a topic at its infancy in the current literature. With the recent availability of a larger data repository, we expect to have the possibility to explore more complex deep architectures, studying how to learn shared representation and how to combine the unimodal networks. Nevertheless, more complex models exacerbate the problem of understanding what the predictions rely on, and also which modalities and features hold an important role [6].

Finally, the third direction we take leads toward ensuring trustworthiness and reliability specifically in the medical domain. Indeed, this is a field where XAI has a major impact, allowing both AI designers and medical experts to rely on meaningful explanations of the black box inner workings and reasoning, so that the decision of AI-based systems can be properly understood and adequately considered when applied in the real-world clinical context. However, in the medical field, identifying anatomical structures or tissue features that can be defined as relevant on an abstract scale is much more challenging and these elements may not be unambiguously defined. Therefore, it is essential to develop methods that can bridge this gap and provide more human-like explanations that users can trust and rely on.

2. Translating XAI to Multivariate Time Series

We take into account the challenge of explaining a real-world multimodal task of anomaly detection on telem-

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy

*Corresponding author.

r.sicilia@unicampus.it (R. Sicilia)
0000-0002-2513-0827 (R. Sicilia)

atics data from vehicles' black-box, where the available modalities are acceleration MTS and velocity univariate TS (UTS). Moreover, in this application there is also a supervised classifier trained to recognise if a crash event occurred or not. The peculiarity of MTS is that they are characterised by complex non-linear temporal dependencies between their attributes, i.e. the points of each UTS are connected with the other sequences via the time dimension. This key issue makes the development of an XAI approach for MTS-based anomaly detection rather challenging. The analysis of the literature shows that the study on XAI for MTS is limited: indeed, more efforts have been directed towards data diverse from TS, i.e. images and tabular data, not taking into account the complex relationship retained in multivariate time series. Thus as first contribution, we studied how to employ three XAI methods suited to explain models working on images and how to extend their application to a multimodal architecture working on telematics MTS acquired by car's black-boxes [7].

A further issue that we tackled in this work is evaluating the provided explanation for the MTS. Indeed, as highlighted in [8], different kinds of explanations produced after interpreting machine learning models may not be equally explainable, so a pressing need is emerging for quantifying the quality of the explanations produced, a topic that is only in its infancy in the current literature [8, 9].

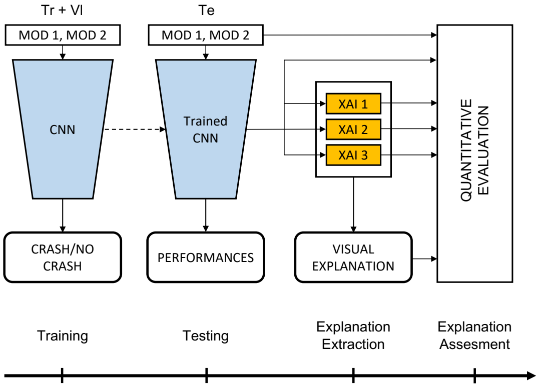

Figure 1: An overview of our approach. The classifier (CNN) is first trained using the training (Tr) and validation (Vl) sets of the dataset encompassing acceleration (MOD 1) and speed (MOD 2) signals of the vehicle black box. Then the trained model (Trained CNN) and its predictions on the test set (Te) are employed for extracting the explanations with the three different XAI methods.

2.1. Methods

Figure 1 shows an overview of the proposed approach that is able to explain a multimodal anomaly detection architecture identifying car crash events from telematics data of vehicles. We first train the black-box multimodal model for classification, which consists of a Convolutional Neural Network (CNN) that can learn local temporal-spatial patterns that are intuitive to visualise for the end-user. Then, the trained CNN, the test samples and the performance on the test set are used as inputs of the XAI framework to generate visual explanations of the decision provided by the model and to evaluate the quality of such explanation.

To consider both temporal and spatial relationships between each dimension of a MTS, we represent each sample as a 2D image where each pixel does not retain only visual features, such as the shape, the intensity and the texture, but also temporal features across each UTS included in the MTS. In this way we gain the advantage of analysing the MTS as a whole, leveraging the rich literature about the explainability on images, and of maintaining the flexibility to search for the best architecture and performance without the constrain of designing a model for explicitly achieving explainability.

Therefore, given the multimodal CNN initially designed, we investigated its explainability by employing three well-known XAI algorithms originally designed for images, providing a saliency map, which is an efficient way of pointing out what causes a certain outcome. The three approaches are: (i) an *agnostic solution* that is generalisable by definition to any model and returns a comprehensible local predictor, namely LIME [10]; (ii) a *model-specific solution*, namely Grad-CAM [11], designed explicitly for convolutional architectures that consider the CNN activation on the specific input sample; (iii) a *model-inspection approach*, namely IG [12], a method that derives explanation by examining the internal model behaviour when modifying the input sample. Indeed, as LIME aims to approximate the decision surface of a complex model using an interpretable one, the explanations from the surrogate models cannot be perfectly faithful with respect to the original model [13]. For this reason we also customized an XAI method suited for CNNs and another more general suited for deep architectures, namely Grad-CAM and IG, respectively. We therefore present how to customise and employ them to deal with a multimodal architecture working on multivariate telematic data [7].

Regarding the evaluation procedure, there is no consensus in the literature on methods to assess explainability [14]. In this respect, we customised the two strategies presented in [9] for UTS, by adapting them to the MTS domain. We exploited a perturbation-based XAI evaluation, measuring the performance drop Δ of the anomaly

Table 1
Each box reports the performance drop per XAI method and perturbation type. The results are in bold if $\Delta_{XAI} > \Delta_{random}$.

		Perturbation Type		
		Zero	Swap	Mean
Drop	**Drop Grad-CAM**	**20.1 %**	6.5 %	2.7 %
	Drop IG	**58.2 %**	**15.2 %**	**5.5 %**
	Drop LIME	**54.3 %**	13.8 %	**10.0 %**
	Drop Random	0.7 %	14.3 %	3.6 %

detection system respectively disrupting the time points identified by the XAI method as relevant (Δ_{XAI}), and random regions of the signal (Δ_{random}). The assessment is based on the assumption that if relevant/random features (time points) get changed, the model's performance should decrease/stagnate. Finally, as we aimed to conduct an exhaustive evaluation, we performed the assessment procedure considering all test set samples in an hold-out setup.

2.2. Main results

Overall, the explanations obtained reasonably find the crucial features used from the model to perform the anomaly detection task, casting light on the black-box's decision process. IG and Grad-CAM are able to exploit the cross-correlation in UTS learned from the CNN, namely defining which UTS is most valuable for the prediction. Instead, LIME explanation does not present this evidence since it uses a surrogate model to approximate the CNN, being an agnostic method, so it does not directly inspect the convolutional architecture's inner workings. To provide an exhaustive comparison between the three methods, Table 1 shows the results emerging from XAI quantitative evaluation. From the results, IG-based perturbations account for the most significant drop in performance and it is the only method that always exceeds the random drop. Hence, the quantitative analysis suggests that IG is the more informative explainability method for the anomaly detection task as it is able to detect the time points valuable for the CNN to perform the prediction. In contrast, Grad-CAM results in the least reliable algorithm from the quantitative evaluation as two out of three times it is overtaken by random drop. Moreover, we find the temporal trend to have a minor influence for the CNN to perform the classification task as we observe from the drops in the values of *swap* and *mean* metrics compared to the *zero* one.

In general, this study provides insight into the quality of explanation and sheds light on the most significant features that are exploited by the CNN when it performs the crash detection task.

2.3. Challenges and perspectives

This work represents a first attempt to tailor XAI algorithms to the multimodal nature of the data, suggesting further research in this field. As a first direction, we will investigate XAI methods able to provide a more human-interpretable representation, since the saliency maps provided by all the XAI methods are hard to be interpreted by users. The second direction will focus on developing a multimodal XAI method able to explain both signals available in the telematics data at hand (i.e. acceleration and speed).

2.4. Papers and available resources

This work was published in [7].

3. Multimodal XAI

Multimodal Deep Learning (MDL) studies how deep neural networks can learn shared representations between different modalities that, used together, may offer deeper insights into the data. It investigates when to fuse the different modalities and how to obtain more powerful data representations. Linking information coming from various modalities can leverage the understanding of a field of study. These considerations are of particular relevance for the field of bio-medicine, where MDL has proven to be useful [15].

Among the different challenges of MDL, we focus here on supervised multimodal fusion applied to early identify patients at risk of the severe outcome, like intensive care or death, among those affected by SARS-CoV-2, and using chest X-ray (CXR) scans and clinical data. Instead of using manually designed or handcrafted modality-specific features, via DL we can automatically learn and extract an embedded representation for each modality, which is then fused with the others in an end-to-end training process that exploits the loss backpropagation.

It is well-known that the major disadvantage of neural networks is their lack of interpretability. In spite of the importance of MDL, to the best of our knowledge, no work has studied the XAI methods for fused modalities, particularly in the medical field. Hence, we developed a deep architecture, explainable by design, which jointly learns modality reconstructions and sample classifications of the aforementioned biomedical multimodal data, i.e., imaging and tabular data.

3.1. Methods

The proposed architecture is shown in Figure 2. The explanation of the decision taken is computed by applying a latent shift that simulates a counterfactual prediction revealing the features of each modality that contribute the

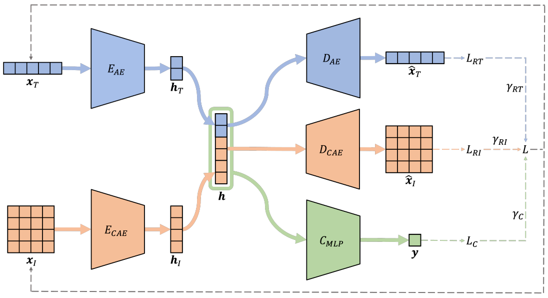

Figure 2: Schematic view of the multimodal deep architecture.

most to the decision and a quantitative score indicating the modality's importance. The latent shift is applied on an embedded representation of the data, h in Figure 2, created by exploiting a Convolutional Autoencoder (CAE) for the imaging modality and an Autoencoder (AE) for the tabular modality, which is connected in an end-to-end manner to a multi-layer perceptron which performs the classification task for the prognosis of the severity of the COVID-19 virus. Exploiting the nature of the model's architecture the variations on the embedded feature vector help us understand how the classifications are performed by the MLP, via the echoed perturbations on the embedded vector and on the CAE's and AE's reconstructions. The explanation of the decision taken is computed by applying a latent shift that, simulates a counterfactual prediction revealing the features of each modality that contribute the most to the decision and a quantitative score indicating the modality's importance.

3.2. Explanation assessment

To study the validity of the proposed method, we conducted a reader study with four radiologists assessing the prognosis of a subset of patients. Each radiologist observed both data modalities simultaneously for each patient and performed the prognosis task. Afterward, the radiologists attributed an importance score, on a scale from 1 to 5, indicating how significant each modality was for the prognosis task. Then, to understand the most important features for each modality, we asked the radiologist to select the clinical variables and to segment the areas of interest in the X-ray image, most useful to stratify the patient. The sanity check, although very time-consuming, was very useful since it showed a high intersection between the explanations provided by the method and those of the radiologists, both for the modality and the feature importance. By reducing the model's opacity which makes it difficult for doctors and regulators to trust the MDL models, we are able to improve trust and transparency.

3.3. Challenges and perspectives

In this work we presented a first attempt to build a transparent end-to-end multimodal architecture that jointly learns modality reconstructions and multimodal classification using tabular clinical and imaging data. The still open challenges in this domain suggest a few directions for our future work: (i) to test our architecture and Multi-modal XAI method on other datasets; (ii) to extend and validate the proposed latent-shift approach on more modalities, e.g. text of clinical records; (iii) to tackle the problem of missing modalities especially from the explanation view point.

3.4. Papers and available resources

This work is available in [16].

4. Towards eXplainable Medical Concepts

Despite the fact that XAI has revolutionized the way we see machine learning models, there are still several limitations with respect to providing explanations closer to human perceptiveness, resulting in users not being able to fully trust the model working [6]. In this context, concept attribution methods have emerged as a new paradigm, providing interpretations of the inner mechanisms of DL methods by measuring the relevance of human-friendly concepts directly defined by the user [17]. In Computer Vision (CV) field the definition of semantic concepts is simple and intuitive: it is based on key aspects contained in images that users can easily relate to real-life contexts (e.g. ear shapes that are much more similar to a dog than a cat). However, in the medical field, identifying anatomical structures or tissue features that can be defined as relevant on an abstract scale is much more challenging and it may not be easy to define them unambiguously.

Therefore we are tackling this open issue focusing on exploring unsupervised approaches to automatically extract meaningful medical semantic concepts. Specifically, we are investigating the effectiveness of deep clustering methods that perform representation learning through Convolutional Autoencoders and clustering simultaneously, operating on the latent space learned from the raw data distribution [18]. This approach has the advantage of producing features that are highly correlated with the image structures, providing semantically meaningful groups of images in the dataset. The key idea is to gain new insights about the data, studying a novel approach able to extract from the deep latent space of the Autoencoder the set of features with the highest possible conceptual expressiveness.

4.1. Methods

In the first stage, concept extraction is performed using the Deep Clustering algorithm. This algorithm constructs a low-level representation of the dataset (called latent space H) and clusters the samples into *k* clusters in the same H-space. To obtain the best spatial representation of the hidden concepts within the dataset, we implemented an iterative process, involving training the algorithm for several initialization of the number of clusters into which to separate the samples of the dataset.

The overall goal is to obtain a concept extraction process based on images structural features (patterns of pixels) common to the largest possible number of patients available in the dataset, so achieving a clustering configuration that most generalizes over the patients distribution. Hence, to determine the optimal value of *k* we employed two sets of metrics: (i) cluster-based metrics, which validate the clustering conditions (Silhouette-score, Davies-Bouldin, Calinski-Harabasz); (ii) patient-based metrics, which are custom metrics designed to capture how the data splits up according to each patient's distribution. Patient-based metrics take into account the global variations between each different clustering configuration and are used to identify the ideal conditions for semantic clustering of patients. The optimal value of *k* is, then, selected based on the best results obtained for these metrics.

4.2. Validation strategy

In order to validate the effectiveness of the proposed concept extraction model we will adopt a comparative strategy applied on a binary task for overall survival prediction on a cohort of 191 patients with Non-small cell lung cancer, including both CT volumetric images (total of 22384 slices) and clinical data. First, we will build a baseline for comparison: we will train a set of classical machine learning classifiers on the clinical data computing their performance for the desired task. Second, we will train the proposed deep clustering model on the set of CT scans, extracting a latent vector for each patient. Third, this vector will be used in early fusion with the clinical data to train the same set of classifiers used as baseline. Comparing the performance with and without the deep clustering concept vectors will allow to check the effectiveness of the approach, testing its ability to extract meaningful information.

4.3. Challenges and perspectives

In the analysis of biomedical datasets, this work shifts the focus from dataset construction towards maximizing data information content. This is accomplished by identifying patterns of pixel structures within the images that define similarity at the semantic level, employing unsupervised DL models. The ultimate goal is to develop an automated system for the unbiased evaluation of images, using XAI tools to explain why images were grouped together. This approach represents a new perspective for the analysis of biomedical data sets and it has the potential to significantly advance the field. Using unsupervised DL models, it is possible to discover patterns within the data that may not be evident to human analysts. Furthermore, by using XAI tools to explain clustering results, the system can provide insights into underlying biology and pathology that are not readily available through traditional image analysis methods. Overall, this work represents an exciting development that has the potential to have a significant impact on basic and clinical research.

Acknowledgments

The work detailed in section 2 was realized in cooperation with Consorzio ELIS (Consel) and Generali Italia S.p.A.

References

[1] T.-c. Fu, A review on time series data mining, Engineering Applications of Artificial Intelligence 24 (2011) 164–181.

[2] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, P.-A. Muller, Deep learning for time series classification: a review, Data mining and knowledge discovery 33 (2019) 917–963.

[3] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of Methods for Explaining Black Box Models, ACM computing surveys (CSUR) 51 (2018) 1–42.

[4] M. Du, N. Liu, X. Hu, Techniques for Interpretable Machine Learning, Communications of the ACM 63 (2019) 68–77.

[5] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining Explanations: An Overview of Interpretability of Machine Learning, in: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA), IEEE, 2018, pp. 80–89.

[6] G. Joshi, R. Walambe, K. Kotecha, A review on explainability in multimodal deep neural nets, IEEE Access (2021).

[7] L. Tronchin, R. Sicilia, E. Cordelli, L. R. Celsi, D. Maccagnola, M. Natale, P. Soda, Explainable ai for car crash detection using multivariate time series, in: 2021 IEEE 20th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC), IEEE, 2021, pp. 30–38.

[8] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608 (2017).

[9] U. Schlegel, H. Arnout, M. El-Assady, D. Oelke, D. A. Keim, Towards a rigorous evaluation of xai methods on time series, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), IEEE, 2019, pp. 4197–4201.

[10] M. T. Ribeiro, S. Singh, C. Guestrin, "Why Should I Trust You?" Explaining the Predictions of Any Classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

[11] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[12] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 3319–3328.

[13] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206–215.

[14] C. Molnar, Interpretable machine learning, Lulu. com, 2020.

[15] D. Ramachandram, G. W. Taylor, Deep multimodal learning: A survey on recent advances and trends, IEEE signal processing magazine 34 (2017) 96–108.

[16] V. Guarrasi, L. Tronchin, D. Albano, E. Faiella, D. Fazzini, D. Santucci, P. Soda, Multimodal explainability via latent shift applied to covid-19 stratification, arXiv preprint arXiv:2212.14084 (2022).

[17] M. Graziani, V. Andrearczyk, S. Marchand-Maillet, H. Müller, Concept attribution: Explaining cnn decisions to physicians, Computers in biology and medicine 123 (2020) 103865.

[18] E. Aljalbout, V. Golkov, Y. Siddiqui, M. Strobel, D. Cremers, Clustering with deep learning: Taxonomy and new methods, arXiv:1801.07648 (2018).