Interpreting Deep Patient Stratification Models with Topological Data Analysis

Anna Jurek-Loughrey, Richard Gault, Baharak Ahmaderaghi, Muhammad Fahim, Lu Bai
School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, UK,
{a.jurek, richard.gault, b.ahmaderaghi, m.fahim, l.bai}@qub.ac.uk

***Abstract*— Patient stratification is a crucial task aimed at categorizing individuals with a specific disease into more homogeneous subgroups based on critical disease-related characteristics. This process enables personalized interventions, optimized care management, and tailored treatments. Patient stratification plays a significant role in drug development and clinical practice for many diseases. However, with the increasing availability of biomedical data, such as gene expression data, clinical records, and lifestyle/environmental factors, the analysis of this vast and multimodal data becomes highly challenging. Machine learning offers methods that can help address the challenges of transforming this extensive and diverse data into usable decision-support tools. Deep learning methods, in particular, have shown impressive results in tasks such as risk stratification and treatment response prediction. However, their impact on data-driven medicine remains limited due to their 'black-box' nature and their inability to provide human-interpretable outputs. In this study, we propose applying topological data analysis to enhance the interpretability of deep learning patient stratification models. Specifically, we suggest using the Mapper algorithm to visualize the latent space learned by the models through the lens of its predictions. We apply the Mapper algorithm to various architectures of recently developed deep patient stratification models and demonstrate how it helps reveal relationships among different patient subgroups. Furthermore, we adapt the Normalized Mutual Information measure to identify the Mapper's parameters that yield the most optimal graph-based representation of the latent space. This approach aims to enrich the power of deep learning with interpretable results in the field of patient stratification.**

Keywords — patient stratification, deep learning, interpretability, topological data analysis, explainable AI

I. Introduction

Precision medicine (PM) is an innovative healthcare approach that takes into consideration an individual's unique genetic makeup, environmental factors, and lifestyle when making medical decisions. By aggregating and analyzing various types of biomedical data, such as omics data, clinical images, electronic health records, and lifestyle factors, PM aims to provide more personalized preventive measures, accurate diagnoses, and effective treatments for severe diseases like cancer and Alzheimer's. The recent advancements in deep learning (DL) and its breakthroughs in diagnostic, prognostic, and predictive tasks have highlighted the potential of machine learning (ML) in PM [1]. The key strength of DL lies in its ability to learn task-specific data representations by performing multiple non-linear transformations on input variables, facilitated by thousands of adjustable network parameters. Some studies suggest that DL models can perform healthcare tasks as well as, or better than, human experts, such as disease detection from medical imaging [2]. Nevertheless, the constrained adaptation of DL-based solutions in clinical settings primarily stems from their inability to provide interpretable outputs that healthcare practitioners can understand.

DL methods have shown immense promise in patient stratification tasks [3], where patients are grouped based on predefined characteristics. Such models project high-dimensional, heterogeneous data into a relatively low-dimensional latent space optimized for clustering based on a defined objective function. The objective function is formulated to address specific research questions. These models produce patient groupings that can be further analyzed in terms of their biological differences. However, assigning patients to a single subgroup may not fully capture the complex patterns in the data. Moreover, distinct categorical groups defined by the model's output may be too coarse to capture nuances within the data. In this study, we introduce an innovative approach to augment the output of DL patient stratification models, facilitating more profound analysis and interpretation of observed patterns. Our method leverages topological data analysis (TDA) to construct and visualize a graph-based representation of the latent space acquired by the model. To tackle a central challenge of the TDA method, which pertains to the selection of its parameters, we propose an adaptation of the Normalized Mutual Information (NMI) measure to identify the parameters that yield the most optimal graph-based representation of the latent space. This approach enables practitioners to gain a global understanding of the structures within the identified clusters and explore relationships between different patient subgroups, potentially leading to further stratification.

The remainder of the paper is structured as follows: we first provide background information on TDA and DL-based patient stratification models. Following this, we discuss relevant work related to the application of TDA in healthcare. Finally, we

present our proposed approach and apply it using different architectures of an existing DL patient stratification model.

II. Preliminaries

A. Topological Data Analysis

TDA has recently emerged as a powerful and interpretable framework for extracting valuable information from high-dimensional data [4]. It provides tools that are based on computational geometry and topology to summarize the inherent shape and structure present in multidimensional data. While many datasets have certain shapes that carry essential information, the application of this fundamental concept in contemporary data science and ML is often limited to tasks like regression (assuming data follows a linear or hyperplane shape) and cluster analysis (assuming data is divided into distinct clusters). However, topology can reveal more intricate and meaningful structures within data, such as loops, flares, or voids, which can be highly relevant for data analysis.

One of the pivotal advancements in TDA is the Mapper algorithm first introduced by Singh and Carlsson in 2007 [5]. Mapper is used to construct graph-based representations of high-dimensional data, capturing both topological and geometric information at a specified level of detail or resolution. The workflow of the Mapper algorithm is illustrated in Fig. 1 [6].

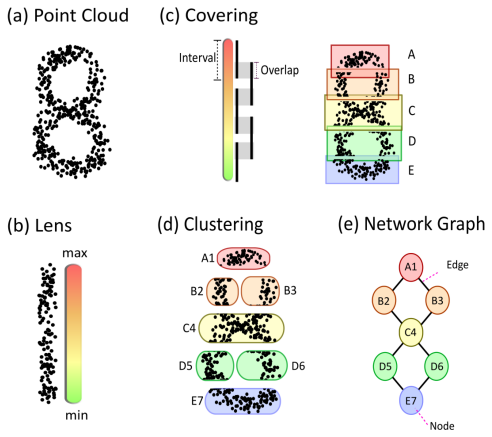

Fig. 1. Ilustration of the Mapper algorithm workflow.

The input data $X \in R^n$ from a high-dimensional space is first projected onto a one-dimensional space using a lens function (Fig. 1 (b)). This lens function can be chosen based on the specific properties of the data that you want to study or highlight (e.g. patient survival). The projected data, now in one-dimensional space, is divided into overlapping intervals or bins of equal length (Fig. 1 (c)). The overlap and the number of bins are crucial parameters that influence the level of detail and granularity in the analysis. Apart from visual inspection, there is no established way to assess the quality of a Mapper graph. Construction of a relevant graph usually requires an exhaustive search through a parameter space followed by manual validation. Within each of these bins, the data points are clustered together based on their original representations in the high-dimensional space (i.e. on their inverse image) as illustrated in Fig. 1 (d). Clustering is performed separately within each bin producing a collection of clusters per bin. Finally, a network graph is constructed, where each node in the graph represents a single cluster (Fig. 1 (e)). If two clusters contain overlapping data points, an edge is created to connect the corresponding nodes in the graph. The result of this process is a Mapper graph that provides a structured representation of the high-dimensional data, capturing both local and global relationships among data points. This graph can be analyzed to gain insights into the underlying structure of the data, identify subgroups, and reveal important topological features.

B. DL Based Stratification Models

The key strength of DL models is their capacity to learn task-specific data representations. This ability to automatically extract relevant features and representations from raw data is a fundamental reason behind the success of DL in various domains, including patient stratification in healthcare [7]. In the context of patient stratification, DL models are utilized to find a data representation, which helps to identify clusters or subgroups of patients that share certain characteristics or exhibit particular behaviors. The DL model learns to transform the high-dimensional, and often complex, input data into a lower-dimensional representation where these clusters or patterns become more apparent.

To demonstrate our proposed approach, we will use our DL-based patient stratification model recently developed for the identification of prognostic liver cancer subgroups [8]. The model has an Autoencoder architecture (Fig. 2), which is commonly used for data dimensional reduction. Training an Artificial Neural Network (ANN), such as an Autoencoder, is typically an iterative process that uses an objective function; commonly known as a loss function. The loss function is designed for the specific task of interest (e.g. clustering) and is used to assess how well the network is performing at each iteration of the learning process. The loss guides the network updates for the next iteration to help arrive at the optimum solution for the task at hand. For an Autoencoder, where the goal is reconstruction of the data, the loss function is used to evaluate how well the original (input) data can be retrieved from the learnt (latent) data representation (referred to as bottleneck). The latent space learnt by the model is represented by the middle layer of the Autoencoder model presented in Fig. 2. In order to adapt an Autoencoder to a patient stratification task and hence incentivize the latent space with patient survival and clustering relevance, a new loss function was introduced in [8]. As presented in Fig. 2, the loss function incorporates three different function which are: (1) reconstruction loss (L_R), (2) clustering loss (L_C), and (3) survival loss (L_S). The reconstruction loss ensures that the data can be effectively projected into a lower-dimensional space while preserving the essential information. The clustering loss is used to ensure homogeneity of each cluster and it was driven from the Silhouette score [9], a well-known cluster evaluation metric. Finally, the survival loss was implemented based on the Cox proportional hazards model to make the identified clusters distinct in terms of prognosis or patient survival outcome. Overall, this approach combines dimensionality reduction,

clustering, and survival analysis in a DL framework to perform patient stratification and identify subgroups that exhibit specific characteristics with clinical relevance.

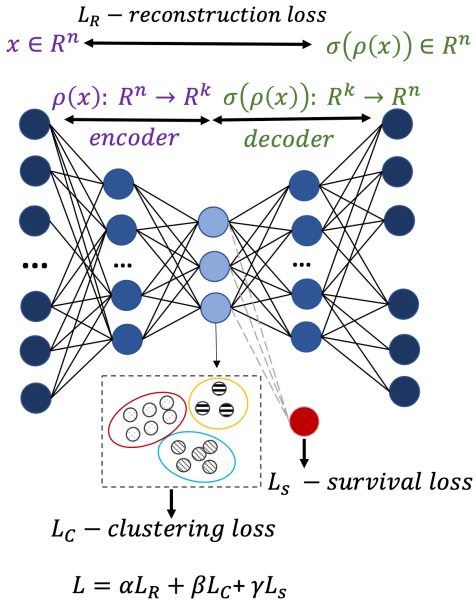

Fig. 2. Deep patient stratification model.

III. Relevant Work

The issue of explainable AI has garnered considerable attention over the past decade [10]. This matter is particularly crucial for DL models, owing to their reliance on a multitude of abstract parameters and complex calculations that render their decisions challenging to elucidate [11]. Contemporary literature discusses various prevalent techniques for interpreting DL models, primarily revolving around the identification of pivotal features contributing to model decisions. Examples encompass training surrogate models, which aim to approximate the 'black box' model's behavior using intrinsically explainable ML methods such as decision trees or linear regression [12], or employing layer wise relevance propagation to discern the most influential features from the input vector on the output vector of a DL model [13].

In this study, we propose a novel approach to address the interpretability challenge of DL models, with a specific focus on the unsupervised task of patient stratification within the exemplar context of liver cancer prognosis. Diverging from existing methods that elucidate model decisions, our technique strives to provide a more profound understanding of the underlying structure of the data within the patient stratification output. We achieve this by visualizing the latent representation of the input data, learned by the DL model, using the Mapper algorithm. This methodology allows us to offer deeper insights into the configuration and composition of the identified patient clusters, unveiling diverse relationships among patients.

TDA has already been employed as a tool to explain predictions generated by ML models. For instance, Saul et al. [14] utilized the Mapper algorithm to visualize predicted probabilities from a trained ML model, enabling the identification of patterns learned by the model and the comprehension of interinstance relationships. In a similar way, Xenopoulos *et al.* [15] proposed a topology-based framework to model and compare various explainability methods, aiming to establish a stable representation of explanations. Elhamdadi *et al.* [16] harnessed TDA to visualize the topological shape of facial landmarks over time in affective computing, providing the capability to derive explanations for identified features. Additionally, in the work of Carlsson *et al.* [17], the authors suggested utilizing the output of ML models as a filter function for the Mapper algorithm to classify different types of prediction errors. In our study, we introduce a novel approach for interpreting the outputs of DL-based patient stratification models. This approach involves applying the Mapper algorithm to the latent data representation learned by the models. To overcome the challenge of selecting Mapper's parameters, we propose to assess the quality of the Mapper graph using the NMI measure.

IV. Method

The central concept of the proposed approach revolves around constructing a Mapper graph within the latent space learned by the DL patient stratification model. This construction serves the purpose of visualizing the structural organization of the identified patient groups and the relationships that exist among various patient subgroups. Depending on the architecture of the stratification model, it is possible to finetune the Mapper graph to align with the specific stratification task at hand.

The model depicted in Fig. 2 operates by constructing the latent space while optimizing three distinct loss functions. The survival loss plays a critical role by incentivizing the latent space representation of samples with survival relevance. Technically, this loss quantifies how accurately we can predict a patient's survival based on the latent representation of their input data. In our approach, we propose to employ the survival prediction values as the lens values for the Mapper algorithm. This lens selection allows us to encode the survival aspect into the graph's structure, thus visualizing relationships in terms of survival probabilities among patients, both within and across prognostic groups.

In cases involving more conventional DL stratification models, where only clustering and reconstruction losses are considered, a lens that computes geometric properties of the data, such as the L2-norm, can be employed. The framework we propose, which integrates the Mapper algorithm with DL stratification models, is illustrated in Fig. 3. Here, a dataset denoted as $X \in R^n$ serves as input to the DL stratification model. After forming patient groupings, the latent representation of the entire patient cohort, denoted as $\rho(X) \in R^k$, is derived from the model and subsequently used as input data for the Mapper algorithm. Based on predicted survival values, samples from X are categorized into overlapping bins and internally clustered based on their original representations in R^n. Using the resulting clusters and their compositions, a graph is constructed, which can be color coded according to the stratification subgroups. It should be noted that this coloring could be modified to align with any other clinically relevant features if further analysis were required.

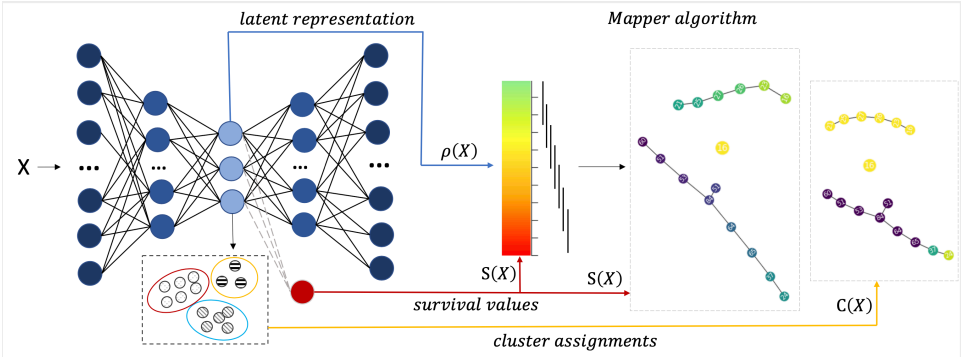

Fig.3. Deep patient stratification model.

A. Selecting Mapper Parameters

One of the primary challenges encountered when applying the Mapper algorithm pertains to the selection of its parameters. Existing literature highlights the sensitivity of the Mapper graph's shape to factors, such as the number of bins and the degree of overlap utilized, and the lack of a universally accepted method for determining their optimal values [18]. Research in this domain has been relatively limited. Consequently, it has become common practice to run the Mapper algorithm with various parameter combinations, subsequently assessing the output graphs by human interpretation.

In our current work, our objective is to identify the Mapper graph that best represents the latent space and the patient groupings discerned by the deep neural network. To achieve this, we propose an adaptation of the NMI measure [19]. The NMI is a metric commonly employed to assess the quality of clustering algorithms in supervised settings where ground truth labels are available. In our context, all patients are categorized as belonging to one of the discovered clusters, effectively serving as our ground truth labels. Once the Mapper graph is constructed, we consider each of its individual components (i.e. Mapper graph can consist of more than one disconnected components/parts) as a cluster. For a set of stratifying clusters denoted as C and the components of the Mapper graph represented as M, we calculate the NMI metric as defined in Equation 1. NMI results in values ranging from *0* to *1*, with *1* signifying a perfect correlation between C and M.

$$NMI(C, M) = \frac{2 \times I(C;M)}{H(C)+H(M)} \quad (1)$$

In this context, the symbol C denotes the ground truth labels, while M signifies different components of the Mapper graph. We calculate the entropy of the clusters, denoted as $H(C)$, and the entropy of the Mapper's components, represented as $H(M)$, according to the formulations provided in Equations 2 and 3 respectively.

$$H(C) = -\sum_{c \in \{C\}} P(c) \log_2 P(c) \quad (2)$$

$$H(M) = -\sum_{m \in \{M\}} P(m) \log_2 P(m) \quad (3)$$

The symbols $P(c)$ and $P(m)$ denote the probabilities of a data point being classified as c and m, respectively. The mutual information between the cluster labels and the Mapper's components labels, denoted as $I(C; M)$, is computed following the formula presented in Equation 4.

$$I(C; M) = H(C) - H(C \| M) \quad (4)$$

Where $H(C \parallel M)$ is the conditional entropy, and calculated as per equation 5.

$$H(C \| M) = \sum_{m \in M} -P(m) \times \sum_{c \in C} P(c \| m) \times log_2(P(c \| m)) \quad (5)$$

In our approach, we execute the Mapper algorithm using various combinations of its two parameters, specifically the number of bins (2 to 30) and overlap values (0.1 to 0.5). We then choose the combination that yields the highest NMI value. This selected graph is regarded as the most optimal representation of the identified subgroups and serves as the final output.

B. Data

To explore the proposed parameter selection method, we consider the problem of stratifying liver cancer patients. For this we use publicly available data involving complex, high-dimensional multi-omics data, including miRNA, RNA- Seq, methylation, and survival information for primary liver tumor samples from Hepatocellular Carcinoma (HCC) patients obtained from The Cancer Genome Atlas (TCGA). The TCGA data was acquired and subjected to preprocessing using TCGA-assembler 2 [20], following a methodology akin to that outlined in [21].

To ensure data quality and relevance, we selected only those samples that possessed all three types of omics data, a non-negative survival value, and a histologic diagnosis of HCC. Subsequently, for each omics type, we removed features that

exhibited either missing values or zero values in more than 20% of the samples. Additionally, we eliminated samples that had more than 20% of their features missing or containing zero values. For imputing missing values, we employed the *impute.knn* function within the R package for imputation.

Following this preprocessing phase, we retained a total of 352 samples for subsequent analyses. The three distinct omic data types were concatenated into a unified vector for each patient, thereby constructing the multi-omics matrix that served as the input for the proposed model. The final dataset encompassed 35,024 features across 352 patients.

V. Resulte

The application of the DL patient stratification model to the multi-omics dataset revealed the existence of two distinct patient subgroups. More comprehensive details regarding the characterization of these groups can be found in [8].

In Fig. 4, we present the Mapper graph with the highest NMI score, which stands at 0.61. The graph is color coded by predicted survival values (on the left) and group labels (on the right). It is noteworthy that both graphs share the same underlying structure but are visualized slightly differently. Each node within the graphs represents a group of patients, with the edges denoting the similarity among these patients. The numbers on the nodes indicate their respective sizes.

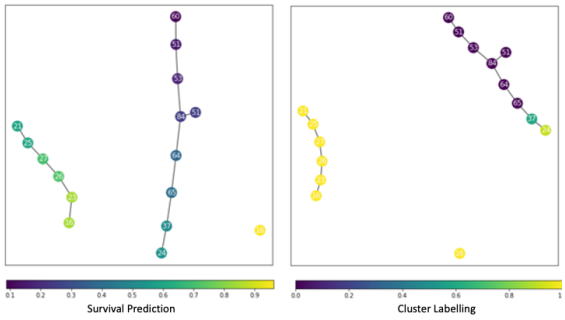

Fig.4. Mapper graph constructed on the latent representation of the data learnt by the patient stratification model. The graphs are colored by the survival prediction (left) and the cluster labels (right).

Analyzing the graph colored by group labels (right), we can discern that the Mapper algorithm adeptly separated the two groups. One of the primary components exclusively comprises samples from one of the groups (yellow). Additionally, a single node containing 16 patients from the same group can be observed. The second major component (purple) retrieved by the Mapper primarily comprises patients from the second discovered group. As indicated by their colors, two nodes within this component contain a mixture of patients from both groups. This implies that some patients cannot be definitively assigned to either group, suggesting potential characteristics that are common to both. Further exploration of this phenomenon could lead to the identification of a new subgroup.

Upon examining the colors of the nodes within the left-hand graph, we notice that the two discovered groups exhibit distinct survival characteristics. However, the presence of two components representing the yellow group may suggest some internal heterogeneity within this group. The single node appears to represent patients with potentially higher survival compared to the rest of the group. This information opens avenues for further exploration of the biological differences among patients within these two components. Such nuances may not be readily apparent solely based on grouping labels provided by the DL stratification model.

In Fig. 5, we present a series of different Mapper graphs constructed within the latent space on the same dataset which differ only by their parameter selection of the number of bins and the size of the overlap. This illustrates how important it is to have appropriate parameter settings to enable a meaningful interpretation. It is noteworthy that the NMI metric penalizes the division of discovered subgroups into multiple components, even when those components exhibit homogeneity in terms of the group labels. This characteristic aligns with our objective of retrieving, to the greatest extent possible, the same number of groups as indicated by the DL stratification model. However, it's important to consider that relaxing the selection criteria (i.e., exploring Mapper graphs with lower NMI values) may potentially lead to the discovery of intriguing subgroups within the overarching stratification groups. This is particularly relevant as each stratification group is represented by multiple components of the Mapper graph.

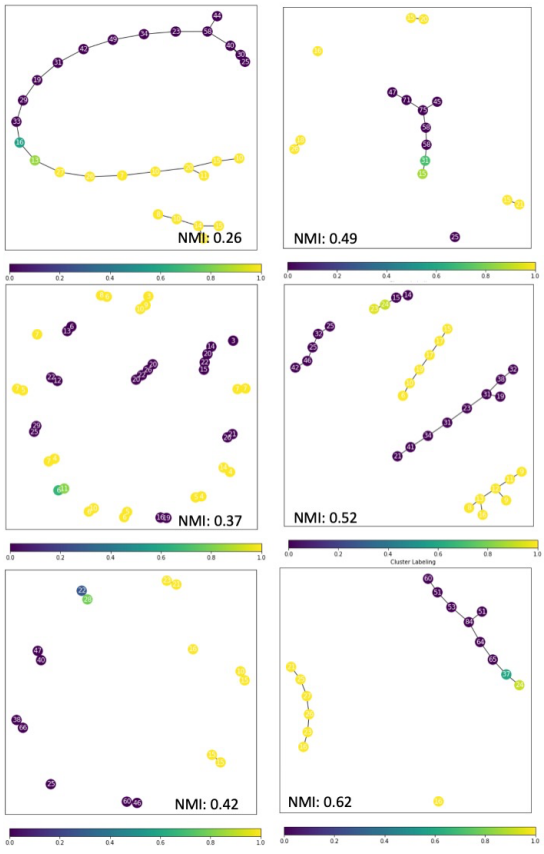

Fig. 5. Mapper graphs constructed with different parameters values on the latent representation of the data. Each Mapper graph is colored by the cluster labels and has its NMI value calculated.

To showcase the versatility of our proposed approach, we applied it with DL stratification models employing different architectures. In this scenario, we modified the model depicted in Fig. 2 to exclude the inclusion of survival loss during the training process. In such a configuration, the model learns the latent space, enabling the detection of biologically homogeneous clusters without the necessity of incentivizing the latent space with survival information. Similar to our previous approach, we utilized the latent representation of the multi-omics data as the input for the Mapper graph. In this instance, we employed the $L2$-norm as the lens function. The resultant Mapper graph (with the greatest NMI), color coded by the group label, is presented in Fig. 6. Notably, the Mapper algorithm once again effectively discerned the two groups learned by the DL stratification model. As observed previously, we can identify subgroups of patients who may not clearly belong to just one of the groups. Additionally, the shapes of each group (resembling branches) provide indications of potential subgroups that warrant further investigation.

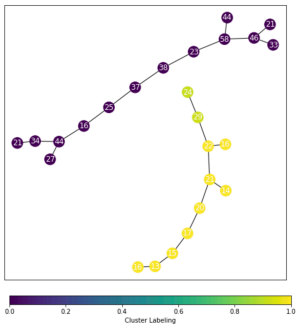

Fig. 6. Mapper graph constructed on the latent representation of the data learnt by the deep clustering model, colored by cluster labels.

VI. Conclusions

We introduce a novel approach for interpreting the results generated by DL patient stratification models. Specifically, we propose the utilization of topological tools to represent the latent space learned by the model as a similarity graph. To facilitate this, we adapt the NMI metric to serve as an evaluation method for Mapper graphs concerning their alignment with the outputs provided by the DL stratification model. Through the application of this approach to real-life data, we illustrate that visualizing the stratification groups as a similarity graph can unveil phenomena that might remain undetectable when solely examining the labels assigned to each patient. In our future work, we intend to further enhance this approach by integrating feature selection and AI explainability techniques into the Mapper graph. This will enable the identification of key features responsible for shaping the overall structure or specific portions of the graph.

References

[1] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang, "Artificial intelligence in healthcare: past, present and future," *Stroke and vascular neurology*, 2017.

[2] X. Liu, L. Faes, A. U. Kale, S. K. Wagner, D. J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern *et al.*, "A compar- ison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta- analysis," *The lancet digital health*, 2019.

[3] B. K. Beaulieu-Jones, W. Yuan, G. A. Brat, A. L. Beam, G. Weber, M. Ruffin, and I. S. Kohane, "Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians?" *NPJ digital medicine*, 2021.

[4] G. Carlsson, "Topology and data," *Bulletin of the American Mathematical Society*, 2009.

[5] G. Singh, F. Me´moli, G. E. Carlsson *et al.*, "Topological methods for the analysis of high dimensional data sets and 3d object recognition." *PBG@ Eurographics*, 2007.

[6] C. Loughrey, P. Fitzpatrick, N. Orr, and A. Jurek-Loughrey, "The topology of data: Opportunities for cancer research," *Bioinformatics*, 2021.

[7] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *TPAMI*, 2013.

[8] A. R. Owens, C. E. McInerney, K. M. Prise, D. G. McArt, and A. Jurek-Loughrey, "Novel deep learning-based solution for identification of prognostic subgroups in liver cancer (hepatocellular carcinoma)," *BMC bioinformatics*, 2021.

[9] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, 1987.

[10] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, 2020.

[11] P. Angelov and E. Soares, "Towards explainable deep neural networks (xdnn)," *Neural Networks*, 2020.

[12] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *ACM SIGKDD*, 2016.

[13] O. Csisza´r, G. Csisza´r, and J. Dombi, "Interpretable neural networks based on continuous-valued logic and multicriteria decision operators," *Knowledge-Based Systems*, 2020.

[14] N. Saul and D. L. Arendt, "Machine learning explanations with topological data analysis," in *VISxAI Workshop*, 2018.

[15] P. Xenopoulos, G. Chan, H. Doraiswamy, L. G. Nonato, B. Barr, and C. Silva, "Topological representations of local explanations," *arXiv preprint arXiv:2201.02155*, 2022.

[16] H. Elhamdadi, S. Canavan, and P. Rosen, "Affectivetda: Using topological data analysis to improve analysis and explainability in affective computing," *IEEE TVCG*, 2021.

[17] L. S. Carlsson, M. Vejdemo-Johansson, G. Carlsson, and P. G. Jo¨nsson, "Fibers of failure: Classifying errors in predictive processes," *Algorithms*, 2020.

[18] M.Carriere,B.Michel,andS.Oudot,"Statistical analysis and parameter selection for mapper," *JMLR*, 2018.

[19] T. O. Kvalseth, "On normalized mutual information: Measure derivations and properties," *Entropy*, 2017.

[20] L. Wei, Z. Jin, S. Yang, Y. Xu, Y. Zhu, and Y. Ji, "Tcga-assembler 2: software pipeline for retrieval and processing of tcga/cptac data," *Bioinformatics*, 2018.

[21] K. Chaudhary, O. B. Poirion, L. Lu, and L. X. Garmire, "Deep learning based multi-omics integration robustly predicts survival in liver cancer," *Clinical Cancer Research*, 2018.