# Explain and Interpret Few-Shot Learning*

Andrea Fedele[1,2,*]

[1]Computer Science Department, University of Pisa, Italy
[2]KDD Laboratory, ISTI, National Research Council, Pisa, Italy

**Abstract**

Recent advancements in Artificial Intelligence have been fueled by vast datasets, powerful computing resources, and sophisticated algorithms. However, traditional Machine Learning models face limitations in handling scarce data. Few-Shot Learning (FSL) offers a promising solution by training models on a small number of examples per class. This manuscript introduces FXI-FSL, a framework for eXplainability and Interpretability in FSL, which aims to develop post-hoc explainability algorithms and interpretable-by-design alternatives. A noteworthy contribution is the SIamese Network EXplainer (SINEX), a post-hoc approach shedding light on Siamese Network behavior. The proposed framework seeks to unveil the rationale behind FSL models, instilling trust in their real-world applications. Moreover, it emerges as a safeguard for developers, facilitating models fine-tuning prior to deployment, and as a guide for end users navigating the decisions of these models.

**Keywords**

Few-Shot Learning, Explainable Artificial Intelligence, Interpretable Machine Learning, Siamese Networks

## 1. Context

In recent years, Artificial Intelligence (AI) has made significant progress due to the availability of large datasets, powerful computing devices, and the development of sophisticated algorithms [1]. Machine learning (ML) models are commonly used in AI systems because of their success in various fields such as image processing, time series analysis, and audio signal processing [2, 3]. However, traditional ML systems have a major limitation in that they rely on large-scale datasets, while real-world applications often have constraints that result in limited data availability [4, 5]. Technical issues may limit the collection of training data, while ethical or privacy concerns may restrict data access [6]. Furthermore, traditional ML systems struggle to generalize from few samples and their performance is often better for classes with more training samples and worse for classes with fewer samples [7]. As a result, these systems are limited in their ability to expand their knowledge beyond the scope of the data they were trained on. In contrast, in [8] the authors show that humans can recognize a Segway despite the fact that they have seen it only once prior in their life.

To overcome these limitations, recent studies proposed the use of *few-shot learning* (FSL), where a ML model must learn to predict the class of a given instance when only a small number

*Late-breaking work, Demos and Doctoral Consortium, colocated with The 1st World Conference on eXplainable Artificial Intelligence: July 26–28, 2023, Lisbon, Portugal*

*Corresponding author.
andrea.fedele@phd.unipi.it (A. Fedele)
0009-0007-0467-0967 (A. Fedele)

of examples of that specific class are available in the training set [9]. One-shot learning is a specific case of FSL where there is only one labeled sample per class, while zero-shot learning aims to predict the class of an instance without prior exposure to that class. Unfortunately, the biggest downside of such systems is the lack of explainability. Understanding the reason why a model takes a specific decision is hugely important to developers, organizations, and end-users on which such decision falls upon. While end-users may prioritize understanding an outcome's explanation over the outcome itself, developers can use explanations to identify potential issues with a model and repeat training procedures in a controlled environment. In recent years, researchers examined the eXplainable Artificial Intelligence (XAI) topic from various perspectives [10, 11]. The relevance of XAI in the context of FSL is easy to see since it could expose whether or not machine think as humans, potentially unveiling human reasoning and its brain interactions, that could be implicit, inherited by prior knowledge and possibly based on inductive inference [12, 13]. More realistically, XAI in FSL would reveal how these kind of systems mimic human learning.

## 2. Related Work

Only recently in [9], the authors introduced a taxonomy of FSL methods dividing them in: *(a)* model-based, *(b)* data-based and *(c)* algorithm-based approaches. While model-based approaches use prior knowledge to constrain the complexity of the hypothesis space so to operate on a smaller one, data-based approaches use it to augment the training dataset increasing the number of training samples. Algorithm-based approaches, on the other hand, use prior knowledge to search for the best parameter configuration settings which defines the best hypothesis in the hypothesis space. FSL paradigm has been employed in various fields ranging from image recognition [14], time series forecasting [15], object recognition [16] and short-text sentiment classification [17]. Considering the convolution structure of different FLS methods [14, 18, 19, 20] and the dataset availability, *few-shot image classification* applicability has been prolific.

Due to the different natures of methodologies, the literature of research projects with the aim of producing interpretable FSL models is quite limited. One research direction points at re-designing the ML systems architecture, so to result in models *interpretable by design* [21, 22]. Such models are natively interpretable in the sense that they come equipped with an accurate global and/or local view of the model behaviour, which is typically learned during the training phase of the system itself. A compelling example in this context is presented in [23], where the authors extend the capabilities of the explainable classifier SCOUTER [24] by introducing a novel *interpretable-by-design* technique for FSL classification. Their approach involves identifying shared patterns between an unseen image and a previously seen set of images. This is achieved through a self-attention mechanism that learns discriminative patterns, which can then be utilized for pairwise matching during classification. Other approaches seem to employ traditional XAI techniques to explain the convolutional networks behaviour commonly built in different *embedding-learning* models [25, 26]. Such techniques can result very limited since they are tied to the architecture and tend to explain only a portion and not the entire architecture of the system. A different approach explored in [27] trains an additional special auto-encoder on the training dataset and uses it as core to the explanation algorithm. However, such an

approach ties the model interpretability to the dataset availability. It is important to note that current state-of-the-art works on XAI in the context of FSL often lack a comparison with well-known and widely used agnostic explainability techniques such as SHAP [28] or LIME [29]. This omission may be attributed to the challenges in a direct application of these techniques due to the specific constraints that FSL imposes.

**Challenges.** Due to the recent introduction of the FSL paradigm, the literature is full of application which may or may not follow a standard formulation of a FSL problem. Many authors refer to Few-Shot Learning, indicating that *few* are the samples of a given class available in the training set. Others, frame the problem considering that *few* are the samples available at inference time instead. The lack of common standards in the problem definitions obviously leads to various FSL application resulting in the need of different designs in interpretability and explainability methodologies. A huge preliminary effort is required in defining the setting where FSL paradigm is applied, in order to then define a standardized toolkit that can help explain and understand the different approaches of FSL. Moreover, it is crucial to conduct a technical assessment in order to compare new proposals with existing agnostic explainers. One hypothesis is that available explainers could be employed in *data-based* models, as these methods involve augmenting the training data in various ways. While current techniques may be applicable out-of-the-box, their effectiveness in the FSL context still needs to be demonstrated. On the other hand, *model-based* methods, ranging from *embedding-learning* to *generative modeling*, present distinct challenges. A concrete example is the limitation of using LIME out of the box with *embedding-learning* Siamese Networks, as noted by the authors of the tool itself[1]. This limitation arises from the lack of default support for multiple inputs, requiring workarounds to concatenate and separate inputs in the prediction function while correctly segmenting them through additional implementations. Although current techniques may prove to be ineffective in *algorithm-based* FSL learning, unable to distinguish whether a decision relies more on preliminary model knowledge or the fine-tuning process itself, experiments are still necessary to verify this.

## 3. Research questions and Approach

This project aims to propose a Framework to eXplain and Interpret Few-Shot Learning methods, namely FXI-FSL, considering the 3-tier taxonomy partition of *model-based*, *data-based* and *algorithm-based* given in [9]. Such framework will analyze the FSL problem w.r.t. two different perspectives: *(i)* develop accessible *post-hoc* explainability algorithm to explain case-specific, inner-family and potentially cross-family FSL methods, and *(ii)* propose novel *interpretable-by-design* alternatives to common FSL architecture by keeping in mind the trade-off between minimization of adjustments and performance.

1. Can explainability and interpretability technique be developed for FSL?
    a) Are FSL approaches open to adjustments that introduce the concept of intrpretability by design?
    b) Are there specific FSL methodologies which tend to be post-hoc explained?

[1] https://github.com/marcotcr/lime/issues/459

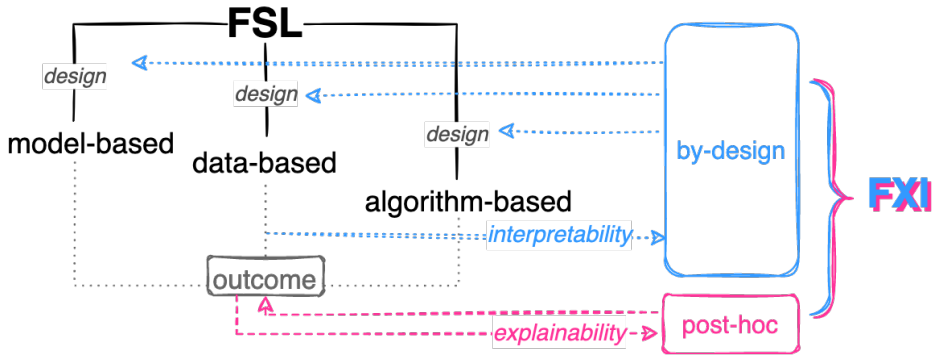


**Figure 1:** FXI-FSL Scheme. Interpretability by-design operates during the design phase of FSL. Post-hoc explainability does not involve architectures re-design, but mostly aims at explaining the outcomes given by FSL traditional models. By-design and post-hoc approaches both cover all FSL methods, regardless of what they are based upon.

   c) What consists of a comprehensive and meaningful explanation in FSL?
2. Can traditional explainability techniques be suitable for FSL applications?
3. How to implement explainable post-hoc and interpretable-by-design approaches?

The FXI-FSL overall scheme is depicted in Figure 1. Interpretability *by-design* operates during the design phase of FSL methods, looking for either conceptual or implementation small changes to attach the element of interpretability to the actual FSL methodology. *Post-hoc* explainability, on the other hand, does not involve any change to the methodology since it aims at explaining the outcomes given by FSL traditional models only after they have been design, trained and potentially deployed. It should be noted that while both *by-design* and *post-hoc* approaches address all FSL methods regardless of the component they are based upon (i.e., model-data or algorithm), they do not overlap and they are not necessarily consecutive. Some scenarios might in fact benefit from the usage of both approaches on a given FSL method, consolidating further FXI-FSL results. Further assessment is required for currently available explainability techniques before their inclusion in the proposed framework. For instance, it is necessary to investigate whether the limitations of LIME are specific to *model-based* FSL models, such as the absence of support for multiple inputs in Siamese Networks, necessitating workarounds in both prediction and segmentation functions. Additionally, questions such as "what background set would SHAP utilize in FSL to learn shapely values if the training set is unavailable?" need to be addressed to ensure the complete integration of traditional XAI techniques in FXI-FSL.

## 4. Preliminary contributions

The first contribution to the FXI-FSL framework is a *SIamese Network EXplainer*, namely SINEX[2], that was introduced in [30] in the context of FSL on audio input data. Siamese Networks (SNs) [14] belong to the *embedding-learning*, which is a sub-family of *model-based* FSL, where

[2]https://github.com/andreafedele/SINEX

traditional ML architectures reduce the input in a smaller embedding space in which similar and dissimilar samples can be easily discriminated. SNs are composed of two or more identical encoding sub-networks that map inputs into an embedding space, where a distance function is applied to calculate the distance between the resulting embedded representations. A similarity score is then calculated based on this distance.

SINEX is a post-hoc local explanation method which computes explanations by means of a perturbation-based approach evaluating instance's *segment-weighted-average* contribution values to the final outcome. These contribution values can be visualized as heatmaps, providing an intuitive representation of the behavior of Siamese Networks. In this work we experimented the effectiveness of the explainer on *AudioMNIST* and *ESC-50* audio dataset, considering the log-mel spectrogram representation of the signal. Morover, the FSL problem setting was framed so to have training, validating and test set as disjoint set of classes. The SNs were therefore asked to classify instances of classes that were never seen neither during training nor during the validating phase, relying only on one sample of each class.

Results on *AudioMNIST* illustrate that the correct classification of female speaker recordings is mainly due to medium-high frequency segments, while their miss-classification depends primarily on segments that reside at the very bottom of the frequency range. A symmetrical behavior is observed for male speakers' audios: a correct classification is usually based on lower frequency values, while incorrect classifications are generally due to segments higher in the frequency spectrum. The application of our method on *ESC-50* extracts different insights, first and foremost the SN inability to discriminate on medium-high frequencies between problematic classes. Experiments also led us to think that the decay between sound events plays a big role, especially if such events are repeated frequently in the overall recording.

## 5. Research Direction and Next steps

Current research direction is moving towards: (i) improve SINEX perturbation technique, (ii) extend the supported data types to images and time-series, (iii) extend compliance with 3-branched SNs architectures and (iv) extend support for distance-based SNs, other than similarity-based ones. The perturbation approach previously described used in [30], involves measuring the contribution of a specific segment by keeping it active while "silencing" all the other segments, i.e., replacing them with non-informative values. On the other hand, the second approach we aim at introducing, measures the contribution of a specific segment by "silencing" it, while keeping all other segments active. Extensions on the data-type support would allow us to offer a data-agnostic explainer, easily twickable via parameter configuration. This would ensure different application and experimentation. For example, the replacing value to use while "silencing" out part of the input during perturbation, might be grey, violet or any other color for RGB images. Similarly, one could choose between different decibel value for audio-inputs. Furthermore, distance-based architecture can be supported with minor adjustment both during the segment-contribution computation and the heatmap generation. Finally, to ensure the complete coverage of Siamese Networks, 3-branched architecture (3SNs) need a little more integration in SINEX. Preliminary experiments demonstrated that the explainer can help discover limitations that SNs might encounter, such as the erroneous dependence on specific

colors (on RGB images) or pixels (on grayscale images) that should not be considered important. Therefore, SINEX provides an effective tool to highlight such limitations and guide a subsequent model re-training phase.

Future research directions will focus on various aspects. First, since a limitation of SINEX is that it can only study the SN behavior locally on each few-shot task, requiring human oversight in multiple analyses of different tasks to get a comprehensive understanding of the network's global behavior, inspired by [31] we aim at proposing a local-to-global abstraction of the logic learned by SN. Also, since many of the other existing FSL methods belonging to the *embedding-learning* sub-family use CNNs as embeding functions [18, 20, 19, 32, 33, 34], the extention of our proposed methodology might be studied. By doing so, it would be possible to understand whether or not a common post-hoc methodology can be an inner-family explainability solution for the model-based FSL branch, regardless of small algorithm adjustments and different input data types.

Two other research directions of this project are aimed at exploring how to interpret and explain *data-based* and *algorithm-based* FSL approaches. While the former employs prior knowledge to typically transform samples and augment the training data, the latter alters the search strategy in the hypothesis space. Since both of these families mainly operate during the training phase, the hypothesis is that introducing some form of *interpretability-by-design* would be a prudent choice whenever training data is accessible.

FXI-FSL will help scientist explore to which extent machine think as humans answering questions like: *Why is the model classifying a query input character as belonging to a specific alphabet instead of another? [14] What is the importance of a motorbike image's specific shape to be recognized as such? [16].* The framework will help gaining trust in FSL models so to deploy them safely in real-world applications. In scenarios where technical or privacy issues lead to few training labeled samples [4, 5, 35], huge benefits could come from both local or global explanations that highlight discriminative inner-class and intra-class features. Interpretable FSL would be valuable for various subjects with possibly different backgrounds (i.e., developers, end-users). In FSL intrusion detection in railway video surveillance [36], FXI-FSL would help developers fine-tune the algorithm when mistaking a bird as an intruder. Similarly, a railway employee would be assisted when deciding whether or not to close a specific route-segment based on the tool insight. To conclude, FXI-FSL would benefit medical experts analysis in FSL medical image classification [37] and, in general, such a framework might be both a safeguard and a guideline whenever FSL is involved.

## Acknowledgments

This work is supported by the European Community under the Horizon 2020 programme: 'G.A. 871042 *SoBigData++*, G.A. 952026 *HumanE AI Net*, ERC-2018-ADG G.A. 834756 *XAI*, G.A. 952215 *TAILOR*, CHIST-ERA grant *CHIST-ERA-19-XAI-010 SAI*, FWF (I 5205), EPSRC (EP/V055712/1), NCN (2020/02/Y/ST6/00064), ETAg (SLTAT21096), BNSF (KP-06-AOO2/5), and the NextGenerationEU programme under the funding schemes PNRR-PE-AI scheme (M4C2, investment 1.3, line on AI) *FAIR* (Future Artificial Intelligence Research), and "*SoBigData.it* - Strengthening the Italian RI for Social Mining and Big Data Analytics" - Prot. IR0000013.

## References


[1] M. Haenlein, et al., A brief history of artificial intelligence: On the past, present, and future of artificial intelligence, Cal. Manag. Review 61 (2019) 5–14.

[2] I. H. Sarker, Machine learning: Algorithms, real-world applications and research directions, SN Comput. Sci. 2 (2021) 160.

[3] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, T. Sainath, Deep learning for audio signal processing, IEEE Journal of Selected Topics in Signal Processing 13 (2019) 206–219. doi:10.1109/JSTSP.2019.2908700.

[4] L. Jiang, D. Meng, T. Mitamura, A. G. Hauptmann, Easy samples first: Self-paced reranking for zero-example multimedia search, in: Proceedings of the 22nd ACM international conference on Multimedia, 2014, pp. 547–556.

[5] J. Fries, S. Wu, A. Ratner, C. Ré, Swellshark: A generative model for biomedical named entity recognition without labeled data, arXiv preprint arXiv:1704.06360 (2017).

[6] M. Ienca, et al., On the responsible use of digital data to tackle the covid-19 pandemic, Nature medicine 26 (2020) 463–464.

[7] S. Rahman, S. H. Khan, F. Porikli, Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts, CoRR abs/1803.06049 (2018). URL: http://arxiv.org/abs/1803.06049. arXiv:1803.06049.

[8] S. Roy, S. Herath, R. Nock, F. Porikli, Machines that learn with limited or no supervision: A survey on deep learning based techniques (2017).

[9] Y. Wang, et al., Generalizing from a few examples: A survey on few-shot learning, ACM Comput. Surv. 53 (2020) 63:1–63:34.

[10] R. Guidotti, et al., A survey of methods for explaining black box models, ACM Comput. Surv. 51 (2019) 93:1–93:42.

[11] A. Adadi, et al., Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160.

[12] D. Hassabis, D. Kumaran, C. Summerfield, M. Botvinick, Neuroscience-inspired artificial intelligence, Neuron 95 (2017) 245–258.

[13] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, N. D. Goodman, How to grow a mind: Statistics, structure, and abstraction, science 331 (2011) 1279–1285.

[14] G. Koch, R. Zemel, R. Salakhutdinov, et al., Siamese neural networks for one-shot image recognition, in: ICML deep learning workshop, volume 2, Lille, 2015.

[15] T. Iwata, A. Kumagai, Few-shot learning for time-series forecasting, arXiv preprint arXiv:2009.14379 (2020).

[16] L. Fei-Fei, R. Fergus, P. Perona, One-shot learning of object categories, IEEE transactions on pattern analysis and machine intelligence 28 (2006) 594–611.

[17] M. Yu, X. Guo, J. Yi, S. Chang, S. Potdar, Y. Cheng, G. Tesauro, H. Wang, B. Zhou, Diverse few-shot text classification with multiple metrics, arXiv preprint arXiv:1805.07513 (2018).

[18] O. Vinyals, et al., Matching networks for one shot learning, in: NIPS, 2016.

[19] F. Sung, et al., Learning to compare: Relation network for few-shot learning, in: CVPR, Computer Vision Foundation, 2018, pp. 1199–1208.

[20] J. Snell, et al., Prototypical networks for few-shot learning, in: NIPS, 2017.

[21] W. Tang, et al., Interpretable time-series classification on few-shot samples, in: IJCNN,

IEEE, 2020, pp. 1–8.
[22] S. Lee, J. Gonzalez, M. Wright, Interpretable few-shot image classification with neural-backed decision trees (2020).
[23] B. Wang, L. Li, M. Verma, Y. Nakashima, R. Kawasaki, H. Nagahara, Mtunet: Few-shot image classification with visual explanations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2294–2298.
[24] L. Li, B. Wang, M. Verma, Y. Nakashima, R. Kawasaki, H. Nagahara, Scouter: Slot attention-based classifier for explainable image recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1046–1055.
[25] Y. Zhang, et al., Siamese style convolutional neural networks for sound search by vocal imitation, IEEE ACM Trans. Audio Speech Lang. Proc. (2019) 429–441.
[26] M. Acconcjaioco, et al., One-shot learning for acoustic identification of bird species in non-stationary environments, in: ICPR, IEEE, 2020, pp. 755–762.
[27] L. V. Utkin, et al., Explanation of siamese neural networks for weakly supervised learning, Comput. Informatics 39 (2020).
[28] S. M. Lundberg, S. Lee, A unified approach to interpreting model predictions, CoRR abs/1705.07874 (2017). URL: http://arxiv.org/abs/1705.07874. arXiv:1705.07874.
[29] M. T. Ribeiro, S. Singh, C. Guestrin, "why should I trust you?": Explaining the predictions of any classifier, CoRR abs/1602.04938 (2016). URL: http://arxiv.org/abs/1602.04938. arXiv:1602.04938.
[30] A. Fedele, R. Guidotti, D. Pedreschi, Explaining siamese networks in few-shot learning for audio data, in: Discovery Science: 25th International Conference, DS 2022, Montpellier, France, October 10–12, 2022, Proceedings, Springer, 2022, pp. 509–524.
[31] M. Setzu, R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, F. Giannotti, Glocalx - from local to global explanations of black box AI models, Artif. Intell. 294 (2021) 103457.
[32] J. Choi, J. Krishnamurthy, A. Kembhavi, A. Farhadi, Structured set matching networks for one-shot part labeling, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3627–3636.
[33] Y. Liu, J. Lee, M. Park, S. Kim, Y. Yang, Transductive propagation network for few-shot learning (2018).
[34] N. Mishra, M. Rohaninejad, X. Chen, P. Abbeel, A simple neural attentive meta-learner, arXiv preprint arXiv:1707.03141 (2017).
[35] W. Naudé, Artificial intelligence vs covid-19: limitations, constraints and pitfalls, AI & society 35 (2020) 761–765.
[36] X. Gong, X. Chen, Z. Zhong, W. Chen, Enhanced few-shot learning for intrusion detection in railway video surveillance, IEEE Transactions on Intelligent Transportation Systems (2021).
[37] A. Cai, W. Hu, J. Zheng, Few-shot learning for medical image classification, in: International Conference on Artificial Neural Networks, Springer, 2020, pp. 441–452.