ML Model Coverage Assessment by Topological Data Analysis Exploration

Ayman Fakhouri, Faouzi Adjed, Martin Gonzalez, Martin Royer

ML Model Coverage Assessment by Topological Data Analysis Exploration

Ayman Fakhouri[1,2], **Faouzi Adjed**[1], **Martin Gonzalez**[1], **Martin Royer**[1,3]

[1]IRT SystemX, 2 Boulevard Thomas Gobert 91120 PALAISEAU, France.
[2]Faculté des Sciences et Techniques, 123 Av. Albert Thomas, 87000 Limoges, France.
[3]DATASHAPE - Inria Saclay 1, rue Honoré d'Estienne d'Orves 91120 Palaiseau, France.
{ayman.fakhouri, faouzi.adjed, martin.gonzalez, martin.royer}@irt-systemx.fr

Abstract

The increasing complexity of deep learning models necessitates advanced methods for model coverage assessment, a critical factor for their reliable deployment. In this study, we introduce a novel approach leveraging topological data analysis to evaluate the coverage of a couple dataset & classification model. By using tools from topological data analysis, our method identifies underrepresented regions within the data, thereby enhancing the understanding of both model performances and data completeness. This approach simultaneously evaluates the dataset and the model, highlighting areas of potential risk. We report experimental evidence demonstrating the effectiveness of this topological framework in providing a comprehensive and interpretable coverage assessment. As such, we aim to open new avenues for improving the reliability and trustworthiness of classification models, laying the groundwork for future research in this domain.

Introduction

The coverage assessment of Artificial Intelligence based (AI-based) models becomes more and more complicated due to the recent architectures of Deep Learning (DL). Therefore, it becomes also an emerging research subject. This research theme has the objective to evaluate in the same time the model stability and generalizability applied on some domain represented by the dataset used. The AI coverage concept, such as code coverage which is based on code and test examples, is based on the couple the ML model and the dataset evaluation.

The deployment of DL models in real-world applications highlights further the issues related to the lack of thorough coverage evaluation (Sun et al. 2019). Indeed, the present available approaches, which use software coverage assessment, such as Neuron Coverage (Guo et al. 2018) based on Modified Condition/Decision Coverage (MC/DC), are not useful to handle correctly the misbehavior of DL models (Zohdinasab et al. 2023). Thus, this study underscores the necessity of thorough coverage evaluation as a cornerstone for high-quality performance assessment. By leveraging topological methods, we aim to detect and rectify data-sparse regions within the model's scope. Ensuring comprehensive coverage evaluation is essential for mitigating risks and ensuring the ethical and reliable deployment of DL-based systems, thereby enhancing overall model performance and trustworthiness.

Our study outlines three main contributions and novelties:

1. **Introduction of the Trust Rips Complex**: We propose a novel method for evaluating classification model coverage by combining topological features from persistence diagrams with model confidence levels, creating the "Trust Rips Complex" that links data structure with model performance.
2. **Impact of Data Augmentation on Persistence**: We demonstrate that expanding datasets within the model's expected domain decreases *persistence*, leading to a more homogeneous data distribution. Statistical and visual analyses show that data augmentation reduces the number of sparse regions or "holes" in the dataset, improving the evaluation of model's coverage.
3. **TDA Framework for Enhanced Model Coverage**: We present a framework that integrates topological data analysis (TDA) with model confidence to effectively enhance model coverage, as evidenced by a significant global reduction in persistence. The work emphasizes the importance of thorough coverage evaluation for reliable deployment of deep learning systems.

The rest of the paper is structured as follows: **Related work** section treats the current state of the art concerning DL coverage approaches. This is followed by the **Proposed approach** section, detailing our methodology, including the use of persistence diagrams to analyze the lifetime of each topological hole. In the **Results and discussion** section, we present our findings with comprehensive visualizations and comparisons of results. We interpret the implications of our results, discuss potential applications and address their limitations. Finally, in the **Conclusion and perspectives** section, we provide a synthetic account our work and suggest future research directions to enhance and improve our approach.

Related Work

To our knowledge, the first recorded contribution on coverage assessment of AI models is based on the intrinsic evaluation close to code coverage. Indeed, Pei et al. (Pei et al. 2017) introduced the neural coverage based on non-activated and activated neurons verification. Thus, the estimated coverage is presented by the rate of activated neurons over the total number of neurons ($NC = AC/N$) where NC defines

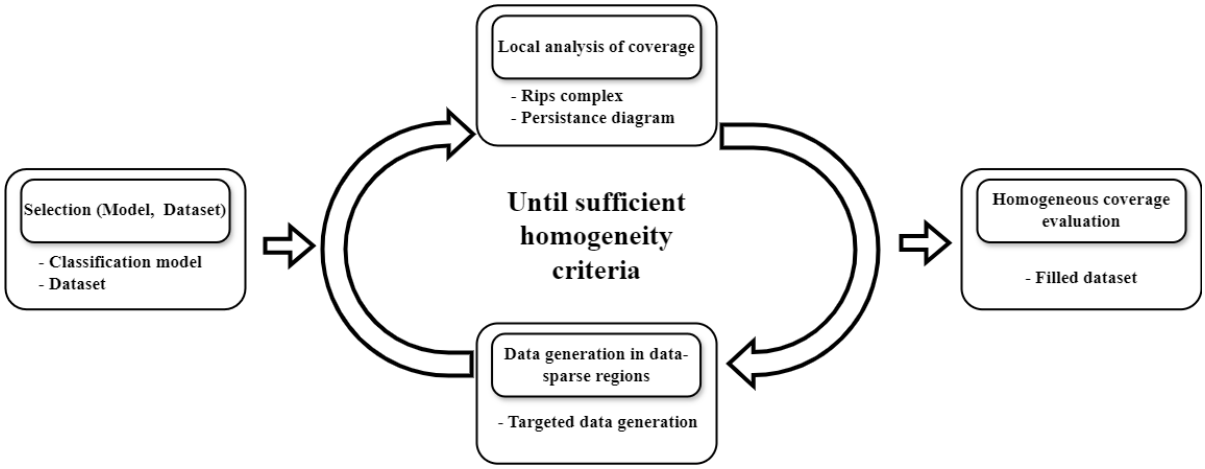


Figure 1: Workflow of the proposed approach summarized in a four-steps iterative process.

neural coverage rate, AC defines activated neurons number and N defines the total number of neurons). Based on this approach, Tian et al. (Tian et al. 2018) proposed an automated testing tool applied on autonomous driving and extended the neural coverage to other DL architectures rather than CNN developed by Pei et al.. Ma et al. (Ma et al. 2018a) proposed a set of criteria for dataset to evaluate the coverage. Following the same methodology, (Sun et al. 2018) proposed four criteria adapted from the Modified Condition/Decision Coverage (MC/DC) process developed in (Hayhurst 2001). Kim, Feldt, and Yoo (Kim, Feldt, and Yoo 2019) added the diversity criteria of the dataset evaluating the coverage. However, as mentioned by (Harel-Canada et al. 2020), the neural coverage is evaluating the software coverage and not the coverage of the AI decisions model. On the other hand, Xie et al. (Xie et al. 2019) proposed fuzzy logic based approach. The authors estimate the intermediary non covered zones to then estimate the coverage of the model. All these approaches were compared by (Yan et al. 2020) by extracting the correlations between the evaluation criteria of the coverage and the robustness of the given model. The authors concluded that these approaches based on neural coverage are evaluating the coverage partially. Odena et al. (Odena et al. 2019) proposed a coverage approach named TensorFuzz. This approach is developed to evaluate the black-box models by its analyzer function. Furthermore, Mani et al. (Mani et al. 2019) integrated the classification boundaries of a test dataset to ensure coverage. the authors propose four metrics; i) equivalence partitioning, ii) centroid positioning, iii) boundary conditioning and iv) pairwise boundary conditioning. However, all these methods do not consider the robustness of the model. In other words, coverage is computed only for the test dataset despite some integration of adversarial attack techniques such as DeepMutation proposed by (Ma et al. 2018b).

Another approach is developed by Adjed et al. (Adjed et al. 2022) introduced a new coverage metric mixing the robustness of the model and the spatial representation of the test dataset. The authors use abstract interpretation adapted from deep learning robustness (Gehr et al. 2018) and TDA (Chazal and Michel 2021) for spatial representation by building simplicial complexes directly from the dataset, using the persistent homology theory.

Despite the developed approaches in the literature, the coverage challenge of AI models remains complex and difficult. This difficulty can be summarized into two main aspects which are the application environment which is huge and the deep learning architecture with billions of parameters. The use of topological data analysis for coverage, as used for telecommunication (De Silva and Ghrist 2006), can be a promising research focus to handle in the same time the environment and model stability. Thus, in the current work, an approach using TDA is proposed to highlight the coverage of ML models by considering the spatial representation in one side and the performance model in the other.

Proposed Approach

The coverage definition considered in the current work is extracted from Confiance.ai[1] program: *The coverage of a couple "Dataset + ML Model" is the ability of the execution of the ML Model on this dataset to generate elements that match the expected space.* (Adjed et al. 2023).

The proposed approach relies the exploration of the homological features of the evaluation dataset and the DL model performances mapped with the application domain. In addition, a new concept is proposed named Trust Rips Complex which extends the Vietoris-Rips simplicial complex by incorporating an additional parameter $\alpha \in [0, 1]$ which is the confidence level of the classification model M. Specifically, the Trust Rips complex is a Rips complex built by considering only those data points for which the model M predicts with a confidence level greater than or equal to α. This selective approach allows for a focus on the regions where the model's predictions are most reliable. Then the Trust Rips Complex is characterized by two parameters, α

[1] A French community dedicated to the design and industrialisation of trustworthy critical systems based on artificial intelligence www.confiance.ai

described above and r from the radius of simplicial complex filtration.

This approach is divided into four steps, as illustrated in Figure 1 and detailed below.

1. **Selection of a couple (Model, Dataset):** This step can be considered as the initiation or selection. It includes the selection of the learned model to be evaluated and the evaluation dataset.
2. **Local analysis of the coverage of the expected space:** In this step, a Trust Rips Complex is constructed by combining the evaluation dataset with the trained model, for each given confidence level α and radius r. Subsequently, a persistence diagram is generated for each class and each α based on the constructed simplicial complexes. By analyzing these persistence diagrams, we identify the existence of topological holes by computing their lifetimes, defined as persistence $(d - b)$, where d and b represent the death and birth of each hole, respectively. Persistences that exceed a certain threshold are flagged as indicators of heterogeneous and underrepresented regions in the dataset.
3. **Generation data:** A data generation/collection is achieved in this step for all identified holes. Thus, this added data can be performed by data collection, augmentation and/or generation.
4. **Homogeneous local coverage evaluation:** Based on the threshold for accepted holes, we assess that the model performs equivalently in the given topology if its evaluation on the newly augmented dataset yields an accuracy of at least α. If this condition is met, we conclude that the model covers the expected domain to a threshold of α.

Implementation

The implementation of the proposed approach is summarized in the Algorithm 1. Therefore, the identification of not filled areas is performed by analyzing the persistence diagrams by identifying the farthest point from the diagonal line and by analyzing visually the grid of Trust Rips complexes. Then, a visual localization and a filling of the related areas on the real data cloud is achieved. Several statistics computed from the persistence are used to discriminate the non filled areas. The application of the whole approach is performed on the half-moon dataset explained in the next section.

We emphasize that visual analysis is only feasible for datasets with up to three dimensions and serves as a heuristic tool to support our statistical analysis by highlighting topological patterns. For higher-dimensional data, a new algorithm based on distribution densities is needed to detect sparse regions, as visual analysis is not applicable in those cases.

Regarding the data generation step for the presented case, once the data-sparse regions were identified, a uniform generation of data within the corresponding rectangular areas was carried out.

Concerning the homogeneity criteria, several options are possible. In the experiment presented, the chosen criterion was achieving a sufficient reduction in the initial mean persistence. However, this criterion must be adjusted according to the tools used and the data distribution, as the extent of reduction may be limited by either of these factors.

Algorithm 1: Coverage Analysis of Deep Learning Model

Input: Dataset D, Classification model M
Output: Local evaluation of M's coverage
repeat
Step 1: Local Coverage Analysis
- Construction of Trust Rips complexes with different trusts α and different radius r arranged in a grid
- Construction of persistence diagrams associated to each class
- Identification of data-sparse regions
- Compute statistical measures of persistence: mean, max, first quartile, third quartile...
- Plot box plots
Step 2: Data Generation
- Generate synthetic data in identified data-sparse regions
- Augment dataset D to form D'
- Update D with D'
Step 3: Final Analysis and Comparison
- Compare visually the grid before data generation with the last grid after data generation
- Compare the last statistical measures of persistence computed with those of the first analysis
- Compare the last box plot with the first box plot to assess significance of improvements
until Homogeneity criteria is met or until convergence
Step 4: Homogeneous coverage evaluation
- Evaluate M with the last augmented dataset D'

Experiments and Discussions

In this section, we present the findings from our analysis of DL model coverage. We start by presenting the dataset and the classification model M used in the whole experiment, followed by the presentation of the obtained results.

Dataset and Model Preparation

Two half-moons are generated to implement the whole workflow presented in Figure 1. This dataset was generated using the *make_moons* function from the **scikit- learn** library, with added Gaussian noise ($\sigma = 0.2$) and consisting of 960 points. Figure 2 illustrates the dataset generated. The choice of this dataset is motivated by its simplicity in terms of model implementation and results visualization and interpretation.

The Deep Neural Network (DNN) M is a sequential model defined as follows:

- **Input Layer:** Accepts an input with shape $(2, 1)$.
- **Hidden Layer 1:** A dense layer with 128 units and `ReLU` activation function.
- **Hidden Layer 2:** A dense layer with 64 units and `ReLU` activation function.

- **Hidden Layer 3:** A dense layer with 32 units and `ReLU` activation function.
- **Output Layer:** A dense layer with 1 unit and `Sigmoid` activation function.

This basic DNN was trained on the presented dataset and tested in an another generated half-moons dataset of 240 points obtaining a test accuracy of around 95%.

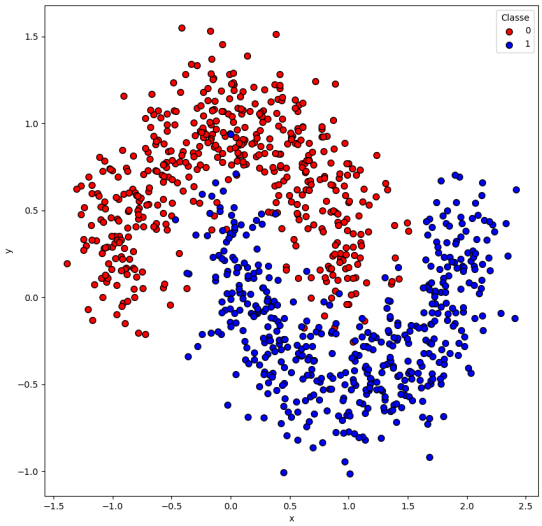


Figure 2: A visualization of the original data: two half-moons with Gaussian noise $\sigma = 0.2$

Obtained Results

Visual Coverage Analysis. To shed evidence of the significance of the proposed approach, we report a series of plots that are created and arranged in a grid format, where each row represents a different radius parameter of the Rips complex, and each column represents a different confidence level α as illustrated in Figures 3 and 4 comparing before data generation (BDG) and post data generation (PDG) step for model confidence $\alpha = 0.8$ and $\alpha = 0.98$, respectively. Each grid of figures can be interpreted in two distinct ways. Firstly, by reading horizontally, the changes in the Trust Rips complexes by increasing the radius r incrementally. This vertical reading provides insights into how the topological structure of the data evolves with varying levels of granularity. Secondly, by reading vertically, the impact of data generation for each radius. Together, these two perspectives offer a comprehensive view of the coverage for a given model's confidence level.

In Figures 3 and 4, a local analysis of the coverage of the model M with a confidence level of 0.8 and 0.98, respectively, across different filtration radius ($r \in \{0.05, 0.15, 0.25\}$) to identify areas with insufficient data to proceed. Depending on the radius, this visual analysis allowed us to clearly identify regions where the data coverage is sparse throughout the grid of Trust Rips complexes as illustrated in Figure 5. The second line of sub-figures (d), (e) and (f) of the two figure 3 and 4 illustrate the visual results after data generation step, where we can see that several topological holes are reduced or totally filled. The visual comparison of holes between Figure 5 and Figure 6 highlights the impact of the data generation step. It can be seen easily that the data in Figure 6 is more homogeneous than the original data illustrated in Figure 5. However, in the first and the third columns of figures 3 and 4, the impact the generation is less visible due to too small (0.05) or too large (0.25) radius value used.

The same analysis can be seen in figures 7 and 8 describing the persistence diagrams for model's confidence level of 0.8 and 0.98, respectively. The first line of these figures illustrates two persistence diagrams of the original data (in BDG step) for class 0 and class 1. Whereas the second line represents the persistent diagrams of the augmented data (in PDG step). The generation of the data is based on simplicial complexes of radius 0.15. Thus, it can be seen that several holes around 0.15 to 0.20 of radius value are moved closer to identity curve. Therefore, the holes with birth value much greater than 0.15 are not impacted. This is resulted from Algorithm 1, where the data generation is based on visual analysis, by treating only appeared holes.

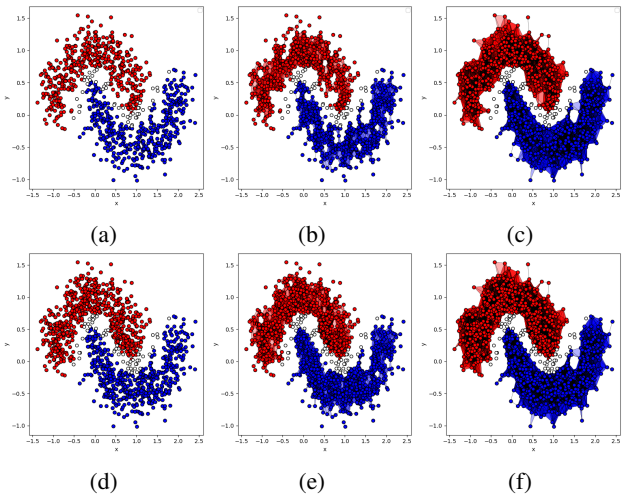


Figure 3: Trust simplicial complexes with $\alpha = 0.8$ (a), (b) and (c) represent simplicial complexes for radius 0.05, 0.15 and 0.25 for BDG, respectively, and (d), (e) and (f) represent simplicial complexes for radius 0.05, 0.15 and 0.25 for PDG

Statistical Coverage Analysis. As a second set of experiments, we conduct a statistical analysis by computing several key statistical characteristics from the persistence diagrams presented in Figures 7 and 8. The statistical measures are number of holes, mean persistence and maximum persistence. This statistical analysis allows us to assess the contribution of the data generation, and the effort still required to further reduce data sparseness. Table 1 provides these statistical results for two confidence level α values (0.8 and

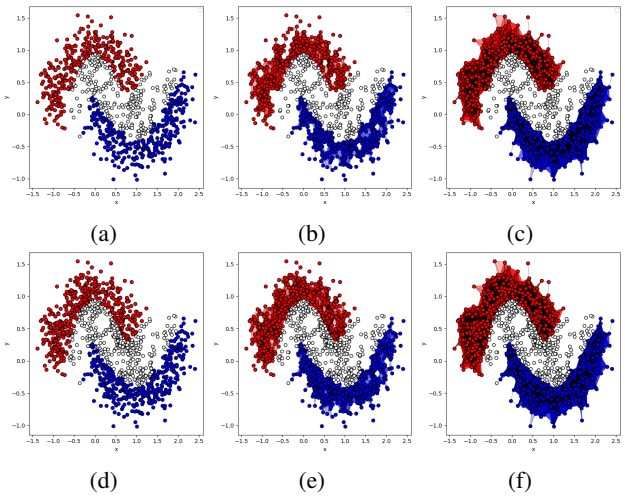


Figure 4: Trust simplicial complexes with $\alpha = 0.98$ (a), (b) and (c) represent simplicial complexes for radius 0.05, 0.15 and 0.25 for BDG, respectively, and (d), (e) and (f) represent simplicial complexes for radius 0.05, 0.15 and 0.25 for PDG

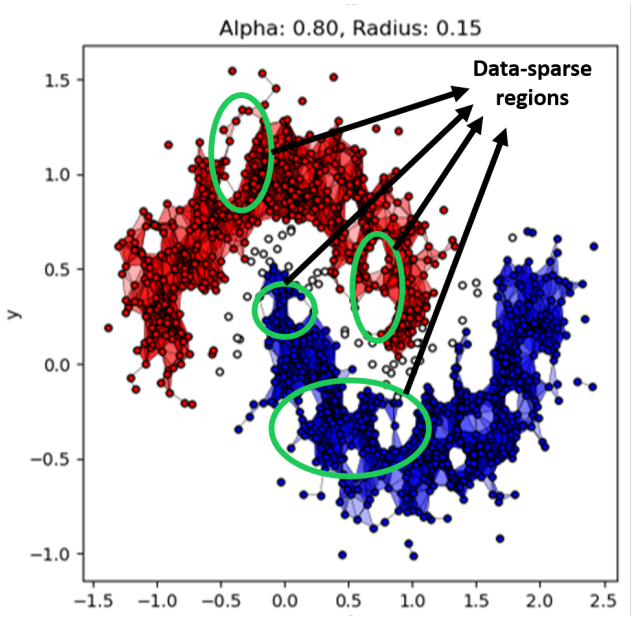


Figure 5: Data-sparse regions identified visually in the Trust Rips complex $r = 0.15$ $\alpha = 0.8$.

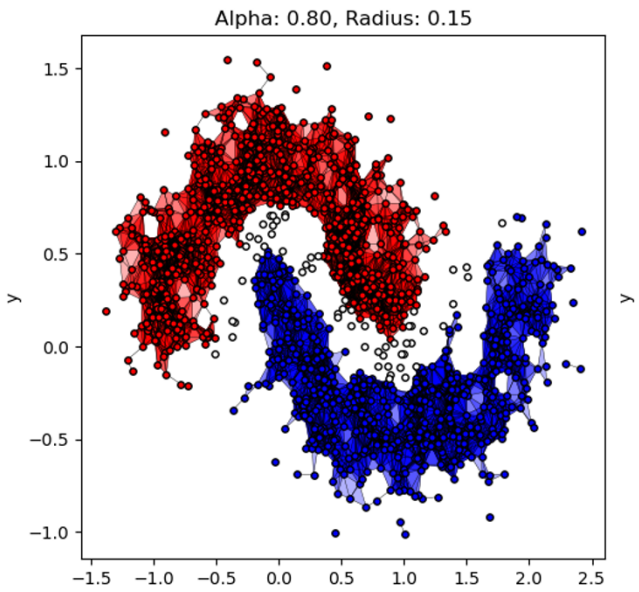
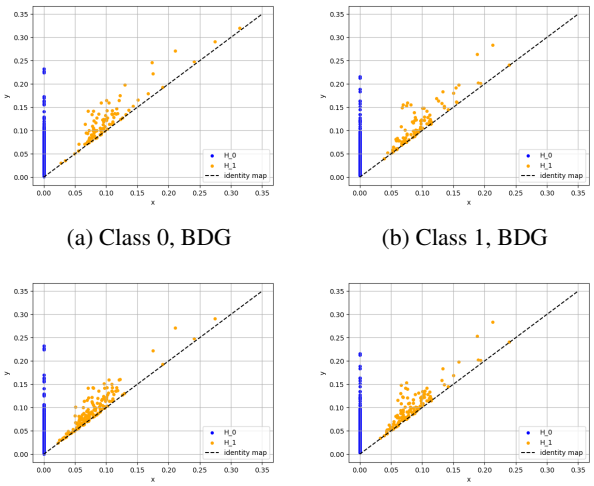


Figure 6: Trust Rips complex post data generation $r = 0.15$ $\alpha = 0.8$.



(a) Class 0, BDG (b) Class 1, BDG

(c) Class 0, PDG (d) Class 1, PDG

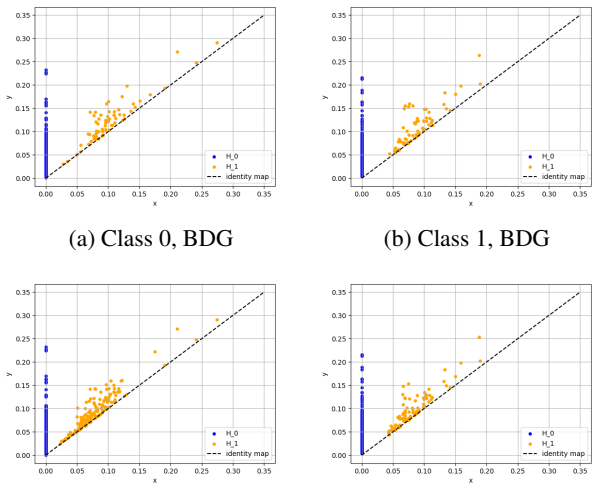
Figure 7: Persistent diagrams of trust simplicial complexes for $\alpha = 0.8$ for class 0 before and after data generation represented by (a) and (c), respectively, and class 1 for before and after data generation presented (b) and (d), respectively.

0.98). A slight improvement for the the two statistics, mean and maximum persistence, can be seen, although it is difficult to interpret them easily. In the other side, by mixing these results with the number of holes, we can see that the data generation decrease the mean and maximum and increase the number of holes. Each data generated splits the hole on two or more holes and reduces its persistence simultaneously. Furthermore, Figure 9 and Figure 10 illustrate more statistics by box plots for before and after data generation of both classes 0 and 1 and for two model's confidence levels 0.8 nad 0.98, respectively.

Discussions

This section discusses the implications and significance of the statistical analysis results, highlighting how the persistence characteristics reveal the data structure and potential applications.

As a sanity-check, we report the persistence diagrams associated to the Trusts Rips complexes built from the half-



(a) Class 0, BDG (b) Class 1, BDG

(c) Class 0, PDG (d) Class 1, PDG

Figure 8: Persistent diagrams of trust simplicial complexes for $\alpha = 0.98$ for class 0 before and after data generation represented by (a) and (c), respectively, and class 1 for before and after data generation presented (b) and (d), respectively.

class	α	number of holes		max		mean	
		BDG	PDG	BDG	PDG	BDG	PDG
0	0.8	91	154	0.0710	0.0643	0.0203	0.0153
	0.98	65	110	0.0680	0.0643	0.0202	0.0146
1	0.8	96	128	0.0836	0.0794	0.0190	0.0165
	0.98	65	87	0.0836	0.0794	0.0203	0.0162

Table 1: Persistence statistics comparison for before and Post data generation.

moons dataset in order to verify the correlation between TDA coverage and the signal-to-noise ratio, which distinguishes meaningful information in the data from non-significant noise on it.

The persistence diagrams reported in Figures 7 and 8 show a local improvement of by an increase in point density near the identity line and a decrease further away. It should be noted that the two figures are illustrating data generation for a trust Rips complex with radius of 0.15, then holes appearing with radius greater than this value would not be filled by the data generation step due to the local improvement. i.e., From the two figures, it can be seen that the persistence of holes appearing in the persistence diagram with high radius ($\gg 0.15$) are not getting an important change. Then, an iterative approach to fill gradually these holes by varying simplicial complex radius could be a solution (local identification).

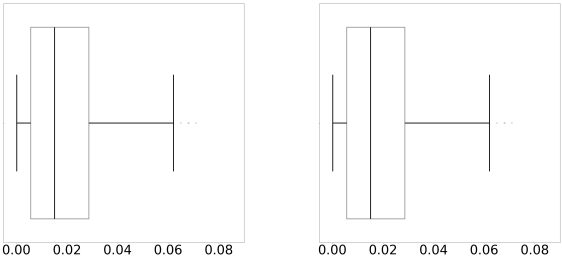
An additional evaluation based on persistence statistics is performed. In order to shed evidence in the behavior described concerning the density of points near the identity line. Table 1 carry over additional results by computing reductions of 25% and 28% for class 0, and 13% and 21% for class 1 in the mean persistence of the associated diagrams for $\alpha = 0.8$ and $\alpha = 0.98$, respectively. Then, despite the local data generation, a significant global improvement is obtained by the proposed approach.

The density with respect to points furthest from the identity line is measured using a metric based on the number of points within "confidence regions" as introduced in the work of (Chazal and Michel 2021). Specifically, for a given $\beta \in \mathbb{R}^+$, the confidence region of a persistence diagram is defined as the half-space $\{(x, y) \in \mathbb{R}^2 \mid y \geq x + \beta\}$. The results obtained using this metric are presented in Table 2. A reduction in the number of points furthest from the identity line is sufficiently evident for class 0. This reduction is less pronounced for class 1. This can be explained that class 1 is less sparse than class 0. We can see from the table that even after data generation, the class 1 is less sparse than 0. The data-sparse regions identified in Figure 5 are characterized independently to holes detected in the persistence diagram. Intuitively, this means that these regions contain only small holes that are quickly filled during the filtration process. This also highlights the limitations of visual analysis, emphasizing the necessity for Cartesian identification of holes and adaption of the filling parameters, in terms of number of generated data for a given hole, is discussed in a subsequent section.

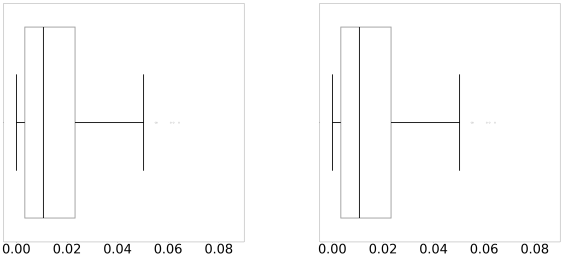
Another important point to consider regarding the proposed approach is the limitation imposed by the data distribution within the expected domain of the model M. Specifically, the expected domain of a given classification model adheres to a certain distribution, which may inherently contain numerous holes of various dimensions that are theoretically not possible to fill due to the nature of the data distribution. This implies that if we initiate Algorithm 1 with a sufficiently varied dataset, the mean persistence is likely to intuitively converge to a constant during the iterative process of the proposed approach. In many cases, the expected domain is infinite, making the exact mean persistence challenging to determine. However, it is possible to estimate it empirically by applying the law of large numbers. Thus, with a sufficiently large number of iterations, an empirical estimation of the mean persistence can be obtained at the end of the loop in the ideal case (considering other limiting factors such as the data generation algorithm, for instance). Furthermore, the workflow of proposed approach can be applied to other ML models, such as regression and object detection.

	α	Confidence regions BDG	Confidence regions PDG
Class 0	0.80	16	15
	0.98	11	9
Class 1	0.80	12	12
	0.98	9	8

Table 2: Number of points of the persistence diagrams in confidence region with $\beta = 0.04$ for before and post data generation.

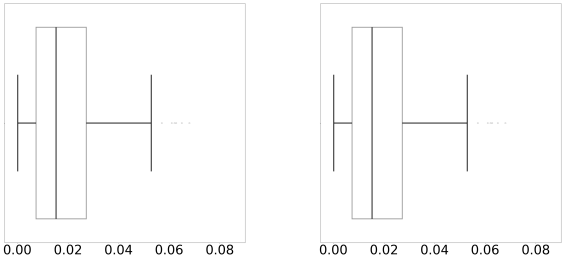


(a) Class 0, BDG (b) Class 1, BDG

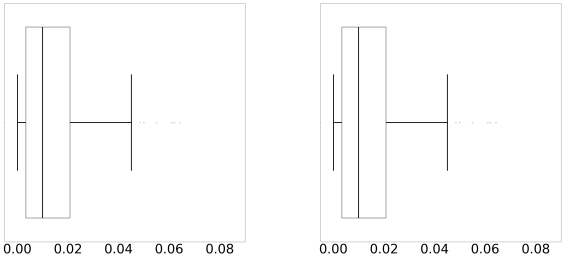


(c) Class 0, PDG (d) Class 1, PDG

Figure 9: Box plots representing statistics of persistence of trust simplicial complex with $\alpha = 0.8$ for before and after data generation (vertically) and class 0 and class 1 (horizontally), respectively



(a) Class 0, BDG (b) Class 1, BDG



(c) Class 0, PDG (d) Class 1, PDG

Figure 10: Box plots representing statistics of persistence of trust simplicial complex with $\alpha = 0.98$ for before and after data generation (vertically) and class 0 and class 1 (horizontally), respectively

Conclusions and Perspectives

Conclusions

In this work, we proposed, using topological concepts, to evaluate classification model coverage. By combining topological features from persistence diagrams with model confidence levels, we introduced the Trust Rips Complex, linking data structure with model performance.

Expanding datasets within a model's expected domain generally decreases persistence, filling sparse regions and improving coverage. Our statistical analysis showed that data augmentation led to a significant decrease in persistence statistics, indicating a more homogeneous data distribution. Visual analysis supported this, revealing that data augmentation reduced or eliminated holes in the initial dataset, as seen in the Trust Rips Complexes and persistence diagrams. Box plot analysis demonstrated a significant global reduction in persistence, highlighting the practical impact of our approach. Our framework effectively combines TDA with model confidence to enhance model coverage evaluation.

This study underscores the importance of thorough coverage evaluation for reliable DL-based system deployment. While our findings are promising, further research is needed to refine the data generation process and validate our approach on more complex datasets and higher-dimensional spaces, ensuring scalability and broader applicability.

Perspectives

The current exploratory study lays the groundwork for several promising future research directions. The perspectives discussed here outline the next steps for enhancing and expanding the scope of our topological coverage analysis for deep learning models.

Tests on Larger Datasets. To validate the robustness and scalability of our approach, it is essential to conduct tests on larger datasets. For instance, the MNIST dataset, which is widely recognized in the field of image classification, provides a more extensive and diverse set of samples. This will allow us to evaluate the performance of our method in handling a broader range of data points.

Data-Sparse Identification Algorithm. One limitation of the proposed approach is the inability to identify data-sparse regions. An interesting avenue for future research could focus on this topic, specifically on finding the Cartesian coordinates of points that define the characteristics associated with any point on a given persistence diagram.

Data Generation algorithm for Data-Sparse Regions. Addressing data-sparse regions is crucial for improving the reliability of the coverage assessment of a classification model. We propose developing a new topological diffusion algorithm specifically designed to generate data that correspond to these sparse regions.

Topological Metric for Local Coverage. Another promising avenue is the development of a topological metric for local coverage, based on statistics extracted from persistent

homology. This metric can be combined with other existing metrics to provide a more comprehensive assessment of model coverage. By integrating topological insights with conventional performance metrics, we can achieve a more nuanced understanding of the model's behavior and ensure more thorough coverage evaluation.

Distinguishing Noise from Signal. The distinction between noise and signal in persistence diagrams is often subjective, especially with small persistence data. To address this issue, a study can help identify objective boundaries between noise and signal, we aim to establish clearer criteria for distinguishing significant topological features from irrelevant noise. This will enhance the precision of our evaluation and contribute to more reliable interpretations of the results.

These perspectives not only build on the findings of our current study but also pave the way for innovative advancements in the field of topological data analysis and classification model evaluation. By addressing these areas, we aim to develop more robust, comprehensive, and reliable methods for assessing and improving the performance of AI models in real-world applications.

References

Adjed, F.; Chaouche, S.; Randon, Y.; Le Coz, A.; Herbin, S.; Karaliolios, N.; Winckler, N.; and Feuilleaubois, E. 2023. Data Quality Assessment Metrics for Machine Learning Process. Technical report, Confiance.ai program - IRT SystemX.

Adjed, F.; Mziou-Sallami, M.; Pelliccia, F.; Rezzoug, M.; Schott, L.; Bohn, C.; and Jaafra, Y. 2022. Coupling algebraic topology theory, formal methods and safety requirements toward a new coverage metric for artificial intelligence models. *Neural Computing and Applications*, 34(19): 17129–17144.

Chazal, F.; and Michel, B. 2021. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *Frontiers in artificial intelligence*, 4: 667963.

De Silva, V.; and Ghrist, R. 2006. Coordinate-free coverage in sensor networks with controlled boundaries via homology. *The International Journal of Robotics Research*, 25(12): 1205–1222.

Gehr, T.; Mirman, M.; Drachsler-Cohen, D.; Tsankov, P.; Chaudhuri, S.; and Vechev, M. 2018. Ai2: Safety and robustness certification of neural networks with abstract interpretation. In *2018 IEEE symposium on security and privacy (SP)*, 3–18. IEEE.

Guo, J.; Jiang, Y.; Zhao, Y.; Chen, Q.; and Sun, J. 2018. Dlfuzz: Differential fuzzing testing of deep learning systems. In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 739–743.

Harel-Canada, F.; Wang, L.; Gulzar, M. A.; Gu, Q.; and Kim, M. 2020. Is neuron coverage a meaningful measure for testing deep neural networks? In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 851–862.

Hayhurst, K. J. 2001. *A practical tutorial on modified condition/decision coverage*. DIANE Publishing.

Kim, J.; Feldt, R.; and Yoo, S. 2019. Guiding deep learning system testing using surprise adequacy. In *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*, 1039–1049. IEEE.

Ma, L.; Juefei-Xu, F.; Zhang, F.; Sun, J.; Xue, M.; Li, B.; Chen, C.; Su, T.; Li, L.; Liu, Y.; et al. 2018a. Deepgauge: Multi-granularity testing criteria for deep learning systems. In *Proceedings of the 33rd ACM/IEEE international conference on automated software engineering*, 120–131.

Ma, L.; Zhang, F.; Sun, J.; Xue, M.; Li, B.; Juefei-Xu, F.; Xie, C.; Li, L.; Liu, Y.; Zhao, J.; et al. 2018b. Deepmutation: Mutation testing of deep learning systems. In *2018 IEEE 29th international symposium on software reliability engineering (ISSRE)*, 100–111. IEEE.

Mani, S.; Sankaran, A.; Tamilselvam, S.; and Sethi, A. 2019. Coverage testing of deep learning models using dataset characterization. *arXiv preprint arXiv:1911.07309*.

Odena, A.; Olsson, C.; Andersen, D.; and Goodfellow, I. 2019. Tensorfuzz: Debugging neural networks with coverage-guided fuzzing. In *International Conference on Machine Learning*, 4901–4911. PMLR.

Pei, K.; Cao, Y.; Yang, J.; and Jana, S. 2017. Deepxplore: Automated whitebox testing of deep learning systems. In *proceedings of the 26th Symposium on Operating Systems Principles*, 1–18.

Sun, Y.; Huang, X.; Kroening, D.; Sharp, J.; Hill, M.; and Ashmore, R. 2018. Testing deep neural networks. *arXiv preprint arXiv:1803.04792*.

Sun, Y.; Huang, X.; Kroening, D.; Sharp, J.; Hill, M.; and Ashmore, R. 2019. Structural test coverage criteria for deep neural networks. *ACM Transactions on Embedded Computing Systems (TECS)*, 18(5s): 1–23.

Tian, Y.; Pei, K.; Jana, S.; and Ray, B. 2018. Deeptest: Automated testing of deep-neural-network-driven autonomous cars. In *Proceedings of the 40th international conference on software engineering*, 303–314.

Xie, X.; Ma, L.; Juefei-Xu, F.; Xue, M.; Chen, H.; Liu, Y.; Zhao, J.; Li, B.; Yin, J.; and See, S. 2019. Deephunter: a coverage-guided fuzz testing framework for deep neural networks. In *Proceedings of the 28th ACM SIGSOFT international symposium on software testing and analysis*, 146–157.

Yan, S.; Tao, G.; Liu, X.; Zhai, J.; Ma, S.; Xu, L.; and Zhang, X. 2020. Correlations between deep neural network model coverage criteria and model quality. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 775–787.

Zohdinasab, T.; Riccio, V.; Gambi, A.; and Tonella, P. 2023. Efficient and effective feature space exploration for testing deep learning systems. *ACM Transactions on Software Engineering and Methodology*, 32(2): 1–38.