Exploring and analyzing high dimensional data using the Extremum Graph

A PROJECT REPORT
SUBMITTED IN PARTIAL FULFILMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
Master of Technology
IN
Faculty of Engineering

BY

Meenaly Yadav

Computer Science and Automation
Indian Institute of Science
Bangalore – 560 012 (INDIA)

July, 2023

Declaration of Originality

I, **Meenaly Yadav**, with SR No. **04-04-00-10-51-21-1-19265** hereby declare that the material presented in the thesis titled

Exploring and analyzing high dimensional data using the Extremum Graph

represents original work carried out by me in the **Department of Computer Science and Automation** at **Indian Institute of Science** during the years **2021-2023**.
With my signature, I certify that:

- I have not manipulated any of the data or results.
- I have not committed any plagiarism of intellectual property. I have clearly indicated and referenced the contributions of others.
- I have explicitly acknowledged all collaborative research and discussions.
- I have understood that any false claim will result in severe disciplinary action.
- I have understood that the work may be screened for any form of academic misconduct.

Date: 16 July 2023 Student Signature:

In my capacity as supervisor of the above-mentioned work, I certify that the above statements are true to the best of my knowledge, and I have carried out due diligence to ensure the originality of the report.

Advisor Name: Vijay Natarajan Advisor Signature

DEDICATED TO

My Parents

Acknowledgements

First, I would like to thank Prof. Vijay Natarajan of the Visualization and Graphics Lab (VGL), under the Department of Computer Science and Automation, IISc Bangalore, for his involvement in this journey from the beginning and providing the necessary support and direction throughout. He was gracious and enthusiastic to accept to mentor and guide me. Regular discussions with him, and his inquisitive questioning at every detail of my work, has propelled me to be a good researcher.

I would like to give my gratitude to Dr. Raghavendra G S. His supportive nature throughout the journey, and his guidance, whenever I got stuck in my project, helped me to keep pace and complete the project.

I would finally like to express my gratitude to my parents and my beloved friend and all the loved ones who were always willing to provide a patient ear when I needed it, and constant encouragement to bring this work to the finish line.

Abstract

Here goes your crisp little abstract. The challenges faced in visualization using machine learning techniques are - the utilization of the black box models in interpreting the model behaviors and the growth in computing produced millions of datasets that needed techniques to handle it. In this, we used the scalable solution to explore and analyze the high - dimensional functions encountered in scientific data analysis. We tried the interactive exploration of the topological and geometric aspects of the high-dimensional data by combining the neighborhood graph construction, corresponding topology computation, and data aggregation. We used the NDDAV – N-Dimensional data analysis and visualization. It is an interactive tool combination of dimension reduction, clustering, neighborhood graphs, and topological analysis. The extremum graphs become important in topology, as they allow for the dimensionality reduction and exploratory analysis of high dimensional scalar fields while preserving the geometric structure. Mostly the extrema are the exciting features of scalar fields, making the extremum graphs an appealing choice for high-dimensional analysis. We provided two use cases from computation biology and high energy physics to show how this setup have produced the findings similar to the other method in both the fields.

Contents

List of Figures

Chapter 1

Introduction

This chapter describes the goal and motivation for the project.

1.1 Motivation

Topological analysis has developed into a crucial tool in all areas of scientific study. A lot may be learned about the inner workings of many natural events from the topology of scalar fields. Scalar fields can be studied using a variety of methods, including extremum graphs, Morse-Smale complexes, and contour trees. Each of these abstractions offers several angles from which to see and engage with the same scalar field.

Extremum graphs have emerged as one of the most popular techniques in the topological analysis community. While maintaining extrema connection, it offers an abstract yet understandable representation of the underlying scalar field. This fundamental characteristic of extremum graphs makes them a very useful tool with a wide range of applications. For instances where many intriguing properties of a scalar field are extrema related, use feature tracking.

A popular tool for topological analysis is now the topological spine. It is the extremum graph in its extended form. An extremum graph can be enhanced with geometric data such as volume under the descending manifold and contour nesting data; this enhanced graph is known as a topological spine. These visual representations of scalar fields maintain their topological and geometric properties. The topological spine connects the chain of extrema using the traditional visual representation while preserving the topology, localization of extrema, and nesting structure of contours.

The visualization includes the design of the data and operation abstraction, characterization of the problem domain, design of the visual encoding of the data, and interaction and algorithm design. We focused on the analysis of the data and domain problem characteri- zation. The statistical, machine learning, topological and geometrical analyses are performed combined. The purpose is to provide ease of use, combining them and other techniques.

1.2 Project Goal

One of the valuable tools in the field of topological analysis is NDDAV (N-Dimensional Data Analysis and Visualization). NDDAV offers a powerful approach to explore high-dimensional data using the extremum graph. It provides an alternative method for analyzing data that replicates the results and improve the results obtained from other techniques such as Mapper or other high-dimensional visualization methods.

The extremum graph, utilized by NDDAV, enables the analysis of complex high-dimensional scalar fields while preserving their local geometric structure. By leveraging the topology of scalar fields, NDDAV extracts essential properties such as the number of connected components on an iso-surface and the identification of critical points.

NDDAV's strength lies in its ability to provide an intuitive and interactive visualization of high-dimensional data. It offers researchers and data analysts an easy and quick way to analyze their datasets. Through NDDAV, users can navigate and explore the extremum graph, gaining insights into the underlying structure of their data.

By replicating analysis results obtained from other techniques, such as Mapper, NDDAV provides a valuable alternative that can be used to validate findings and ensure robustness in data analysis. This allows researchers to employ different tools and compare results, enhancing the reliability and credibility of their analyses.

Overall, NDDAV plays a significant role in facilitating the exploration and analysis of high-dimensional data. Its availability as a user-friendly tool contributes to the advancement of topological analysis and enables researchers to gain deeper insights into complex datasets.

The purpose of the application of cancer dataset [9] and inertial confinement fusion dataset [1] is to understand the results but also to provide other insights. We have the high-dimensional domain, which is the set of the input parameters or the latent space, and the scalar function on

the domain to analyze it. We use the system to use the topological techniques to address the problem. The topology of a function provides information about its global behavior and local features. Topology provides an abstraction for visualization and analysis, whose complexity depends on the function. It needs to provide more insight. In this system, topology information is combined and linked with geometric information, which provides exploration among parallel coordinates, topological features, and scatter plots.

The topological and geometrical analysis is done on the genomic dataset with the help of N-Dimensional Data Analysis and Visualization tool(N-DDAV) and found the found the survival percentage of the various tumors and whose molecular structure is different from the normal tissues. The NDDAV is the interactive software to explore, analyze and visualize high-dimensional data. The layout has several analysis techniques, like dimension reduction and clustering, with SOTA(state-of-the-art) techniques in topological and high-dimensional neighborhood graphs. It combines the structure of a particular quantity of interest and insights into the shape and structure of the domain of interest. The layout has a drag-and-drop technique, and modules of the system are cross-linked so that the user can explore the influence of any parameter on any other result. The modules contain combination of machine learning and both the topological and geometrical features visualization like the topological spine, scatter plots, parallel coordinate, clustering, peel ploting, dimenstional reduction and many more.

The scientist studying inertial confinement fusion at high energies, to investigate a broad ensemble of simulations, physicists constructed a sophisticated surrogate model for the application. Scientists are interested in comprehending and comparing the sample distributions since the latter is driven by a sampling procedure in a high-dimensional latent space. In this scenario, the objective is to build trust in the model itself while also providing insight into the overall simulation findings, such as the reliance of fusion yield on design factors. Therefore, we want methods for assessing model errors and uncertainties with a focus on how they vary across the model domain. None of the currently available methods are especially suitable for analyzing these scientific models because they either cannot scale to the requisite sample numbers or do not give the essential information.

It is still necessary to use analytical techniques that take into account the high dimensional characteristics of high-throughput biological data, whether it is produced by sequencing, transcriptional microarrays, proteomic analysis, or other ways. The mathematical theory of form identification in high dimensions continues to be a key component of data analysis since the

computational aspect of data analysis finally finds shape features in the organization of data sets. In contrast to existing analytical methods, the method described collects information from high throughput microarray data and uses topology to deliver deeper insight.

Chapter 2

Related Work

This chapter discusses about the paper which introduces extremum graphs, existing works for Morse-Smale complexes, NDDAV, and some other past visualization tools.

2.1 Extremum Graphs

The concept of extremum graphs was initially introduced by Correa et al. as a method to visualize high-dimensional data in a two-dimensional representation while preserving the topology and geometric properties of the original data. In their work, they also presented a sequential algorithm for simplifying extremum graphs. The algorithm proposed by Correa et al. involves computing the gradient flow graph for the scalar field under analysis. It simultaneously performs two tasks: persistence simplification and identification of gradient paths from saddles to extrema. By simplifying the extremum graphs based on persistence, they are able to reduce complexity while preserving the essential features of the data. To enrich the extremum graphs with geometric information, Correa et al. introduced additional characteristics such as the volume under the descending manifold of adjacent extrema. This augmented version of extremum graphs is referred to as topological spines. By incorporating geometric attributes, topological spines provide enhanced insights into the underlying structure of the data. The work of Correa et al. represents a significant advancement in the field of topological analysis by introducing extremum graphs and their simplification algorithm. These techniques enable effective visualization and analysis of high-dimensional data while retaining important topological and geometric properties. The augmentation of extremum graphs into topological spines further enhances their utility by incorporating additional geometric information, enabling a more comprehensive understanding of the data.

2.2 Morse-Smale Complex

Shivashankar et al. developed a parallel computation algorithm for the MS complex, which analyzes scalar fields. Their method parallelizes the identification of critical points and gradient paths, improving computational efficiency. They also proposed techniques for performing external memory computations to handle large datasets that don't fit in memory. Their research addresses the need for faster analysis of increasingly large scientific datasets and contributes to the field of topological analysis.

Machine learning solves problems in many fields, such as bioinformatics, physics, and many others. With recent advances, deep learning models coupled with data analysis increases scientist interest in these models for further discovery.

Many approaches have been proposed in machine learning to probe into the model's mechanism. Like machine learning, visual analytics focuses on the interactive exploration of internal models. These techniques are primarily model specific and tied to specific setups or architectures. Recently, there have been systems focusing on developing the generic engine.

Compared to machine learning and statistical analysis tools, topological data analysis allows to pick up of the important outliers that would get ignored in standard other analysis. Topological analysis has been utilized in various previous works, but scientists are also interested in studying the information related to the high-dimensional data domain. So the HDVis[10] was proposed for computing the topology for a high-dimensional scalar-valued function. However, this contains a small number of samples. There needs to be more visualization of the mismatch between the large datasets and the topology for the high-dimensional scalar function.

As the dataset size increases, the system needs to cope with rendering and handle the visual encoding. Many previous works proposed to address these challenges, designing visual encoding to modeling and rendering the data.

In previous works, many tools guide users through high-dimensional data analysis in an interactive visual environment. XmdvTool[12] helps in the visual exploration of multivariate data and visualization techniques focusing on n-dimensional projection, like hierarchical clustering, brushing, and graphical representations. VisuMap[2] is developed to analyze the high-dimensional dataset, including dimension reduction, clustering, and linked data views for ad-

vanced applications. Dimstiller[5] focuses on the analysis and dimensionality reduction, providing global guidance to the users through expression and operator abstraction. In topological analysis techniques, Mapper[11] and persistence-based clustering[3] have been proposed to construct the represents for visualization and data analysis of high-dimensional datasets.

This system combines the visualization of both the features of topological information and geometric information. The NDDAV [8]includes graphical representation (examples – parallel coordinates, scatter plots), dimensionality reduction, clustering, and topological analysis techniques.

The usage of Mapper has previously been successful in revealing special features of RNA folding patterns[7]. In this research, mapper is used to analyze transcriptionally genomic data from diseases, with the help from Disease-Specific Genomic Analysis filtering function. By specifying a transformation that gauges how much sick tissue deviates from healthy tissue, the Disease-Specific Genomic Analysis technique of mathematical analysis of genomic data exposes the component of data relevant to illness. When used in conjunction with Mapper, Disease-Specific Genomic Analysis transformations offer a way to specify the guiding filter function by, basically, unravelling the data in accordance with the degree of overall departure from a healthy condition.

Chapter 3

Background and Definitions

Given a scalar field $f : R^n \rightarrow R$, a point $x \epsilon R^n$ is critical iff $\nabla f(x) = 0$ and regular otherwise.

$$\nabla f = (\frac{\delta f}{\delta x_1}, \frac{\delta f}{\delta x_2}, ..., \frac{\delta f}{\delta x_n})$$

The critical points of function f, can be categorized based on their Morse index. The Morse index of a critical point x corresponds to the count of negative eigenvalues in the Hessian matrix H_f evaluated at x, where

$$H_f = \begin{bmatrix} \frac{\delta^2 f}{\delta {x^2}_1} & \frac{\delta_2 f}{\delta x_1 \delta x_2} & .....\frac{\delta^2 f}{\delta x_1 x_n} \\ \frac{\delta^2 f}{\delta x_2 \delta x_1} & \frac{\delta^2 f}{\delta {x^2}_2} & .....\frac{\delta^2 f}{\delta x_2 x_n} \\ . & . & . \\ . & . & . \\ \frac{\delta^2 f}{\delta x_n \delta x_1} & \frac{\delta^2 f}{\delta x_n \delta x_2} & .....\frac{\delta^2 f}{\delta {x^2}_n} \end{bmatrix}$$

The Morse index of critical points ranges from 0 to n, where critical points with an index of 0 represent the minima of function f, and critical points with an index of n represent the maxima. Critical points with an index of k, where k is between 1 and n-1, are referred to as k-saddles. Consequently, function f exhibits (n-1) different types of saddles.

3.1 Morse-Smale Complex

The Morse-Smale Complex of a scalar field f can be represented as a directed graph. Its vertex set consists of all the critical points of f, while the edge set consists of gradient paths that connect critical points with a difference in index of 1.

At regular points in the scalar field, we can define a gradient vector that indicates the direction of the steepest increase in function value. A gradient path between two critical points, p and q, differing in index by 1, is a sequence of points where each adjacent pair (r, s) represents the

terminal point of the gradient vector at r.
For example, in a 2D scalar field, the Morse-Smale complex would include minima, 1-saddles, and maxima. The complex would consist of gradient paths connecting maxima and 1-saddles, as well as gradient paths connecting 1-saddles and minima.
Computing the Morse-Smale complex for scalar fields with high dimensionality becomes computationally challenging due to the exponential growth in the number of critical points and gradient paths. Additionally, the presence of inter-saddle gradient paths can lead to significant occlusion problems when visualizing features related to extrema.

3.2 Extremum Graph

Extremum graphs are a specific type of the Morse-Smale complex. For any scalar field $f : R^n ßR$, there are two extremum graphs:

1. Maximal extremum graph: This graph consists of maxima and (n-1)-saddles, along with the connecting gradient paths between them.
2. Minimal extremum graph: This graph includes minima and 1-saddles, along with their connecting gradient paths.

Based on the definition of critical points, the maxima of the scalar field -f correspond to the minima of f. Similarly, the (n-1)-saddles of -f correspond to the 1-saddles of f. Hence, an algorithm designed to compute the Maximal extremum graph can be utilized to compute the Minimal extremum graph. To simplify terminology, the collection of both maxima and minima is commonly referred to as extrema.

However, in order to address occlusion problems in visualization, we introduce the concept of an edge-bundled extremum graph. This subset of extremum graphs ensures that for each nonempty intersection of adjacent saddles belonging to a pair of extrema, only one representative saddle is chosen, discarding the rest. In our implementation, the representative saddle is selected based on the highest function value.

In top show an extremum graph of simple 2D terrain connects critical points along steepest ascending(or descending) lines, which join adjacent extrema and therefore better preserve locality. Bottom show an abstract representation of the extremum graph.

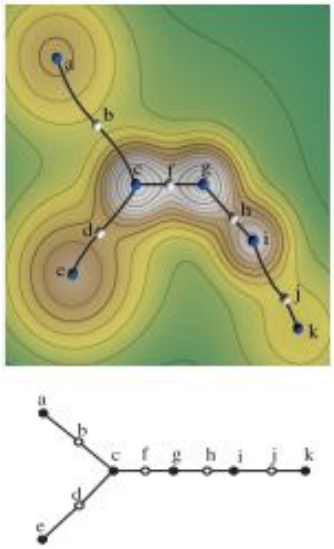

Figure 3.1: Extremum Graphs

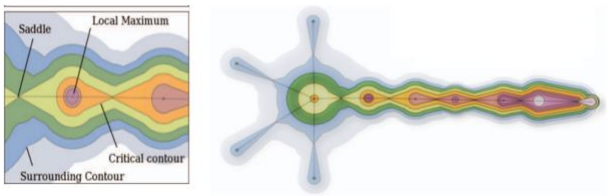

Figure 3.2: Topological spines

3.3 Topological Spine

Correa et al. first introduced the concept of topological spines, which serve as a canonical visual representation that connects a sequence of critical points while preserving the topology, locality of maxima, and nesting structure of surrounding contours. Topological spines retain essential geometric information and are particularly useful when numerous existing features are associated with extrema or when exploring the presence of neighborhood critical points in a dataset.

To visualize topological spines, the first step is to compute the extremum graph. This graph serves as the foundation for constructing the topological spines, enabling a comprehensive representation of the data's structure and relationships between critical points. By utilizing the extremum graph, topological spines provide a visual means of understanding the connectivity and hierarchy of extrema and their associated contours.

3.4 Scatter PLot

Parallel coordinates is a visualization technique commonly used to explore multivariate data. It involves representing each data point as a polyline in a coordinate system, where each dimension is represented by a vertical axis. These polylines are then displayed in parallel, allowing for the simultaneous visualization of multiple variables.

In a parallel coordinates plot, each axis represents a different variable or dimension of the data. The data points are connected by lines, creating a polyline that spans across all the axes. This representation enables the examination of patterns, trends, and relationships among the variables.

By visually comparing the shape and behavior of the polylines, one can identify correlations, clusters, and outliers in the data. Parallel coordinates provide an effective way to explore and understand the interactions between variables, as well as to detect patterns that might not be easily discernible in traditional scatter plots or histograms.

Parallel coordinates are particularly useful for visualizing high-dimensional datasets, as they provide a compact and intuitive representation of the data's structure. Additionally, interactivity can be incorporated into parallel coordinates plots, allowing users to dynamically filter, highlight, and explore specific subsets of the data based on their interests and research questions.

3.5 Parallel Coordinate

A scatter plot is a visual representation of the relationship between two numerical variables. Each data point is represented by a dot on a two-dimensional coordinate system, with one variable on the x-axis and the other on the y-axis. Scatter plots help identify patterns, trends, and correlations in the data. They are useful for exploring and analyzing numeric data, detecting outliers, and understanding the relationship between variables.

3.6 Relative Neighbour Graph

The Relative Neighbor Graph (RNG) is a proximity graph that connects points based on their relative distances. It links each point to its nearest neighbors that are not closer to any other point. The RNG is effective for handling non-uniformly distributed data and can reveal local structure and clustering patterns. It is useful in areas such as pattern recognition, data

clustering, and spatial analysis.

3.7 Gabriel Graph

The Gabriel Graph is a geometric graph construction method used in computational geometry and spatial data analysis. In the Gabriel Graph, an edge is drawn between two points if and only if the circle defined by their pair of points does not contain any other data point. It captures the proximity relationships between points while avoiding unnecessary connections. The graph provides insights into spatial arrangement, connectivity, and neighborhood structure in the data.

3.8 Diamond Graph

The Diamond Graph is a graph construction method used in computational geometry and network analysis. It connects nodes based on their shared neighbors, forming a diamond-shaped graph. It captures local connectivity patterns and is useful for studying proximity relationships in networks. The Diamond Graph helps identify clusters and analyze structural characteristics in applications such as social network analysis and spatial data analysis.

3.9 B-Skeleton Graph

The B-skeleton Graph is a method used in computational geometry and image processing to simplify the representation of objects or shapes. It constructs a graph based on the connectivity and local characteristics of the shape, emphasizing its main branches or skeleton-like structure. The B-skeleton Graph is commonly used for shape analysis and feature extraction in image processing tasks, providing insights into the shape's structural components and facilitating further analysis.

Chapter 4

Problem Description

This chapter describes the problem statement and the solution for the project.

4.1 Problem Statement

The challenges faced in visualization using machine learning techniques are - the utilization of the black box models in interpreting the model behaviors and the growth in computing produced millions of datasets that needed techniques to handle it. In this, we used the scalable solution to explore and analyze the high- dimensional functions encountered in scientific data analysis. We tried the interactive exploration of the topological and geometric aspects of the high-dimensional data by combining the neighborhood graph construction, corresponding topology computation, and data aggregation. We used the NDDAV – N-Dimensional data analysis and visualization[4].

4.2 Solution Overview

NDDAV – N-Dimensional data analysis and visualization is an interactive tool combination of dimension reduction, clustering, neighborhood graphs, and topological analysis. The extremum graphs become important in topology, as they allow for the dimensionality reduction and exploratory analysis of high dimensional scalar fields while preserving the geometric structure. Mostly the extrema are the exciting features of scalar fields, making the extremum graphs an appealing choice for high-dimensional analysis. We provided two use cases from computation biology and high energy physics to show how this setup have produced the findings similar to the other method in both the fields.

Chapter 5

Design and Implementation

5.1 System Modules

The initial system in the filter module consists of the function definition and the dataset selection. It provides the convenience of loading the files, scaling and standardizing the data, and defining the function in the domain and range. If needed, the module automatically normalizes the values by scaling each axis to either standard deviation or range.
The clustering module consists of the number of clusters and the type of clustering techniques like DBSCAN, Kmeans, Meanshift, and spectral, which are represented as color-mapped scatter plots. The persistence level is required to achieve the results in the topological decompositions setting of scale parameters. The plateaus in the curve represent the stable threshold, and the user can determine which persistence relates to noise and which to the valuable features.

The scatter plot module is linked and synchronized to the filter module. It correlates between the input pairs with colored points by either output values or topological segmentation and clustering. The table view module is linked with the parallel coordinate and scatter plot, which allows the user to investigate a particular point in the table.

The neighborhood graphs modules have three parameters to select – the number of neighbors, the Beta value, and the type of neighborhood graph. There are different types of neighborhood graphs examples – relative neighbor, ANN, Gabriel, diamond, Bskeleton, relaxed Gabriel, and many others. The parallel, coordinated module is linked with the filter module, scatter plot, and topological module; it helps to see the link in all the graphs when the user wants to investigate any of the extrema or the points with other value input parameters. The Morse complex is computed of the function which provides the maxima or minima of the function or

high dimensional valleys and mountains. The topological spine module encodes the extremum graph of scalar fields by representing the connectivity of the saddles and extrema together with the peaks. There are other modules like the summary scatter plot and summary parallel coordinate plot which can be used when the dataset is too large for the visualization. The table module provides the table information from where the user can see all modules visualization of particular data point. The plot peeling module helps the user to peel out the part from any of the geometric or topological graph and see it's other visualization in other modules.
Before, providing the datasets of both the cases, the high-dimensional datasets are firstly preprocessed and cleaned accordingly to the cases and then provided to the filtering modules and other modules of the system. So when the dataset is given in the fitering module the dataset get preprocessed and then given to the other visualization module.

5.2 Experimental Analysis

We tried to do the visual analysis on the Intertial Confinement Fusion Dataset[6] and the breast cancer tumor dataset. In this application, the physicists focus on the accuracy and behaviors like the area of input parameter that produces more errors. In this application, we focused on the high-dimensional landscapes which produce the errors. Using the NDDAV system tool, we predicted the error as a function in the input domain, which can be analyzed as a high-dimension scalar-valued function. We linked the several visualization graphs to each other modules, giving users the interactive linked view analysis and exploration. As the dataset contain some empty rows and the values are not in the normalized form. So, the dataset have the parameters and scalars values of the high dimensions which are cleaned and preprocessed as we give the dataset to the filtering module.

Using the tool, we are able to present that the system allows the interactive linked visualization analysis and exploration. From the topological spine module, there are two extrema of errors in the 15D scalar space. The number of the local extrema in the spine structure can be changed by changing the range and number of extrema in the plot. The highlighted parallel coordinated can be seen of the particular extrema. To explore the relationship between the parallel coordinates and the topological spine, just user have to focus on the peak of the topological spine which will relate and update in the parallel coordinate plot.

The extrema correspond to the patterns and plots in the parallel coordinates which is ignored by the statistical analysis techniques. Seeing the various pattern of the plots of the error together, we found that there are similarity in the plots of the parallel coordinates, extrema,

but some of the highlights do not reappear. So, in this way it helps to the physicist to explore further in the cause of the errors which were ignored before the visualization analysis. It will make interesting to understand the reason and solve it by fixing it.

In the tumor dataset biological case as the data contain of all the age group with different features like age in years, event death percentage, survival percentage, chemotherapy success, hormonal therapy, amputation breast removal, histological grade, diameter of tumour, positive in lymph nodes, cancer grade, extent of invasion into blood vessels, extent of lymphocytic infiltration, estrogen receptor expression and there are many other features present in the dataset. So, doing the feature engineering and data preprocessing on the dataset by removing the early birth age group which do not have the cancer, normalizing and removing the null dataset in the data. By seeing the relation between the features we removed the features after providing the dataset to the filtering module. After loading the dataset in the filter module and after getting the data to be preprocessed initially with default parameters of every module the system provides the visualization of different topological and geometrical graphs.

With the help of the extremum graph and topological spine visualization to the genomic data, we can create an equally graph from a significantly less accessible one as the dataset is high dimensional dataset. Many objectives can be accomplished by using the disease component deviation from the health data rather than the original data like identifying the extent to which diseased tissue data differ from the healthy tissue data, allowing range of variability within the normal range or incorporating the controls into the analysis. The use of the disease data had been found to outperform than the use of the original data and reveal distinct biology. The use of the illness component of the data has been found to perform better than the use of original data and to reveal distinct biology. In compare to a direct comparison of data from normal and neoplastic tissue, which tends to emphasise the background molecular signature of the tumor's progenitor cell type, this method emphasises how aberrant a tumor's gene expression is.

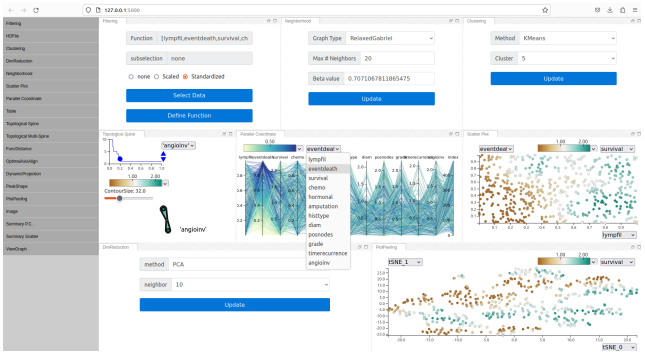

Figure 5.1: Analysis Window

Chapter 6

Results

6.1 Visualization

The topological spine encodes the extremum graph of a scalar field by representing the connectivity of the extrema and saddles together with the size of the peaks. The contours around the saddles and extrema represent the level of the function. The contour size in the module is defined as the number of samples above the contour function value. The number of the extrema in the topological spine is controlled by the plot where the x-axis is the function range, and the y-axis is the number of the local extrema.

The Neighbourhood module contains three parameters in it - the maximum number of neighbourhood, the beta value and the graph type as relative neighbor, ANN, Gabriel, diamond, Bskeleton, relaxed Gabriel, relaxed relative gabriel, relaxed diamon , grid, and relaxed bskelton. This neighbourhood graph is linked with the topological spine module. As we changes the graph type or the number of neighbourhoods or the beta value the extremum graph values get changes and which gives us different topological spin.

The number of the peaks is shown in y-axis for the given persistence values in x-axis. The persistence values at the plateau areas shows more stable topological structures.

The topological spine visualization with the different number of extremas and different type of the graphs.

The cluster module have different clustering method like PCA(principal component analysis), spectral, locally linear, Isomap, and tSNE, with the other parameter as the neighbour.

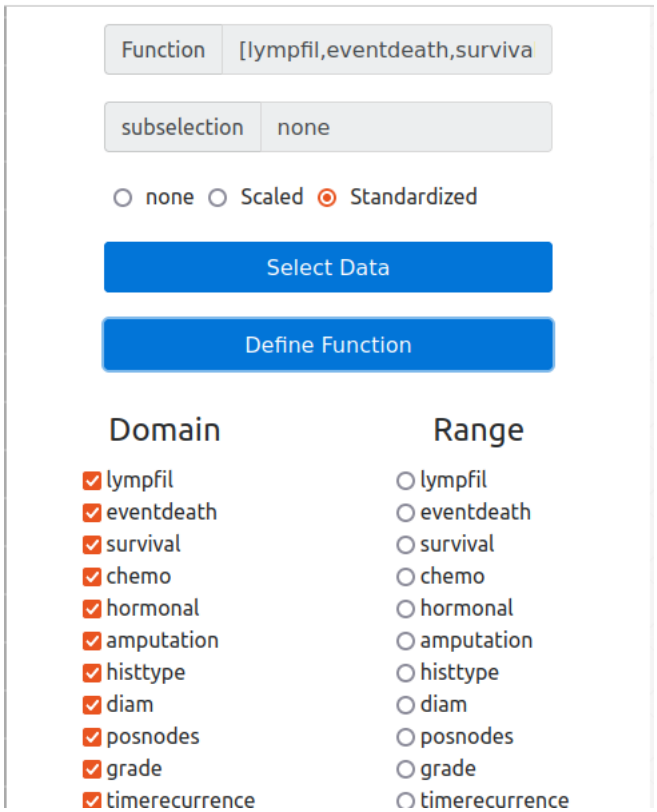

Figure 6.1: The tumor dataset is load in the filtering module with the function having the domain and range in selection and having subselection to be scaled dataset or the standardized dataset.

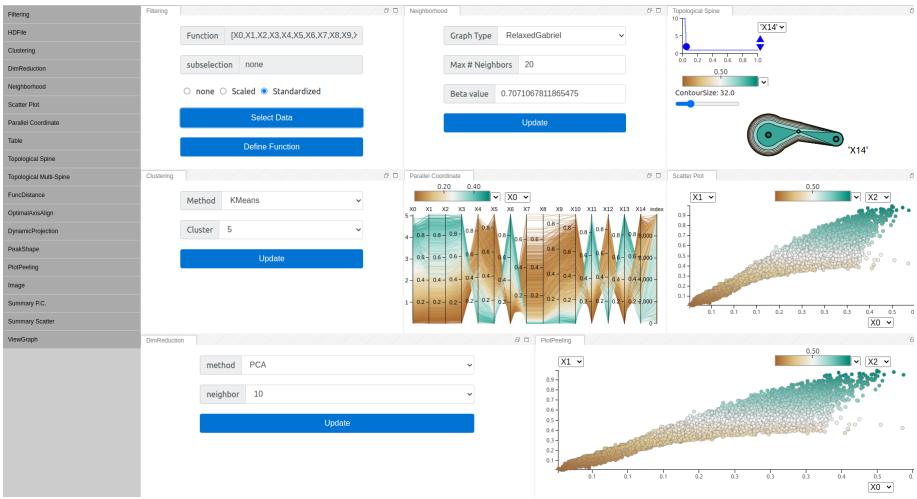

Figure 6.2: The initial layout of NDDAV containing various modules on the left allows user to have new module and drag to right side into the work area and visualize it. It is the initial visulization with the inertial confinement fusion dataset.

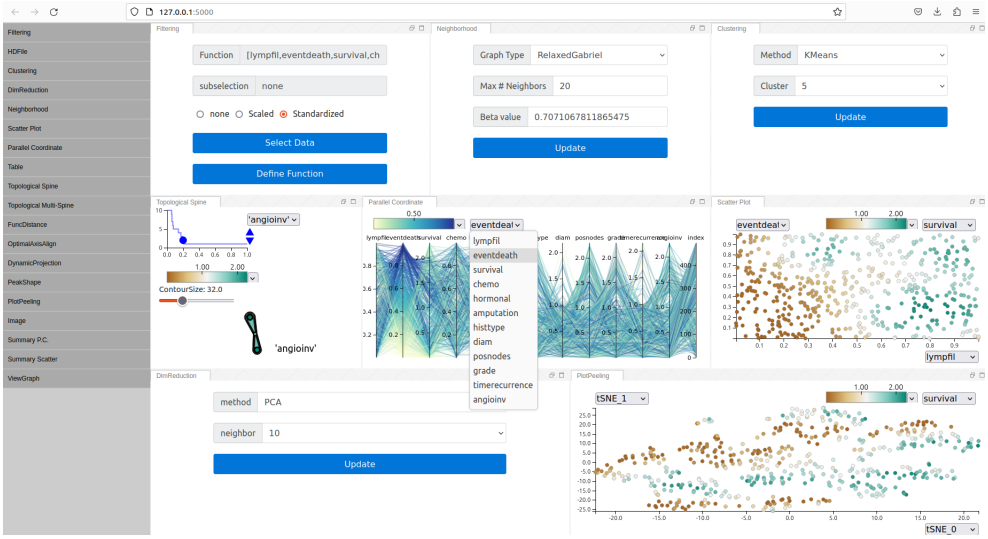

Figure 6.3: The layout shows when the tumor dataset is loaded initially after getting preprocessed, all modules are interlinked with each other.

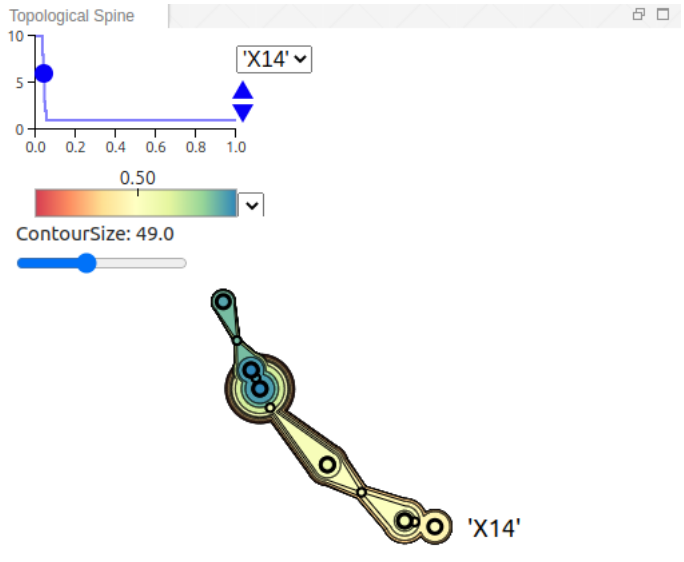

Figure 6.4: Topological Spine of the Inertial confinement fusion.

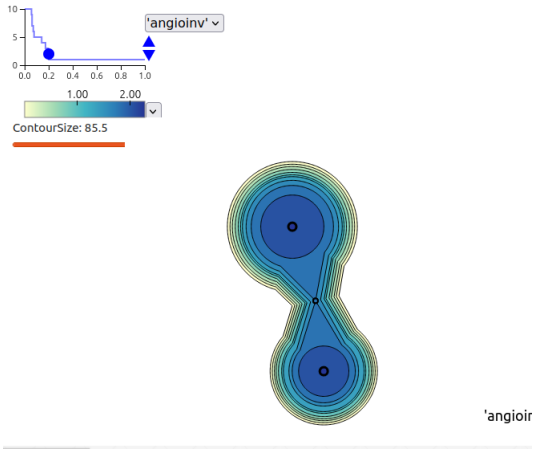

Figure 6.5: Topological Spine of the tumor dataset.

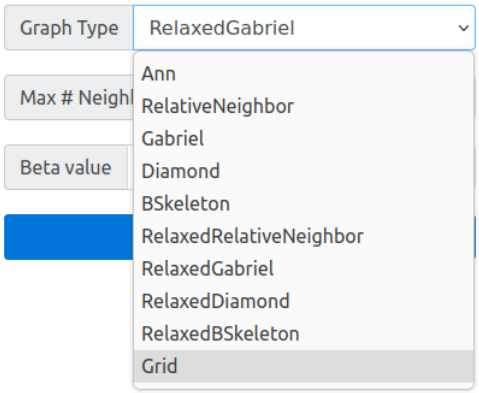

Figure 6.6: Neighbourhood Module

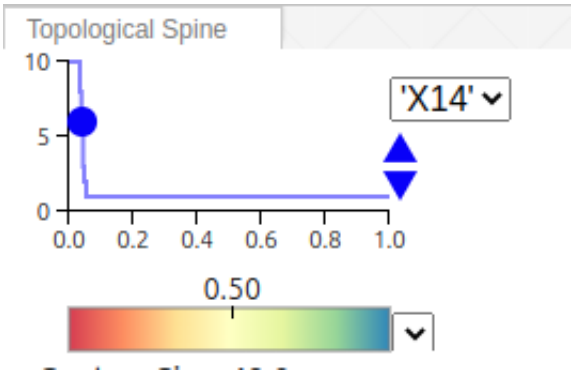

Figure 6.7: Topological Multi-Spine

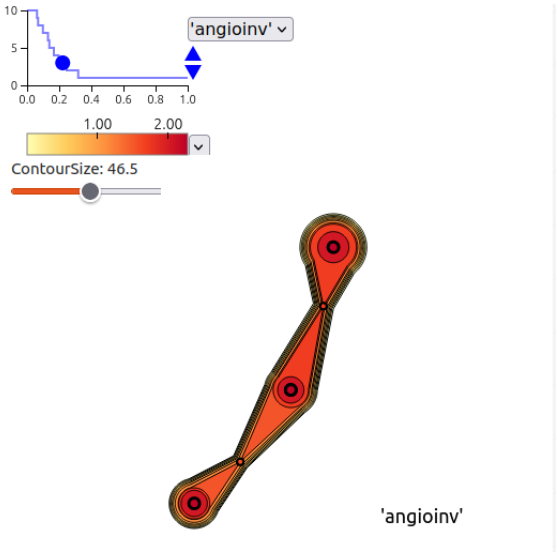

Figure 6.8: The topological spine represent the tumor dataset with the three extremas and the relative neighbour graph type

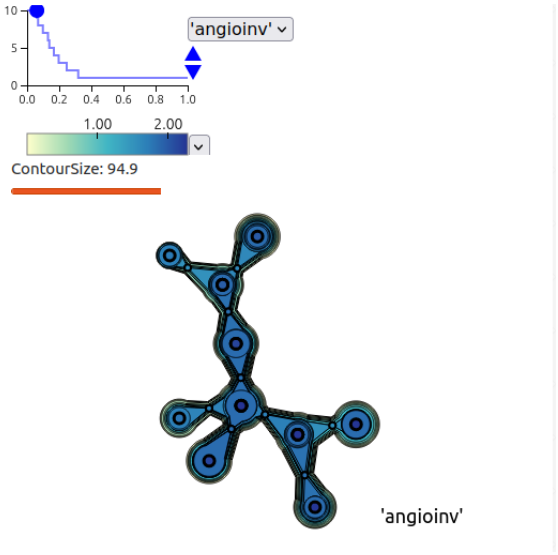

Figure 6.9: The visualization represent the tumor case with the ten extremas and the relaxed gabriel graph type

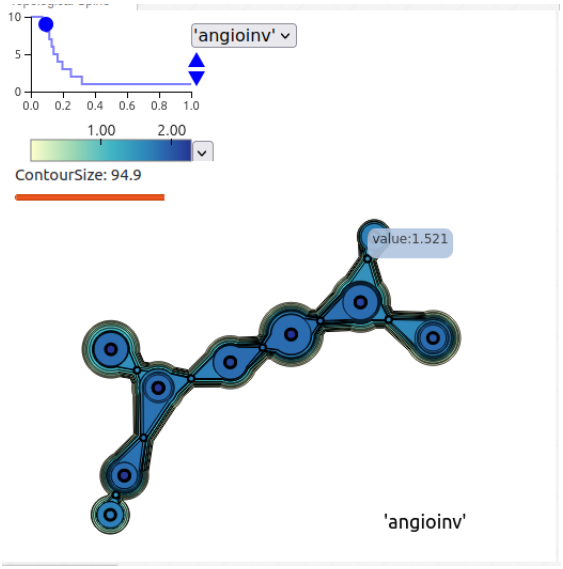

Figure 6.10: The topological spine represent the canser usecase with the nine extremas and the Bskeleton graph type

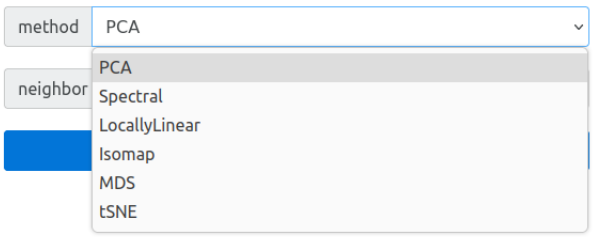

Figure 6.11: Clustering Module

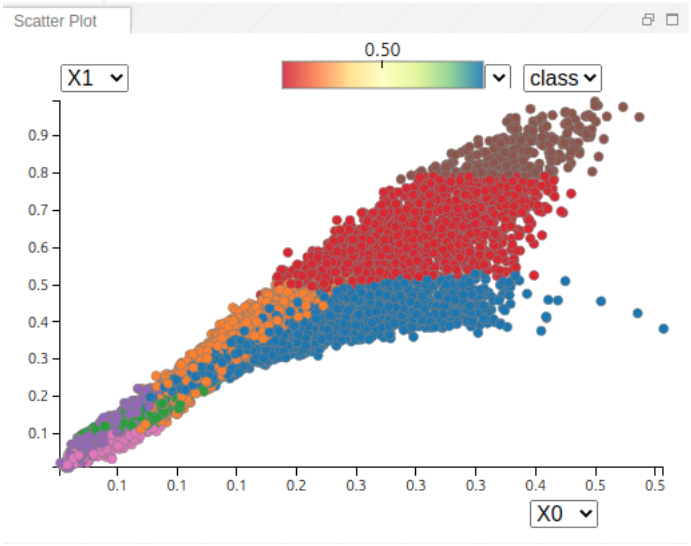

Figure 6.12: Clustering

In the clustering module, it's the inertial confinement fusion use case, it gets linked to the scatter plot and color-point plot view. The user can change the parameters in the clustering techniques like the number of the clusters, the kernel bandwith and the type of the clustering techniques.

The density scatterplots are obtained from the histogram and render the joint distribution density to avoid the overplotting problem when there are large number of datapoints.

The parallel coordinates visualization is drawn from the 2D joint distributions of the datacube. It shows the relation of the input value parameter with each other and synchronized with the other modules in the system.

The scale of the parallel coordinate contains various features for scaling the graph with giving the relation between different parameters of the dataset.

The dimensional reduction module contains two parameters as the method of the clustering and the number of clusters for the clustering method. The module gives different clustering method to the user like DBSCAN,KMeans, Meanshift and spectral clustering.
In this it is the joint exploration of the topological and geometric features of the functions in the input paramter space. Here, as we change the number of the neighbourhood or the neighborhood graph type it directly relates to the topological spine and the extrema in it.

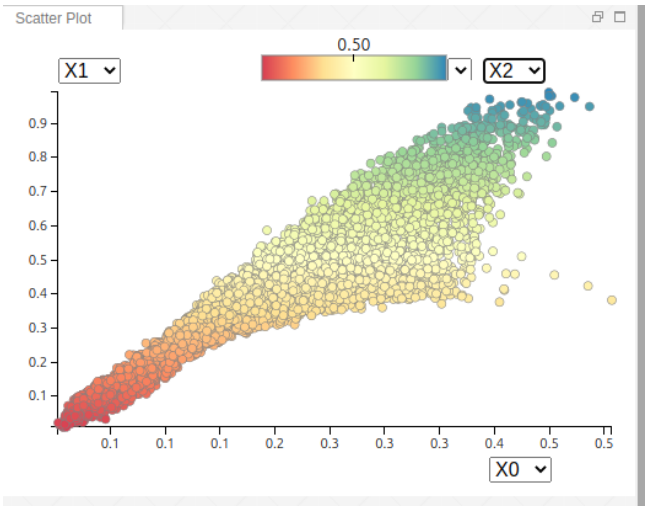

Figure 6.13: Scatter Plot

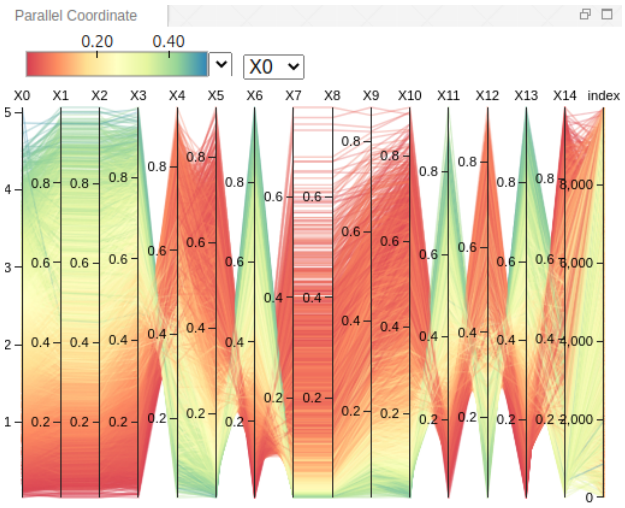

Figure 6.14: Parallel Coordinate of the intertial confinement fusion usecase

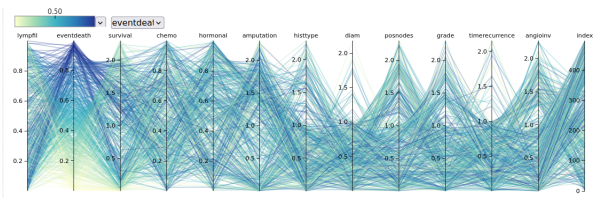

Figure 6.15: Parallel Coordinate of cancer usecase

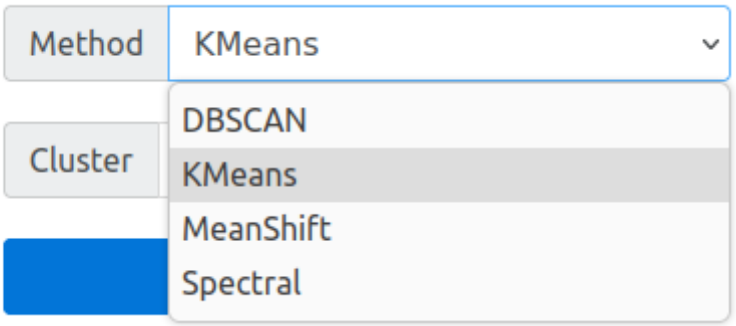

Figure 6.16: Dimension Reduction Module

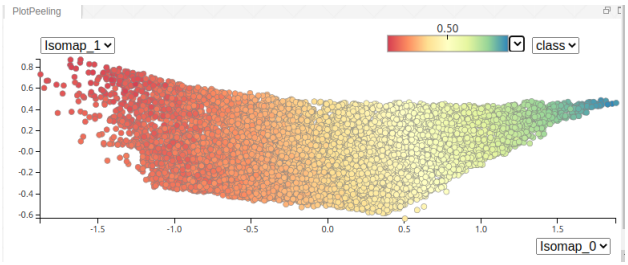

Figure 6.17: Dimension Reduction for the fusion usecase

The structure of the toplogical spine gets changed. As there are four extrema in the topology spine which represent the output error. The other module - the scatter plot and the parallel coordinate are linked to the module. The user select any of the extream point on the topological spine both the plots get peeled off and show the information related to the particular points. The user can focus on any other points of other plots and can see the peeled plot of other modules.

We changed the neighborhood graph by changing either the no. of neighborhoods, the β value or the type of the neighborhood graphs. The topology spine contains the six local extrema and the user can explore the other plots of the particular extrema by focusing on the point.

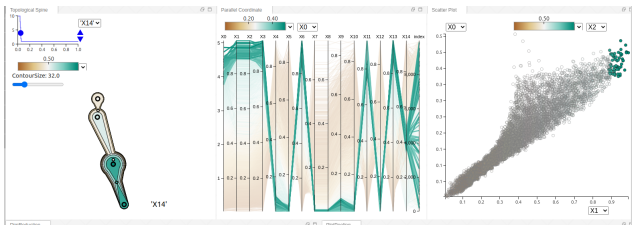

Figure 6.18: The top first extrema is selected in the topological spine and by selecting the data information got clipped in the parallel coordinate and the scatter plot.

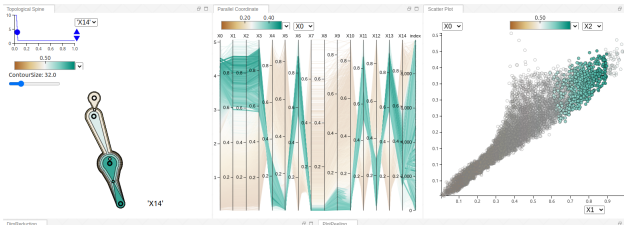

Figure 6.19: The top second extrema is selected and accordingly the views provide complementary information, and the linked selection enables a joint analysis of both geometric and topological features.

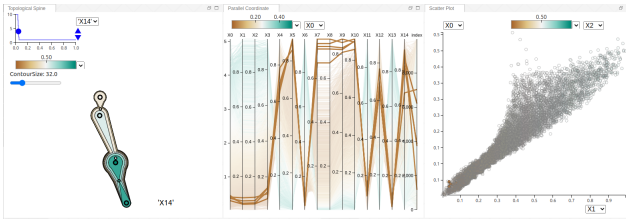

Figure 6.20: The last extrema is selected as it shows the joint exploration of both topological and geometric characteristics of the surrogate's errors as functions in the input parameter space.

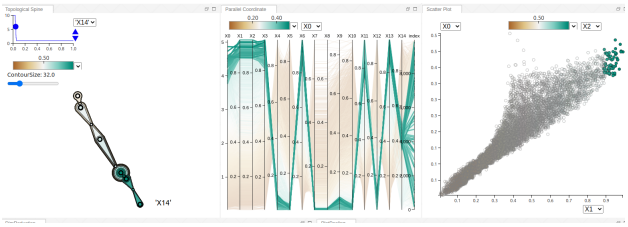

Figure 6.21: This is the flow of the visualization as we changes the parameters of the various different modules like the neighbourhood module by changing the number of neighbourhood and the graph type.

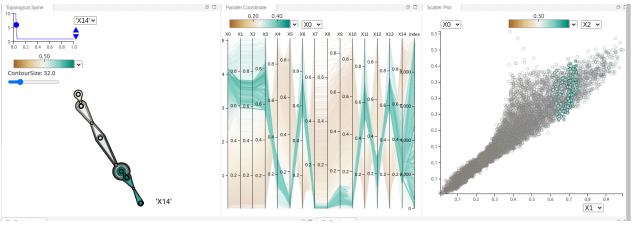

Figure 6.22: The range in the parallel coordinate in on of the features is selected and the other visualization got clipped accordingly.

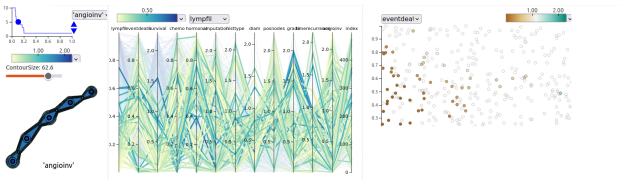

Figure 6.23: The visualization shows the deviation of the five percentage tumors are very large from the normal original tissue of a particular type.

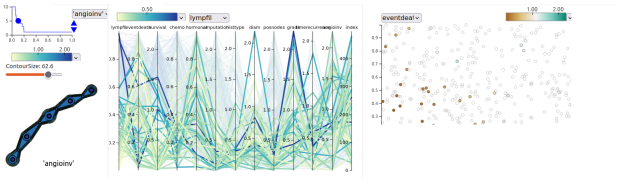

Figure 6.24: The user select the other extrema from the graph the visualization give the joint information of particular features of the tumor with it's topological and geometric features.

Here as we changes the parameters of the neighborhood modules the extremum graph results changes like the function min, function max, the extrema size and the sadddle size output.
The following gives us the visualization and insights of the cancer dataset usecase, the which type of the tumor are ordered by the deviation from the normal,it's fatal or not. The data is being clustered with the mean of the filter on the points.The tumors are being clustered on the basis of the deivation on from the healthy original data. The extrema shows us the particular tumor with it's different features which gives insight about the tumor deviation, the age group fall in that type of the tumor, the death and survival percentage, the percentage of the human tissue affected and many more.

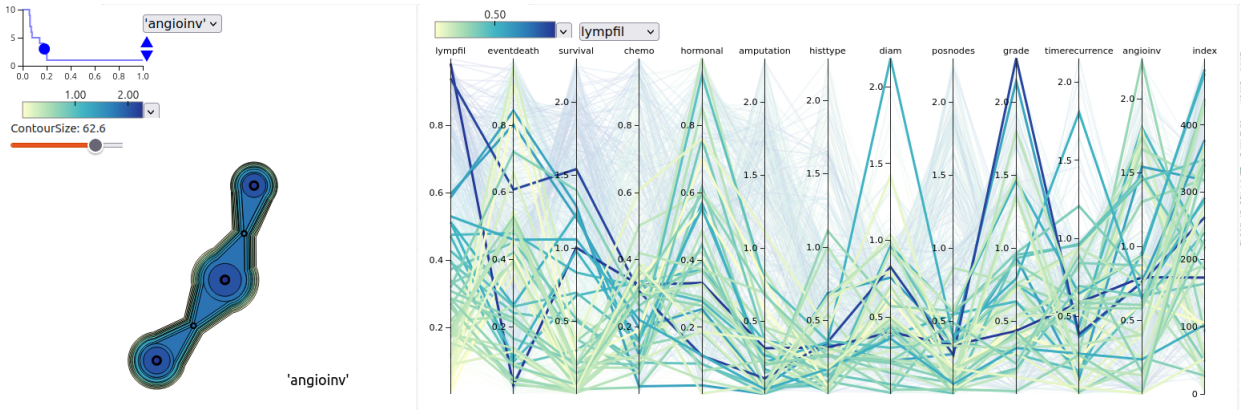

Figure 6.25: Biologist select different age group to find out what type of tumor does that group have and how much it is deviated from the original.

```
---- ExtremumGraphExt::initialize ---- 0x7ffe7617db10 500
  size: 500 attr: 12 dim: 12 func: 11
function min: 0.000602 function max: 2.275067
---- ExtremumGraphExt::computeSegmentation ----
---- ExtremumGraphExt::Reset EdgeIterator ----
---- ExtremumGraphExt::computeHierarchy ----
Before Simplify: mExtrema size: 17, mSaddle size: 96
---- ExtremumGraphExt::simplify ----
---- ExtremumGraphExt::storeLocations ----
After simplification max_segment: 10, mExtrema size: 10, mSaddle size: 86
finish computeSegments
computeHistograms 2
```

Figure 6.26: This gives the biologists regarding the insight with the help of the extremum graph with the visualization of the graphs.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

In this we have setup the system for the topological and geometrical data analysis and visualization. We tried to explore and analyse the Intertial confinement fusion dataset. We tried to identified the set of the tasks in analyzing the data derived from the pipelines for the discovery and will further try to address these challenges with the combination of the both the geometric and topological features of the other dataset of the any organ cancer. The objective is to analyse the extremum graph construction for the high dimensional scalar valued functions and to linked the plots to explore the extrema of the topological spine, to find the insights from it which can help the scientist for further research. The aim of the tool is to give more information related to the domain input parameters which are ignored by the statistical or deep learning techniques. In the scientific discovery it is more important to build the confidence in the model and understand where and why the model is unreliable. The system can help in the evaluating and fine tuning the models of the domain. The NDDAV present the stepping stone towards the visualization and analysis of the high-dimensional functions.

7.2 Future Work

We plan to work upon the visualization of the extremum graph module on the tool itself and the display of the extremum graph information data on the system. We plan to improve the tool in terms of the extremum graph analysis and adding the other modules. The main plan is to work on the other domain of the scientific discovery and explore and analyze the high dimensional dataset of the and how exyremum graph analysis can help to other domain scientist also. By the exploration will like to provide the insights to the domain which was ignored during the statistical or machine learning technique. To work with other high dimensional application

areas and to expand and improve the tool. [**?**].

Bibliography

[1] Rushil Anirudh. Jag inertial confinement fusion simulation dataset. 2019. URL https://github.com/LLNL/macc. 2

[2] Alberta Calgary. Visumap technologies inc: Visumap—a high dimensional data visualizer. 2009. 6

[3] Frédéric Chazal, Leonidas J. Guibas, Steve Y. Oudot, and Primoz Skraba. Persistence-based clustering in riemannian manifolds. page 97–106, 2011. doi: 10.1145/1998196.1998212. URL https://doi.org/10.1145/1998196.1998212. 7

[4] VGL IISc. Nddav-extremum. 2023. URL https://bitbucket.org/vgl_iisc/nddav-extremum/src/master/. 13

[5] Munzner T. Irvine V. Tory M. Bergner S. Möller T. Ingram, S. Dimstiller: workflows for dimensional analysis and reduction. 2010. 7

[6] Shusen Liu, Di Wang, Dan Maljovec, Rushil Anirudh, Jayaraman J. Thiagarajan, Sam Ade Jacobs, Brian C. Van Essen, David Hysom, Jae-Seung Yeom, Jim Gaffney, J. Luc Peterson, Peter B. Robinson, Harsh Bhatia, Valerio Pascucci, Brian K. Spears, and Peer-Timo Bremer. Scalable topological data analysis and visualization for evaluating data-driven models in scientific applications. *CoRR*, abs/1907.08325, 2019. URL http://arxiv.org/abs/1907.08325. 15

[7] Ciara F Loughrey, Padraig Fitzpatrick, Nick Orr, and Anna Jurek-Loughrey. The topology of data: opportunities for cancer research. *Bioinformatics*, 37(19):3091–3098, 07 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab553. URL https://doi.org/10.1093/bioinformatics/btab553. 7

[8] Peer-Timo Bremer Dan Maljovec Avishek Saha Bei Wang Jim A Gaffney Brian K Spears Valerio Pascucci. Nddav: N-dimensional data analysis and visualization. pages 326–333, 2015. doi: 10.1007/s00791-015-0241-3. 7

[9] Devi Ramanan. Jag inertial confinement fusion simulation dataset. 2019. URL https://data.world/deviramanan2016/nki-breast-cancer-data. 2

[10] V. Pascucci S. Gerber, P.-T. Bremer and R. Whitaker. Visual exploration of high dimensional scalar functions. 16(6):1271,, 2010. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3099238/. 6

[11] Gurjeet Singh, Facundo Memoli, and Gunnar Carlsson. Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. 2007. ISSN 1811-7813. doi: 10.2312/SPBG/SPBG07/091-100. 7

[12] M.O. Ward. Xmdvtool: integrating multiple methods for visualizing multivariate data. pages 326–333, 1994. doi: 10.1109/VISUAL.1994.346302. 6