# The Shape of Digits

## A Bayesian Topological Data Analytic Approach to Classification of Handwritten Digits

Thomas Reinke  Theophilus A. Bediako  Daniel Lim

August 16, 2025

## Table of contents

# 1 Abstract

*This paper ...*[1]

# 2 Introduction

The motivation for this work was to compare and examine how dimension reduction via topological data analysis can be used with machine learning and classification models. Algebraic topologist, Gunnar Carlsson, has a quote, "Data has shape, shape has meaning, and meaning brings value." The work in this paper follows this idea, that if there is inherent structure present in data, it can be exploited to aid in modeling.

The Modified National Institute of Standards and Technology database is a set of handwritten digits that is frequently used to train and test image processing models. Our goal is to classify handwritten digits to their correct numeric label. We compared model performance primarily on accuracy, and secondly on the number of features or predictors. First, we want to build an accurate model that can discriminate between digits. Similar to the idea of parsimony in model selection, if two models have comparable accuracy, we will favor the model trained on fewer features. We also consider the ability to make inference on the predictors of a model.

# 3 Background and Related Work

## 3.1 MNIST

The development of the MNIST dataset stems from early efforts in optical character recognition (OCR), a field of automating the processing of handwritten information like postal codes and census forms. Its lineage can be traced back to the late 1980s with the creation of the USPS database, a collection of 16×16 grayscale images of handwritten zip codes used to train the LeNet neural network. During the same period, the National Institute of Standards and Technology (NIST) was developing its own Special Databases for OCR research, sourcing images from census forms and high-school student samples. One such collection, SD-7, released in 1992, would later form a core part of the MNIST test set.

Challenges in generalizing models across these different datasets, highlighted during a 1992 NIST/Census Bureau competition, revealed biases within the original NIST data. To address these issues, the MNIST database was created in 1994, providing a cleaner, more standardized benchmark for the machine learning community. The dataset led to the development of several modern variations. These include EMNIST (2017), which extended the character set to include letters; QMNIST (2019), which restored the complete original test set; and Fashion MNIST (2017), a drop-in replacement featuring images of clothing items.

[1] Example footnote

### 3.2 Related Work

Topological Data Analysis is a technique for extracting structural features from datasets, proving effective in classification tasks across various domains. Work in this area includes that of Nicolau et al., who successfully used a topology-based approach to identify a distinct subgroup of breast cancers with excellent survival rates, demonstrating the method's ability to uncover patterns invisible to other methods (Nicolau, Levine, and Carlsson 2011). Researchers have developed more generalized TDA-based classification methods, applying them to problems involving multiple measurements and other complex data structures (Riihimäki et al. 2019; Kindelan et al. 2021).

Many traditional machine learning models have been successfully applied to MNIST (Yeboah 2025), but it has also been a subject of interest for topological methods. Garin and Tauzin provide a "Topological 'Reading' Lesson" by applying TDA specifically to classify MNIST digits, showing that topological features can serve as effective predictors (Garin and Tauzin 2019).

We try an integration of a Bayesian framework to the TDA methodology. This is inspired by the Bayesian framework for persistent homology by Maroulas et al. (Maroulas, Nasrin, and Oballe 2020). Here, we can quantify uncertainty in topological features, in an attempt to increase the predictive power of TDA-based classification models.

## 4 Methodology

### 4.1 Traditional Machine Learning

#### 4.1.1 Neural Networks

**Dropout**

**Ridge**

**Lasso**

**Multinomial Logistic Regression as ML**

### 4.2 Bayesian TDA Methodology

*latex flowchart*

#### 4.2.1 Cubical Complexes & Persistent Homology

#### 4.2.2 Marked PPP

#### 4.2.3 Bayes Update & Guassian representation

### 4.3 TDA + ML

#### 4.3.1 Filtering

## 5 Experiments

### 5.1 Data Summary

- **Data Composition**:
  - **Training set**: 60,000 images (50% SD-3, 50% SD-7).
  - **Test set**: 10,000 images (originally 60,000, later reduced to 10k).
- **Image Processing**:
  - Resized to **28x28 pixels** with anti-aliasing.
  - Normalized to center of mass alignment.

#### 5.1.1 EDA

### 5.2 NN dropout

### 5.3 NN Ridge

### 5.4 NN Lasso

### 5.5 Multinomial Logistic

### 5.6 Bayes TDA

### 5.7 TDA + ML

## 6 Discussion and Analysis

*Comparing the models*

## 7 Conclusion

## 8 References


Garin, Adélie, and Guillaume Tauzin. 2019. "A Topological "Reading" Lesson: Classification of MNIST Using TDA." *CoRR* abs/1910.08345. http://arxiv.org/abs/1910.08345.

Kindelan, Rolando, José Frías, Mauricio Cerda, and Nancy Hitschfeld. 2021. "Classification Based on Topological Data Analysis." *CoRR* abs/2102.03709. https://arxiv.org/abs/2102.03709.

Maroulas, Vasileios, Farzana Nasrin, and Christopher Oballe. 2020. "A Bayesian Framework for Persistent Homology." *SIAM J. Math. Data Sci.* 2 (1): 48–74.

Nicolau, Monica, Arnold J. Levine, and Gunnar Carlsson. 2011. "Topology Based Data Analysis Identifies a Subgroup of Breast Cancers with a Unique Mutational Profile and Excellent Survival." *Proceedings of the National Academy of Sciences* 108 (17): 7265–70. https://doi.org/10.1073/pnas.1102826108.

Riihimäki, Henri, Wojciech Chachólski, Jakob Theorell, Jan Hillert, and Ryan Ramanujam. 2019. "A Topological Data Analysis Based Classification Method for Multiple Measurements." *CoRR* abs/1904.02971. http://arxiv.org/abs/1904.02971.

Yeboah, Felix. 2025. "Classification and Evaluation of Machine Learning Algorithms on the MNIST Dataset."