Finding Patterns in Ambiguity: Interpretable Stress Testing in the Decision Boundary

Inês Gomes
ines.gomes@fe.up.pt

Luís F. Teixeira
luisft@fe.up.pt

Jan N. van Rijn
j.n.van.rijn@liacs.leidenuniv.nl

Carlos Soares
csoares@fe.up.pt

André Restivo
arestivo@fe.up.pt

Luís Cunha
up201706736@fe.up.pt

Moisés Santos
mrsantos@fe.up.pt

Abstract

The increasing use of deep learning across various domains highlights the importance of understanding the decision-making processes of these black-box models. Recent research focusing on the decision boundaries of deep classifiers, relies on generated synthetic instances in areas of low confidence, uncovering samples that challenge both models and humans. We propose a novel approach to enhance the interpretability of deep binary classifiers by selecting representative samples from the decision boundary — prototypes — and applying post-model explanation algorithms. We evaluate the effectiveness of our approach through 2D visualizations and GradientSHAP analysis. Our experiments demonstrate the potential of the proposed method, revealing distinct and compact clusters and diverse prototypes that capture essential features that lead to low-confidence decisions. By offering a more aggregated view of deep classifiers' decision boundaries, our work contributes to the responsible development and deployment of reliable machine learning systems.[1]

1. Introduction

Nowadays, Deep Learning (DL) models are broadly used in various domains, but their lack of interpretability due to their black-box nature poses a significant challenge [1]. Recent efforts explore DL models' decision-making processes, particularly around decision boundaries, where models often struggle to make correct predictions. Research initiatives such as DeepDIG [12], GASTeN [2] and AmbiGuess [26] study the decision boundary in a data-driven way by generating borderline instances, *i.e.* synthetic low-confidence examples, using techniques like Generative Adversarial Networks (GANs) or Variational Auto-encoders (VAEs). While many borderline instances consist of noisy data with patterns undetectable to the human eye, similar to adversarial examples [25], prior work has shown that a well-selected subset of such samples resembles genuinely hard-to-classify images, even for humans [2, 26]. Building on this, we propose a novel approach to enhance the interpretability of deep binary classifiers by selecting representative samples from the decision boundary — prototypes — and applying post-model explanation algorithms.

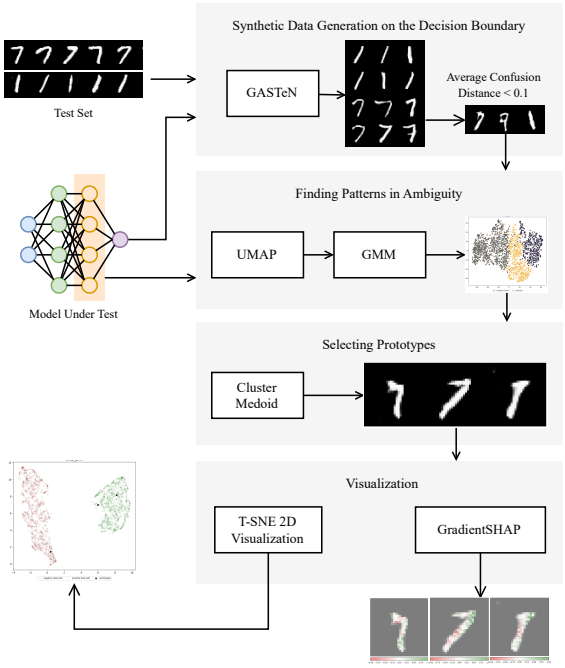


Figure 1. Schematic overview of the proposed method for improving the decision boundary interpretability of the Model Under Test by combining synthetic image generation and deep clustering.

Our method, illustrated in Fig. 1, comprises four steps: 1. generate synthetic data near the decision boundary with GASTeN; 2. detect patterns in these examples using UMAP [17] and Gaussian Mixture Models (GMM); 3.

[1] https://github.com/inesgomes/db-patterns

choose a representative prototype from each cluster; and 4. visualize the prototypes and the decision boundary using 2D space visualization and GradientSHAP [15]. We empirically evaluated the method using three Convolutional Neural Networks (CNN) of different complexity on binary subsets of MNIST and Fashion-MNIST. Results show the potential of the method, revealing distinct, compact clusters and diverse prototypes that embody the features contributing to low-confidence decisions.

Ultimately, our method aims for a more responsible use of AI models by supporting development and auditing. During development, it can spot potential model limitations by identifying and explaining key examples the model struggles with. Labelling and using these examples for further training can improve the model through active learning or data augmentation. These examples are also valuable for developing models with a reject option [8] — good models should refrain from predicting on these prototypes. Moreover, prototypes can support deployment by providing information about the data types for which a model is expected to make low-confidence predictions, serving as a semi-automatic tool to generate model cards [19].

2. Related Work

2.1. Stress Testing Machine Learning Models

Stress testing is an evaluation process to assess system robustness, limitations, and overall performance under challenging conditions. When applied to ML models, it involves testing models on adverse conditions, including out-of-distribution [9], adversarial [6, 21] or ambiguous inputs solely for the model [2, 14] or both the model and humans [26]. Recent work has explored model decision boundaries to understand the limits of ML models. Weiss et al. [26] generate ambiguous data points to train and test DNN supervisors; Heo et al. [10] use samples near the boundary for knowledge distillation; Liu et al. [14] and Demir et al. [4] study the decision boundary to comprehend models on safety-critical fields.

When considering only data-driven approaches, techniques such as DeepDIG [12] and DeepBoundary [14] generate adversarial examples combined with binary search to find the closest points to the decision boundary; AmbiGuess [26] leverages autoencoders to target specific latent space distributions; GASTeN [2] introduces a GAN-based methodology that incorporates the output of a classifier as part of the generator's loss function; and Demir et al. [4] combine state-of-the-art methods, including image transformations, GANs and adversarial attacks, followed by ML models that select those with highest uncertainty.

Few studies explore how to use samples that fall within a model's low-confidence region for responsible AI. Demir et al. [4] suggest using post-model explanations on error-prone class samples, while Cunha et al. [2] incorporates such samples into model cards.

2.2. Slice Discovery Methods

Slice discovery consists of methods that identify semantically meaningful subgroups within unstructured data, particularly where models perform poorly [7]. Our approach, while similar, diverges by focusing on low-confidence instances. These methods leverage deep clustering, a technique that uses neural networks to capture relevant input features, followed by traditional clustering algorithms [18].

The slice discovery methods state-of-the-art show numerous techniques for representing image data. GEORGE [24] and PlaneSpot [20] extract the embeddings from the penultimate layer of their model, whereas DOMINO [7] uses pre-trained embeddings, more specifically CLIP [22] and ConVIRT [28]. These methods then fine-tune GMMs for clustering.

These methods employ dimensionality reduction when facing challenges with high-dimensional data, such as inefficient similarity measures [18] and the curse of dimensionality. DOMINO uses PCA with 128 components, PlaneSpot opts for *scvis* [5] with 2 dimensions, and GEORGE selects UMAP with 1 or 2 components based on the dataset. Given that UMAP helps preserve the essential structure of the data, combining UMAP with GMM, as seen in studies like N2D [16], demonstrates that manifold learning techniques can significantly improve clustering quality by considering the local data structure.

3. Borderline Prototype Generation

Figure 1 summarizes the proposed method. The model that is being subjected to the stress-test must be a deep binary image classification, and it is referred to as the Model Under Testing (MUT). First, we populate the decision boundary by generating synthetic images close to the MUT decision boundary, *i.e*. our borderline instances. To that end, we employ GASTeN, a GAN-based technique trained with the MUT predictions, to approximate its decision boundary [2]. We chose GASTeN as it generates realistic challenging borderline examples for a specific classifier. Then, we filter the synthetic images using the Average Confusion Distance (ACD) Cunha et al. [2], that measures the closeness of a sample to the decision boundary. We filter the images by $ACD < 0.1$ to ensure only low-confidence predictions. We assess the quality of the borderline images by calculating the Fréchet Inception Distance (FID) scores [11].

In our second step, we apply deep clustering to find patterns in the borderline instances generated in the previous step. To that end, we extract the high-level feature embeddings from the MUT's penultimate layer. Then, we apply UMAP for dimension reduction, followed by GMM clustering to group visually similar images. We selected

this combination given the favourable findings in the literature review [7, 20, 24]. Particular hyperparameters are tuned, considering the specific characteristics of the low-confidence region, as we explain in Sec. 4.2. Finally, the quality of the resulting clusters is assessed through the silhouette score [23] and the Davies-Bouldin index [3]. We use these measures of cluster definition and separation to ensure the formation of distinct and coherent groups.

In the third step, we select the medoid from each cluster to represent the cluster. The medoid is calculated by minimizing the sum of distances to all other objects in that cluster. As a centrally located sample, it ensures a robust representation of each identified pattern.

In the fourth step, we evaluate the representativeness of the selected prototypes through visual inspection. Our goal is to generate prototypes that demonstrate greater feature diversity, more dispersed distribution across the 2D space, and enhanced interpretability through GradientSHAP maps. With this in mind, we train UMAP on the test set to capture the structure of the original data. Then, we analyze the positioning of the prototypes within the 2D space created. For an in-depth analysis of why these images are close to the decision boundary, we use GradientSHAP — a technique that explains the contribution of each pixel to the model's output by integrating gradients with SHAP values [15].

4. Experimental Setup

4.1. Dataset

We use MNIST [13] and Fashion-MNIST [27] datasets to evaluate our method. We chose these datasets for their interpretability without needing expert knowledge, simplicity in size, and lack of color. We created binary subsets from these datasets for binary classification, focusing on similar concepts: *7 vs 1*, *8 vs 0* and *5 vs 3* for MNIST and *dress vs top* and *sneaker vs sandal* for Fashion-MNIST.

4.2. Model Architecture

To obtain more general conclusions, we evaluated MUTs architectures of varying complexities. Following the GASTeN study, we utilized a CNN architecture with two convolutional blocks, where the complexity is adjusted by varying the number of filters [2]. For the MNIST dataset, we tested CNN models with 1, 2, and 4 filters, while for Fashion-MNIST, we used 4, 8, and 16 filters.

To train GASTeN, we adapted its setup based on previous findings by Cunha et al. [2], tailoring its hyperparameters towards our stress-testing objectives. Training GASTeN required choosing two specific hyperparameters beyond the standard DCGAN parameters: the confusion distance weighting (α) and the pre-training epochs. The α value critically influences GASTeN's loss function, while the pre-training duration affects image realism. Based on the original authors' suggestions, we opted for 5 pre-training epochs for MNIST and 10 for Fashion-MNIST, each with an α weight of 25. We determined GASTeN's optimal training duration by optimizing the FID-ACD minimization [2], leading to selecting 10 epochs for MNIST and 15 for Fashion-MNIST. We generated 15,000 synthetic images for each task.

For deep clustering, we optimized the silhouette score using the Bayesian hyperparameter optimization method with 25 iterations. With UMAP, we investigated the optimal number of neighbours to balance local versus global data structures. Given our focus on classifying similar concepts and analyzing regions of low confidence, where features are less distinct, our analysis prioritized local structures. Therefore we explore a range of 5 to 25 neighbours. We also varied the minimum distance between 0.01 and 0.25 to control embedding compactness and set the components between 10 and 60 to ensure detailed clustering without losing critical information. For GMM, we varied the cluster count from 3 to 15 to maintain a practical number of prototypes for analysis. We also selected the covariance as *full*, as it allows for each cluster to have its covariance matrix.

5. Results and Discussion

5.1. Synthetic Data Generation

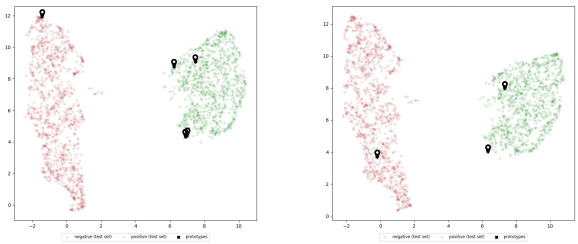
We tested three MUTs on the five binary subsets, The MNIST *5 vs 3* subset using a CNN with one filter attained the lowest accuracy of 92.53%, while the *8 vs 0* subset with a four-filter CNN reached the highest accuracy of 98.92%.

After training GASTeN for each classifier-dataset subset combination, we generated 15,000 synthetic images and subsequently filtered those with $ACD < 0.1$. This process resulted in an average FID score increase of 250 points for images near the decision boundary and an average 86% reduction in image count post-filtering. The significant FID score rise and the substantial image count decrease after filtering suggest GASTeN's limited efficiency in generating decision boundary-near samples.

During this process, we observed some correlation between the model complexity and GASTeN FID scores ($\rho_{nf,FID} = 0.52$). We expected this outcome, as lower classifier capacities lead to more challenging classifications, resulting in less confident images. However, a contrary example is our least accurate classifier, which produced realistic (low FID scores) synthetic images that were unrealistic (high FID scores) near the decision boundary.

5.2. Finding Patterns in Ambiguity

With the resulting synthetic images from the previous step, we optimized UMAP and GMM hyperparameters to achieve the highest silhouette score. UMAP frequently select hyperparameters that highlight the local structure and



(a) Baseline (b) Prototypes

Figure 2. UMAP 2D space for MNIST *7 vs 1* and four-filter CNN. The test set is marked by stars with *1* in red, and *7* in green. Black pins indicate the prototypes or baseline positions.

preserve image ambiguity. GMM clustering resulted in either a small (*e.g.* 3) or large (*e.g.* 15) number of clusters without a discernible pattern.

Evaluation metrics indicated modest clustering quality, including the silhouette score (0.26 — 0.52) and the Davies-Bouldin Index (0.7 — 1.35) on the 15 classifier-dataset subset pairs. These metrics suggest that while clusters are reasonably distinct and compact. There is room for improvement, possibly due to some overlap or sparseness in clusters. The best performance occurred on the MNIST *7 vs 1* subset with a four-filter CNN and the poorest on the *5 vs 3* subset with a four-filter CNN.

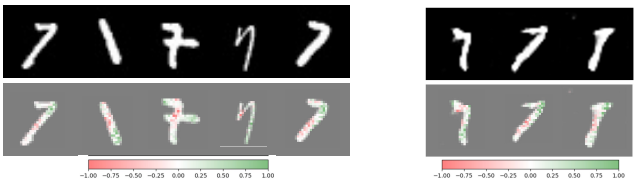
We noticed a negative correlation between the quantity of low-confidence images and the silhouette score ($\rho_{\#images,SIL} = -0.62$), indicating that fewer images generally lead to more effective clustering. We suspect this could be due to noise in the generated images, suggesting that exploring alternative dimensionality reduction and clustering techniques could enhance our clustering results.

5.3. Selecting and Visualizing Prototypes

After clustering, we select the medoid of each cluster as our prototype and assess its representativeness through a 2D visualization and with GradientSHAP. To illustrate the usefulness of the proposed method, we chose the best-performing subset based on the silhouette score for this section analysis: MNIST subset *7 vs 1* with four-filter CNN.

Figure 2 shows the distribution of prototypes versus the baseline in the UMAP 2D space. In this example, we conclude that prototypes are more dispersed than the baseline which even includes overlapping images.

Observing the baseline in Fig. 3a, two images appear remarkably similar, likely those clustered together in the 2D space, with a fourth image resembling noise. In contrast, in Fig. 3b, the prototypes display distinct features, particularly regarding rotation. We also note that all images lack part of the seven's upper section. We hypothesize that this feature is an intrinsic attribute of the low-confidence region in this MUT, and that the prototypes effectively capture it.



(a) Baseline (b) Prototypes

Figure 3. Selected images and the corresponding GradientSHAP maps for MNIST *7 vs 1* and four-filter CNN. Features contributing to the classification of *1* are red, and *7* are green.

GradientSHAP maps indicate that features leading to classification as the positive class (7) include the top elbow and the middle line of the seven. On the other hand, attributes favoring the negative class (1) typically relate to pixels in the center of the image. Remarkably, these maps also express uncertainty regarding the missing portion of the seven's upper section, validating our hypothesis that the model has difficulty learning this feature.

6. Conclusions

In this work, we investigate borderline instances that contain visual properties that make predictions complex, even for humans. We study the impact of combining synthetic image generation and deep clustering on the interpretability of deep binary classifiers' decision boundary.

Further research includes improving the generation of ambiguous images, exploring other clustering and embedding techniques for improved performance, and developing quantitative metrics to validate our prototypes statistically. Additionally, as we tested our approach on simplified datasets, where it is possible to assess ambiguity visually, the performance in complex scenarios needs to be assessed.

Nevertheless, we obtained promising results that show that it is possible to uncover patterns in borderline images. By visual inspection, we can study the representativeness of our prototypes and the features associated with the low-confidence region. Such insights are invaluable for auditing machine learning models, identifying and mitigating potential weaknesses during development, or documenting the limitations of classifiers in model cards upon deployment.

Acknowledgments AISym4Med (101095387) through the Horizon Europe Cluster 1: Health, ConnectedHealth (n.º 46858); Competitiveness and Internationalisation Operational Programme (POCI) and Lisbon Regional Operational Programme (LISBOA 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF); NextGenAI - Center for Responsible AI (2022-C05i0102-02), supported by IAPMEI; FCT plurianual funding for 2020-2023 of LIACC (UIDB/00027/2020 UIDP/00027/2020).

References

[1] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020. 1

[2] Luís Cunha, Carlos Soares, André Restivo, and Luís F Teixeira. GASTeN: Generative Adversarial Stress Test Networks. In *Advances in Intelligent Data Analysis {XXI} - 21st International Symposium on Intelligent Data Analysis*, pages 91–102, 2023. 1, 2, 3

[3] David L Davies and Donald W Bouldin. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1:224–227, 1979. 3

[4] Demet Demir, Aysu Betin Can, and Elif Sürer. Distribution Aware Testing Framework for Deep Neural Networks. *IEEE Access*, 11:119481–119505, 2023. 2

[5] Jiarui Ding, Anne Condon, and Sohrab P Shah. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nature Communications*, 9: 2002, 2018. 2

[6] Isaac Dunn, Hadrien Pouget, Tom Melham, and Daniel Kroening. Adaptive Generation of Unrestricted Adversarial Inputs. *CoRR*, 2019. 2

[7] Sabri Eyuboglu, Maya Varma, Khaled Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Ré. Domino: Discovering Systematic Errors with Cross-Modal Embeddings. *The Tenth International Conference on Learning Representations*, 2022. 2, 3

[8] Kilian Hendrickx, Lorenzo Perini, Dries Van der Plas, Wannes Meert, and Jesse Davis. Machine Learning with a Reject Option: A survey. *CoRR*, 2021. 2

[9] Jens Henriksson, Christian Berger, Markus Borg, Lars Tornberg, Sankar Raman Sathyamoorthy, and Cristofer Englund. Performance Analysis of Out-of-Distribution Detection on Various Trained Neural Networks. In *2019 45th Euromicro Conference on Software Engineering and Advanced Applications*, pages 113–120, 2019. 2

[10] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge Distillation with Adversarial Samples Supporting Decision Boundary. In *The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, pages 3771–3778, 2019. 2

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637. Curran Associates, Inc., 2017. 2

[12] Hamid Karimi and Tyler Derr. Decision Boundaries of Deep Neural Networks. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1085–1092, 2022. 1, 2

[13] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278–2324, 1998. 3

[14] Yue Liu, Lichao Feng, Xingya Wang, and Shiyu Zhang. DeepBoundary: A Coverage Testing Method of Deep Learning Software based on Decision Boundary Representation. In *2022 IEEE 22nd International Conference on Software Quality, Reliability, and Security Companion*, pages 166–172, 2022. 2

[15] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 2, 3

[16] Ryan McConville, Raul Santos-Rodriguez, Robert J Piechocki, and Ian Craddock. N2D: (Not Too) Deep Clustering via Clustering the Local Manifold of an Autoencoded Embedding. In *25th International Conference on Pattern Recognition*, pages 5145–5152, 2020. 2

[17] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *CoRR*, 2018. 1

[18] Erxue Min, Xifeng Guo, Qiang Liu, Gen Zhang, Jianjing Cui, and Jun Long. A Survey of Clustering with Deep Learning: From the Perspective of Network Architecture. *IEEE Access*, 6:39501–39514, 2018. 2

[19] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229, 2019. 2

[20] Gregory Plumb, Nari Johnson, Angel Cabrera, and Ameet Talwalkar. Towards a More Rigorous Science of Blindspot Discovery in Image Classification Models. *Transactions on Machine Learning Research*, 2023. 2, 3

[21] Viraj Uday Prabhu, Sriram Yenamandra, Prithvijit Chattopadhyay, and Judy Hoffman. LANCE: Stress-testing Visual Models by Generating Language-guided Counterfactual Images. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8746–8763, 2021. 2

[23] Peter J Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. 3

[24] Nimit Sohoni, Jared A Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No Subclass Left Behind: Fine-Grained Robustness in Coarse-Grained Classification Problems. In *Advances in Neural Information Processing Systems 33*, 2020. 2, 3

[25] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR 2014)*, 2014. 1

[26] Michael Weiss, André García Gómez, and Paolo Tonella. Generating and detecting true ambiguity: a forgotten danger in DNN supervision testing. *Empirical Software Engineering*, 28:146, 2023. 1, 2

[27] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, 2017. 3

[28] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Proceedings of the Machine Learning for Healthcare Conference*, pages 2–25, 2022. 2