Topology Meets Truth: Leveraging Persistent Homology in Deepfake Identification

Srikar Kunapuli

October 19, 2025

0.0.1 Motivation

The proliferation of AI has dramatically transformed the landscape of digital media, with deepfake technology and automated content playing pivotal roles in shaping online discourse. In recent years, social media platforms have witnessed an influx of hyperrealistic images, videos, and text produced by generative AI. While these advances offer new modes of artistic and informational expression, they present profound challenges regarding authenticity, misinformation, and public trust. Deepfakes—AI-generated synthetic media that can convincingly emulate real individuals—pose a unique threat by enabling the creation of deceptive content that can be difficult to distinguish from human-generated media.

Modern approaches to deepfake detection rely on deep learning and computer vision techniques to identify subtle inconsistencies in manipulated videos and images. For instance, a deepfake detector may analyze pixel-level features, facial landmarks, lighting mismatches, eye-blinking patterns, and audio-visual synchronization to distinguish between authentic and synthetic media. Common approaches include convolutional neural networks (CNNs) trained on large datasets of real and fake faces, frequency-domain analysis detecting abnormal signal patterns, and multimodal fusion models assessing both visual and auditory cues. Despite these advances, modern detectors face several limitations: they are often vulnerable to adversarial attacks, struggle to generalize across unseen datasets or new generation techniques, and may produce high false-positive rates when detecting low-quality or compressed media. Moreover, as generative AI models continue to improve realism, the differences that detectors rely on become increasingly imperceptible.

Topological Data Analysis (TDA) offers a promising solution to address these challenges by focusing on intrinsic structural features of data rather than pixel-level cues. Through methods like persistent homology, TDA can capture the "shape" of data distributions across layers of deep networks. By analyzing homology, TDA reveals topological patterns that remain invariant under transformations such as rotation, lighting, or compression. By integrating TDA with deepfake detection pipelines, researchers can develop models that are more robust to distribution shifts and adversarial manipulations. Essentially, TDA can provide a higher-level, noise-resistant representation of visual and temporal coherence in facial motion or texture patterns, enabling detectors to generalize better to new types of deepfakes while maintaining interpretability and resilience.

For this project, I will be implementing a methodology inspired by Nathan Weaver, Max Logalbo, and Jonathan Pipping, who proposed a TDA framework for classifying AI-generated faces in their publication *Topological Data Analysis for Classification of AI Generated Faces (2024)*. Their study demonstrates how persistent homology and cubical filtrations can capture geometric and structural differences between authentic and synthetic images. I will adapt and extend their approach by applying similar filtrations (height, radial, density, and dilation) to my own dataset, extracting persistence diagrams and landscapes, and integrating these features into machine learning classifiers for evaluating real versus AI-generated facial images.

0.0.2 Data and Resources

For this project, I will utilize the *Deepfake Face Image Classification Dataset* by Pujan Paudel, available on Hugging Face: https://huggingface.co/datasets/pujanpaudel/deepfake_face_classification. This dataset contains 16,060 AI-generated images and 16,060 authentic images in JPEG format. The images are organized by class, and each fake image was generated using a range of modern deepfake

synthesis techniques that manipulate facial features and expressions to mimic authentic visuals. The dataset provides a balanced distribution between real and synthetic samples, making it well-suited for training and evaluating deepfake detection models.

In preprocessing, I plan to resize all images to a uniform resolution, grayscale pixel values before applying TDA to study geometric structures in feature space. Expected challenges include dealing with variability in image resolution, compression artifacts, and noise introduced by generation methods, which can affect feature consistency and topological stability across samples.

0.0.3 Methods

In this research, I will adopt and extend the methodological framework introduced in *Topological Data Analysis for Classification of AI Generated Faces (2024)*, applying persistent homology-based analysis of facial images to distinguish between AI-generated and authentic faces. The dataset used will be the Deepfake Face Image Classification dataset. All images will be resized to a uniform resolution and converted to grayscale for consistency. Let $\mathcal{P}$ define a pixel space.

$$\mathcal{P} = \{1, 2, \ldots, W\} \times \{1, 2, \ldots, H\},$$

where W and H denote the image width and height, respectively. Each grayscale image I_n will then be treated as a function mapping this pixel space $\mathcal{P}$ to intensity values $\mathbb{R}$. To discretize the continuous grayscale images, I will apply a binarization function with a chosen threshold t:

$$B(p) = \begin{cases} 1, & \text{if } p \geq t, \\ 0, & \text{if } p < t. \end{cases}$$

After binarization, I obtain $\mathcal{B}(I_n)$, which can be used to construct a set of filtrations based on cubical complexes. In this work, I will construct and analyze four filtrations: *height*, *radial*, *density*, and *dilation*.

$$H(i) = \begin{cases} \langle i, u \rangle, & \text{if } B(i) = 1, \\ \max\{\langle p, u \rangle : B(p) = 1\}, & \text{if } B(i) = 0. \end{cases}$$

and similarly define:

$$R(i) = \begin{cases} \|c - i\|, & \text{if } B(i) = 1, \\ \max\limits_{j \in I}\{\|c - j\|\}, & \text{if } B(i) = 0. \end{cases}$$

I will also compute the density filtration:

$$D_r(i) = \big|\{\, p \in I : B(p) = 1 \text{ and } \|p - i\| \leq r \,\}\big|.$$

and the dilation filtration:

$$D(i) = \min\{\, \|i - p\| : B(p) = 1 \,\}.$$

After generating filtrations, I will compute cubical persistence diagrams using `Giotto-TDA` and extract persistence landscapes, persistence entropy, and Wasserstein amplitudes. I will then standardize these features and classify them as real or fake using a Random Forest Classifier and a Support Vector Classifier (polynomial kernel). Then I will run my classifier on the dataset to the test the validity of my classifier.

0.0.4 Expected Outcomes

Each image I_n is expected to produce persistence diagrams for both H_0 and H_1 topological features across four filtrations, yielding a collection of $4n^2$ diagrams. These diagrams will visualize connected components and loops, illustrating geometric and structural complexity. Each diagram will then be transformed into a persistence landscape for use in machine learning. Additional descriptors—*Wasserstein amplitude* and *persistence entropy*—will quantify magnitude and diversity of persistence features.

After feature extraction, PCA visualizations will assess class separability between real and AI-generated faces. I expect partial clustering of classes in PCA space and stronger nonlinear separation via SVC. Overall, topological signatures derived from persistence landscapes and entropies should yield measurable accuracy in differentiating synthetic from authentic images.

0.0.5 Timeline

Week	Dates	Milestones and Objectives
1	Oct 20 – Oct 26	Set up Python environment; preprocess dataset; implement binarization.
2	Oct 27 – Nov 2	Implement height, radial, density, and dilation filtrations.
3	Nov 3 – Nov 9	Compute persistence diagrams and landscapes with `Giotto-TDA`.
4	Nov 10 – Nov 16	Extract entropy and Wasserstein features; normalize data; perform PCA.
5	Nov 17 – Nov 21	Train and evaluate Random Forest and SVC models with cross-validation.
6	Nov 22 – Nov 25	Interpret results, analyze filtrations, and finalize report for submission.

Table 1: Timeline and milestones for the Topological Deepfake Detection Project.

0.0.6 References

https://www.causeweb.org/usproc/sites/default/files/usresp/2024-1/usresp%203357%20-%20topological%20data%20analysis%20for%20classification%20of%20ai%20generated%20faces.pdf

Weaver, N., Logalbo, M., & Pipping, J. (2024). *Topological data analysis for classification of AI-generated faces.* Undergraduate research project, USPROC/USRESP, University of Florida. https://www.causeweb.org/usproc/sites/default/files/usresp/2024-1/usresp%203357%20-%20topological%20data%20analysis%20for%20classification%20of%20ai%20generated%20faces.pdf

Tauzin, G., Lupo, U., Tunstall, L., Pérez, J. B., Caorsi, M., Reise, W., Medina-Mardones, A., Dassatti, A., & Hess, K. (2021). *Giotto-TDA: A topological data analysis toolkit for machine learning and data exploration.*

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society, 6*, 205630512090340. http://dx.doi.org/10.1177/2056305120903408