ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Evaluation of Rashomon Sets for Determination of Stable and Plausible Model Explanations

M. Kaden[1], M. Karimi[1], S. Panda[1], T. Pfaff[1], and T. Villmann[1,2] *

1- Saxon Institute for Computational Intelligence and Machine Learning (SICIM), Mittweida University of Applied Sciences, Mittweida, Germany

2- Technical University Bergakademie Freiberg, Freiberg, Germany

Abstract. Training of machine learning models for classification frequently yields several different solutions although the performance remains approximately the same, i.e. one observes many close-to-optimum solutions with only marginal performance differences which are, however, qualitatively well-distinguishable. This behaviour is known as the Rashomon effect and may be dedicated to the stochastic in the training process, different learning strategies or various initial settings. Hence, model explanations may become difficult and have to be related to a given configuration. Therefore, stable and plausible explanations are required based on the evaluation of the Rashomon set. Yet, the consistency of the resulting explanations remained largely unexplored so far.

Here we propose to evaluate the Rashomon set qualitatively by means of a cluster analysis based on the determination of the feature importance. Feature importance of a model gives insights about the decision making process and, hence, provides an appropriate criterion to distinguish model decision realizations. Clustering of them reveal stable and plausible classification strategies and, hence, contribute to reliable explanations.

1 Introduction

Nowadays complex deep networks become the standard to solve successfully difficult and challenging classification and regression tasks frequently realized by deep multi-layer perceptrons (MLP). However, frequently these models work as black-box approaches. To understand the behavior and the decision process of these models, a well-accepted approach in explainable artificial intelligence (XAI) is to analyze input-output-relations with respect to the feature influence. There are many approaches available with different strategies for feature relevance quantification like correlation analysis, Shapley values obtained according to methods of cooperative game theory or feature relevance propagation. These approaches are primarily but not exclusively designed for deep feed-forward networks to analyze model decision making by providing the importance of input features through the decomposition of the prediction output backward [1, 2, 3, 4]. Further, respective feature evaluations are used to derive reasoning models to explain causal inference for decision making for complex models [5, 6, 7].

However, multiple different models can achieve nearly identical performance close to an generally unknown optimum solution, which is known as the Rashomon effect [8, 4, 9, 10, 11]. This effect may lead to conflicting evaluations

*This work is financially supported by European Social Fund (ESF, 100734114, 100715238).

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

of the models and also influences the rating, which variables are most important. A variable deemed important in one model may be irrelevant in another equally accurate model.

These findings motivate a key question: if many different models achieve similar accuracy, do they also agree on feature importance, i.e. which models can be explained similarly and yield stable and plausible interpretations of the decision making process? This challenge requires a reliable assessment of the corresponding Rashomon set [12]. Hence, we propose to analyze the Rashomon set by means of cluster analysis with respect to feature classification importance. Thereby, we concentrate in this initial study to very shallow classifier models, which, however, often provide sufficiently accurate models [13].

2 Rashomon sets and Model Explanations

We consider data $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^n \times \mathbb{R}$ following an unknown distribution $\mathscr{D}$. A multilayer perceptrons (MLP) realizes a function $f_{\mathcal{P}} : \mathcal{X} \to \mathcal{Y}$ depending on the parameter set $\mathcal{P}$ consisting of the set of weights $\mathcal{W}$ and biases $\mathcal{B}$. Following [13], we define the hypothesis space as $\mathcal{F} = \{f_{\mathcal{P}} | \mathcal{W}, \mathcal{B}\}$. A model $f_{\mathcal{P}}$ is evaluated by a loss function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$. The respective empirical risk is given as

$$E_{\mathcal{Z}}\left(f_{\mathcal{P}}\right) = \mathbb{E}_{z=(\mathbf{x},y)\in\mathcal{Z}\propto\mathscr{D}}\left[L\left(\hat{y} = f_{\mathcal{P}}\left(\mathbf{x}\right), y\right)\right].$$

It is well-known that complex machine learning models like deep MLP deliver a large subset of the hypothesis space depending on the parameter configuration, learning strategies and regularization constraints [4, 14]. Let $f_{\mathsf{opt}} \in \mathcal{F}$ be the optimum solution for the given task. The so-called *Rashomon-set*

$$\mathcal{R}_{\varepsilon}\left(L, \mathcal{F}, \mathcal{Z}\right) = \{f_{\mathcal{P}} \in \mathcal{F} | E_{\mathcal{Z}}\left(f_{\mathcal{P}}\right) \leq E_{\mathcal{Z}}\left(f_{\mathsf{opt}}\right) + \varepsilon\}$$

is understood as the set of close-to-optimum-solutions (CTOS).

It is claimed in [13] that for challenging tasks the Rashomon set becomes large giving rise that there is a high chance to tackle the task successfully also by shallow models [15]. Frequently, respective shallow approaches are better to explain or to interpret [16]. Yet, also for these simpler models the Rashomon effect frequently is inevitable [9]. Yet, we can suppose that the corresponding Rashomon sets are smaller compared to those of complex models.

Several attempts were made to characterize the Rashomon sets: It is assumed that significantly deviating models in this set reflect qualitatively different approaches to solve the task, i.e. the data evaluation perspective of the models differ substantially providing insights regarding data feature dependencies. Unfortunately, the evaluation of *empirical Rashomon sets* obtained by training different models for a given task, did not gained great research attraction so far [12, 4, 17]. A promising idea is to investigate Rashomon importance distributions of data features, which are related to classification importance profiles (CIP) of features in prototype based classification learning and XAI [18, 19, 20]. Hence, we propose using those feature importance profiles to identify groups of qualitatively similar models (clusters) in the empirical Rashomon set deciding models in the discriminative classification learning.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

3 Shallow MLP and Rashomon Set Cluster Analysis

In this initial investigation we restrict ourselves to a shallow MLP for a classification task $\mathbf{x} \mapsto \mathbf{y}$ with one-hot class coding of m classes. More specifically, we study a two-layered feed-forward network (shallow MLP) given as

$$\hat{\mathbf{y}} = \mathsf{softmax}\left(a\left(\mathbf{W} \cdot \mathbf{x} + \mathbf{b}\right)\right) \tag{1}$$

whereas standard perceptron layer with $a(\cdot)$ being the ReLU-activation is followed by a softmax-layer [21, 22]. Both, the weight matrix $\mathbf{W} \in \mathbb{R}^{m \times n}$ and the bias vector $\mathbf{b} \in \mathbb{R}^m$, constitute the model parameter set $\mathcal{P}$. For a trained model, feature importance can be evaluated by the so-called Shapley-profile vector $\mathbf{s} \in \mathbb{R}^n_+$ [23]. The vector entries s_i judge the importance of the data features x_i for solving the classification task by the model. Yet, the calculation of the profile may require expensive computational costs due to the combinatorial complexity for high-dimensional data $\mathbf{x}$ but it gained large attraction in XAI using sophisticated approximation methods [2]. A simpler approach to evaluate feature importance in the above shallow model is known from relevance learning in learning vector quantization: A feature importance profile $\boldsymbol{\lambda} \in \mathbb{R}^n_+$ is obtained via $\boldsymbol{\Lambda} = \mathbf{W}^\top \mathbf{W}$ and $\lambda_i = \sum_{j=1}^n |\Lambda_{i,j}|$, which reflects the relevance of a feature to class separation [24].

As pointed out in [9, 25, 15], the Rashomon effect is inevitable also in case of shallow or strongly reguralized models if uniqueness is not guaranteed. Further, the Rashomon set becomes larger when data or model noise (in consequence of initialization dependence, learning strategies etc.) is involved in the training process. Thus, clustering of the Rashomon set can help to detect qualitatively deviating models. Thereby, the models follow different decision strategies by means of the parameter sets $\mathcal{P}$ usually realizing different weighting strategies for the data features known as *feature importance* [26].

As already mentioned, the Rashomon set for a given task usually covers different aspects and strategies realized by the models for the problem solving. Hence, the intrinsic structure of this set probably reflects these strategies by topological counterparts in a corresponding representation space, which then can be detected by a respective cluster analysis. Yet, here the type of representation as well as the choice of an appropriate dissimilarity measure are essential ingredients for the demanded cluster analysis.

In the context of shallow MLP, the strategically different task solution can be identified regarding their feature information profiles [27, 4], i.e. we can represent the several strategies by the corresponding Shapley profiles $\mathbf{s}$ or the classification importance profiles $\boldsymbol{\lambda}$. Thus, a given Rashomon set of the shallow MLP can be represented by these profiles and, hence an intrinsic structure analysis (clustering) of this representation space should reveal the intrinsic structure original Rashomon space. For the profile similarity in the respective cluster analysis of the representation space, we prefer the Spearman-rank-correlation *corr* to deal with non-linear correlations and to mitigate scaling effects [28]. Because correlations are non-metric proximities, a consistent cluster method has to be chosen. A robust approach capable of handling proximities appropriately, is affinity propagation (AP) [29]. AP is based on message passing between po-

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

tential cluster representatives (here denoted as exemplars), which remain to be data samples.

4 Numerical Experiments – Results

We apply shallow MLP as explained in Section 3 on three small datasets: Pima Indians Diabetes (PIMA), Wheat Seed (SEED) and reduced Adrenal Tumor (AT) dataset. PIMA and SEED are classical UCI-datasets [30, 31], one from medicine and one from quality assurance in the food industry. PIMA contains eight clinical values for 768 female patients of Pima heritage with a binary outcome indicating the presence or absence of diabetes. SEED includes 210 data points with information on seven geometric measurements of wheat seeds, the classification task being to distinguish between species among three varieties. For the AT dataset, urine steroid–metabolite profiles are measured. The raw cohort comprises 147 adrenal–tumor patients (102 adrenocortical adenomas, ACA; 45 adrenocortical carcinomas, ACC) and 88 healthy controls (total 235 subjects). Following the curation used by the data provider, we removed variables with wide-spread missingness and highly collinear features, producing a panel reduced to 11 metabolites [32]. The data are all z-score normalized. It should be noted that the bias b in 1 is set to zero and the only parameters to learn are the weight matrix from input to the first layer. The three datasets are low dimensional such that the Shapley values can be explicitly calculated without any approximation.

The results in Table 1 show that for shallow settings such as (1), the importance profiles $\mathbf{s}$ and $\boldsymbol{\lambda}$, the Spearmann-Rank-correlation as well as the adjusted rank index (ARI) between the cluster solutions is not very high. Thus, several explanation methods using different strategies provide different conclusions. In addition, we receive explanations with about 2–3 clusters, i.e., 2–3 different feature importance profiles, which in turn means that 2–3 explanations of the feature's significance are available and can now be forwarded to experts for further discussion. Due to the lack of space, we cannot provide a more in-depth analysis here. However, since we only have a small number of cluster resulting from AP, we can give this information to experts for meaningful interpretation and manageable explanatory approaches, which finally helps to decide on the usability or usefulness of the models. Further, we have verified the existence of the non-trivial Rashomon set already in this shallow setting as predicted in [9].

5 Conclusions and Future Work

In this contribution we consider the evaluation of Rashomon sets to determine qualitatively distinguishable classification models, which realizes different decision strategies rated by data feature importance profiles. In doing so, we result in stable and plausible model explanations/interpretations for further evaluation by domain experts. Here, we limited ourselves to shallow two-layered feed-forward networks that already yield promising results. Future work will include the comparison of these simple models with easy to interpret prototype-based classifiers that are a preferred alternative to deep networks for complex tasks [33].

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

			$\boldsymbol{\lambda}$		**s**		
dataset	accuracy	SP($\boldsymbol{\lambda}$,**s**)	corr	#Clusters	corr	#Clusters	ARI
PIMA	87.9% ±3.4%	0.880 ±0.009	0.915 ±0.009	3	0.907 ±0.012	2	0.26
SEED	99.7% ±1.1%	0.529 ±0.012	0.892 ±2.3	3	0.964 ±0.5	2	0.092
AT	87, 9% ±3.3%	0.804 ±0.006	0.948 ±0.004	2	0.955 ±0.007	3	0.07

Table 1: Different mean results for three analysed data sets (ARI - adjusted rank index between cluster solutions). We have tested 100 models, and the ones we are taking into account are the ones that perform about the same as the average.

References

[1] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layerwise relevance propagation. *PLOS One*, 10(7):e0130140, 2015.

[2] S.M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 4768–4777. Curran Associates, Inc., 2017.

[3] E. Štrumbelj and I. Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3):647–665, 2014.

[4] A. Fisher, C. Rudin, and F. Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

[5] D. Janzing, L. Minorics, and P. Bloebaum. Feature relevance quantification in explainable AI: A causal problem. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (AISTAT)*, volume 108 of *Proceedings of Machine Learning Research*, pages 2907–2916, 2020.

[6] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000.

[7] J. Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications ACM*, 62(3):54–60, 2019.

[8] L. Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.

[9] L.M. Paes, R. Cruz, F.P. Calmon, and M. Diaz. On the inevitability of the Rashomon effect. In *Proc. of IEEE International Symposium on Information Theory (ISIT), Taipei*, pages 559–554, 2023.

[10] M.T. Ribeiro, S. Sing, and 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 (KDD 2016)*, pages 1135–1144, 2016.

[11] W. Samek, G. Montavon, S. Lapuschkin, C.J. Anders, and K.-R. Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278, 2021.

[12] Sebastian Müller, Vanessa Toborek, Katharina Beckh, Matthias Jakobs, Christian Bauckhage, and Pascal Welke. An empirical evaluation of the Rashomon effect in explainable machine learning. In D. Koutra, C. Plant, M.G. Rodriguez, E. Barallis, and F. Bonchi, editors, *European Conference on Machine Learning (ECML PKDD: Machine Learning and Knowledge Discovery in Databases: Research Track*, volume 14171 of *Lecture Notes in Computer Science (LNAI)*, pages 462–478. Springer, 2023.

[13] L. Semenova, C. Rudin, and R. Parr. On the existence of simpler nachine learning models. In *FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1827–1858, 2022.

[14] S. Hochreiter and J. Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997.

[15] C. Rudin, C. Zhong, L. Semenova, M. Seltzer, R. Parr, J. Liu, S. Katta, J. Donnelly, H. Chen, and Z. Boner. Amazing things come from having many good models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, number Art. No 1742, pages 42783 – 42795, 2024.

[16] P. Lisboa, S. Saralajew, A. Vellido, R. Fernández-Domenech, and T. Villmann. The coming of age of interpretable and explainable machine learning models. *Neurocomputing*, 535:25–39, 2023.

[17] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Survey*, 16:1–85, 2022.

[18] J. Donnelly, S. Katta, C. Rudin, and E.P. Browne. The Rashomon importance distribution: getting RID of unstable, single model-based variable importance. In *Proceedings of the 37th International Conference on Neural Information Processing System (NeurIPS)*, number Art. No 274, pages 6267 – 6279, 2023.

[19] J. Voigt, S. Saralajew, M. Kaden, K. Bohnsack, L. Reuss, and T. Villmann. Biologically-informed shallow classification learning integrating pathway knowledge. In *Proceedings of the 17th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC'2024) – Bioinformatics Workshop*, volume 1, pages 357–367. SCITEPRESS – Science and Technology Publications, Lda., 2024.

[20] J. Voigt, M. Kaden, L. Reuss, and T. Villmann. Reliable classification learning for medical data analysis using prototype-based models. In L. Rutkowski, R. Scherer, M. Korytkowski, W. Pedrycz, R. Tadeusiewicz, and J.M. Zurada, editors, *Proceedings of the 24th International Conference on Artificial Intelligence and Soft Computing - ICAISC'2025, Zakopane*, volume Part I of *LNCS/LNAI*, pages 206–2018, Cham, 2025. Springer International Publishing, Switzerland.

[21] C.M. Bishop and H. Bishop. *Deep Learning – Foundations and Concepts*. Springer International Publishing, 2024.

[22] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, Cambridge, MA, 2016.

[23] L.S. Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

[24] M. Kaden, K.S. Bohnsack, M. Weber, M. Kudla, K. Gutowska, J. Blazewicz, and T. Villmann. Learning vector quantization as an interpretable classifier for the detection of SARS-CoV-2 types based on their RNA-sequences. *Neural Computing and Applications*, 34(1):67–78, 2021.

[25] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16:1–85, 2022.

[26] A. Shrikumar, P. Greenside, and A. Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70, pages 3145–3153, 2017.

[27] M. Kaden, S. Saralajew, and T. Villmann. Domain knowledge integration in machine learning systems. In M. Verleysen, editor, *Proceedings of the 32nd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2024), Bruges (Belgium)*, pages 405–412, Louvain-La-Neuve, Belgium, 2024. i6doc.com.

[28] L. Sachs. *Angewandte Statistik*. Springer Verlag, Heidelberg Berlin, 7-th edition, 1992.

[29] B.J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 215(5814):972–976, 2007.

[30] A. Asuncion and D.J. Newman. Indian diabetes data set (PIMA). http://archive.ics.uci.edu/ml/.

[31] Niewczas Jerzy Kulczycki Piotr Kowalski Piotr Charytanowicz, Magorzata and Szymon Lukasik. Seeds. UCI Machine Learning Repository, 2010. DOI: https://doi.org/10.24432/C5H30K.

[32] M. Biehl, P. Schneider, D. Smith, H. Stiekema, A. Taylor, B. Hughes, C. Shackleton, P. Stewart, and W. Arlt. Matrix relevance LVQ in steroid metabolomics based classification of adrenal tumors. In M. Verleysen, editor, *20th European Symposium on Artificial Neural Networks (ESANN 2012)*, pages 423–428. d-side publishing, 2012.

[33] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.