# Topological interpretability for deep learning

Adam Spannaus[1*], Heidi A. Hanson[1], Lynne Penberthy[3], Georgia Tourassi[2]

**1** Computational Science and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA
**2** National Center for Computational Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, USA
**3** National Cancer Institute, Bethesda, MD, USA

* spannausat@ornl.gov

## Abstract

With the increasing adoption of AI-based systems across everyday life, the need to understand their decision-making mechanisms is correspondingly accelerating. The level at which we can trust the statistical inferences made from AI-based decision systems is an increasing concern, especially in high-risk systems such as criminal justice or medical diagnosis, where incorrect inferences may have tragic consequences. Despite their successes in providing solutions to problems involving real-world data, deep learning (DL) models cannot quantify the certainty of their predictions. And are frequently quite confident, even when their solutions are incorrect.

This work presents a method to infer prominent features in two DL classification models trained on clinical and non-clinical text by employing techniques from topological and geometric data analysis. We create a graph of a model's prediction space and cluster the inputs into the graph's vertices by the similarity of features and prediction statistics. We then extract subgraphs demonstrating high-predictive accuracy for a given label. These subgraphs contain a wealth of information about features that the DL model has recognized as relevant to its decisions. We infer these features for a given label using a distance metric between probability measures, and demonstrate the stability of our method compared to the LIME interpretability method. This work demonstrates that we may gain insights into the decision mechanism of a DL model, which allows us to ascertain if the model is making its decisions based on information germane to the problem or identifies extraneous patterns within the data.

## Introduction

Understanding the process through which any machine learning (ML) algorithm makes a decision is crucial for high-risk domains: this is especially true in both criminal justice and healthcare. Whether humans either make a decision with the assistance of an algorithmic process or delegate the entirety of the decision-making process, the need for insight is paramount. Indeed, the desiderata for assurance and reliability across AI was identified as a grand challenge for establishing foundational scientific AI advancements [1]. Due in part to the black-box nature of many AI-based algorithms [2] the field of machine-assisted medical decision-making has been slow to adopt this paradigm. Even when their solutions turn out to be incorrect, AI-based models provide high confidence levels in their predictions [3]; the medical professional's hesitancy to bring machine assisted decision-making into the clinical setting is understandable. Typical deep learning (DL) models cannot either reliably quantify the confidence in their predictions, nor provide an explanation for any particular decision.

From a given set of inputs, medical professionals must be able to trust the decisions of the DL model and understand why it made a specific prediction to reliably use the predictions in their practice. However, as noted in [4], trusting the predictions of a model is not sufficient; one must have trust in the model as well. The idea of trusting predictions is straightforward, it answers the question of whether or not the end-users have sufficient confidence in a prediction to take action. Trust in a model is more subtle; to engender it, one must demonstrate that a prediction is made for the right reasons.

To increase the assurance in an AI model's predictions, we propose an interpretability method that is based on the topological and geometric properties of the model's prediction space. We distinguish salient features in the input data that both inform the classification of each input, and identify clusters of inputs with similar features and classification performance. Moreover, as defined by [5], our proposed methodology is *justified*. The authors of [5] argue that explanations from any interpretability method must to be tied to the training data set, ensuring that the explanations are not the result of a modeling artifact, but rather an inherent feature of the data. Our proposed methodology includes training data in our topological construction of the prediction space, and both global and local explanations are given with respect to the training data.

We create a topological representation of the AI model's prediction space through the Mapper algorithm [6]. Our construction clusters inputs by their ground truth class into regions of high and low predictive accuracy. We then engage in an unsupervised semantic analysis of the documents in these clusters to better understand the trained model's decision mechanism. This analysis reveals important words to our classification model both for documents it classifies correctly, and those it predicts incorrectly. As a direct result of this analysis, we obtain insight into similarities between correctly classified documents, and why others are incorrectly classified.

Previous approaches for post-hoc interpretability, such as LIME [7] or SHAP [8], perturb the original data, then sample from these perturbations to make their inferences. These methods then either fit a proxy model which is more interpretable [7], or ranks the feature importance, which inform a particular classification [8]. However, for similar inputs, these methods can provide very different results. As shown in [9], the explanations provided by LIME on a collection of UCI datasets demonstrated both instability and high variance. Moreover, these methods infer counter-factual explanations. Such explanations are a minimal perturbation necessary to change the prediction of the proxy model, and may not generate explanations that are connected to the ground truth [10]. Consequently, to establish both meaningful interpretations and assurances in the decisions from an AI model, perturbation and sampling based approaches are insufficient.

Here we present an interpretability method for any AI model that yields a topologically correct construction of the model's prediction space. Our method yields both global and local views of the features informing a decision. For the global view, we extract features associated with samples classified with high predictive accuracy and create a low-dimensional representation of our model's prediction space. We provide a method for inferring the features related to the classifying any input to the model. We demonstrate the stability of our method as compared with LIME [7] and present results on two datasets. First, we consider a multi-task convolutional neural network (MTCNN) taking cancer pathology repots as input and predicting cancer phenotype [11].

The second model is a CNN trained on the publicly available 20newsgroups dataset [12]. Through our analysis of these datasets, we see that our method is able to identify relevant features in the training data which inform the trained model's decision mechanism.

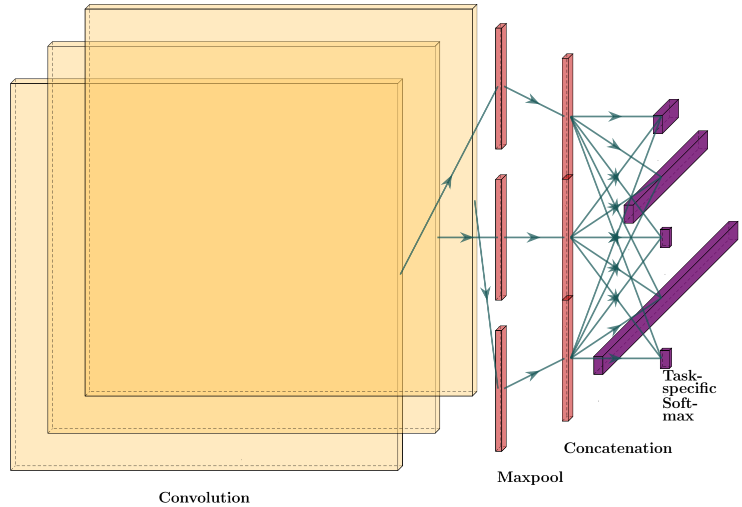


**Fig 1. Multi-task convolution neural network architecture** Our multi-task convolutional neural network architecture used for cancer pathology report information extraction. The network consists of three parallel convolutional filters, followed by a maxpooling and concatenation layer. To each maxpooling layer, we employ a dropout rate of 50%. After applying the dropout, we concatenate the remaining vectors and pass them to a task-specific softmax classification layer.

# Materials and methods

## Multi-task Deep Learning Model

The MTCNN is a deep learning (DL) architecture for information extraction from cancer pathology reports [11, 13]. It takes as input a pathology report in which the words have been tokenized, i.e., mapped to a unique integer value, and outputs simultaneous predictions about five cancer phenotypes (tasks): primary site, subsite, histology, laterality, and grade. The parallel convolutional layers in the MTCNN model extracts features, which are concatenated into a single feature vector (the document embedding), which is subsequently passed to task-specific softmax classification layers. See Figure 1 for a visual representation of the neural network's architecture.

## Topological Data Analysis

Topological Data Analysis (TDA) is a field of mathematics used to extract valuable information about connections within large and high-dimensional datasets, by applying the abstract ideas and techniques from topology. Three key ideas that differentiate topology from other geometric methods for data analysis are: (a) *coordinate invariance*; (b) *deformation invariance*; and (c) *compressed representation*.

One of the primary building blocks of topology is the concept of a metric space. It is a set of points equipped with a notion of distance between any pair, and gives rise to the notion of coordinate invariance. Traditional methods of data analysis such as Principal Component Analysis (PCA) or multidimensional scaling embeds the data into Euclidean space, and thus swaps any data-centric notion of similarity for the traditional Euclidean metric. If the data is indeed non-Euclidean, this embedding distorts relevant features. Topological methods are insensitive to the choice of distance metric and within the data-space, measures of closeness between a point and a subset.

The property of deformation invariance is central to any discussion of topological methods, as this approach allows for shapes to be stretched, contracted, and/or twisted in such a way that no new holes are created, nor are existing holes closed. Any continuous transformation preserves the topological features present within the data.

Lastly, the idea of compressed representation is best illustrated through an example. Consider a circle, which is composed of an infinite number of points. Topologically speaking, its essential feature is that it forms a closed loop. Now if we instead consider a hexagon, it too forms a closed loop and is topologically equivalent to the circle. Although geometric curvature information is lost by adopting this point of view, such trade-offs are common throughout computational mathematics. Indeed, viewing the circle as an hexagon, we may now use it in computations as the hexagon requires six points to encode, as opposed to an infinite number for the circle.

We apply techniques from the mathematical study of shape to extract fundamental information about the intrinsic relationships present within our data. Relying on the topological notion of ’closeness’, we may extract insights from our data that are unavailable through rigid geometric notions of shape and distance. The end result of our analysis is a low dimensional representation of our data that is amenable to both further mathematical analysis and visualization.

## Mapper Graph

We will now discuss the Mapper graph, the primary tool through which we create a topologically consistent construction of an AI model’s prediction space.

The Mapper graph was proposed as an algorithmic method to study high-dimensional data via simplicial complexes through chosen filter function(s) [6]. These functions are used to indicate properties of the data, such as a kernel density estimate, distance, or eccentricity. This construction has found applications in such diverse fields as spinal cord injuries, breast cancer omics studies, and voting trends [14–16]. The method has been used as a feature selection tool for identifying subpopulations within a larger group [6, 16], and recently the algorithm has been applied in an error analysis and activation functions of a DL model [17, 18].

The methodology creates a low-dimensional representation of high-dimensional data while preserving the original data’s topology. Similar to PCA or a spectral embedding, the Mapper construction may be thought of as an unsupervised clustering algorithm. In both PCA or a spectral embedding though, the data is assumed to be in Euclidean space; the Mapper construction creates a graph, or higher dimensional simplicial complex, yielding a combinatorial representation of the data without any assumptions on the data space [19]. This provides a scaffold for a topological construction by reducing the problem of approximating a shape to a linear algebra problem. This process is analogous to reconstructing a complex shape from a suitably chosen combination of points, lines, triangles, and their higher dimensional analogues.

The primary idea of the Mapper construction is to summarize a dataset by creating a neighbor graph of the data. To this end, we create subsets of the filter function’s domain, and apply a clustering method to these subsets. The resulting graph captures clusters of inputs that exist solely in one subset, resulting in a vertex of the graph or those inputs appearing in multiple subsets, which creates an edge of the graph. Through this construction, we are able to realize two types of neighbor relationships in a dataset: (*i*) 0-dim information, those points that are assigned to a specific cluster, resulting in a vertex; and (*ii*) 1-dim information, those points that are assigned to multiple clusters which produce an edge on the graph, i.e., these data points contain some shared attributes. Consequently, the ensuing graph construction is topologically faithful to the original point cloud. Inputs having similar features will form a connected subgraph, and be disconnected from dissimilar inputs, thus shape and neighbor relationships in the original data are preserved through this construction. This result is is non-trivial, we omit the technical details here; see [6, 19–21] and references contained therein.

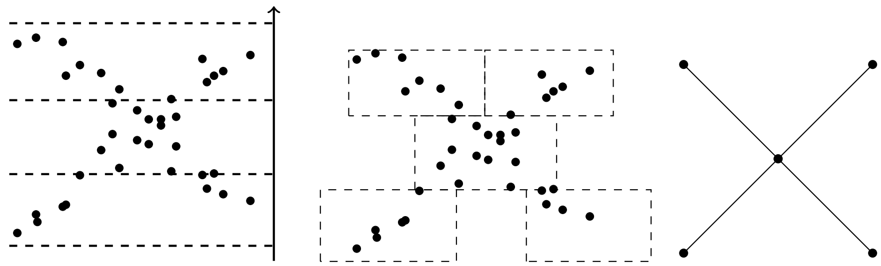


**Fig 2. Illustration of the Mapper algorithm** In (a), observe a point cloud, filter function, chosen as the height function, and an open cover, denoted by the dashed lines. Fig. (b) we have covered the range of values in (a) by consecutive intervals. Clustering on the intervals in (b), and creating the *nerve*. In fig. (c), we have the resulting Mapper construction, demonstrating the compressed representation of the original data.

## Mapper Algorithm

We will now describe the process of creating a Mapper graph. The construction begins with point cloud $\mathbb{X}_P = \{X_i\}_{i=1}^{P}$, $X_i \in \mathbb{R}^n$ for $n \in \mathbb{N}$, for a sample of $P$ points, $\mathbb{X}_P$. Letting $f : \mathcal{X} \to \mathbb{R}$ be continuous, which is called a *filter function*. Through this function, one may show the equivalence of the Mapper reconstruction and the underlying topological space in question by observing that vertices of a graph that are connected by an edge, of $f^{-1}(y)$ for some $y$ in the image of $f$.

Any collection of continuous functions that form an open cover of their image may be chosen as the filter in a Mapper construction. The preimages of this collection then form an open cover of the data space. On this open cover of the data space, we apply a clustering algorithm, identifying the connected components of the data space. The clustering algorithm is iteratively applied to subsets of the preimage that are connected through the filter functions. Intuitively, one may think of this clustering process as repeated partial clusterings on the preimage of the filter functions. This repeated clustering ensures similarity of features in each vertex of the Mapper graph. Employing different filter functions in a construction may highlight different aspects of a dataset. Indeed, as these functions separate the input data into clusters that have different values under the filter function, one may create multiple Mapper graphs from the same data, each with a different filter to emphasize different aspects of the dataset. Visually, the algorithm proceeds as in Figure 2 and is written:

1. Choose $j$ filter functions such that $f_i : \mathcal{X} \to \mathbb{R}, 1 \leq i \leq j$.
2. For each filter function, choose a number of intervals $S_i$ creating a cover $\mathbb{U}_i = \{U_s\}_{s=1}^{S_i}$ that covers $f_i(\mathbb{X}_N)$.
3. Apply clustering algorithm to the pre-image $f_i^{-1}(U_i)$, defining a *pullback cover* $C$ on $\mathbb{X}_N$.
4. The Mapper graph $\mathcal{M}$ is the *nerve* of $C$; each vertex of $\mathcal{M}$ corresponds to one $c_{s,k} \in C$ and vertices $v_{s,k}, v_{s',k'}$ are connected iff $c_{s,k} \cap c_{s',k'} \neq \emptyset$.

The choice of clustering algorithm is left to the user, as is defining the resolution (the number of intervals), and gain (the percent overlap between intervals). We follow the guidelines described in [20] to set these two hyper-parameters.

## Mapper on an AI Model

Crucial to gaining insight and interpretability of a model is the choice of the filter in the Mapper algorithm. As previously noted, these functions enforce the separation of data points with sufficiently different filter values into separate vertices in the graph. By choosing the ground truth class of the input as one of the filters, we ensure that the datapoints in any of the graph's vertices are homogeneous with respect to the ground truth. Partitioning the input space in such a manner, we may investigate relevant features of our deep learning model by: ($i$) identifying clusters of inputs that are classified with high predictive accuracy sharing the same label; or ($ii$), identifying those features associated with low-predictive accuracy for given class label.

Our construction clusters input features by similarity, ground truth label, and predicted probability of the ground truth label. Furthermore, these clusters contain documents with a similar document embedding layer, i.e., the concatenation layer of our MTCNN model (Figure 1), thus the individual features in the documents are similar. We seek to infer the pertinent features leading to both correct and incorrect classifications with a high probability, gaining insight into the AI model's decision mechanism.

Clustered by similarity of the document embedding layer in our MTCNN model, the nodes of our Mapper graph may contain anywhere from a small handful to hundreds of documents. To provide interpretable results for the end-user, we must identify the relevant features in these documents from regions of high predictive accuracy.

## Geometric Analysis on a Probability Measure

In a classification problem, one seeks to infer the probability that a given collection of features belongs to 1-of-$k$ classes. The inverse question of interpretability may be posed in a similar manner, where the given data is a class label, and we seek a collection of features informing the particular classification.

To infer the relevant features from the data informing a classification, we investigate the probability density

$$\pi(w \in \mathcal{V} \mid \mathcal{H}), \tag{1}$$

where $\mathcal{H} = \{X \mid p(y \mid X) \geq \alpha, y \in \{1, \ldots, k\}, \alpha \in [0,1]\}$, is the set of all inputs that have confidence scores from an AI model of at least $\alpha$. That is a model taking input $X \in \mathbb{R}^{N \times d}$ and outputs a confidence score $p(y \mid X), y \in \{1, \ldots, k\}$ for a $k$-class classification problem, where $\hat{y} = \operatorname{argmax}_{y \in \{1,\ldots,k\}} p(y \mid X)$ is the predicted label of $X$. Lastly, we define $\mathcal{V} = \{(x, n) \mid n \in \mathbb{N}, x = X_{i,*}, X \in \mathcal{H}\}$ as the multiset of all rows of $X$ and $X_{i,*}$ is the $i^{th}$ row of $X$. We investigate the support of Equation (1) via the distance to measure (dtm) function of [22].

We will present a summary of the theory necessary for our proposed methodology; for a thorough treatment, see [22, 23] and references therein. The dtm function was initially proposed as a technique for outlier detection from a noisy sample where the end goal was manifold reconstruction and inference of topological invariants. Formally, the dtm measures the distance from a Dirac mass to a probability measure. We will first define the pseudo-distance $\delta_{\mu,m}$ which is required for the dtm function [22], whose definition follows.

**Definition.** Given any $0 \leq m < 1$ and a probability measure $\mu$ on $\mathbb{R}^d$, define the pseudo-distance $\delta_{\mu,m}$ as

$$\delta_{\mu,m} : x \in \mathbb{R}^d \mapsto \inf\{r > 0 : \mu\left(\overline{B}(x,r)\right) > m\},$$

where $\overline{B}(x,r)$ is the closed ball of radius $r$ centered at $x$. This is the probabilistic interpretation of a distance function that measures the distance from a point to a compact set.

We now define the distance to measure of [22].

**Definition.** Assume that $\mu$ is a probability measure on $\mathbb{R}^n$ and $\hat{m} \in (0,1)$ is a parameter. Then the distance function to $\mu$ with parameter $\hat{m}$ is defined by:

$$\mathrm{d}^2_{\mu,\hat{m}} : \mathbb{R}^d \to \mathbb{R}^+, x \mapsto \frac{1}{\hat{m}} \int_0^{\hat{m}} \delta_{\mu,m}(x)^2 \mathrm{d}m. \tag{2}$$

The function defined in Equation (2) retains properties of a traditional distance function, but is less sensitive to noisy outliers. Indeed, in the case that $m = 0$ we recover the traditional Euclidean distance, but when $m > 0$, we see that the dtm will have greater distances than the traditional Euclidean distance, since some portion of mass $m$ will be included in the closed ball centered at $x$.

Through our Mapper construction, we have an topologically faithful reconstruction of $\mathcal{H}$ and want to study the distance from individual features to the probability measure defined in Equation (1). Our goal is to gain insight into the features that a DL model uses in its classification rules. By examining the distances to the support of this probability measure, we may identify the key features in the learned classification rules. Intuitively, we may think of the dtm process as yielding an estimate of the density [24], where the density is over the distribution of words within nodes of our mapper graph.

## Feature Extraction

While the output from the Mapper algorithm clusters documents with similar predictive accuracies into vertices; it does not yield relevant features present within the documents informing the clusters in the graph's vertices. As our Mapper construction partitions the output space by ground truth labels, we may consider the connected subgraph for a specific ground truth label, and infer the pertinent features informing the classifications, either correct or incorrect, for a specific ground truth label.

To infer the relevant features for a classification we investigate the structure of the conditional probability distribution defined in Equation (1). We compute the dtm values for the subset of the vocabulary present in the documents clustered in the graph's vertices having at least a predictive confidence level of $\alpha$ from the multiset of all words contained in the documents with the same label. In practice this involves finding the $K-$nearest neighbors to each word in the label specific subset from the multiset of words contained in all documents with the same label. To then compute Equation (2) we use the discrete approximation

$$\hat{d^2}_{\mu,\hat{m}}(x) = \frac{1}{K}\sum_{i=1}^{K} \|x - X_i(x)\|^2, \tag{3}$$

where $X_i(x)$ denotes an ordering of the data so that $\|x - X_j(x)\|^2 \geq \|x - X_\ell(x)\|^2$ for $j < \ell$ and then $\hat{d^2}_{\mu,\hat{m}}(x)$ is simply the average squared Euclidean distance to the point $x$. Intuitively, we are interested in the word density in the graph vertices with a high predictive accuracy. Computing this density is not computationally tractable, due to the dimension of the word embedding vectors and the length of the documents. Previous works have noted the connection between computing the dtm is related to the density estimation problem [24], and we leave the problem of density estimation of the word vectors for a future investigation.

We choose $K$ dependent on the size of the vocabulary comprised of all words in the documents having the same label, so that $\hat{m} = [0.05, 0.25]$. This value controls the amount of mass contained in the ball centered at each point in the multiset, so higher values will contain more points, and is dependent on the data itself. Some guidance in choosing this parameter is given in [22], but rules for choosing this parameter optimally remains an open question.

# Results

## Cancer Pathology Reports

The data for our information extraction task is comprised of cancer pathology reports collected by the NCI Surveillance, Epidemiology, and End Results (SEER) program. These cancer pathology reports are collated and analyzed by the Louisiana cancer registry.

The reports maintained by this program are comprised of highly-technical, structured text that describes the characteristics of a cancerous tumor at the time of initial diagnosis. Each tumor diagnosis is assigned a unique tumor ID, and a report is created, documenting these characteristics associated with that specific case. The SEER registries abstract the information contained in these reports relevant to the characteristics associated with cancer site, subsite, histology, laterality, and behavior to monitor cancer incidence in population for cancer control and prevention, inform research, and advise on new and emerging research directions.

Before being used in the information extraction task, each pathology report was preprocessed according to the methodology described in [11]. Each of the remaining words in the cleaned documents were tokenized and padded to a uniform length of 3,000 tokens, and was used as input to the MTCNN [11]. From our trained model, we then made multiple predictions on samples from the training set, allowing the dropout layer to vary between predictions, specifically connections to the subsequent layer were randomly activated with a probability of 0.5; this ensured differences in the network connections, and consequently the confidence scores, between subsequent predictions on each document. We then computed the average softmax score for each document and took the class with the greatest score as the average predicted class for each document. We also computed the average softmax score of the ground truth class for each document. We used these two values, in conjunction with the ground truth label, and two-dimensional embedding [25] of the words in each document as inputs to create our mapper graph. These results are presented in Figure 3.

Additionally, we may examine subgraphs in Figure 3, such as those with primary site originating within the digestive organs, as shown in Figure 4. Within the subgraphs presented in Figure 4, we observe large connected graphs associated with a high predictive accuracy, and smaller, disconnected graphs where the model has low predictive accuracy on the documents clustered into the graph's nodes. This figure also illuminates the difficulty the trained model has with primary sites that are not well-defined. Primarily, we see multiple small disconnected graphs with labelled "C26: other digestive organs" and "C24 : other digestive and biliary tract". The graphs associated with those labels are dispersed throughout the image and are disconnected from other graphs with the same label. The documents clustered into these nodes are often misclassified as originating in an anatomical neighbor, or a range of other primary sites, suggesting that the digestive system tumor has spread from another primary site.

We then queried our mapper construction for each class in our dataset for those nodes in the graph that were classified at better than 97% accuracy. These nodes contain a subset of the vocabulary that are associated with high predictive accuracy for a given label. However, not all the words in this subset are indicative of any particular class. If we consider conjunctive words, they might tie together relevant words for a class but alone do not indicate any particular class label. Thus we want to infer the words with the highest probability of being associated with a specific class; to approximate this probability, we use the distance to measure detailed in the previous section. For primary cancer sites, breast (C50) and prostate (C61), a table of the most prominent, i.e., closest to the high accuracy subset, is provided in Table 1.

The high-probability words in Table 1 contain indicators of disease that have been identified in the literature. Indeed, 'pN' indicates prostate cancer with lymph node involvement [26], although the last identifier in the code, typically an integer or + character, has been stripped away in the data preprocessing, similarly for breast cancer we see the term 'ptnm' which is the pathological tumor, node, metastasis (TNM) staging classification. The specific code is omitted, due to differences between pathology reports, but we can see that this feature was identified as important across the entire collection of pathology reports. Additionally, we note the word 'htn', which is shorthand for 'hypertension' and has been noted as an increased risk factor for both breast [27] and prostate [28] cancers. Additionally, 'kb', in particular nuclear factor-kappa B (NF-$\kappa$B), is a biomarker for both cancers as well [29, 30]. There are also words contained in both columns of this table, such as 'chromatin', 'cellular', and 'demonstrates', which are important to the structure of the reports, but are not indications of a cancer diagnosis by themselves. However,

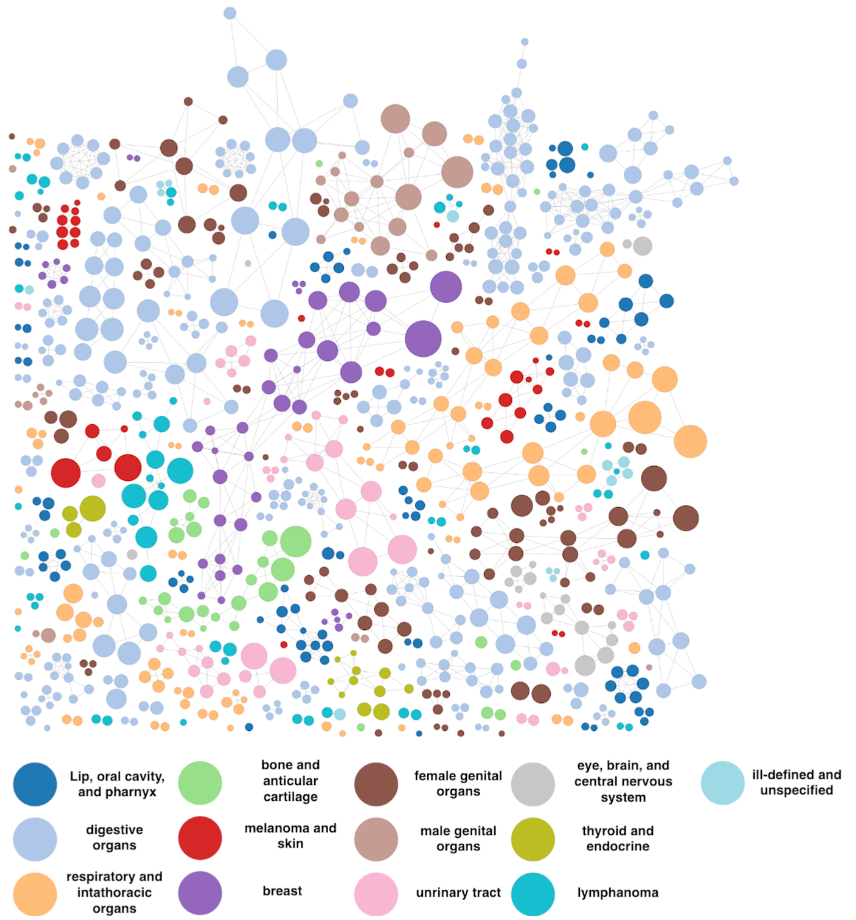


**Fig 3. Visual representation of our mapper graph for the site task.** Colors designate the ground truth label, and the nodes are proportional to the number of documents in the associated node. The purple subgraph at the center is associated with C50 (breast cancer), and the C61 (prostate cancer) cluster is the tan nodes directly above the breast cancer subgraph.

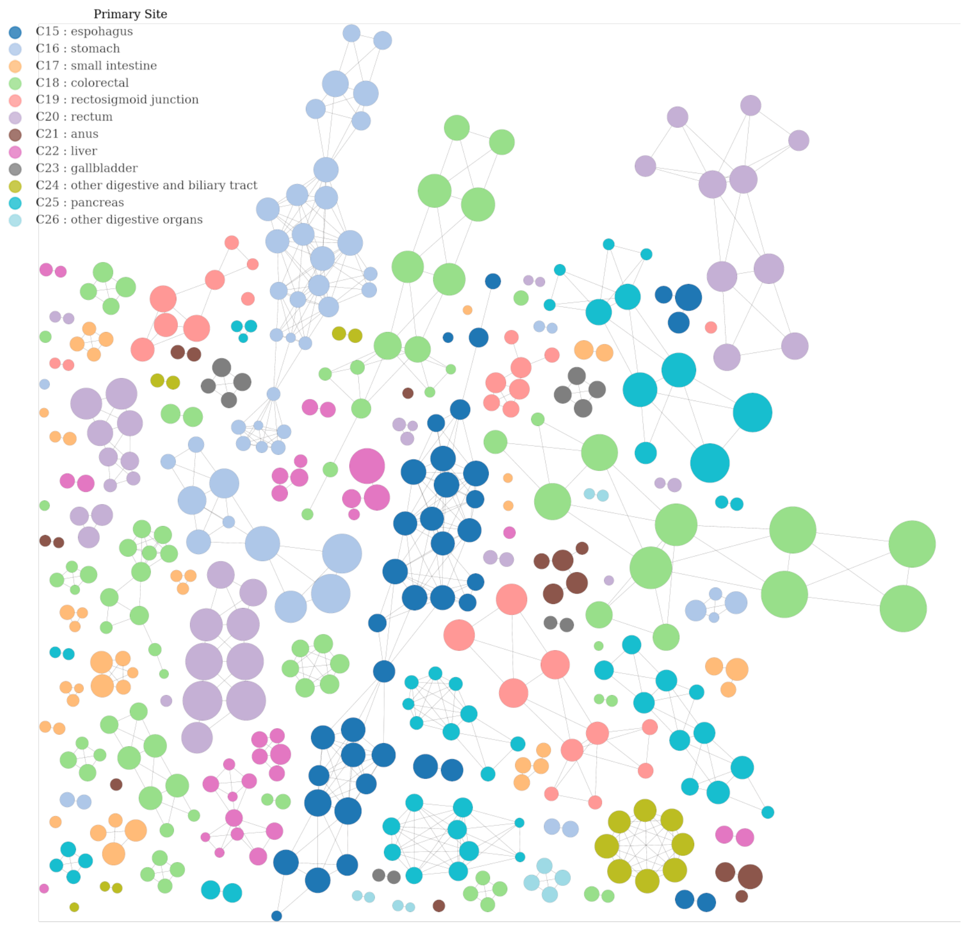


**Fig 4. Digestive system subgraph of Figure 3.** The ground truth labels are given by color, and the node size are proportional to the number of documents in the associated node.

**Table 1.** Label-specific keywords and associated density estimate inferred from our method computed over 2.3 million NCI cancer pathology reports for primary cancer site C50 (breast) and C61 (prostate) with $\hat{m} = 0.25$.

| C50 | | C61 | |
|---|---|---|---|
| Keyword | Density | Keyword | Density |
| absent | 0.545855 | cellular | 0.384322 |
| htn | 0.561171 | firm | 0.388699 |
| individual | 0.568243 | htn | 0.390837 |
| obvious | 0.580716 | marked | 0.392986 |
| marked | 0.586395 | examined | 0.393910 |
| barrett | 0.587200 | deep | 0.394185 |
| hyperplastic | 0.587436 | individual | 0.397800 |
| kb | 0.590752 | hemorrhagic | 0.400443 |
| hemorrhagic | 0.596458 | dense | 0.400927 |
| herpes | 0.598618 | versus | 0.403204 |
| gd | 0.599345 | iron | 0.404446 |
| near | 0.601415 | hyperplastic | 0.405272 |
| condensation | 0.604456 | numerous | 0.407425 |
| follicles | 0.606486 | attempt | 0.407799 |
| white | 0.606775 | degenerated | 0.412005 |
| respectively | 0.608007 | specify | 0.412277 |
| demonstrates | 0.609365 | present | 0.413304 |
| specify | 0.609693 | absent | 0.415616 |
| sparse | 0.609972 | obvious | 0.415673 |
| two | 0.610398 | sparse | 0.416259 |
| firm | 0.612019 | basophilia | 0.416335 |
| cellular | 0.612091 | evaluated | 0.416506 |
| normocellular | 0.612507 | blanchard | 0.417760 |
| name | 0.613017 | minor | 0.417820 |
| behavior | 0.613755 | laboratories | 0.418889 |
| touching | 0.616048 | juvenile | 0.419542 |
| indeterminate | 0.616372 | grams | 0.419809 |
| hyalinized | 0.617308 | pn | 0.420389 |
| yes | 0.617461 | placed | 0.420410 |
| body | 0.617730 | mild | 0.420849 |
| eliminating | 0.618531 | present | 0.420899 |
| intraductal | 0.619408 | respectively | 0.421208 |
| verruca | 0.619526 | demonstrates | 0.422027 |
| fascia | 0.620582 | reactive | 0.422078 |
| random | 0.620797 | found | 0.422473 |
| ray | 0.622538 | follicles | 0.422947 |
| amplified | 0.623354 | condensation | 0.423040 |
| attempt | 0.623893 | kb | 0.424496 |
| like | 0.624322 | random | 0.425921 |
| kimbrell | 0.624984 | hyalinized | 0.425946 |
| simulate | 0.625814 | benign | 0.425954 |
| sb | 0.625973 | fascia | 0.427258 |
| ulceration | 0.626190 | adenomatous | 0.427284 |
| dense | 0.626372 | none | 0.427427 |
| lymph | 0.626979 | roughly | 0.427465 |
| evaluated | 0.627098 | tinged | 0.427717 |
| examined | 0.627652 | somewhat | 0.427889 |
| noncohesive | 0.628100 | free | 0.428067 |
| reg | 0.628291 | kim | 0.428151 |
| ptnm | 0.628346 | fibrosis | 0.428679 |

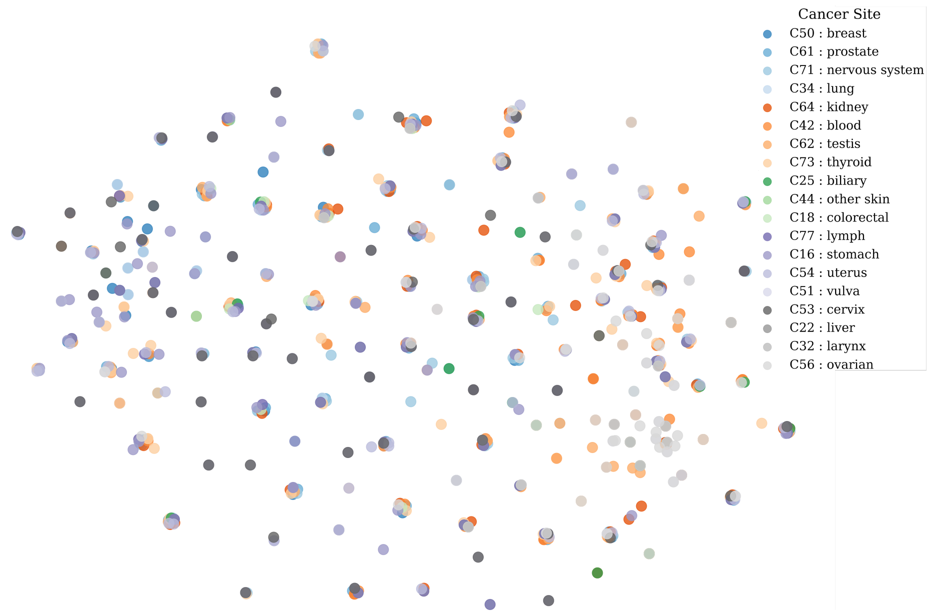


**Fig 5. Two dimensional representation of the learned word embeddings** These embeddings are from regions of high-predictive accuracy in our mapper graph. The clusters of words from multiple classes are common words indicating a cancer diagnosis, but are not specific as to a particular site; high-probability words are singletons typically found in regions with other words associated with the same class.

as a majority of electronic pathology reports contain these words, it follows that they are identified as being part of the set of words contained in documents with high-predictive accuracy. Clustering high-predictive words in two-dimensions across all primary sites in Figure 5 reveals clusters for words indicative of a cancer diagnosis, but not site specific such as: cellular, hemorrhagic, and hyperplastic. Singletons are more specific to a particular site; prostate and lung are primary examples.

## 20 Newsgroups Dataset

Due to the private nature of the information in the electronic health records, we repeated our analysis on the publicly available 20 Newsgroup dataset [12]. Specifically, we created and trained a CNN with an architecture similar to [31] on a subset of the 20 Newsgroups data. We used the ‘alt.atheism’, ‘comp.graphics’, ‘sci.space’, and ‘talk.religion.misc’ classes for our analysis. To preprocess step, we stripped out all of the headers and footers from the messages prior to tokenizing the data, and removed non-alphanumeric characters and words shorter than three characters. The remaining data was split into train, test, and validation sets. The training set was comprised of 3,048 instances, of which 20% were set aside for a validation set, and the test set contained 339 samples. We trained our CNN to achieve nearly 99% accuracy on the training set using a cross-entropy loss function and the Adam optimizer [32].

The results of our Mapper construction on this dataset is shown in Figure 6. For all classes, we observe large connected graphs where the large nodes contain clusters of documents demonstrating high predictive accuracy from the CNN model. There are subgraphs from documents being from the ‘alt.atheism’ or ‘talk.religion.misc’ classes that are disconnected from their associated primary graph. These two classes have overlapping vocabularies, which potentially leads to misclassifications between these two classes, as opposed to the words from the ‘sci.space’ class, which has few overlapping words with either of ‘alt.atheism’ or ‘talk.religion.misc’. One interesting overlap between these two classes and ‘sci.space’ are words

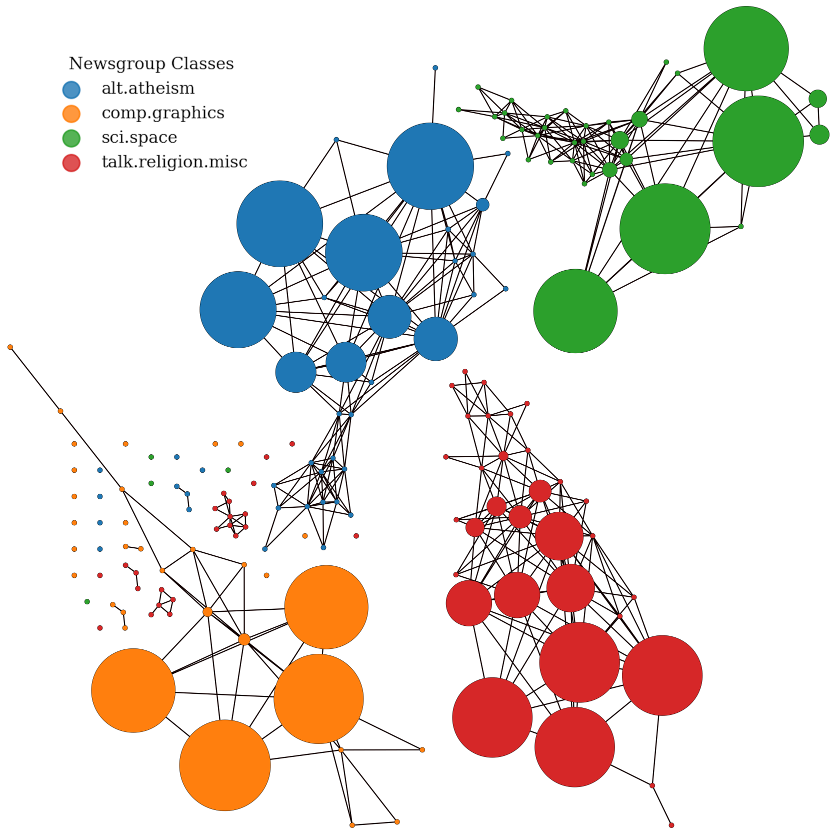


**Fig 6. Visual representation of our mapper graph over the 20 Newsgroups subset.** Colors designate the ground truth label, and the nodes are proportional to the number of documents in the associated node. Note that not all subgraphs associated with a specific class label are connected. These smaller, disconnected regions contain documents that share similar text characteristics and are likely to be misclassified by our model.

associated with planetary names from our solar system. For example, 'Saturn' may refer to either the planet ('sci.space') or the god in Roman mythology, the context was inferred by the different parallel convolutional layers in the CNN model.

Computing the dtm, setting $m = 0.1$, function for the different classes in the 20Newsgroups dataset, we see the results from a t-SNE plot of the words associated with high-predictive accuracy in Figure 7. An interactive version of this plot is available in the git repository[1]. We observe that the words associated with each class are well-separated, except for the alt.atheism' or 'talk.religion.misc' classes. This is expected behavior, as these classes have overlapping vocabulary, and discuss similar topics. Indeed, in Table 2 we see the words 'adultery' and 'sinners' are strongly associated with both 'alt.atheism' and 'talk.religion.misc'.

In Tables 3 to 4, we study how different values of $\hat{m}$ in Equation (2) effect the results of our method, recalling that $\hat{m}$ is a smoothing parameter which controls the amount of mass contained in balls centered at each point. An optimal value for $\hat{m}$ may be chosen in the case of a known target, but a distribution over words in the dataset is unknown. Furthermore, the behaviour of Equation (2) with respect to $\hat{m}$ is not monotonic [23], which precludes typical approaches to model selection. The smaller values of $\hat{m}$ may introduce more variance to the results due to the global nature of the stability results between the empirical and theoretical quantile functions of the distribution in question.

[1]https://code.ornl.gov/3t6/DL-interpretability

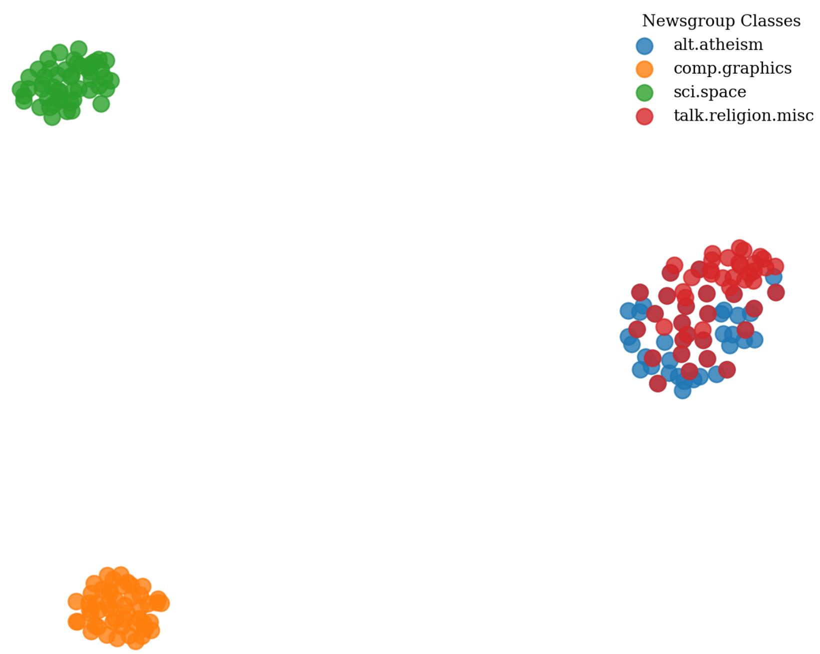


**Fig 7. Two dimensional representation of the learned word embeddings for the newsgroups dataset.** These words are inferred from regions of high-predictive accuracy in our mapper graph. Observe that the classes are separated, except for the words associated with either alt.athesim or talk.religion.misc in the lower right-hand corner; these latter two classes have significant overlap between their high-probability words.

**Table 2.** Label-specific keywords and associated distance, $\hat{m} = 0.1$ approx. 600 nearest neighbors, inferred from our method computed over 2,438 documents in the 20Newsgroups dataset from the four classes: alt.atheism, comp.graphics, sci.space, and talk.religion.misc. Words in red ink indicate changes from the previous table.

| alt.atheism | | comp.graphics | | sci.space | | talk.religion.misc | |
|---|---|---|---|---|---|---|---|
| Keyword | Distance | Keyword | Distance | Keyword | Distance | Keyword | Distance |
| unwilling | 0.13562 | autodesk | 0.09517 | moons | 0.04566 | sinners | 0.12811 |
| hatred | 0.13575 | halftone | 0.09757 | jupiters | 0.04756 | sins | 0.12915 |
| victims | 0.13611 | sunview | 0.09804 | jovian | 0.04831 | adultery | 0.12942 |
| practised | 0.13616 | printers | 0.10066 | rings | 0.05272 | preached | 0.12989 |
| oppression | 0.13650 | pcs | 0.10120 | landers | 0.05305 | infant | 0.13036 |
| adultery | 0.13655 | xview | 0.10139 | neptune | 0.05345 | meat | 0.13058 |
| attacks | 0.13711 | fbm | 0.10313 | flybys | 0.05439 | mormonism | 0.13200 |
| revenge | 0.13790 | nodes | 0.10337 | neptunes | 0.05450 | wicked | 0.13335 |
| imprisonment | 0.13826 | metafile | 0.10436 | nucleus | 0.05564 | victory | 0.13419 |
| sinners | 0.14130 | kernel | 0.10486 | crust | 0.05759 | selfrighteous | 0.13444 |
| protestant | 0.14164 | font | 0.10530 | volcanic | 0.05804 | souls | 0.13606 |
| religiously | 0.14377 | modeler | 0.10585 | volcanoes | 0.05844 | torture | 0.13627 |
| inappropriate | 0.14512 | pixmap | 0.10629 | ecliptic | 0.05964 | grave | 0.13657 |
| meat | 0.14516 | polyray | 0.10652 | twin | 0.05967 | hatred | 0.13690 |
| insulting | 0.14521 | atari | 0.10697 | icy | 0.06003 | denied | 0.13807 |
| conflict | 0.14646 | gis | 0.10739 | lander | 0.06072 | insulting | 0.13855 |
| selfrighteous | 0.14658 | macintosh | 0.10782 | callisto | 0.06078 | servant | 0.13923 |
| irrelevant | 0.14711 | fortran | 0.10792 | vega | 0.06117 | goddess | 0.13959 |
| behaviors | 0.14726 | standalone | 0.10796 | methane | 0.06129 | attacks | 0.13971 |
| shares | 0.14915 | graphical | 0.10818 | polar | 0.06143 | midst | 0.13979 |
| darksideosrheuoknoredu | 0.14941 | rtrace | 0.10839 | orbiters | 0.06160 | husband | 0.13995 |
| racial | 0.14989 | userdefined | 0.10840 | ammonia | 0.06175 | prophets | 0.14042 |
| brutal | 0.15080 | vogle | 0.10843 | saturns | 0.06296 | oppression | 0.14046 |
| unbelievers | 0.15105 | publicdomain | 0.10863 | pioneer | 0.06304 | inappropriate | 0.14058 |
| motives | 0.15121 | optional | 0.10947 | uranus | 0.06326 | abraham | 0.14070 |
| buddhists | 0.15190 | siggraphorg | 0.10949 | plutos | 0.06367 | demons | 0.14116 |
| simondcswarwickacuk | 0.15236 | histograms | 0.10990 | balloons | 0.06399 | atone | 0.14140 |
| dictator | 0.15246 | pubrtrace | 0.10995 | poles | 0.06407 | buried | 0.14158 |
| condemn | 0.15248 | terminals | 0.11001 | cassini | 0.06455 | lords | 0.14181 |
| tactics | 0.15292 | portable | 0.11151 | crater | 0.06507 | conservative | 0.14199 |
| believers | 0.15314 | polygonal | 0.11167 | ocean | 0.06623 | valiant | 0.14275 |
| criminal | 0.15320 | modeller | 0.11220 | venerean | 0.06696 | brutal | 0.14300 |
| convicted | 0.15333 | fft | 0.11222 | galilean | 0.06720 | praise | 0.14311 |
| saddam | 0.15345 | flc | 0.11242 | meteor | 0.06729 | enemies | 0.14342 |
| questioned | 0.15529 | ascii | 0.11247 | explored | 0.06817 | moses | 0.14349 |
| sins | 0.15536 | mirrorsmsdosgraphics | 0.11278 | craf | 0.06822 | forgiven | 0.14365 |
| idolatry | 0.15569 | converter | 0.11285 | ulysses | 0.06885 | sinner | 0.14391 |
| infant | 0.15572 | licensing | 0.11288 | triton | 0.06950 | motive | 0.14430 |
| allpowerful | 0.15573 | scn | 0.11292 | voyager | 0.06954 | zarathushtra | 0.14450 |
| mistaken | 0.15573 | disimp | 0.11313 | flyby | 0.06962 | repentance | 0.14467 |
| supernatural | 0.15626 | ultrix | 0.11319 | jun | 0.06997 | unwilling | 0.14486 |
| ousrvroulufi | 0.15647 | emulation | 0.11319 | huygens | 0.07011 | trials | 0.14531 |
| defend | 0.15652 | interactively | 0.11384 | helium | 0.07037 | misrepresentation | 0.14535 |
| implications | 0.15659 | renderers | 0.11424 | halley | 0.07075 | fathers | 0.14559 |
| childish | 0.15701 | armymil | 0.11427 | decay | 0.07085 | bethlehem | 0.14560 |
| martyrs | 0.15712 | workstation | 0.11433 | cometary | 0.07132 | possession | 0.14564 |
| kempmpphoenixoulufi | 0.15742 | plotters | 0.11512 | asteroid | 0.07219 | faithful | 0.14603 |
| consequences | 0.15747 | wsmrsimtel | 0.11533 | venusian | 0.07300 | thirst | 0.14615 |
| timmbakemclucsbedu | 0.15748 | xmgr | 0.11540 | ganymede | 0.07382 | martyrs | 0.14633 |
| beef | 0.15772 | pdb | 0.11556 | viking | 0.07398 | forsake | 0.14762 |

**Table 3.** Label-specific keywords and associated distance, with $\hat{m} = 0.05$ approx. 300 nearest neighbors, inferred from our method computed over 2,438 documents in the 20Newsgroups dataset from the four classes: alt.atheism, comp.graphics, sci.space, and talk.religion.misc.

| alt.atheism | | comp.graphics | | sci.space | | talk.religion.misc | |
|---|---|---|---|---|---|---|---|
| Keyword | Distance | Keyword | Distance | Keyword | Distance | Keyword | Distance |
| mcl | 0.07152 | autodesk | 0.06834 | moons | 0.02539 | midst | 0.08606 |
| healtasaturnwwcedu | 0.07256 | pcs | 0.06951 | jupiters | 0.03036 | wicked | 0.08652 |
| timmbake | 0.07324 | userdefined | 0.07074 | jovian | 0.03071 | servant | 0.08712 |
| newnham | 0.07691 | sunview | 0.07246 | landers | 0.03218 | preached | 0.09280 |
| timmbakemclucsbedu | 0.07727 | pixmap | 0.07274 | rings | 0.03221 | demons | 0.09455 |
| nanci | 0.07742 | histograms | 0.07311 | flybys | 0.03290 | souls | 0.09469 |
| kempmpphoenixoulufi | 0.07772 | printers | 0.07335 | neptunes | 0.03427 | lords | 0.09650 |
| darksideosrheuoknoredu | 0.07798 | font | 0.07337 | neptune | 0.03450 | fulfil | 0.09675 |
| pihko | 0.07809 | ultrix | 0.07370 | nucleus | 0.03484 | praise | 0.09679 |
| ousrvroulufi | 0.07832 | macintosh | 0.07387 | lander | 0.03579 | cupportalcom | 0.09704 |
| saturnwwcedu | 0.07904 | xview | 0.07402 | crust | 0.03614 | abraham | 0.09762 |
| cvillewamumdedu | 0.07909 | halftone | 0.07402 | pioneer | 0.03673 | infernal | 0.09768 |
| lnewnhambradfordacuk | 0.07931 | polyray | 0.07420 | cassini | 0.03703 | buried | 0.09775 |
| bradfordacuk | 0.08026 | nodes | 0.07633 | ammonia | 0.03723 | usenetinscwruedu | 0.09839 |
| westnext | 0.08055 | modeler | 0.07686 | callisto | 0.03754 | lanphier | 0.09869 |
| mimsyumdedu | 0.08081 | atari | 0.07688 | volcanoes | 0.03763 | behold | 0.09884 |
| agrinoenkidumiccl | 0.08139 | fft | 0.07696 | saturns | 0.03841 | denied | 0.10022 |
| andres | 0.08171 | plots | 0.07697 | vega | 0.03851 | spoken | 0.10043 |
| bilokcforumosrheedu | 0.08244 | polygonal | 0.07735 | voyager | 0.03856 | alicea | 0.10090 |
| deankaflowitz | 0.08269 | aix | 0.07778 | uranus | 0.03860 | lanph | 0.10112 |
| grino | 0.08384 | platforms | 0.07782 | flyby | 0.03879 | prophets | 0.10116 |
| stilgar | 0.08448 | siggraphorg | 0.07795 | twin | 0.03908 | adultery | 0.10117 |
| mangoecsumdedu | 0.08555 | armymil | 0.07871 | orbiters | 0.03999 | mormonism | 0.10122 |
| qie | 0.08764 | pubrtrace | 0.07882 | methane | 0.04020 | bethlehem | 0.10158 |
| warren | 0.08817 | rtrace | 0.07891 | crater | 0.04037 | sins | 0.10165 |
| simondcswarwickacuk | 0.08825 | metafile | 0.07896 | volcanic | 0.04054 | victory | 0.10172 |
| decaycbnewsjcbattcom | 0.08871 | publicdomain | 0.07908 | polar | 0.04107 | isaiahs | 0.10210 |
| umar | 0.08874 | fortran | 0.07926 | icy | 0.04128 | zion | 0.10228 |
| ann | 0.08889 | kernel | 0.07944 | viking | 0.04203 | goddess | 0.10267 |
| madhausnetcomcom | 0.08924 | wsmrsimtel | 0.08052 | ocean | 0.04211 | milomathscarolinaedu | 0.10278 |
| conner | 0.08928 | scn | 0.08065 | ecliptic | 0.04222 | meat | 0.10282 |
| jsn | 0.08943 | phong | 0.08066 | craf | 0.04283 | sinners | 0.10284 |
| schaertel | 0.08962 | mirrorsmsdosgraphics | 0.08103 | poles | 0.04301 | moses | 0.10293 |
| hens | 0.08974 | converter | 0.08124 | mariner | 0.04356 | forsake | 0.10296 |
| wamumdedu | 0.09120 | fbm | 0.08143 | venerean | 0.04429 | righteous | 0.10344 |
| buddha | 0.09153 | graphical | 0.08161 | meteor | 0.04441 | forgiven | 0.10350 |
| cbnewsjcbattcom | 0.09181 | ascii | 0.08174 | ganymede | 0.04448 | speaks | 0.10364 |
| newscsouiucedu | 0.09183 | interactively | 0.08176 | plutos | 0.04460 | thirst | 0.10367 |
| dcswarwickacuk | 0.09201 | flc | 0.08179 | galilean | 0.04481 | apostles | 0.10393 |
| inn | 0.09245 | portable | 0.08180 | asteroid | 0.04520 | wine | 0.10410 |
| khanitditdnrlnavymil | 0.09287 | gis | 0.08184 | balloons | 0.04535 | qie | 0.10412 |
| dpsnasakodakcom | 0.09295 | msdos | 0.08193 | sakigake | 0.04584 | satans | 0.10427 |
| fkthorusapmchpsnide | 0.09311 | ppm | 0.08225 | triton | 0.04585 | infant | 0.10437 |
| dfullerportalhqvideocartcom | 0.09333 | standalone | 0.08227 | helium | 0.04593 | zarathushtra | 0.10468 |
| brandt | 0.09560 | vivid | 0.08234 | venusian | 0.04647 | oppression | 0.10512 |
| vonroeschlaub | 0.09609 | amigaphysikunizhch | 0.08243 | ulysses | 0.04662 | attacks | 0.10529 |
| wmantiscouk | 0.09646 | gradient | 0.08258 | cometary | 0.04671 | husband | 0.10534 |
| petri | 0.09678 | vogle | 0.08271 | huygens | 0.04682 | faithful | 0.10563 |
| halatpoohbears | 0.09705 | drivers | 0.08281 | jun | 0.04748 | nyikos | 0.10570 |
| cfaehlvestaunmedu | 0.09732 | interpolation | 0.08285 | explored | 0.04857 | hatred | 0.10572 |

**Table 4.** Label-specific keywords and associated distance, $\hat{m} = 0.25$ approx. 1600 nearest neighbors, inferred from our method computed over 2,438 documents in the 20Newsgroups dataset from the four classes: alt.atheism, comp.graphics, sci.space, and talk.religion.misc. Words in red ink indicate changes from the previous table.

| alt.atheism | | comp.graphics | | sci.space | | talk.religion.misc | |
|---|---|---|---|---|---|---|---|
| Keyword | Distance | Keyword | Distance | Keyword | Distance | Keyword | Distance |
| unwilling | 0.20452 | autodesk | 0.16208 | orbiters | 0.12453 | selfrighteous | 0.19492 |
| practised | 0.20817 | printers | 0.16567 | landers | 0.12871 | meat | 0.19690 |
| adultery | 0.20937 | halftone | 0.16764 | jupiters | 0.13149 | adultery | 0.19924 |
| hatred | 0.21039 | nodes | 0.16776 | nucleus | 0.13460 | sinners | 0.20140 |
| imprisonment | 0.21110 | fbm | 0.16980 | ammonia | 0.13558 | inappropriate | 0.20219 |
| victims | 0.21301 | sunview | 0.17079 | volcanic | 0.13576 | infant | 0.20333 |
| attacks | 0.21340 | optional | 0.17108 | flybys | 0.13587 | hatred | 0.20735 |
| inappropriate | 0.21410 | standalone | 0.17226 | explored | 0.13629 | sins | 0.20891 |
| selfrighteous | 0.21551 | terminals | 0.17545 | polar | 0.13712 | insulting | 0.20985 |
| protestant | 0.21644 | metafile | 0.17629 | twin | 0.13753 | attacks | 0.21010 |
| revenge | 0.21885 | kernel | 0.17654 | ecliptic | 0.13787 | unwilling | 0.21275 |
| buddhists | 0.21900 | gis | 0.17674 | ulysses | 0.13944 | victims | 0.21311 |
| insulting | 0.22175 | vogle | 0.17865 | volcanoes | 0.14093 | grave | 0.21357 |
| oppression | 0.22176 | plotters | 0.17973 | wfpc | 0.14125 | preached | 0.21407 |
| meat | 0.22256 | xview | 0.18014 | crust | 0.14140 | buddhists | 0.21562 |
| unbelievers | 0.22297 | gphigs | 0.18046 | cassini | 0.14179 | brutal | 0.21583 |
| behaviors | 0.22304 | font | 0.18126 | jun | 0.14274 | atone | 0.21612 |
| brutal | 0.22322 | polyray | 0.18173 | icy | 0.14364 | torture | 0.21614 |
| shares | 0.22412 | graphical | 0.18265 | balloons | 0.14379 | possession | 0.21652 |
| racial | 0.22425 | colored | 0.18267 | moons | 0.14402 | motive | 0.21664 |
| mistaken | 0.22467 | supercomputers | 0.18285 | jovian | 0.14414 | victory | 0.21676 |
| religiously | 0.22470 | licensing | 0.18337 | transmitter | 0.14429 | valiant | 0.21681 |
| conflict | 0.22539 | atari | 0.18375 | decay | 0.14534 | witness | 0.21722 |
| sinners | 0.22610 | modeler | 0.18419 | swing | 0.14537 | guided | 0.21843 |
| irrelevant | 0.22650 | ibmpc | 0.18426 | meteor | 0.14553 | oppression | 0.21863 |
| dictator | 0.22667 | disimp | 0.18513 | helium | 0.14574 | mormonism | 0.21961 |
| motives | 0.22771 | publicdomain | 0.18562 | cometary | 0.14587 | unbelievers | 0.22050 |
| believers | 0.22818 | pub | 0.18594 | antennas | 0.14657 | goddess | 0.22095 |
| questioned | 0.22847 | nurbs | 0.18615 | neptunes | 0.14728 | persecuted | 0.22105 |
| condemn | 0.22949 | pubrtrace | 0.18665 | giotto | 0.14758 | sinner | 0.22136 |
| beef | 0.23086 | emulation | 0.18711 | mariners | 0.14766 | protestant | 0.22150 |
| implications | 0.23167 | macintosh | 0.18714 | methane | 0.14773 | lawyers | 0.22192 |
| defend | 0.23325 | ascii | 0.18716 | huygens | 0.14883 | imprisonment | 0.22237 |
| rejecting | 0.23370 | pcs | 0.18726 | rings | 0.14928 | trials | 0.22296 |
| tactics | 0.23375 | animator | 0.18727 | callisto | 0.14958 | conservative | 0.22323 |
| contradiction | 0.23398 | siggraphorg | 0.18757 | crater | 0.15090 | martyrs | 0.22400 |
| applicable | 0.23439 | renderers | 0.18799 | basalts | 0.15148 | murdered | 0.22417 |
| allpowerful | 0.23448 | fortran | 0.18820 | craf | 0.15207 | pleasure | 0.22448 |
| disregard | 0.23500 | modeller | 0.18829 | vega | 0.15223 | repentance | 0.22466 |
| attacked | 0.23507 | toolkits | 0.18845 | highspeed | 0.15282 | misrepresentation | 0.22487 |
| consequences | 0.23557 | psl | 0.18853 | lander | 0.15522 | applicable | 0.22497 |
| notions | 0.23592 | windowing | 0.18874 | neptune | 0.15572 | denied | 0.22547 |
| witness | 0.23649 | pixmap | 0.18879 | ranger | 0.15683 | fathers | 0.22596 |
| wishes | 0.23699 | volumetric | 0.18900 | gravitational | 0.15704 | condemn | 0.22681 |
| convicted | 0.23716 | mirrorsmsdosgraphics | 0.18909 | stays | 0.15753 | conflict | 0.22684 |
| naturally | 0.23742 | rtrace | 0.18910 | waterice | 0.15770 | hiding | 0.22685 |
| supernatural | 0.23759 | animations | 0.18913 | venerean | 0.15803 | husband | 0.22702 |
| lawyers | 0.23769 | swivel | 0.18915 | ocean | 0.15809 | enemies | 0.22722 |
| awareness | 0.23835 | miscunixralcgmralcgmtarz | 0.18921 | saturns | 0.15865 | rejecting | 0.22774 |
| spelled | 0.23878 | hanaumastanfordedu | 0.18933 | venusian | 0.15877 | believers | 0.22783 |

Comparing the results for 'alt.atheism' change as the parameter $\hat{m}$ increases from 0.05 to 0.1, we notice a substantial change in the inferred features between the two runs. Most noticeably, in Table 3 the words seemingly have no apparent connection to atheism. If we examine the documents individually, we see that the email addresses are associated with individuals who frequently poster to the message board. Furthermore, some of the messages contain multiple responses, hence the addresses may appear multiple times within a single document and in multiple instances in the training set. In some sense, these words are most strongly correlated with the topic, since these individuals are frequent posters to the message board. However, without looking back at the original data, these results are difficult to interpret. As the value of $\hat{m}$ is increased to 0.1, the most important words inferred through our method for the 'alt.atheism' class are words that require no further interpretation and are readily identified with a discussion of religion and atheism.

If we now consider the classes 'comp.graphics' and 'sci.space', observe that the results are similar between the Tables 2 to 4. In Figure 7 the words most strongly associated with these classes comprised well-separated homogeneous clusters associated with a single class. Hence, increasing the parameter $\hat{m}$, which is the same as increasing number of nearest neighbors in the computation,

As we set $\hat{m} = 0.25$ in Table 4, we observe that the inferred features in each of the columns are relatively similar as compared with the results in Table 2. Further increasing the value of $\hat{m}$ yields similar results with an increased computational cost, and are not presented here.

## Stability

In Table 5 we present the results from computing the label-specific Lipschitz constants for both datasets from our method and LIME. In both cases considered here, we observe that our interpretability results are more stable than those provided by LIME.

**Table 5.** Label-specific Lipschitz constants for our proposed method and LIME interpretability method computed over 500 NCI cancer pathology reports for the site task and 20 Newsgroup dataset taken from the test set.

| | NCI Pathology Reports | | 20newsgropups data | |
|---|---|---|---|---|
| | Our Method | LIME | Our Method | LIME |
| Lipschitz Stability | 0.0013 | 0.570 | 0.202 | 0.561 |

We compare the stability of our method as compared against LIME by computing the label-specific Lipschitz constant which is defined as

$$L(X_i) = \underset{X \in \mathcal{N}(X_i;\epsilon)}{\operatorname{argmax}} \frac{\|g(X_i) - g(X)\|}{\|X_i - X\|}, \tag{4}$$

where $\mathcal{N}(X_i;\epsilon)$ is the neighborhood of radius $\epsilon$ centered at $X_i$ that is homogeneous with respect to the class label of $X_i$, and $g(\cdot)$ is the output from an interpretability method. As the Lipschitz value $L(X_i)$ measures the rate of change about a point $X_i$, we only consider those points in the neighborhood $\mathcal{N}(X_i;\epsilon)$ with the same label as $X_i$, since we would expect to see a greater rate of change between different inputs from different classes.

In this experiment, we selected 500 documents at random from test set of the NCI pathology report dataset and 20Newsgroups dataset, and fed them and input into our and LIME explainability methods. We took the output from each method, computed Equation (4), and reported the maximum value. As in the previous results of Tables 2 to 4, we observe that the explanations are stable with respect to the inputs.

## Discussion

Our interpretability method demonstrates that our MTCNN cancer phenotype prediction model has identified decision rules that contain clinically relevant information. We have inferred a multiset of features associated with high-predictive accuracy for each class in the dataset, and identified those words closest to these multisets from each word in the vocabulary. This technique may also be used to identify those words that are indicative of regions of poor-predictive accuracy, which yields two pieces of information. Primarily, these features may indicate errors in the ground truth class labels, which neither benchmark nor our datasets are immune from [33]. This method allows us to identify both systematic misclassifications by our model and common traits between misclassifications with the same erroneous labels.

As our interpretability method yields important features for each class, as an extension of this work, we may investigate the effects of training a model on a reduced vocabulary. Specifically, the vocabulary for the MTCNN model presented herein contains approximately $178,000$ words, each of which is embedded in a 300-dimensional space that is learned at training time. One could infer the words associated with high-predictive accuracy for each class, and use the collection of these sets as a reduced vocabulary for training subsequent models. One would expect to see similar accuracy metrics, e.g., precision and recall, as a model trained on the full vocabulary but enjoying a reduced time-to-solution. Additionally, the density estimates computed from our method could be used to estimate the mutual information between the vocabulary and labels to reduce the size of the vocabulary.

The translational aspect of our method could be enhanced through the inclusion of longer word combinations in the inferred features from the MTCNN model. This would aid in differentiating between important phrases with different interpretations in different contexts. Moreover, investigating longer windows of text would aid in interpreting the results from our method, and would align with the derived features from the three parallel convolutional filters in the model architecture, recall these windows look at patterns $3, 4$, and $5$ words long in each of the convolutional layers. Other tunings could be easily implemented, for example, one could look at those words with the highest self-attention score in an attention based architecture [13].

One potential limitation of our method is the ability to look at features, words in the cases presented here, within their original context. Many words exhibit polysemy, whereby a word can have multiple meanings depending on the original context. For example, the word 'bass' might refer to: a fish, a musical instrument, or low-frequency output. Only through context can we disentangle the intended meaning. Continuing on with this work, we will extend our method to consider not only single words, but different length combinations which will give context to the inferred words and engender increased trust in the underlying model predictions.

## Conclusions

We have presented a novel interpretability method that is based on the topological and geometric features in a DL model's prediction space. Our method does not assume any form of the learned decision rules and is faithful to the training data. We have demonstrated Lipschitz stability of our method as measured across inputs with the same ground truth label. Lastly, our method infers a low-dimensional structure from a complex and high-dimensional dataset, shedding light on the black-box nature of DL algorithms. We may further develop the idea of low-dimensional representations by considering manifold approximation ideas from TDA coupled with the ideas presented here.

# Acknowledgments

The authors gratefully acknowledge Xiao-Cheng Wu of the Louisiana Tumor Registry for data curation and providing the data.

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725.

The authors would like to acknowledge the contribution to this study from other staff in the participating central cancer registries. These registries are supported by the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program, the Centers for Disease Control and Prevention's National Program of Cancer Registries (NPCR), and/or state agencies, universities, and cancer centers. The participating central cancer registries include the following:

- Louisiana working under contract numbers SEER: HHSN261201800007I/HHSN26100002 and NPCR: NU58DP0063.

# Declarations

## Funding

This work was completed under contract numbers: DE-AC05-00OR22725 and AC02200200100000.

## Availability of data and materials

The code used to create the figures with the 20 Newsgroups data and interactive versions of the figures are in a git repository `https://code.ornl.gov/3t6/DL-interpretability`. Please contact Adam Spannaus, `spannausat@orn.gov` with any questions regarding the code.

## Competing interests

The authors declare that they have no competing interests.

## Authors' contributions

AS designed the research and the experimental study, performed research, analyzed the data, and wrote the manuscript. LP and GT initiated the MOSSAIC project. All authors reviewed the manuscript.

# References

1. Stevens R, Taylor V, Nichols J, Maccabe AB, Yelick K, Brown D. AI for Science: Report on the Department of Energy (DOE) Town Halls on Artificial Intelligence (AI) for

Science. Argonne National Lab.; 2020. Available from: `https://www.osti.gov/biblio/1604756`.

2. Begoli E, Bhattacharya T, Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. Nature Machine Intelligence. 2019;1(1):20–23.
3. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: International conference on machine learning. PMLR; 2017. p. 1321–1330.
4. Ribeiro MT, Singh S, Guestrin C. " 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 mining; 2016. p. 1135–1144.
5. Laugel T, Lesot MJ, Marsala C, Renard X, Detyniecki M. The dangers of post-hoc interpretability: Unjustified counterfactual explanations. arXiv preprint arXiv:190709294. 2019;.
6. Singh G, Mémoli F, Carlsson GE. Topological methods for the analysis of high dimensional data sets and 3d object recognition. SPBG. 2007;91:100.
7. Ribeiro MT, Singh S, Guestrin C. " 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 mining; 2016. p. 1135–1144.
8. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017;30.
9. Alvarez-Melis D, Jaakkola TS. On the robustness of interpretability methods. arXiv preprint arXiv:180608049. 2018;.
10. Rudin C. Please stop explaining black box models for high stakes decisions. Stat. 2018;1050:26.
11. Alawad M, Gao S, Qiu JX, Yoon HJ, Blair Christian J, Penberthy L, et al. Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks. Journal of the American Medical Informatics Association. 2020;27(1):89–98.
12. Lang K. 20 Newsgroups;.
13. Gao S, Qiu JX, Alawad M, Hinkle JD, Schaefferkoetter N, Yoon HJ, et al. Classifying cancer pathology reports with hierarchical self-attention networks. Artificial Intelligence in Medicine. 2019;101:101726. doi:https://doi.org/10.1016/j.artmed.2019.101726.
14. Nielson JL, Paquette J, Liu AW, Guandique CF, Tovar CA, Inoue T, et al. Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury. Nature communications. 2015;6(1):1–12.
15. Lum PY, Singh G, Lehman A, Ishkanov T, Vejdemo-Johansson M, Alagappan M, et al. Extracting insights from the shape of complex data using topology. Scientific reports. 2013;3(1):1–8.
16. Nicolau M, Levine AJ, Carlsson G. 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. 2011;108(17):7265–7270.
17. Carlsson LS, Vejdemo-Johansson M, Carlsson G, Jönsson PG. Fibers of failure: Classifying errors in predictive processes. Algorithms. 2020;13(6):150.

18. Rathore A, Chalapathi N, Palande S, Wang B. TopoAct: Visually Exploring the Shape of Activations in Deep Learning. In: Computer Graphics Forum. vol. 40. Wiley Online Library; 2021. p. 382–397.

19. Carlsson G. Topology and data. Bulletin of the American Mathematical Society. 2009;46(2):255–308.

20. Carriere M, Michel B, Oudot S. Statistical analysis and parameter selection for mapper. The Journal of Machine Learning Research. 2018;19(1):478–516.

21. Hatcher A. Algebraic Topology. Cambridge, UK: Cambridge University Press; 2002.

22. Chazal F, Cohen-Steiner D, Mérigot Q. Geometric inference for probability measures. Foundations of Computational Mathematics. 2011;11(6):733–751.

23. Chazal F, Massart P, Michel B. Rates of convergence for robust geometric inference. Electronic journal of statistics. 2016;10(2):2243–2286.

24. Biau G, Chazal F, Cohen-Steiner D, Devroye L, Rodríguez C. A weighted k-nearest neighbor density estimate for geometric inference. Electronic Journal of Statistics. 2011;5(none):204 – 237. doi:10.1214/11-EJS606.

25. Tenenbaum JB, Silva Vd, Langford JC. A global geometric framework for nonlinear dimensionality reduction. science. 2000;290(5500):2319–2323.

26. Amin MB, Greene FL, Edge SB, Compton CC, Gershenwald JE, Brookland RK, et al. The eighth edition AJCC cancer staging manual: continuing to build a bridge from a population-based to a more “personalized” approach to cancer staging. CA: a cancer journal for clinicians. 2017;67(2):93–99.

27. Han H, Guo W, Shi W, Yu Y, Zhang Y, Ye X, et al. Hypertension and breast cancer risk: a systematic review and meta-analysis. Scientific reports. 2017;7(1):1–9.

28. Navin S, Ioffe V. The association between hypertension and prostate cancer. Reviews in urology. 2017;19(2):113.

29. Biswas DK, Shi Q, Baily S, Strickland I, Ghosh S, Pardee AB, et al. NF-$\kappa$B activation in human breast cancer specimens and its role in cell proliferation and apoptosis. Proceedings of the National Academy of Sciences. 2004;101(27):10137–10142.

30. Jin RJ, Lho Y, Connelly L, Wang Y, Yu X, Saint Jean L, et al. The nuclear factor-$\kappa$B pathway controls the progression of prostate cancer to androgen-independent growth. Cancer research. 2008;68(16):6762–6769.

31. Kim Y. Convolutional Neural Networks for Sentence Classification. CoRR. 2014;abs/1408.5882.

32. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014;.

33. Northcutt CG, Athalye A, Mueller J. Pervasive label errors in test sets destabilize machine learning benchmarks. arXiv preprint arXiv:210314749. 2021;.