REPUBLIC OF TÜRKİYE
YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

INTERACTIVE ANALYSIS AND DATA VISUALIZATION FOR HIGH DIMENSIONAL DATASETS

Parisa SALMANIAN

MASTER OF SCIENCE THESIS
Department of Computer Engineering
Program of Computer Engineering

Assoc. Prof. Ali Can KARACA

July, 2024

REPUBLIC OF TÜRKİYE
YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

INTERACTIVE ANALYSIS AND DATA VISUALIZATION FOR HIGH DIMENSIONAL DATASETS

A thesis submitted by Parisa SALMANIAN in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE THESIS** is approved by the committee on 22.07.2024 in Department of Computer Engineering , Program of Computer Engineering .

Assoc. Prof. Ali Can KARACA
Yildiz Technical University
Supervisor

Approved By the Examining Committee

Assoc. Prof. Ali Can KARACA, Supervisor
Yildiz Technical University _____

Ass. Prof. Furkan ÇAKMAK, Member
Yildiz Technical University _____

Ass. Prof. Nihal ALTUNTAŞ, Member
Yildiz Technical University _____

I hereby declare that I have obtained the required legal permissions during data collection and exploitation procedures, that I have made the in-text citations and cited the references properly, that I haven't falsified and/or fabricated research data and results of the study and that I have abided by the principles of the scientific research and ethics during my Thesis Study under the title of "Interactive Analysis and Data Visualization for High Dimensional Datasets " supervised by my supervisor, Assoc. Prof. Ali Can KARACA . In the case of a discovery of false statement, I am to acknowledge any legal consequence.

Parisa SALMANIAN

Signature

Dedicated to my family
and my beloved husband

ACKNOWLEDGEMENTS

Stepping into this research journey with the focus on improvement of the interpretability of complex data in the captivating field of Visual Analysis, has truly interested me. All through this incredible research, I have not only gained broad knowledge in this era but also developed a passion for exploring it further in the future.

I extend my deepest gratitude to my supervisor, Assoc. Prof. Ali Can KARACA (Yildiz Technical University, Türkiye), my co-supervisor, Dr. Rafael M. MARTINS (Linnaeus University, Sweden), and my co-supervisor, Dr. Angelos CHATZIMPARMPAS (Utrecht University, The Netherlands). Their invaluable guidance, constructive feedback, diligent attention, and endless support were crucial in the development of my web-based visualization tool.

Finally, I wish to express my sincere and heartfelt appreciation to my family and my beloved husband for their endless and invaluable support throughout this incredible journey.

Parisa SALMANIAN

TABLE OF CONTENTS

LIST OF ABBREVIATIONS vii

LIST OF FIGURES viii

LIST OF TABLES xxi

ABSTRACT xxii

ÖZET xxiii

1 INTRODUCTION 1

1.1 Dimensionality Reduction Techniques 1
1.2 Problem Formulation 3
1.3 Research Goals 3
1.4 Document Structure 5
1.5 Ethics 5

2 BACKGROUND AND RELATED WORK 7

2.1 Background 7
2.1.1 Why Uniform Manifold Approximation and Projection (UMAP)? 8
2.1.2 Interpretable Machine Learning 9
2.1.3 Explainable Boosting Machine (EBM) 11
2.2 Related Work 13
2.2.1 Goals and Shortcomings of Similar Works 13
2.3 Overcoming Limitations in DR Visualization Interpretation 17

3 DIMVIS: A VISUAL INTERPRETER FOR PROJECTIONS 18

3.1 Enhancing Data Comprehension through Interactive Visualization . 18
3.2 Goal 1: Interactive UMAP and EBM Visualizations Based on User Selection 20
3.3 Goal 2: Individual and in-depth Feature Assessment 22
3.4 Goal 3: User Support 25
3.5 Algorithms and Implimentation 26

3.5.1 Utilizing Plotly Dash for Dynamic Web-Based Applications 26
3.5.2 Updating Scatter Plot with UMAP in Dash Applications . . 26
3.5.3 Updating EBM Scatter Plot in Dash Application 27
3.5.4 Plotting Feature Importance Graph in Dash Application . . 28

4 APPLICATION 1: BREAST CANCER 30
4.1 DimVis in Action 30
4.2 Use Case: Exploring Data Clusters Without Ground-Truth Labels . 30
4.2.1 Exploring Different Projections 30
4.2.2 Evaluating Cluster Formation 31

5 APPLICATION 2: PIMA INDIAN DIABETES 42
5.1 DimVis in Practice: The Second Case Study 42
5.2 Use Case: Understanding a Projection 42
5.2.1 Overall Accuracy 43

6 APPLICATION 3: HYPERSPECTRAL IMAGES 54
6.1 HyEx in Action 54
6.2 Use Case: Analyzing Hyperspectral Image Data with Ground-Truth Labels 54
6.2.1 Data Prepration 55
6.2.2 Investigating Cluster Formations 62

7 DISCUSSION 84
7.1 Development Insights and Future Directions of DimVis and HyEx . 84
7.1.1 Design Choices 84
7.1.2 Limitations and Future work 86

8 CONCLUSION 89
8.1 Summary and Insights 89

REFERENCES 90

A PROJECT ENVIRONMENT SETUP 93

PUBLICATIONS FROM THE THESIS 94

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
DimVis	Dimensionality Visualization
DR	Dimensionality Reduction
DT	Decision Tree
EBM	Explainable Boosting Machine
HD	High-Dimensional
HSI	Hyperspectral Images
HyEx	Hyperspectral Images Explanation
IML	Interpretable Machine Learning
LD	Low-Dimensional
MDS	Multidimensional Scaling
ML	Machine Learning
PCA	Principal Component Analysis
t-SNE	t-Distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection
XAI	Explainable Artificial Intelligence

LIST OF FIGURES

Figure 2.1 Illustration of UMAP Working Process. Stage 1 (left) involves projecting cells as points on a multi-dimensional plot and determining similarities between cells in multiple dimensions. Stage 2 (right) involves refining the plot by moving points until the similarities in the low-dimensional plot capture the similarities in high dimensions, thereby accurately representing the relationships between data points. [12] 8

Figure 2.2 A comparison between traditional Machine Learning (ML) and Interpretable Machine Learning (IML). The upper part of the figure illustrates the ML process, where a model makes a prediction (e.g., identifying a husky) without providing the reasoning behind it, leaving the user uncertain. The lower part shows the IML process, where an interpretable model not only makes the prediction but also provides an explanation for the decision (e.g., highlighting specific features that led to the identification of the husky which is one of the features of glassbox models), enhancing user understanding and trust. [17] . 10

Figure 2.3 A comparison between Black-Box and Glass-Box models. The Black-Box model (left) provides a solution without showing the steps involved, making the process opaque and hard to understand. In contrast, the Glass-Box model (right) transparently displays each step of the calculation, making it clear how the solution is derived and enhancing interpretability and trust. [18] 11

Figure 2.4 Steps involved in the Explainable Boosting Machine (EBM) interface. The process begins with the input data, followed by binning of features, scoring of feature bins, adding an intercept, summing the scores, and applying softmax and argmax functions to generate the final prediction. This sequence ensures model accuracy while maintaining interpretability. [19] 12

Figure 2.5 Comparison of model interpretability and accuracy, highlighting the Explainable Boosting Machine (EBM) as a model that achieves both high interpretability and high accuracy. While traditional models like Decision Trees, Linear Regression, and Classification Rules are highly interpretable but less accurate, and models like Random Forests, Boosted Trees, and Support Vector Machines are highly accurate but less interpretable, EBM balances both attributes effectively. [20] 12

Figure 2.6 Comparison of visualizations on the Iris dataset using DT-SNE and t-SNE methods in DT-SNE paper. [9] 13

Figure 2.7 IXVC interface overview [8]. The top left scatter plot (A) represents the DR visualization, with instances color-coded based on the cluster selection displayed in the top middle section (B). The decision tree (D) explaining the selected clusters is shown at the bottom. The tree's prediction colors are visualized in the top right section (C). Users can toggle between the country and zoo datasets using the earth and cat icons (E) for evaluation purposes (refer to Section 6). 14

Figure 2.8 Visual analysis of t-SNE results using t-viSNE: (a) interface for uploading datasets, selecting execution modes (grid search or single parameter set), and saving/loading executions; (b) result overview with data-specific labels shown in categorical colors; (c) Shepard Heatmap displaying all pairwise distances; (d) histogram illustrating Density and Remaining Cost distributions; (e) ranked list of available projections by quality; (f) main scatterplot showing neighborhood Density in high-dimensional space and Remaining Cost for each point; (g) Neighborhood Preservation bar chart/line plot; (h) control elements for different interaction modes; (i) visual mapping panel with user options, including an annotation tool for multi-session notes; (j) Dimension Correlation bar chart showing data dimension correlations; and (k) Adaptive PCP plot highlighting the most significant dimensions. [10] 15

Figure 3.1 Components of DimVis: a panel for (a) choosing dataset and (b) adjusting the UMAP hyperparameters. (c) Shows the quality metrics. (d) Illustrates the lasso tool which is used to select data points interactively. (e) Shows the selected data points. (f) Demonstrates the unselected data points. (g) Depicts the EBM ranking graph for the selected data points; and (h) presents the Score of the selection, while (I) shows the distribution for one feature 20

Figure 3.2 Overall view of our proposed tool. (a) Demonstrates the dataset selection options in the dropdown. (b) Displays the hyperparameters adjustments. (c) Shows the quality metrics. (d) Illustrates the lasso tool which is used to select data points interactively. (e) Shows the selected data points. (f) Demonstrates the unselected data points. (g) Depicts the EBM ranking graph for the selected data points. 22

Figure 3.3 Individual feature assessment graph. (a) Shows the selected data points against the unselected ones in the dataset. (b) Displays the selected feature from EBM feature ranking graph. (c) Demonstrates the density of the data points in the whole dataset. (d) Illustrates the distribution of the data points across the whole dataset and also depicts the value range for the clicked feature from EBM feature ranking graph. The light gray background surrounding the linear chart in Score graph represents the uncertainty of the model's prediction for the clicked feature and as this background goes wider, the uncertainty increases. 23

Figure 3.4 Pairwise Feature Assessment Graph. (a) Illustrates the value range for Insulin feature (b) Shows the value range for Glucose feature. (c) Displays the color scale for the intensity of the interaction among these two features. 24

Figure 4.1 Visualization of the three data clusters generated by UMAP (left) alongside a Feature Importance Graph displaying the most significant features as determined by the EBM model trained on the entire dataset (right). Using the lasso tool to select each cluster allows us to observe the changes in the EBM's Feature Importance Graph. 31

Figure 4.2 Selecting C1 retrains the EBM model to highlight the differences between C1 and the other clusters. The top features are: *barenuc*, *sizeun*, *nornuc*, *clumpthic*, and *shapeun*. 32

Figure 4.3 Individual Feature Analysis for *barenuc* in Cluster C1. The Score graph shows the *barenuc* values for data points within C1 compared to those outside the cluster. The Density graph displays the overall distribution of *barenuc* across the entire dataset. 33
Figure 4.4 Individual Feature Analysis for *barenuc* in Cluster C1. The Score graph shows the *barenuc* values for data points within C1 compared to those outside the cluster. The Density graph displays the overall distribution of *barenuc* across the entire dataset. 33
Figure 4.5 Pairwise Feature Graph for the *nornuc* and *sizeun* Combination in Cluster C1. The score graph illustrates the combined impact of these two features on the selected data points. While *nornuc* alone is not among the top six features in the Feature Importance graph, its combination with *sizeun* ranks fourth. . . 34
Figure 4.6 Individual Feature Graph for *clumpthic* in Cluster C1. Initially, the graph displays a linear trend, but for the selected data points, it highlights very high *clumpthic* values. Additionally, the Density graph reveals an almost uniform distribution of this feature across the entire dataset. 34
Figure 4.7 Individual Feature Graph for *shapeun* in Cluster C1. The Score graph displays an almost linear trend for the *shapeun* feature within C1. Additionally, the Density graph indicates a higher concentration of *shapeun* values in the 1 to 1.6 range. 35
Figure 4.8 Selecting C2 retrains the EBM model to highlight the differences between C2 and the other clusters. The top features identified are *clumpthic*, *shapeun*, *sizeun*, and *barenuc*. 35
Figure 4.9 Individual Feature Graph for *clumpthic* in Cluster C2. The Score graph displays the *clumpthic* values for the selected cluster, C2. Additionally, the Density graph reveals the same distribution of *clumpthic* as observed in C1. 36
Figure 4.10 Individual Feature Graph for *shapeun* in Cluster C2. The Score graph represents the second most important feature and indicates high uncertainty for the selected data points. Additionally, the Density graph shows the distribution of *shapeun* values across the entire dataset. 36
Figure 4.11 Individual Feature Graph for *sizeun* in Cluster C2. The Score graph displays a consistent linear trend. The Density graph remains similar to the overall distribution of the *sizeun* feature. . 37

Figure 4.12 Individual Feature Graph for *barenuc* in Cluster C2. The Score graph reveals high uncertainty for the unselected data points. Additionally, the Density graph illustrates the distribution of *barenuc* values across the entire dataset. 37
Figure 4.13 Selecting C3 retrains the EBM model to highlight the differences between C3 and the other clusters. The most important features are *sizeun*, *blandchr*, *marg_adh*, and *sizeun* . 38
Figure 4.14 Pairwise Feature Graph for the combination of *sizeun* and *blandchr* in Cluster C3. The Score graph illustrates the combined effect of these two features on the selected data points, with their intersection representing our selection. 38
Figure 4.15 Individual Feature Graph for *blandchr* in Cluster C3. The Score graph displays an almost linear trend with low uncertainty for both selected and unselected data points. Additionally, the Density graph indicates a nearly uniform distribution of *blandchr* values across the entire dataset. 39
Figure 4.16 Pairwise Feature Graph for the combination of *marg_adh* and *blandchr* in Cluster C3. This combinatorial graph shows the intersection of both features, highlighting low feature values corresponding to our selection. 39
Figure 4.17 Individual Feature Graph for *sizeun* in Cluster C3. The Score graph exhibits a similar pattern to other Individual Feature Graphs for clusters C2 and C3. The Density graph illustrates the distribution of *sizeun* across the dataset. 39
Figure 4.18 This figure displays a projection of the breast cancer dataset with hyperparameters set to 12 for Number of Neighbors and 0.3 for Minimum Distance. It is used to validate our findings, showing the same three clusters as before, now color-coded with the ground-truth labels. 41
Figure 5.1 Illustrates the Pima Indian Diabetes dataset before any selection. (a) Displays the dropdown menu for selecting the dataset. (b) Illustrates the hyperparameter settings. (c) Shows the Precision and Recall values for the EBM classifier trained on the entire dataset. (d) Provides an overview of UMAP clusters, including data clusters and their corresponding class labels. (e) Demonstrates the features sorted by the EBM's importance ranking. (f) Shows the selected features dynamically presented to enhance user understanding and interpretation. 43

Figure 5.2 As illustrated in the figure, C1 is selected as the first cluster and it is depicted that the Feature Importance Ranking Graph has changed significantly. The first four important features in the EBM graph are highly important rather than the rest of them. . . 45

Figure 5.3 As displayed in the figure, C2 is selected cluster. Again, the feature importance graph has changed remarkably and the ranking is different. For this selection, *Insulin*, *Skin Thickness*, *Skin Thickness and Insulin* and *Insulin and DPF* are the most important four features. 46

Figure 5.4 Depicted the figure, C3 is the selection. For this new selection, the feature importance graph has changed again notably and the most four important features are: *Insulin*, *Skin Thickness*, *Blood Pressure*, and *Age*, respectively. The is a huge gap between *Insulin* and other features for this selection also. 46

Figure 5.5 Cluster C4 is selected in this figure. The feature rankings has been changed once more and the most important feature is *Insulin* again which is remarkably distinguished among the rest of the features in the EBM plot. 47

Figure 5.6 The selected cluster C5, is subcluster of a another cluster named C4. This subcluster is the left half of the cluster C4. These clusters are examined separately to find the reason behind their division into to separate clusters. On the other hand, for this selection, the most two important features have extremely different importance values from the rest of the feature in the provided feature ranking list. 47

Figure 5.7 The selected cluster C6 is the right part of the larger cluster C4. In the previous figure, Sara explored the other subcluster, and with this selection, she investigates the C6 subcluster. For this selection, the first two important features are the same as in the other subcluster; however, their order is reversed. 48

Figure 5.8 As illustrated in the figure, *Insulin* Individual Feature Graph for C6 is displayed. It shows very low values for the selection and also a linear trend for the rest of the datapoints. 49

Figure 5.9 The Individual Feature Graph of *Skin Thickness* for C6 is shown in this figure. It has an increasing trend at first which is followed by a linear trend. 49

Figure 5.10 The Individual Feature Graph of *Insulin* for C5 is displayed in this figure. It shows very low values for the selection and also a linear trend for the rest of the datapoints. 50

Figure 5.11 The Individual Feature Graph of *Skin Thickness* for C5 is displayed in this figure. This is the second important feature for C5 selection and has wide intervals of uncertainty for selected data points. 50
Figure 5.12 The Individual Feature Graph of *Insulin* for C4 is illustrated in this figure. It displays the *Insulin* trend for this selection. The value of *Insulin* is so low for the selected datapoints and then it demonstrates a linear trend for the rest of the dataset. This graphs shows very similar trend to the *Insulin* graph of C5 and C6 selection which is another way that supports Sara's hypothesis about the feature which gathered those two subclusters and shaped a bigger one. 51
Figure 5.13 The Individual Feature Graph of *Insulin* for C3 is illustrated in this figure. It displays the Insulin trend for this selection. The selected cluster is so small and that is why the area that corresponds to the selection is so small. 51
Figure 5.14 The Individual Feature Graph of *Insulin* for C2 is illustrated in this figure. It shows a decreasing trend at first, then a linear trend, after that it displays an increasing and again a linear trend. The selection C2 corresponds to Insulin values ≈ 20 up to ≈ 420. Any section out of this interval belongs to the rest of the dataset. 52
Figure 5.15 The Individual Feature Graph of *Insulin* for C1 is illustrated in this figure. The selection covers from *Insulin* values ≈ 430 up to the end and the model shows a high uncertainty for the whole dataset prediction. 52
Figure 6.1 Displays the Indian Pines dataset before any feature reduction. This initial visualization highlights the complexity and overlapping different class labels of the data, underscoring the need for effective dimensionality reduction techniques to simplify the analysis and improve the clarity of the resulting clusters. The original dataset includes hyperspectral images with ground-truth labels for 16 different classes, which will be crucial in validating the effectiveness of the reduction techniques and subsequent clustering. 57
Figure 6.2 Displays the result of applying MNF to the Indian Pines dataset. The transformed data shows a reduction in noise, making the underlying patterns more discernible. 58

Figure 6.3 Shows the result of applying PCA to the Indian Pines dataset. The first few principal components capture most of the variance, providing a simplified yet informative representation of the data. 59
Figure 6.4 Presents the result of applying MNF coupled with Bilateral to the Indian Pines dataset. The data exhibits enhanced signal clarity with well-preserved edges, making it suitable for subsequent clustering analysis. 60
Figure 6.5 Illustrates the result of applying PCA coupled with Bilateral to the Indian Pines dataset. The data retains significant spectral information while minimizing noise, providing a balanced and clear representation for analysis. 61
Figure 6.6 Comparison of UMAP results applied to the Indian Pines dataset after each feature reduction technique has been applied. . 62
Figure 6.7 Depicts the six data clusters generated by UMAP (left) and a Feature Importance Graph for the most important features as identified by the EBM model trained on the entire dataset (right). By selecting each of the clusters using the lasso tool, we can observe the changes in EBM's Feature Importance Graph. 63
Figure 6.8 Selecting C1 retrains the EBM model to highlight the differences between the datapoints of cluster C1 and the rest of the dataset. The top features identified are *band 5* and *band 8*. 64
Figure 6.9 Individual Feature Graph for *5th* band of the hyperspectral image dataset in our first cluster which is C1. Showing very intense effect of this feature on the selected datapoints. Additionally, a Density graph is also displayed in this figure which shows a higher population for *5th* band between 0.4 to 0.6 value range. 64
Figure 6.10 Ground truth of the Indian Pines dataset with class labels. Each color represents a different class, as indicated by the numbers on the image. This visual representation provides the actual class distribution and locations of the data points, serving as a reference for comparing and validating the clustering results. . . 65
Figure 6.11 *5th* band of the Indian Pines hyperspectral dataset. This grayscale image represents one spectral band of the hyperspectral data, highlighting the variations in reflectance across the different areas of the scene. This band is used for detailed analysis and comparison with the ground truth to identify and interpret specific features within the dataset for the selected cluster C1. 66

Figure 6.12 Comparison of the *5th* band of the Indian Pines hyperspectral dataset with the ground truth image. The red rectangles highlight the specific areas in the *5th* band that correspond to the shapes that cover class label *6* seen in the ground truth image. This comparison is used to validate the intensity effect values for our selected points, which are particularly significant due to their low values and intense effects. Other shapes may be visible in this band, but the highlighted areas emphasize the importance of our selected points based on their intensity color. 67

Figure 6.13 Individual Feature Graph for *8th* band of the hyperspectral image dataset in our first cluster which is C1. Showing moderate effect of this feature on the selected datapoints. Additionally, a Density graph is also displayed in this figure which shows a higher population for *5th* band between 0.5 to 0.7 value range. 67

Figure 6.14 Comparison of the *8th* band of the Indian Pines hyperspectral dataset with the ground truth image. The red rectangles highlight the specific areas in the *8th* band that correspond to the shapes that cover class label *6* seen in the ground truth image. This comparison is used to validate the effect values for our selected points. Other shapes may be visible in this band, but the highlighted areas emphasize the importance of our selected points based on their intensity color. 68

Figure 6.15 Selecting C2 retrains the EBM model to highlight the differences between the datapoints of cluster C2 and the rest of the dataset. The top features identified are band *2* and band *1* respectively. 69

Figure 6.16 Individual Feature Graph for band *2* of the hyperspectral image dataset in our second cluster which is C2. Showing moderate effect of this feature on the selected datapoints. Additionally, a Density graph is also displayed in this figure which shows a higher population for *5th* band between 0.3 to 0.5 value range. . 69

Figure 6.17 Comparison of the *2nd* band of the Indian Pines hyperspectral dataset with the ground truth image. The red rectangles highlight the specific areas in the *2nd* band that correspond to the shapes that cover class label *8* seen in the ground truth image. This comparison is used to validate the effect values for our selected points. Other shapes may be visible in this band, but the highlighted areas emphasize the importance of our selected points based on their intensity color. 70

Figure 6.18 Individual Feature Graph for *1st* band of the hyperspectral image dataset in our first cluster which is C2. Showing very close score values to 0 which depicts that these leads do not have very intense effect but carry important factors for the selection between the values 0.5 and 0.65. Also, the Density graph shows almost equally distributed values among the dataset for this band. 71

Figure 6.19 Comparison of the band *1* of the Indian Pines hyperspectral dataset with the ground truth image. The red rectangle highlights the specific areas related to the feature values for class label *8* in the relevant selection and displays the comparison between the ground truth and the specified area. . . 71

Figure 6.20 Selecting C3 retrains the EBM model to highlight the differences between the datapoints of cluster C3 and the rest of the dataset. The top features identified are band *1* and band *11* respectively. There is a huge gap between the first two important features which demosntrates the importance difference between these two important features. 72

Figure 6.21 Individual Feature Graph for band *1* of the hyperspectral image dataset in C3. Showing almost intense effect of this feature on the selected datapoints. Additionally, the Density graph shows almost equally distributed values among the dataset for this band. 72

Figure 6.22 Comparison of the *1st* band of the Indian Pines hyperspectral dataset with the ground truth image. The red rectangles highlight the specific areas in the *1st* band that correspond to the shapes that cover class label *10* seen in the ground truth image. This comparison is used to validates our investigation. . 73

Figure 6.23 Individual Feature Graph for band *11* of the hyperspectral image dataset in C3. Showing very intense effect of this feature on some part of the selected datapoints. Additionally, the Density graph shows the distribution almost equally for selected and unselected datapoints. 73

Figure 6.24 Comparison of the *11th* band of the Indian Pines hyperspectral dataset with the ground truth image. The red rectangles highlight the specific areas in the *1st* band that correspond to the shapes that cover class label *10* seen in the ground truth image. This comparison is used to validate our hypothesis. . . . 74
Figure 6.25 Selecting C4 retrains the EBM model to highlight the differences between the datapoints of cluster C4 and the rest of the dataset. The top features identified are band *1* and band *8* respectively. 75
Figure 6.26 Individual Feature Graph for band *1* of the hyperspectral image dataset in C4. Showing a linear trend through the graph. Additionally, the Density graph shows a dense distribution around the end of the selection interval. 75
Figure 6.27 Comparison of the *1st* band of the Indian Pines hyperspectral dataset with the ground truth image. The red rectangles highlight the specific areas in the *1st* band that correspond to the shapes that cover class label *11* seen in the ground truth image. This comparison is used to validates our investigation. . 76
Figure 6.28 Individual Feature Graph for band *8* of the hyperspectral image dataset in C4. Showing a linear trend through the graph. Additionally, the Density graph shows a dense distribution around the selection interval. 76
Figure 6.29 Comparison of the band *8* of the Indian Pines hyperspectral dataset with the ground truth image. The red rectangle highlights the specific areas in the *8th* band that correspond to the shapes that cover class label *11* seen in the ground truth image. This comparison is used to validates our investigation. . 77
Figure 6.30 Selecting C5 retrains the EBM model to highlight the differences between the datapoints of cluster C5 and the rest of the dataset. The top features identified are band *11* and band *1* respectively. 78
Figure 6.31 Individual Feature Graph for band *11* of cluster C5 that shows a linear trend. Additionally, the Density graph shows a dense distribution around the beginning of the selection interval and the rest of the dataset. 78

Figure 6.32 Comparison of the *11th* band of the Indian Pines hyperspectral dataset with the ground truth image. The red rectangles and the red triangle highlight the specific areas in the image band *11* that correspond to the shapes that cover class label *12* seen in the ground truth image. This comparison is used to validates our investigation. 79
Figure 6.33 Individual Feature Graph for band *1* of cluster C5 that shows a linear trend. Additionally, the Density graph shows a dense distribution around the selection interval. 79
Figure 6.34 Comparison of the image band *1* of the Indian Pines hyperspectral dataset with the ground truth image. The red rectangles and the red triangle highlights the specific areas in the *1st* band that correspond to the shapes that cover class label *12* seen in the ground truth image. This comparison is used to validates our investigation. 80
Figure 6.35 Selecting C6 retrains the EBM model to highlight the differences between the datapoints of cluster C6 and the rest of the dataset. The top features identified are band *1* and band *6* respectively. 80
Figure 6.36 Individual Feature Graph for band *1* of cluster C6 that shows a linear trend until our selection and for the selection it demonstrates a decreasing trend until the end of the graph. Additionally, the Density graph shows an almost equally distributed datapoints in the dataset for this feature. 81
Figure 6.37 Comparison of the image band *1* of the Indian Pines hyperspectral dataset with the ground truth image. The red rectangles highlight the specific areas in the *1st* band that correspond to the shapes that cover class label *14* seen in the ground truth image. This comparison validates our investigation. 81
Figure 6.38 Individual Feature Graph for band *6* of cluster C6 that shows a linear trend for our selection and then an increase and finally again a linear trend for the unselected datapoints. Additionally, the Density graph shows a dense distribution around the end of our selection interval and the beginning of the unselected datapoints interval. 82

Figure 6.39 Comparison of the image band *6* of the Indian Pines hyperspectral dataset with the ground truth image. The red rectangles highlight the specific areas in the *6st* band that correspond to the shapes that cover class label *14* seen in the ground truth image. This comparison validates our investigation. 82

LIST OF TABLES

Table 2.1 Options comparison of our tool with the related proposed tools in the literature. 15

Table A.1 We utilized an Anaconda environment to maintain consistent package management and dependencies. The following table lists the package names along with their respective versions. . . . 93

ABSTRACT

Interactive Analysis and Data Visualization for High Dimensional Datasets

Parisa SALMANIAN

Department of Computer Engineering
Master of Science Thesis

Supervisor: Assoc. Prof. Ali Can KARACA

The complexity of high-dimensional data poses a considerable challenge for researchers and data scientists in Analyzing, interpreting, and understanding these data. Techniques such as t-SNE and UMAP have long been favorites for dimensionality reduction which offer tools to visualize the high-dimensional data in lower dimensions. Aforementioned methods transform complex datasets into a more interpretable and comprehensible form by either preserving the local structure of the data or highlighting and spotlighting the most significant features. This approach facilitates data visualization and thus makes the analysis more understandable. However, interpreting the visualizations produced by these methods remains a noteworthy challenge. accordingly, there is a growing need for more interpretable and understandable data visualizations. This necessity goes beyond convenience, aiming to enhance the interpretability of visualizations in data-driven research.

Keywords: Visualization, interaction, projection interpretability, UMAP, dimensionality reduction, EBM-based explanations, machine learning, AI, hyperspectral images.

YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

ÖZET

Yüksek Boyutlu Veriler için Etkileşimli Analizi ve Veri Görselleştirmesi

Parisa SALMANIAN

Bilgisayar Mühendisliği Anabilim Dalı
Yüksek Lisans Tezi

Danışman: Doç. Dr. Ali Can KARACA

Yüksek boyutlu verilerin karmaşıklığı, bu verileri analiz etmek, yorumlamak ve anlamak konusunda araştırmacılar ve data uzmanları için büyük bir zorluk teşkil etmektedir. t-SNE ve UMAP gibi teknikler, uzun zamandır boyut indirgeme için favori olarak kabul edilen ve yüksek boyutlu verileri daha düşük boyutlarda görselleştirmek için araçlar sunan yöntemlerdir. Bahsi geçen yöntemler, verilerin yerel yapısını koruyarak veya en önemli özellikleri vurgulayarak karmaşık veri setlerini daha anlaşılır ve yorumlanabilir bir forma dönüştürmektedir. Bu yaklaşım, veri görselleştirmesini kolaylaştırır ve böylece analizi daha anlaşılır hale getirir. Ancak, bu yöntemlerle üretilen görselleştirmeleri yorumlamak hala önemli bir zorluk olarak kalmaktadır. Bu nedenle, daha yorumlanabilir ve anlaşılabilir veri görselleştirmelerine olan ihtiyaç giderek artmaktadır. Bu gereklilik, araştırmacılara kolaylık sağlamakla birlikte veri odaklı araştırmalarda görselleştirmelerin yorumlanabilirliğini artırmayı hedefler.

Anahtar Kelimeler: Görselleştirme, etkileşim, projeksiyon yorumlanabilirliği, UMAP, boyut indirgeme, EBM tabanlı açıklamalar, makine öğrenmesi, yapay zeka, hiperspektral görüntüler.

YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

1
INTRODUCTION

1.1 Dimensionality Reduction Techniques

Dimensionality reduction (DR) encompasses a variety of techniques designed to simplify the visualization of high-dimensional data in the field of data analysis. These techniques convert complex and challenging high-dimensional data into a lower-dimensional form which facilitates the simplification of datasets, enhances computational efficiency, and retains the most critical information for analysis. DR allows data scientists and researchers to interpret and explore vast amounts of data more easily and accurately. Therefore, deeper insights across different fields becomes possible.

While traditional and linear DR methods like Principal Component Analysis (PCA) [1] and Multidimensional Scaling (MDS) [2] have been widely used due to their straightforward interpretation and ease of use, there has been a rise in nonlinear DR techniques such as t-distributed Stochastic Neighbor Embedding (t-SNE) [3] and Uniform Manifold Approximation and Projection (UMAP) [4]. These modern methods can recognize and highlight more convoluted patterns in complex datasets, thereby affording them an advantage over linear methods.

The choice between UMAP and t-SNE depends on how they apply dimensionality reduction. t-SNE excels at preserving local structures within data, making it effective for visualizing compact and dense clusters of highly correlated and neighboring data points. However, this approach can sometimes result in the loss of global relationships and is computationally intensive, which limits its scalability for large datasets. In contrast, UMAP balances local and global structures, and by this, it simplifies data analysis and offers faster and more scalable performance. Not only DR techniques address the challenge of interpreting today's complex datasets, but also they handle the explanation of modern and convoluted data analysis pipelines, such as machine learning models. Significant studies in this area include the works of DeGroat et al. [5] and Ru-tao et al. [6]. Methods such as deep

neural networks, random forests, and ensemble models have grown increasingly intricate to manage large-scale and high-dimensional datasets. These sophisticated techniques, while powerful and capable of achieving high accuracy, present several notable challenges. Firstly, the complexity of deep neural networks often results in them being perceived as "black boxes," where the internal workings are not transparent or easily understood. This opacity makes it difficult to comprehend, debug, and control their processes. Similarly, random forests, which aggregate the results of numerous decision trees, and ensemble models, which combine the predictions of multiple algorithms, also suffer from a lack of interpretability. The aggregation process in these models obscures the contribution of individual features, making the outcomes difficult to trust and validate.

The challenge of interpretability is critical in fields where understanding the decision-making process is as important as the prediction itself, such as in healthcare, finance, and legal applications. Consequently, Interpretable Machine Learning (IML) and Explainable AI (XAI) have become prominent research topics. Among recent innovations in this era, the Explainable Boosting Machine (EBM) [7] stands out. EBM is a tree-based, cyclic gradient boosting Generalized Additive Model with automatic interaction detection. It excels at addressing the critical balance between model accuracy and comprehensibility. Unlike black-box models, EBM constructs models progressively, one feature at a time, ensuring both precision and clarity. This transparent method makes the decision-making process of the model easier to understand and comprehend, fostering trust and making it an ideal choice in cases where clear explanations are essential.

EBM is an IML framework that merges the power of gradient boosting with systematic methods to produce highly precise predictive models. In contrast to conventional black-box models, EBM constructs additive models where each feature or combination of features is considered an independent additive input. This approach ensures that each feature's contribution is clear and interpretable, providing insights into the model's decision-making process.

Implementing EBM provides a graphical list of ranked features from the most to the least crucial. Additionally, it enables the individual display of each feature upon selection, offering an opportunity for deeper understanding and interpretation of each feature. EBM offers various options and alternatives to enhance transparency and clarity when prioritizing these features. This ability to visualize and interpret the significance of each feature not only enhances trust in the model but also empowers users to make better-informed decisions based on the model's outputs.

1.2 Problem Formulation

Dimensionality reduction (DR) techniques such as t-SNE and UMAP have been popular in recent years because of their effective process in managing complex patterns in high-dimensional data, outperforming linear methods like PCA. Even though they are highly advantageous over linear techniques, these nonlinear methods pose significant interpretation challenges. One major issue is the lack of easily interpretable axes in visualization graphs produced by t-SNE and UMAP. While linear models often have axes that correspond to specific attributes or combinations thereof, nonlinear models generate visual outputs without inherent explanations. Hence, while t-SNE and UMAP are effective at identifying and recognizing clusters within datasets, they do not inherently offer insights into the unique characteristics or the formation processes of these clusters. This results in the challenge of interpreting, exploring and understanding the identified clusters to determine what makes each of them unique and what are the most important charactirisrics that have gathered them together as a cluster, as there is no widely accepted method for extracting meaningful insights from these complex visualizations.

In summary, this thesis aims to develop a tool to investigate and interpret the clusters and the factors influencing their formation recognizes by nonlinear DR methods. This work introduces HyEx, which utilizes UMAP to create dimensionality reduction visualizations and then, it employs Explainable Boosting Machines (EBM) as an interpretative step. The specifics of this approach will be elaborated upon in the subsequent sections.

1.3 Research Goals

Drawing inspiration from the work of Bibal et al. [8] and the existing researches and literature on data interpretation and analysis using dimensionality reduction (DR) techniques [9] and [10], we introduce a tool designed to address the problem of interpretibility of the visualizations provided by nonliear DR methods , as outlined in the previous section. This tool facilitates interactive data assessment by combining UMAP and EBM and also leverages UMAP's two most important hyperparameters.

Building on the foundational studies mentioned, our tool adopts an enhanced approach to elucidate the relationship between patterns in the data visualized using the UMAP algorithm in Low Dimensional (LD) spaces and their corresponding High Dimensional (HD) forms. The proposed tool aims to improve the

interpretability and understandibility of the visualized data.

To solve the aforementioned problem and accomplish this goal, we integrate the EBM technique into the interactive exploration of DR visualizations. The core concept is that the patterns recognized and identified in the dataset have unique and hidden characteristics within the distribution of their data points and features. A pattern is considered to be a group of points that visually stands out from the rest of the dataset, forming a distinct visual feature in the displayed visualization. Common patterns include dense clusters of points separated by white space. Our approach does not rely on a formal definition of patterns, as they are determined and selected by users through interactive visualization rather than automatic extraction.

By learning to separate and isolate these patterns based on the features of the dataset, we can interpret this isolation and understand the shaped pattern itself. To delve deeper, we train a supervised machine learning classifier in real-time to differentiate a user-selected pattern from the rest of the dataset. We use the EBM technique for this task due to its inherent explainability. Once the EBM model accurately isolates the pattern, it provides an explanation of the most influential features and their values as a ranking graph, offering insights into the pattern itself.

This tool is integrated into a web application that offers a menu with options for dataset selection, hyperparameter adjustments, and displays for precision and recall. Additionally, it features three graphs: two displayed initially and a third that appears when a specific feature is selected for deeper assessment.

In summary, our contributions include:

- A new pipeline which proposes interactive data analysis utilizing non-linear DR with IML to explore and investigate patterns in high-dimensional data.
- The implementation of this pipeline as a user-friendly, web-based interactive tool using UMAP and EBM.
- A detailed application of this tool on three case studies from the 'Wisconsin Breast Cancer', 'Pima Indian Diabetes' and 'Indian Pines' datasets, illustrating the reasoning behind the visualizations created by UMAP through interactive cluster selection and feature inspection. These case studies validate the effectiveness of DimVis and HyEx and demonstrate its capability to uncover patterns in complex real-world datasets.

To conclude, this thesis presents a new approach to interactive data visualization, addressing gaps left by previous works in this field such as t-SNE and decision

trees. The results from the two case studies highlight our tool's ability to enhance the interpretability and understanding of complex and high-dimensional datasets and suggest a promising direction for future improvements and works.

1.4 Document Structure

The structure of this Master's thesis is as follows:

Section 2 covers the background and related work. It aims to provide an overview of the technologies employed in the tool.

Section 3 introduces DimVis, detailing the design, functionalities, goals, visualization capabilities of the tool and outlines the algorithms and methodologies utilized in the development of the tool.

Section 4 presents the first use case that demonstrates the practical application of DimVis in real-world contexts and discusses the tool's advantages.

Section 5 presents another distinct use case that depicts the practical application of DimVis in real-world health domain and shows the tools capabilities in real-world different datasets.

Section 6 illustrates a totally different use case that shows the practical application of HyEx in a very different domain which is hyperspectral image dataset.

Section 7 discusses the limitations, user feedback, and potential future work of the tool.

Finally, Section 8 summarizes the entire process, highlights key achievements, and discusses expectations for future developments of the tool.

1.5 Ethics

Ethical values and considerations were of the greatest importance throughout the development and implementation of DimVis and HyEx. Since both tool aim to help users and enhance their comprehension of complex datasets, it is essential that the data processed and visualized respects user privacy and security. The datasets utilized in our applications are publicly accessible and have been widely used in previous research. Furthermore, no personally identifiable information was included in the case studies, ensuring that the insights derived from our research do not compromise the integrity, privacy, or security of the original data. Additionally,

DimVis and HyEx guarantee that the interpretations and insights generated from the visualizations are conducted ethically and without bias, providing transparent views into the datasets and their relationships as it always should be.

2
BACKGROUND AND RELATED WORK

2.1 Background

Dimensionality reduction is a set of crucial techniques in data analysis and machine learning. The primary objective of DR is to decrease the number of input features or dimensions in a dataset while retaining essential and crucial information such as distances, neighborhoods, and clusters. This transformation converts high-dimensional data into a more manageable lower-dimensional representation. DR techniques are particularly valuable when working with complex datasets, where the high number of dimensions complicates data processing and interpretation. By using mathematical procedures, DR aims to simplify the data while preserving its relevant structures and vital data as much as possible to make data analysis more comprehensible and understandable. Several traditional DR techniques have been widely used, each with its strengths and limitations:

Principal Component Analysis (PCA): PCA is one of the most well-known DR techniques. It works by identifying the principal components that capture the maximum variance in the data. PCA is efficient and straightforward but has limitations. It assumes linear relationships between features, which can be restrictive for more complex data structures. Moreover, PCA can struggle with noise and outliers, potentially leading to misleading representations of the data [1].

Multidimensional Scaling (MDS): MDS is a technique that aims to place each data point in low-dimensional space while preserving the pairwise distances between points as much as possible. While MDS can be effective, it is computationally expensive and can be sensitive to noise and outliers, similar to PCA. Moreover, it does not scale well with very large datasets [2].

t-Distributed Stochastic Neighbor Embedding (t-SNE): t-SNE is another popular DR technique, particularly known for its ability to create visually interpretable

two-dimensional maps from high-dimensional data. It focuses on preserving the local structure of the data by converting high-dimensional Euclidean distances into conditional probabilities representing similarities. However, t-SNE has several drawbacks. It is computationally intensive, making it unsuitable for large datasets. It also suffers from the "crowding problem," where significant clusters may be squeezed together, making it difficult to interpret the global structure of the data [3].

2.1.1 Why Uniform Manifold Approximation and Projection (UMAP)?

Uniform Manifold Approximation and Projection (UMAP) has emerged as one of the leading techniques in dimensionality reduction (DR). UMAP is designed to preserve the inherent structure of the data, reflecting the original arrangement and relationships between data points. Despite traditional methods that often focus on maintaining and preserving distances between points, UMAP intends to capture the underlying manifold structure of the data, drawing on principles from Riemannian geometry and algebraic topology [4]. UMAP operates on the assumption that data is uniformly distributed on a Riemannian manifold. It also leverages the topological structure of the data to accurately depict it in lower dimensions.

To achieve dimensionality reduction, UMAP constructs a high-dimensional graph to identify the nearest neighbors for each data point and weights the edges based on the local metric in the high-dimensional space. It then creates a low-dimensional graph that approximates the high-dimensional graph as closely as possible, preserving both local and global data structures (See Figure 2.1) [11].

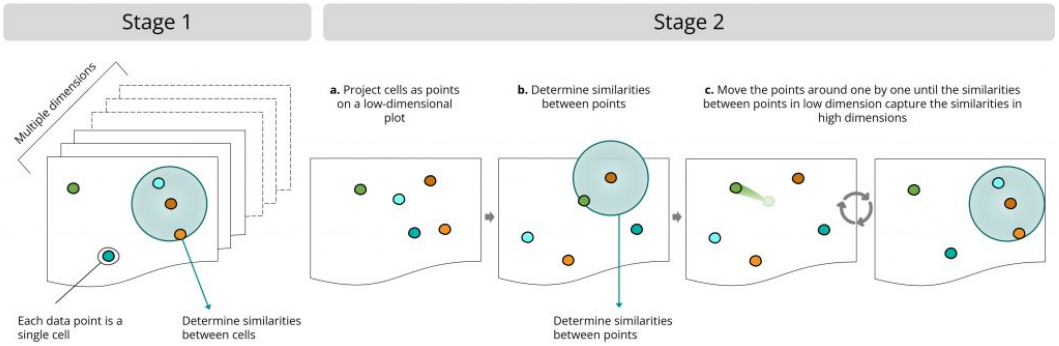


Figure 2.1 Illustration of UMAP Working Process. Stage 1 (left) involves projecting cells as points on a multi-dimensional plot and determining similarities between cells in multiple dimensions. Stage 2 (right) involves refining the plot by moving points until the similarities in the low-dimensional plot capture the similarities in high dimensions, thereby accurately representing the relationships between data points. [12]

UMAP offers several advantages. One of these advantages is its high adaptability, making it suitable for both tabular and complex structured data types. Another benefit is its foundation in algebraic topology, which provides robust representations and avoids common issues like the "crowding problem" seen in methods such as t-SNE. Additionally, UMAP is computationally efficient, making it well-suited for larger datasets [13].

UMAP's flexibility and effectiveness have been demonstrated in a wide range of applications across different areas. In the field of bioinformatics, UMAP has been used for visualizing high-dimensional single-cell RNA sequencing data, providing clear, interpretable and understandable clustering results that facilitate the identification of cell types and states [14]. In genomics, UMAP has been influential in analyzing complex gene expression data, enabling researchers to uncover and reveal meaningful patterns and relationships among genes [4].

UMAP has demonstrated its utility in various fields, particularly in image processing and computer vision. It has been employed for feature extraction and dimensionality reduction in large image datasets, significantly enhancing the performance of downstream tasks such as image classification and object detection [15]. This usefulness is further exemplified in our third use case, which involves a large image dataset. Additionally, UMAP has been adopted in natural language processing (NLP) to visualize word embeddings and document representations, facilitating the understanding of semantic relationships and topic modeling [16].

These examples underscore UMAP's high adaptability, usefulness and computational efficiency, making it a preferred choice for handling complex and large-scale datasets which is very important factor for our tool. Its ability to preserve both local and global structures in the data while providing clear and interpretable visualizations has made UMAP an invaluable tool in various research and application areas, including image processing, computer vision, and natural language processing. By presenting the limitations of other DR techniques, the strengths of UMAP become more apparent. This highlights its suitability for your research and justifies its selection as the primary method for dimensionality reduction in this study.

2.1.2 Interpretable Machine Learning

Interpretable Machine Learning (IML) is a field within machine learning and related areas that focuses on creating decision-making models that are transparent, comprehensible, and understandable to human experts which is illustrated in

Figure 2.2.

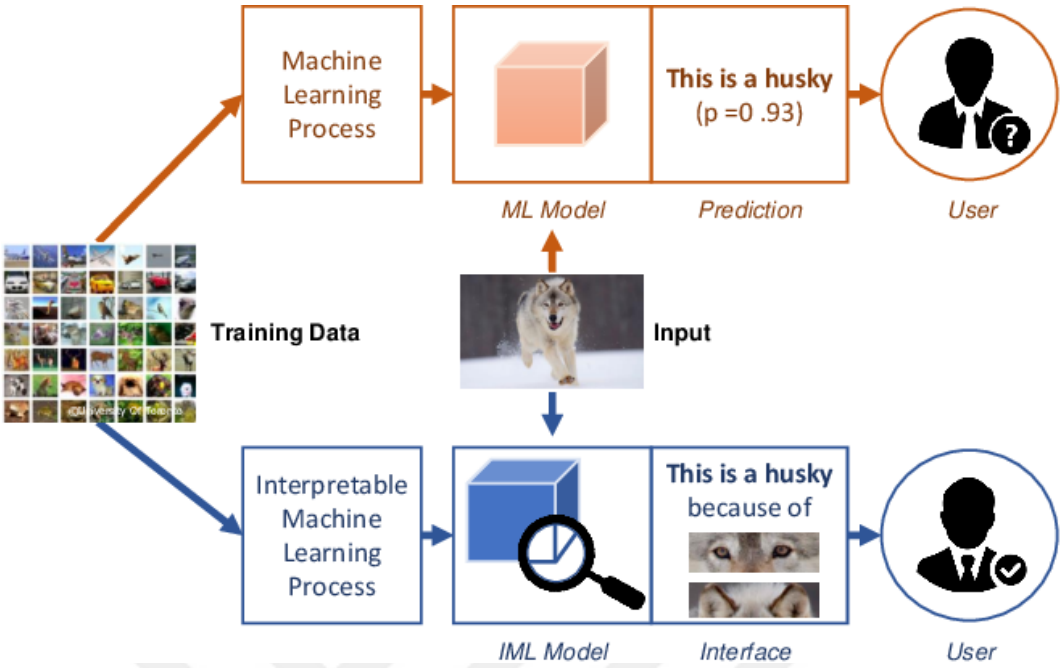


Figure 2.2 A comparison between traditional Machine Learning (ML) and Interpretable Machine Learning (IML). The upper part of the figure illustrates the ML process, where a model makes a prediction (e.g., identifying a husky) without providing the reasoning behind it, leaving the user uncertain. The lower part shows the IML process, where an interpretable model not only makes the prediction but also provides an explanation for the decision (e.g., highlighting specific features that led to the identification of the husky which is one of the features of glassbox models), enhancing user understanding and trust. [17]

IML ensures that the internal workings of algorithms are clear and interpretable, eliminating "black box" models. This transparency is essential in critical areas such as healthcare and finance, where opaque decision-making processes can lead to severe consequences as shown in Figure 2.3. The importance of IML has led to the development of various methods that balance prediction accuracy with human interpretability.

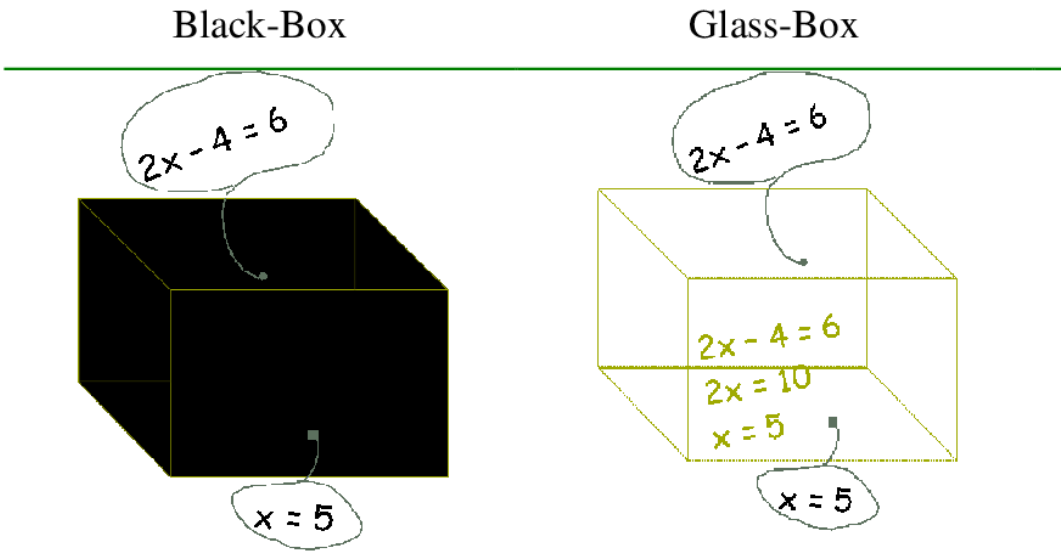


Figure 2.3 A comparison between Black-Box and Glass-Box models. The Black-Box model (left) provides a solution without showing the steps involved, making the process opaque and hard to understand. In contrast, the Glass-Box model (right) transparently displays each step of the calculation, making it clear how the solution is derived and enhancing interpretability and trust. [18]

2.1.3 Explainable Boosting Machine (EBM)

The Explainable Boosting Machine (EBM) is an approach within Interpretable Machine Learning (IML) that combines the transparency and interpretability of machine learning models with the methodologies of boosted tree algorithms (See Figure 2.4). EBM excels at addressing the critical balance between model accuracy and comprehensibility. In contrast to advanced models that often operate as black boxes, EBM constructs models progressively, one feature at a time, ensuring both precision and clarity. This transparent method makes the decision-making process of the model easier to understand and comprehend, which fosters trust and makes it an ideal choice in cases where clear explanations are essential.

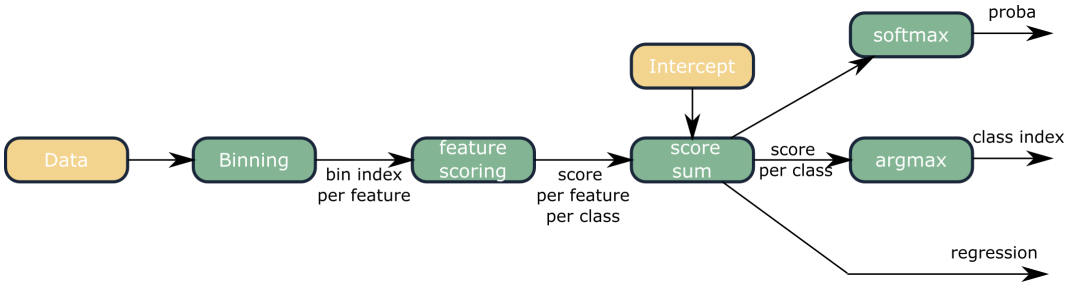


Figure 2.4 Steps involved in the Explainable Boosting Machine (EBM) interface. The process begins with the input data, followed by binning of features, scoring of feature bins, adding an intercept, summing the scores, and applying softmax and argmax functions to generate the final prediction. This sequence ensures model accuracy while maintaining interpretability. [19]

Additionally, EBM is a tree-based, cyclic gradient boosting Generalized Additive Model (GAM) with automatic interaction detection. EBMs are often as accurate as state-of-the-art black-box models while remaining completely interpretable. This combination of high accuracy and full interpretability makes EBMs particularly valuable in scenarios where understanding the model's predictions is crucial. See Figure 2.5

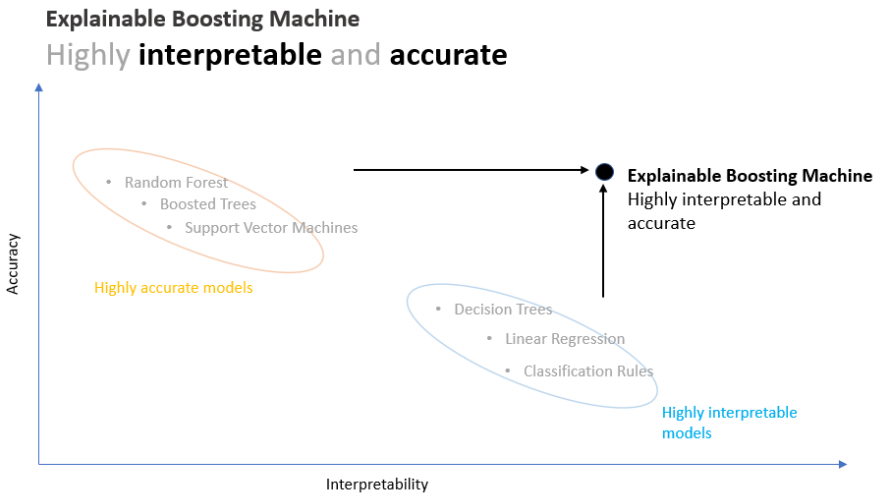


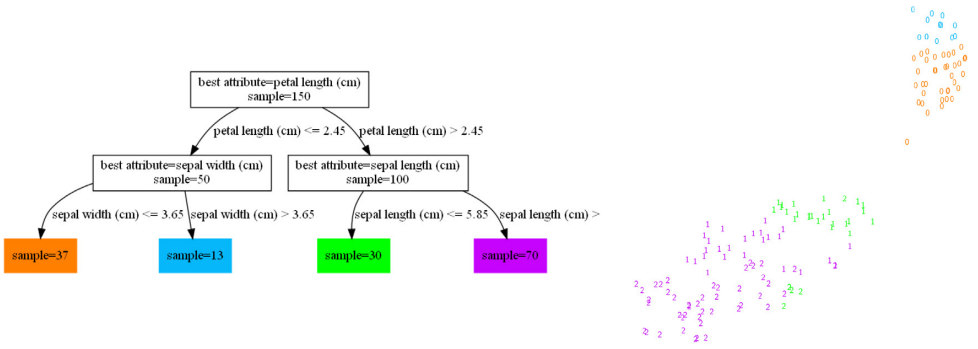
Figure 2.5 Comparison of model interpretability and accuracy, highlighting the Explainable Boosting Machine (EBM) as a model that achieves both high interpretability and high accuracy. While traditional models like Decision Trees, Linear Regression, and Classification Rules are highly interpretable but less accurate, and models like Random Forests, Boosted Trees, and Support Vector Machines are highly accurate but less interpretable, EBM balances both attributes effectively. [20]

2.2 Related Work

In this section, we provide an overview of similar researches aimed at addressing data interpretation after the application of dimensionality reduction (DR), and compare these techniques to the functionality of our tool. The comparison is based on several criteria listed below: whether the tool employs the t-SNE algorithm; whether it uses the UMAP algorithm; the extent of user interaction, including options for dataset selection, data manipulation, selection of subsets or specific data points, and adjustment of visualization parameters; the use of ranked dimensions, from the most important to least, to highlight and organize data features by importance; the tool's capacity for isolated feature exploration to enhance understanding of the data and identified clusters; and finally, providing quality metrics such as precision/recall to assess accuracy and reliability. Over the past few years, several tools have been proposed that incorporate DR techniques and attempt to interpret the data post-DR. In the following section, we will examine their objectives, strengths, and limitations.

2.2.1 Goals and Shortcomings of Similar Works

DT-SNE [9] integrates decision trees with t-SNE visualizations to improve data comprehensibility by combining the methodologies of both techniques. However, this approach has several limitations. For instance, it does not support user engagement and lacks interactive capabilities, leading to less intuitive data interpretation. Additionally, the combination of decision trees with t-SNE is not pleasing for larger datasets. Please see Figure 2.6.



(a) Results using DT-SNE with a perplexity of 5.

(b) Visualization obtained using *t*-SNE with a perplexity of 50 and colored w.r.t. the leaves in (a). The labels 0 (Iris Setosa) , 1 (Iris Versicolour) and 2 (Iris Virginica) are also provided.

Figure 2.6 Comparison of visualizations on the Iris dataset using DT-SNE and t-SNE methods in DT-SNE paper. [9]

IXVC [8] also utilizes t-SNE for dimensionality reduction to provide visualizations structured through decision trees, similar to DT-SNE. The key difference is that

IXVC proposes user interaction, enhancing data perception through the simplicity of decision trees. Despite this, it still faces the inherent limitations of decision trees when dealing with complex datasets. Decision trees can be too simplistic for explaining intricate nonlinear patterns and tend to fail when working with more complex datasets. See the interface of IXVC in Figure 2.7.

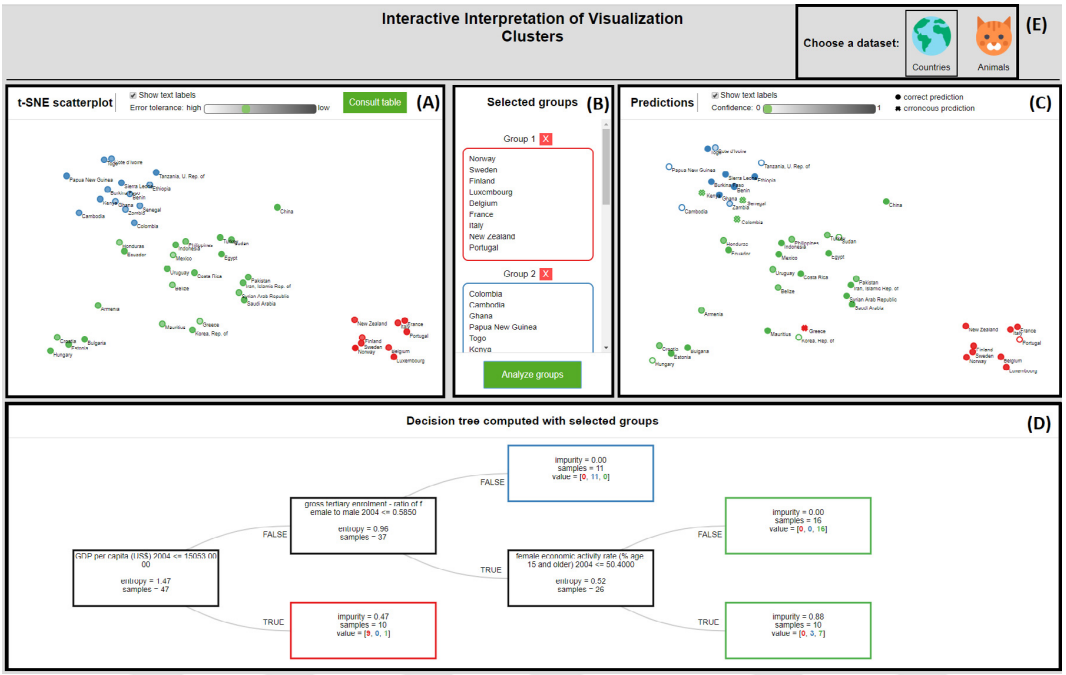


Figure 2.7 IXVC interface overview [8]. The top left scatter plot (A) represents the DR visualization, with instances color-coded based on the cluster selection displayed in the top middle section (B). The decision tree (D) explaining the selected clusters is shown at the bottom. The tree's prediction colors are visualized in the top right section (C). Users can toggle between the country and zoo datasets using the earth and cat icons (E) for evaluation purposes (refer to Section 6).

t-viSNE [10] provides an interactive platform for evaluating t-SNE visualizations. While it supports users in better understanding t-SNE results and visualizations, its focus only on t-SNE and lack of built-in features for interpreting clusters (beyond shapes) are significant drawbacks. Interface of t-viSNE is provided in Figure 2.8.

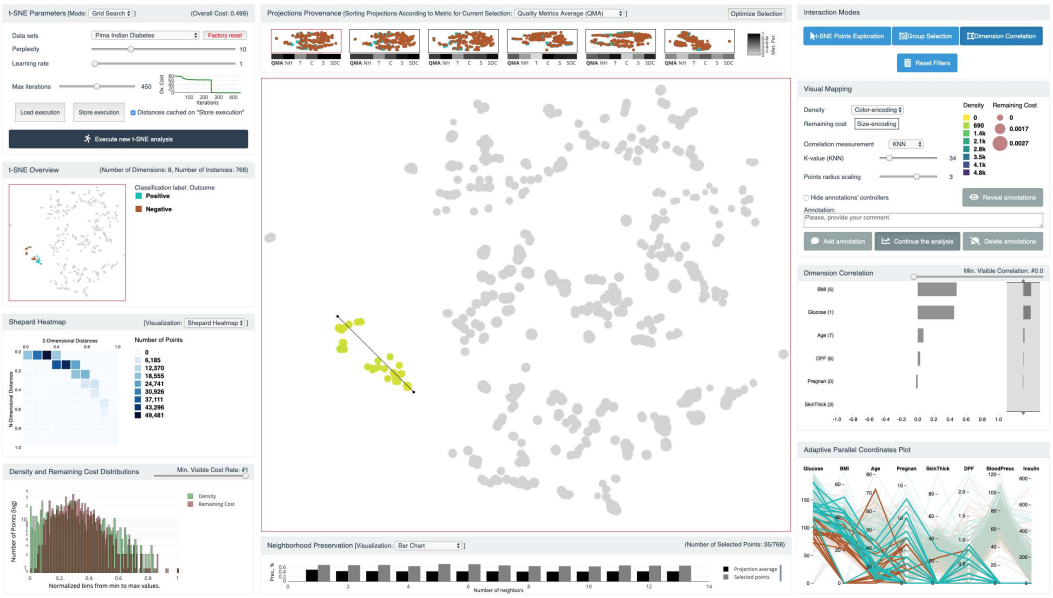


Figure 2.8 Visual analysis of t-SNE results using t-viSNE: (a) interface for uploading datasets, selecting execution modes (grid search or single parameter set), and saving/loading executions; (b) result overview with data-specific labels shown in categorical colors; (c) Shepard Heatmap displaying all pairwise distances; (d) histogram illustrating Density and Remaining Cost distributions; (e) ranked list of available projections by quality; (f) main scatterplot showing neighborhood Density in high-dimensional space and Remaining Cost for each point; (g) Neighborhood Preservation bar chart/line plot; (h) control elements for different interaction modes; (i) visual mapping panel with user options, including an annotation tool for multi-session notes; (j) Dimension Correlation bar chart showing data dimension correlations; and (k) Adaptive PCP plot highlighting the most significant dimensions. [10]

Furthermore, a summarized comparison of these tools with our proposed tool is presented in Table 1.1. A tick indicates that the relevant tool supports the mentioned feature. This table includes the main tools proposed in recent years that share common goals with our tool.

Table 2.1 Options comparison of our tool with the related proposed tools in the literature.

Tools	Options					
	t-SNE	**UMAP**	**User Interaction**	**Ranked Dimensions**	**Individual Feature Assessment**	**Precision / Recall**
Our Tool		✓	✓	✓	✓	✓
DT-SNE	✓					
IXVC	✓		✓			
t-viSNE	✓		✓	✓	✓	

DR Algorithm Preference In the referenced literature, the other three tools use t-SNE for their dimensionality reduction (DR) process. However, our tool utilizes UMAP, which demonstrates remarkable performance with complex datasets and offers faster execution compared to t-SNE.

User Interaction Both IXVC and t-viSNE tools provide interactive user engagement through a scatterplot view, allowing users to select clusters or adjust hyperparameters to create the visualization that best fits their needs, although IXVC does not support the latter feature. Similarly, our tool offers user interaction in the same manner as t-viSNE.

Ranked Dimensions Among the four papers discussed, only t-viSNE and our tool enable users to explore and investigate dimensions in depth ranked by importance, enhancing data interpretation.

Individual Feature Assessment Again, only t-viSNE and our tool allow users to assess features individually among the four discussed papers, thereby significantly improving data interpretability.

Quality Assessment Our proposed tool is unique among the discussed papers in providing precision and recall metrics. These quality metrics were provided in order to enable users to have a better assessment over the quality of the interpretation model itself, not just the DR visualization.

Precision (2.1) measures the fraction of relevant instances among retrieved instances, also known as "positive accuracy". Here, TP (True Positive) denotes the number of correctly identified positive instances, while FP (False Positive) indicates the number of instances that were incorrectly identified as positive.

$$Precision = \frac{TP}{TP + FP} \tag{2.1}$$

Recall (2.2) also known as sensitivity, calculates the ratio of relevant instances retrieved over the total number of relevant instances. This helps users understand the accuracy of the interpretation model and its ability to identify positive cases (or we can say target cases). Here, FN (False Negative) denotes the number of relevant instances that were not retrieved, while TP (True Positive) represents the number

of correctly identified positive instances. A higher recall indicates that the model is effective at capturing most of the relevant instances, minimizing the occurrence of false negatives.

$$Recall = \frac{TP}{TP + FN} \tag{2.2}$$

2.3 Overcoming Limitations in DR Visualization Interpretation

Despite the insights provided by the methods and papers discussed previously, achieving a thorough interpretation of DR visualizations is still a big challenge. These approaches often fall short, especially when it comes to complex datasets, in delivering a comprehensive understanding. With recognizing this gap, we saw the necessity for more robust and powerful interactive interpretations. Consequently, we developed our tool, named DimVis and HyEx (based on its usage type), which integrates UMAP, a robust DR technique, with EBM, a powerful method for data explanation.

3
DIMVIS: A VISUAL INTERPRETER FOR PROJECTIONS

3.1 Enhancing Data Comprehension through Interactive Visualization

The studies discussed in Section 2 primarily depend on t-SNE technique for dimensionality reduction (DR) and employ non-interactive methods. In addition, some of these works have utilized decision trees for interpreting and understanding clusters and the reason behind their shapes, which can become especially complex and challenging to interpret when it comes to work with high-dimensional data. When dealing with large-scale and high-dimensional datasets, decision trees often grow deep, resulting in intricate and hard-to-understand structures.

Although the related works in Section 2 offer certain benefits, there remains a significant need for tools which provide interactive user engagement to interpret high-dimensional datasets effectively. This thesis introduces DimVis, a tool designed for interactive data exploration and interpretation, particularly suited for complex, high-dimensional datasets without the drawbacks associated with deep and complex decision trees.

DimVis integrates a state-of-the-art Interpretable Machine Learning (IML) model known as Explainable Boosting Machine (EBM) to explain and interpret the patterns in DR visualizations. EBMs are Fundamentally interpretable and adapted to handle complex, large-scale, high-dimensional datasets and patterns.

DimVis offers several advantages over the previous works discussed in Section 2. To start with, we can highlight its dynamic model training capabilities using advanced DR techniques. Essentially, DimVis trains the model in real-time based on user interactions, which are achieved by selecting desired clusters or sub-clusters, ensuring that data visualization and clustering are tailored to the user's needs and preferences.

This section provides an overview of DimVis as a web-based tool, designed with multiple views and interactive sections to facilitate better data understanding and interpretation, especially for complicated datasets. We aim to improve user interaction by enabling data point selections using a lasso tool in the provided visualizations and hyperparameter adjustments, thereby providing a comprehensive analysis of the data. The tool's overall layout is shown in Figure 3.1, and each distinct view is described below, focusing on three main design goals: Interactive EBM Visualization Based on User Selection (G1), Individual Feature Assessment (G2), and User Support (G3).

- Goal 1 (G1) aims to increase and enhance interactive user engagement by enabling interaction with the tool, particularly through selecting any desired groups of data points and creating different clusters visualized in lower dimensions. This approach simplifies the exploration, comprehension and interpretation of complex datasets visualized by UMAP technique.
- Goal 2 (G2) focuses on exploring the ranking of features from the most important to the least, their impact on the dataset, and the detailed examination of any specific feature. This goal highlights individual feature assessment, allowing users to easily investigate and explore any specific feature in depth with a simple click, thereby enhancing the understanding of each feature's contribution to the overall patterns and insights provided by the UMAP.
- Goal 3 (G3) is centered around providing user support by offering a flexible and user-friendly environment. This goal addresses the need for a tool that accommodates a wide range of datasets across various fields and the user can work with a variety of options in a wide range of domains. Additionally, it aims to enable users to create their desired visualization using UMAP with simple adjustments to UMAP's hyperparameters that fits their needs the most.



Figure 3.1 Components of DimVis: a panel for (a) choosing dataset and (b) adjusting the UMAP hyperparameters. (c) Shows the quality metrics. (d) Illustrates the lasso tool which is used to select data points interactively. (e) Shows the selected data points. (f) Demonstrates the unselected data points. (g) Depicts the EBM ranking graph for the selected data points; and (h) presents the Score of the selection, while (I) shows the distribution for one feature

The implementation of these goals—G1, G2, and G3—in the design of DimVis and the relevant algorithms will be elaborated later in this section.

3.2 Goal 1: Interactive UMAP and EBM Visualizations Based on User Selection

To achieve our first goal (G1), users can select a dataset from a dropdown menu (Figure 3.2(a)). If they find the clusters in the visualization too dense, too separated, or unsatisfactory for any reason, users can adjust two hyperparameters (Figure 3.2(b)) until they achieve a desirable visualization. This interaction capability is designed to enhance user engagement and understanding. Validation metrics, Precision, and Recall, which measure the efficacy and accuracy of the visualization model, are calculated and displayed (Figure 3.2(c)).

When a user selects a subset of data points in the visualization plot provided by UMAP using the lasso tool (Figure 3.2(d)), a backend process begins. Selected data points are labeled as "selected" (Figure 3.2(e)), while the remaining data points

are labeled as "unselected" (Figure 3.2(f)). This helps to have better and focused analysis, in line with G1. The EBM visualization graph, which is crucial for both G1 and G2, is displayed on the right side of the main view (Figure 3.2(g)). Initially, this graph provides a ranked list of features based on their importance in the dataset, from the most important to the least.

When a user selects a specific cluster or group of data points using the lasso tool, the EBM model is re-trained and re-calculated to display the characteristics of the selected data points against the rest of the selection. Initially, the EBM model is trained on the entire dataset to provide a ranked list of important features. This helps users understand the global structure of the dataset and feature importance. Users can dynamically select any groups of data points, and the original class labels are discarded for this specific inquiry. The EBM model is re-trained with the new labels, providing descriptions and ranked dimensions specific to the selected data points. This retraining occurs dynamically with each selection.

Whether it is displaying the visualization for the entire dataset or providing the visualization for a specific selection, the callback function *"update_ebm_plot"* handles data point selection, labeling, and model retraining (see Algorithm 2). It recognizes the chosen dataset and retrains the EBM model based on user selections in the UMAP graph. Precision and recall are computed as performance metrics and displayed in the top right corner of the main view (Figure 3.2(c)), aiming to help users assess the model's performance on the selected data points.

The Feature Importance Graph visually ranks features based on their influence on the model's predictions. Features located higher on the list have more influence on the model's decisions, while those lower on the list have less effect. This visualization provides insights into which features are vital for the model's predictions.

Finally, the Feature Importance Graph is calculated and plotted in the EBM graph as a ranked list relevant to the selected data points by the user. The EBM model is stored for future reference in order to avoid repeated calculations in case the cluster selections repeats.

With the EBM visualization and backend logic, our tool offers a comprehensive and interactive way to explore and understand complex datasets. By providing both general analysis on the entire dataset and specific analysis on selected data points, DimVis offers deeper insights into the data. This capability enhances the interpretability of data visualized using the UMAP dimensionality reduction technique, especially for complex and high-dimensional datasets, which are

increasingly prevalent.



Figure 3.2 Overall view of our proposed tool. (a) Demonstrates the dataset selection options in the dropdown. (b) Displays the hyperparameters adjustments. (c) Shows the quality metrics. (d) Illustrates the lasso tool which is used to select data points interactively. (e) Shows the selected data points. (f) Demonstrates the unselected data points. (g) Depicts the EBM ranking graph for the selected data points.

3.3 Goal 2: Individual and in-depth Feature Assessment

To achieve Goal 2 (G2), our tool includes the dynamic EBM Feature Plot. This feature is activated when a user clicks on any feature in the main EBM visualization graph (Figure 3.2(g)). This user interaction is crucial for G2, allowing users to explore and understand the specific impact of features. The graph appears below the UMAP and EBM plots and disappears when a new interaction, such as a different selection or a click on another feature, occurs. This enhances the user's ability to interpret how individual features influence either a selected subset of data points or the entire dataset.

When a feature is clicked in the EBM plot, the callback function picks out the feature's name and checks if it matches the most recently clicked feature of the plot. If it does, the relevant plot's visibility toggles. If not, the function retrieves the EBM explainer object from storage, identifies the index of the clicked feature, and generates a detailed plot using the EBM explainer's visualize method (Figure 3.3). This process, essential for G2, creates a dynamic selected feature-specific visualization that includes two components: the Score graph and the Density graph.

The Score graph shown in Figure 3.3(a) illustrates the feature's values and their corresponding scores on the y-axis, highlighting each feature's value contribution

to the prediction. The x-axis (Figure 3.3) displays the range of possible values that the specific feature can take in the dataset, empowering users to observe and explore the impact of various value intervals on the model's output. Furthermore, the light gray background on the Score graph indicates the model's confidence in its predictions; a thicker background signifies a wider confidence interval and greater uncertainty.

The Density graph (Figure 3.3(c)) shows the distribution of data points for the selected feature across the dataset, revealing the distribution and range of feature values.

Examining these graphs together in the same view helps users to better and more deeply understand the dataset and its underlying patterns, addressing the needs of G2. This is particularly valuable for data scientists and analysts working with complex datasets, as it aids in deriving meaningful insights and understanding the data's complicated patterns more effectively.

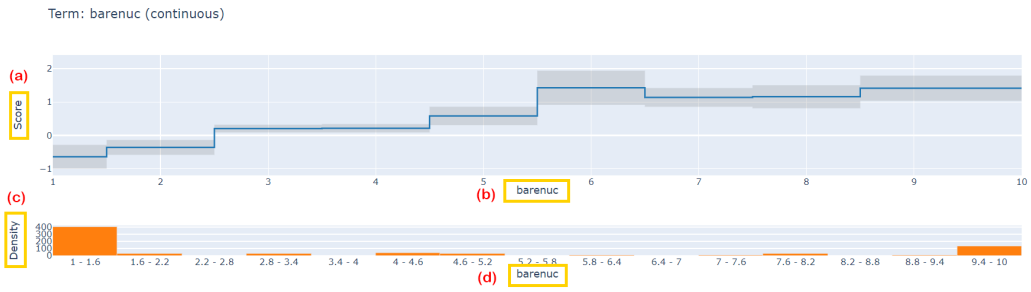


Figure 3.3 Individual feature assessment graph. (a) Shows the selected data points against the unselected ones in the dataset. (b) Displays the selected feature from EBM feature ranking graph. (c) Demonstrates the density of the data points in the whole dataset. (d) Illustrates the distribution of the data points across the whole dataset and also depicts the value range for the clicked feature from EBM feature ranking graph. The light gray background surrounding the linear chart in Score graph represents the uncertainty of the model's prediction for the clicked feature and as this background goes wider, the uncertainty increases.

The EBM visualization includes a Pairwise Feature Graph to illustrate how two features together influence the model's predictions. This graph is especially useful when a single feature does not provide a complete understanding on its own. By combining two features, the Pairwise Feature Graph offers a more detailed interpretation on how the combination of these two features affect the model's outcome and provides a comprehensive view of how real-world data behaves and impacts the model's outcomes.

The Pairwise Feature Graph is a visualization tool designed to analyze and interpret

the interaction effects between two features. For example, as shown in Figure 3.4, it demonstrates how Glucose and Insulin features together affect the model's prediction. The x-axis (Figure 3.4(b)) displays the range of Glucose values across the entire dataset, while the y-axis (Figure 3.4(a)) shows the range of Insulin values among the whole dataset. Each cell in the graph represents the interaction score between these two features. The color intensity within each cell indicates the strength and the intensity of the interaction effect, with a color scale provided on the right side of the graph (Figure 3.4(c)) for reference.

Darker colors signify stronger interaction effects, indicating a significant influence on the model's predictions. For instance, the dark purple and dark pink areas in Figure 3.4 indicate that specific combinations of high Glucose and Insulin levels have a considerable impact on the model's output. By this, we can understand that these areas could indicate that high levels of both features together strongly predict an outcome, such as a higher likelihood of diabetes.

Contrariwise, lighter colors in the graph depict weaker interaction effects, showing less influence on the model's predictions. These variations in color are vital for interpreting how the interaction between Insulin and Glucose affects the model's behavior.

The Pairwise Feature Graph is critical for understanding complex datasets by identifying and interpreting the interactions between features. It provides users with insights into how combinations of features impact outcomes, enhancing their comprehension of the model's behavior and outcome.

In summary, the Pairwise Feature Graph uses heatmaps to visualize the combined effect and interaction of two features, while the Individual Feature Graph employs histograms to show the distribution and impact of single features on the model's predictions.

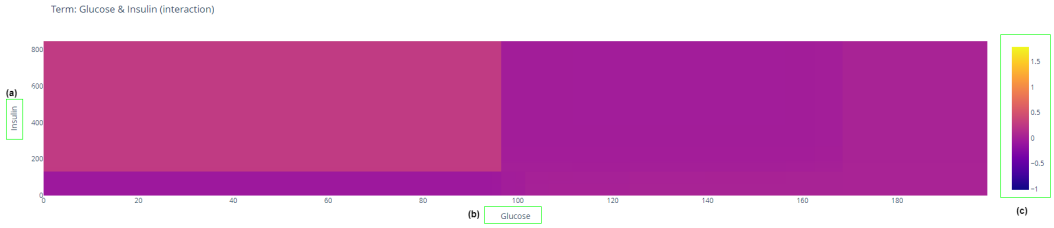


Figure 3.4 Pairwise Feature Assessment Graph. (a) Illustrates the value range for Insulin feature (b) Shows the value range for Glucose feature. (c) Displays the color scale for the intensity of the interaction among these two features.

3.4 Goal 3: User Support

To fulfill Goal 3 (G3), the primary design focus of our tool is its flexibility in handling various datasets. This critical feature of G3 allows researchers and data analysts to upload and analyze multiple datasets from different fields or with varying dimensions, without being limited to a specific type or a specific domain. As shown in Figure 3.2(a), users can effortlessly switch between datasets to compare patterns, analyze features, and conduct in-depth analysis of each feature. This flexibility supports the key aspect of G3: enhancing user support.

The ability to select any desired dataset or switch between them, offers numerous benefits. It minimizes the need for multiple analytical tools, thereby allowing analysts to focus more on their analyses. Furthermore, it provides a framework for identifying and recognizing patterns across diverse datasets, aligning with the user support objectives of G3.

To achieve the goals of G3, DimVis offers the capability to work with a variety of datasets and leverages the UMAP algorithm for data visualization. The UMAP algorithm proposes a range of visualization options for the same dataset through hyperparameter adjustments. DimVis assists users in discovering the optimal visualization that meets their needs by allowing hyperparameter adjustments, as shown in Figure 3.2(b). Once a user selects their preferred dataset, the system provides a set of clusters based on default hyperparameter settings—"Number of Neighbors" set to 15 and "Minimum Distance" set to 0.1, as shown in Figure 4.1. This initial visualization helps users start their exploration and investigation from an appropriate point and understand how hyperparameters influence data clustering with slight adjustments.

Our tool enables users to explore and test up to 1,000 different UMAP projections simply by clicking and modifying and adjusting these hyperparameters. This wide array of possibilities is calculated as 100 options for "Number of Neighbors" multiplied by 10 choices for "Minimum Distance." DimVis generates one of these 1,000 clustering options once at a time, and if users do not find the current visualization satisfactory, they can interactively change the hyperparameters until they find the best alternative that meets their needs. This dynamic adjustment capability ensures that DimVis fulfills the user support goal of G3.

3.5 Algorithms and Implimentation

Our proposed tool, DimVis, is designed to help data scientists and researchers interpret and understand complex visualizations. It integrates a series of algorithms that handle various tasks, from visualizing UMAP data projections to plotting EBM feature importance. The core methodologies and logic that form the backbone of our system are as follows.

3.5.1 Utilizing Plotly Dash for Dynamic Web-Based Applications

Plotly Dash, commonly referred to as Dash, is a fundamental component of the DimVis platform. Dash is a Python framework for building web applications for analytical purposes without the need for JavaScript. This framework makes the development process faster and easier. By combining Plotly's visualization capabilities with Flask's web serving potentials, we have created interactive web-based visualizations and applications. Using Python, developers can implement dynamic and advanced user interfaces with Dash, facilitating the creation of visual interfaces tailored to the tool's complex algorithms and visualizations. For more information, refer to the official Dash documentation and resources [21].

3.5.2 Updating Scatter Plot with UMAP in Dash Applications

The foundation and understructure of our application is a UMAP graph illustrating the entire dataset. This design ensures proper handling of data from various datasets, such as 'Breast Cancer Wisconsin', 'Pima Indian Diabetes' and 'Indian Pines'. A notable feature of this algorithm is its ability to store and retrieve UMAP results, reducing unnecessary computations and optimizing performance and response time. Detailed procedures are provided in Algorithm 1. This algorithm outlines the process of initializing parameters, handling dataset inputs, applying UMAP, storing results, and generating the scatter plot for visualization.

Algorithm 1 :Update Scatter Plot with UMAP in Dash Application.

Initialize `umap_status` as `False` and `fig_title` as an empty string
if dataset is `None` **then**
    Raise `PreventUpdate`
end if
Load input file into `data_df`
Set `fig_title` with the dataset's title
Prepare hyperparameters from `n_neighbors` and `min_dist`
if `umap_status` exists and dataset and hyperparameters match **then**
    Retrieve stored UMAP data
else
    Apply UMAP to `data_df` using hyperparameters
    Store UMAP result in `umap_status`
    Update `data_df` with UMAP embeddings in columns `umap_1` and `umap_2`
end if
Generate scatter plot figure `fig` with UMAP data
Update `fig` layout and aesthetics
Return `fig` and other relevant Output values

3.5.3 Updating EBM Scatter Plot in Dash Application

Simultaneously with the visualization of the UMAP projection either for the entire dataset or for the selected data points, the EBM scatter plot is displayed in the main view to provide actionable insights into the data's features, as shown in Algorithm 2. This algorithm utilizes the *ExplainableBoostingClassifier*, which uses a fixed random state to ensure consistency and predictability. Based on the selected data points, the algorithm recomputes validation metrics for the fitted EBM technique. Additionally, it calculates feature importance scores and ranks attributes based on these calculations, serving as an essential tool for detailed data analysis. This algorithm details the steps for initializing parameters, handling data selection, applying the ExplainableBoostingClassifier, computing feature importance scores, and generating the EBM plot for visualization.

Algorithm 2 : Update EBM Plot in Dash Application.

Initialize class column
Initialize data based on selected dataset
if no data is selected or all data is selected **then**
    Use the original class labels as y
else
    Use 'selected' and 'unselected' as new class labels
end if
Drop ID and Class columns to form feature matrix X
Initialize ExplainableBoostingClassifier with random state 42
Fit the model with X and y
Predict class labels using the same X
if all data points or no data points are selected **then**
    Calculate weighted precision and recall
else
    Calculate precision and recall using 'selected' as positive label
end if
Create a global explainer for the model
Generate feature importance scores and names
Sort feature importance in descending order
Update and return the EBM plot figure with relevant settings

3.5.4 Plotting Feature Importance Graph in Dash Application

This algorithm facilitates interactive exploration of ranked individual EBM features. It is activated when the user clicks on any attribute listed in the EBM Feature Importance plot, delivering real-time insights into the selected feature for the chosen data points (see Algorithm 3). The main purpose of this graph is to provide a dynamic and responsive interface that addresses the user's immediate queries for deeper exploration. The algorithm outlines the process of initializing parameters, handling dataset inputs, applying UMAP, storing results, and generating the scatter plot for visualization.

Algorithm 3 : Update Scatter Plot with UMAP in Dash Application.

if clickData and feature_names are not None **then**
    Extract the clicked feature name
    Retrieve last clicked feature for toggling
    if clicked feature is the same as the last clicked feature **then**
        Toggle the visibility of the plot
        Return updates without changing the figure
    else
        Load the explainer model from file
        Find the index of the clicked feature
        Generate and return the new feature score plot
    end if
else
    No update for the figure, plot remains collapsed
end if

In conclusion, our offered tool benefits from these algorithms to provide a platform which simplifies interpretation and understanding the visualization of complicated and high-dimensional datasets. It is designed to enable users to derive useful insights from their chosen datasets or selected data points with minimal effort, regardless of their technical background in data visualization.

4
APPLICATION 1: BREAST CANCER

4.1 DimVis in Action

To illustrate the applicability and effectiveness of our tool, we present a use case involving real-world healthcare dataset. The use case demonstrates DimVis's ability to assist users, such as data analysts, in evaluating the formation of data clusters in an unsupervised setting within the healthcare domain. In the following section, we will detail where and how these goals have been achieved.

4.2 Use Case: Exploring Data Clusters Without Ground-Truth Labels

For this use case, we select the Breast Cancer (Wisconsin) dataset from the Kaggle repository [22] to demonstrate how DimVis assists users in exploring and investigating data clusters based on various dimensions. Initially, we will interpret the data clusters without using the ground-truth labels. Subsequently, we will include the ground-truth labels to evaluate DimVis's effectiveness in explaining the formation of these clusters (Figure 3.2).

The dataset comprises 699 breast cancer cases labeled as either *Benign* or *Malignant*. The *Benign* label includes 458 instances, while the *Malignant* label includes 241 samples. Each data instance in this dataset has nine dimensions/features: *clumptick*, *sizeun*, *shapeun*, *marg_adh*, *epithsize*, *barenuc*, *blandchr*, *nornuc*, and *mitoses*. These features have values ranging from 1 to 10, where higher values indicate a greater likelihood of being *Malignant*. At this step, Goal 3 is achieved by selecting an appropriate dataset from the available options.

4.2.1 Exploring Different Projections

Upon deploying DimVis, the initial UMAP projection forms two compact data clusters with default values set for its hyperparameters (Neighbors: 15; Min Dist:

0.1). Although these initial clusters are distinct, we proceed to experiment with different UMAP hyperparameters to verify the reliability of the projection. To make this happen, we adjust the Neighbors (number of neighbors) parameter to 12 and the Min Dist (minimum distance) parameter to 0.3. These changes result in a new UMAP projection that reveals three distinct and more dispersed data clusters (see Figure 4.1). By adjusting the hyperparameters and achieving a satisfactory UMAP visualization, we further support the accomplishment of Goal 3.

4.2.2 Evaluating Cluster Formation

As the variation in the number of data clusters (two versus three) between the two projections is perplexing we are motivated to prompt an investigation into the reasons behind these cluster formations. We focus on the projection with three well-separated clusters to further examine the influence of UMAP hyperparameters. The clusters, labeled C1, C2, and C3, differ in size and are depicted in Figure 4.1. Without access to ground-truth labels, we utilize DimVis to interpret and understand these data clusters.

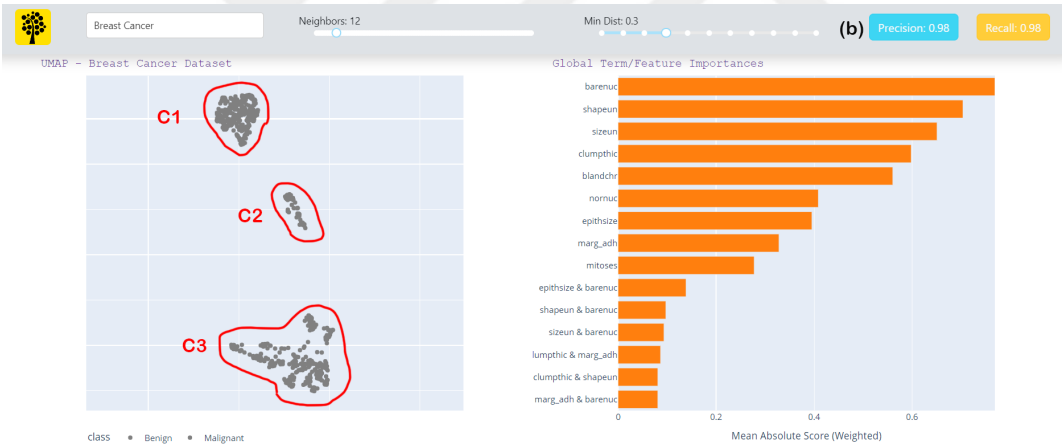


Figure 4.1 Visualization of the three data clusters generated by UMAP (left) alongside a Feature Importance Graph displaying the most significant features as determined by the EBM model trained on the entire dataset (right). Using the lasso tool to select each cluster allows us to observe the changes in the EBM's Feature Importance Graph.

4.2.2.1 Analyzing Cluster C1

We begin our analysis by selecting cluster C1. The EBM plot re-trains and updates immediately, reflecting the difference between the selected data points and the rest of the dataset. This selection significantly alters the ranking of the features (see Figure 4.2). The new order for the top five most important features is: *barenuc*,

sizeun, *nornuc*, *clumptick*, and *shapeun*. The ability to select different clusters and retrain the EBM model to generate various visualizations of ranked features aligns with the objectives of Goal 1.

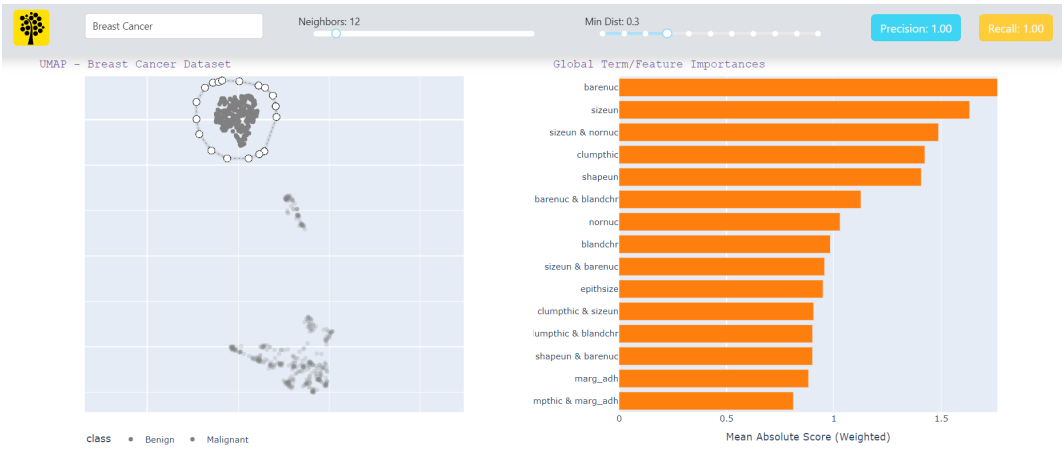


Figure 4.2 Selecting C1 retrains the EBM model to highlight the differences between C1 and the other clusters. The top features are: *barenuc*, *sizeun*, *nornuc*, *clumpthic*, and *shapeun*.

Upon selecting C1, we turn our attention to the validation metrics. The EBM proxy model performs exceptionally well in explaining the data points within this cluster, achieving perfect scores of 1 (or 100%) for both precision and recall. These optimal performance metrics underscore the model's robustness and reliability.

To delve deeper and investigate more, we begin with the most important feature, *barenuc*, which is located at the top of the Feature Importance Graph. Clicking on the *barenuc* bar reveals a Score graph and a Density graph below, allowing us to analyze this feature in isolation, as illustrated in Figure 4.2. This in-depth exploration satisfies Goal 2 by enhancing our understanding of the dataset.

The Score graph (y-axis: contribution score; x-axis: feature values; light gray background: model's uncertainty) indicates that the majority of the selected data points have *barenuc* values above 5 (see Figure 4.3, negative score line). Given that higher values suggest a greater likelihood of being *Malignant*, with *barenuc* values approaching 10, we hypothesize that a significant portion of the selected data points likely belong to the *Malignant* class.

The Density graph (y-axis: density of data instances; x-axis: histogram of values) shows the distribution of *barenuc* values across the entire dataset. Peaks on the x-axis indicate a high number of data points within specific value ranges, while lower bars signify fewer instances. This graph consistently reveals the distribution

of the *barenuc* feature for the entire dataset, remaining unchanged across different cluster selections since it represents the overall dataset distribution.

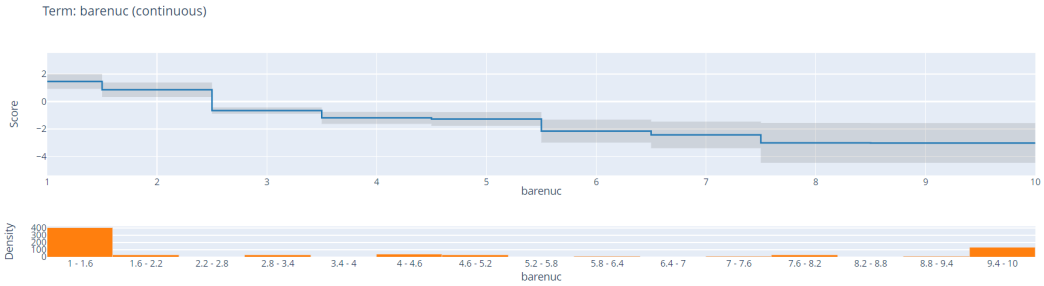


Figure 4.3 Individual Feature Analysis for *barenuc* in Cluster C1. The Score graph shows the *barenuc* values for data points within C1 compared to those outside the cluster. The Density graph displays the overall distribution of *barenuc* across the entire dataset.

Following this engrossing result, we proceed by clicking on the second most important feature, *sizeun*. The Individual Feature graph for *sizeun* closely resembles the previous one, reinforcing the notion that the majority of the selected data points are likely to have *Malignant* class label (See Figure 4.4). However, a notable difference between these two Feature graphs is the increased uncertainty for the selected data points, evident from the wider band in the light gray background.

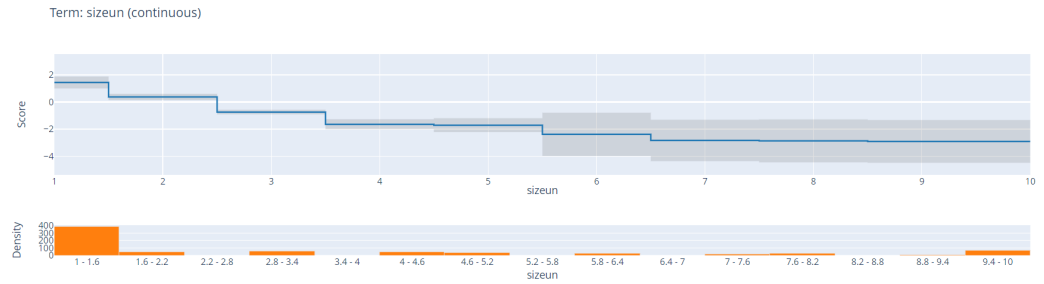


Figure 4.4 Individual Feature Analysis for *barenuc* in Cluster C1. The Score graph shows the *barenuc* values for data points within C1 compared to those outside the cluster. The Density graph displays the overall distribution of *barenuc* across the entire dataset.

To validate our initial hypothesis, we examined the third, fourth, and fifth most important features, as shown in Figure 4.5, Figure 4.6, and Figure 4.7, respectively. Figure 4.5 demonstrates the interaction between the *nornuc* and *sizeun* features. Unlike the previous individual feature graphs, this graph highlights the combined effect of these two features on cluster formation, which cannot be captured when examining each feature separately.

In this combined graph, the selected data points are indicated by negative scores in dark purple. Higher values for both *nornuc* and *sizeun* make Cluster C1 stand out from the other clusters. Interestingly, low *sizeun* values still characterize the selected cluster when *nornuc* values are sufficiently high (above 8).

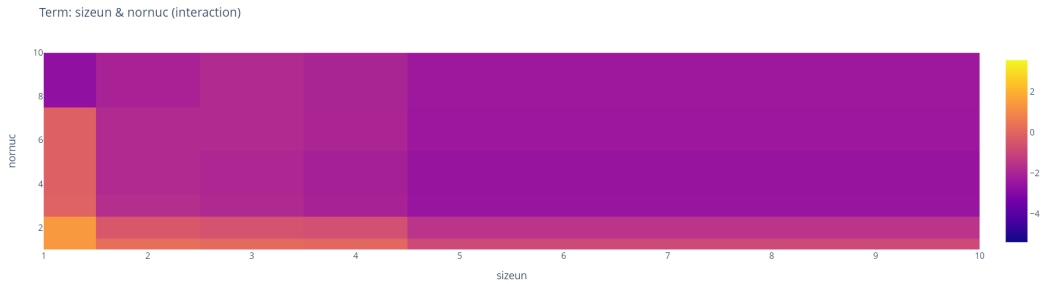


Figure 4.5 Pairwise Feature Graph for the *nornuc* and *sizeun* Combination in Cluster C1. The score graph illustrates the combined impact of these two features on the selected data points. While *nornuc* alone is not among the top six features in the Feature Importance graph, its combination with *sizeun* ranks fourth.

In Figure 4.6, the selected data points show significantly higher values, approaching 10. Given that higher values are typically associated with the *Malignant* class label, this finding supports our hypothesis. However, there remains considerable uncertainty for values above 8 in this graph.

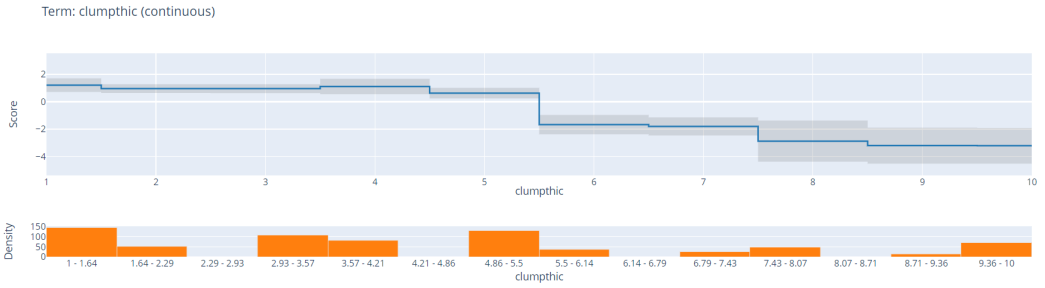


Figure 4.6 Individual Feature Graph for *clumpthic* in Cluster C1. Initially, the graph displays a linear trend, but for the selected data points, it highlights very high *clumpthic* values. Additionally, the Density graph reveals an almost uniform distribution of this feature across the entire dataset.

In Figure 4.7, the selected data points show an increasing trend in the *shapeun* value. Similar to previous graphs, this observation supports our initial hypothesis that Cluster C1 is associated with the *Malignant* class. However, this graph also clarifies why points in this cluster differ from the remaining points.

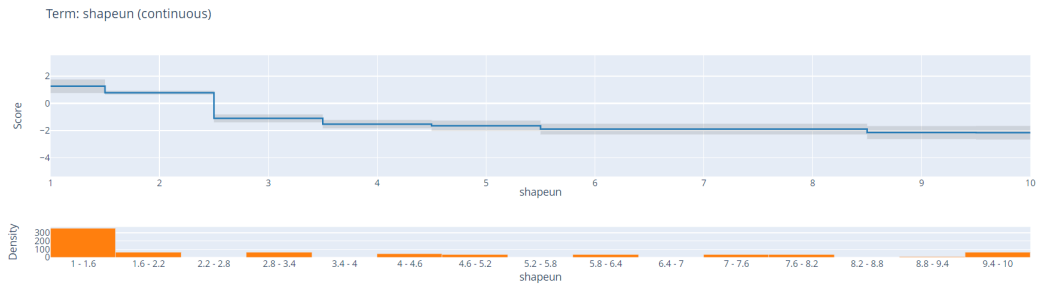


Figure 4.7 Individual Feature Graph for *shapeun* in Cluster C1. The Score graph displays an almost linear trend for the *shapeun* feature within C1. Additionally, the Density graph indicates a higher concentration of *shapeun* values in the 1 to 1.6 range.

4.2.2.2 Analyzing Cluster C2

Having thoroughly investigated cluster C1 and obtained crucial insights and ideas into the features that play a significant role in its formation, we now turn our focus to the interactive exploration of C2. Selecting C2 reveals notable changes in the EBM's Feature Importance Graph, as shown in Figure Figure 4.8.

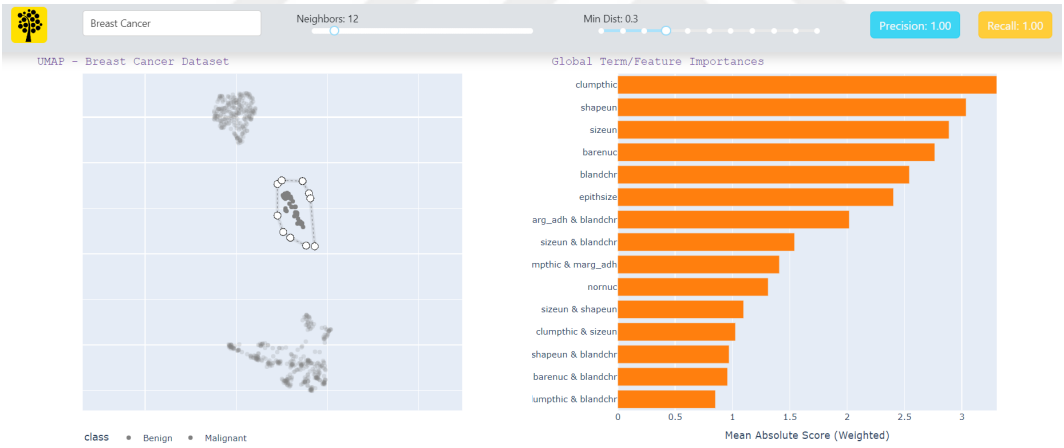


Figure 4.8 Selecting C2 retrains the EBM model to highlight the differences between C2 and the other clusters. The top features identified are *clumpthic*, *shapeun*, *sizeun*, and *barenuc*.

According to Figure 4.9, the most important features for C2 are *clumpthic*, *shapeun*, *sizeun*, and *barenuc*. To delve deeper and investigate more, we begin with the top feature, *clumpthic*. By analyzing its individual graph, we observe that *clumpthic* influences the cluster in a steadily increasing manner. A closer look reveals that the selected region (where the score value is below 0) has very low *clumpthic* values. This observation leads to our second hypothesis: these low values suggest that the selected data points are likely Benign cases. Additionally, we note a high degree of

uncertainty for the unselected data points (the rest of the dataset), whereas there is no uncertainty for our selected data points.

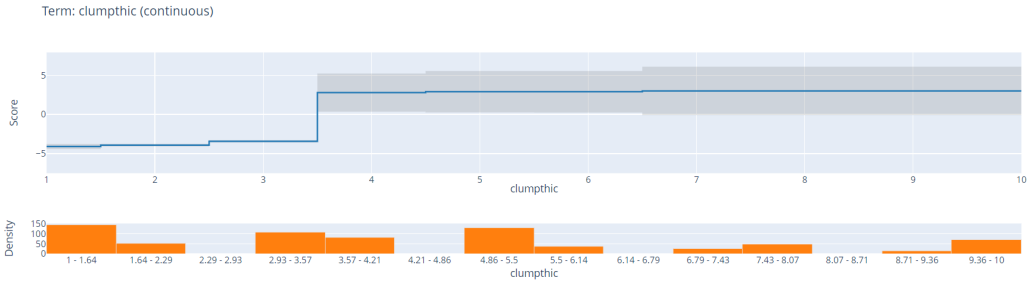


Figure 4.9 Individual Feature Graph for *clumpthic* in Cluster C2. The Score graph displays the *clumpthic* values for the selected cluster, C2. Additionally, the Density graph reveals the same distribution of *clumpthic* as observed in C1.

To validate our hypothesis about the data points in C2, we proceed to analyze *shapeun*, the second most important feature for this cluster. Upon selecting this feature, we observe results similar to those of *clumpthic*. However, the selected data points exhibit much lower *shapeun* values, tending toward 0, as shown in Figure 4.10. This finding further supports our second hypothesis.

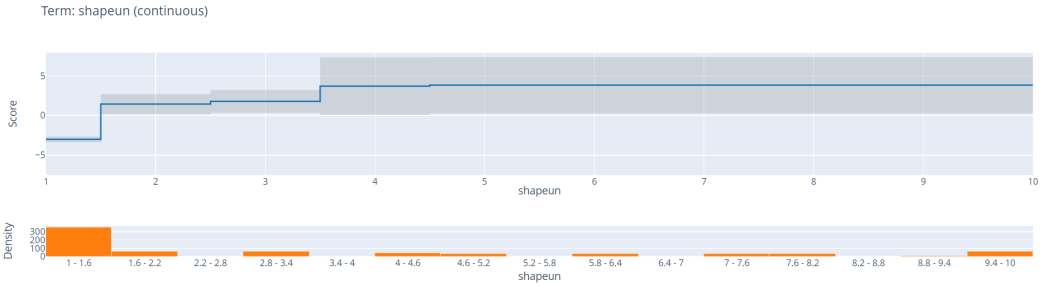


Figure 4.10 Individual Feature Graph for *shapeun* in Cluster C2. The Score graph represents the second most important feature and indicates high uncertainty for the selected data points. Additionally, the Density graph shows the distribution of *shapeun* values across the entire dataset.

Next, we examine *sizeun* to gain further insights into C2 and its most important features (see Figure 4.11). The selected data points exhibit very low *sizeun* values, reinforcing our hypothesis about C2. A notable aspect of this graph is the constant straight trend of the selected data points. A consistent score (y-axis in the line plot) indicates that the instances for this feature have values of either 1 or 2. This observation supports our hypothesis that the majority of data points in C2 belong to the Benign class.

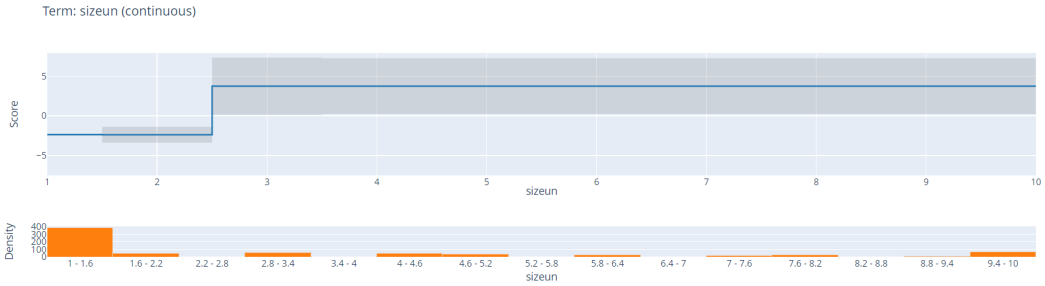


Figure 4.11 Individual Feature Graph for *sizeun* in Cluster C2. The Score graph displays a consistent linear trend. The Density graph remains similar to the overall distribution of the *sizeun* feature.

To conclude our examination of C2, we analyze the fourth most important feature, *barenuc*. Clicking on this feature reveals the graph shown in Figure 4.12. This graph shares similar characteristics with the Individual Feature graph for *clumpthic*, though with slight differences in feature values. Both graphs display lower values for their respective features, which further supports our hypothesis.

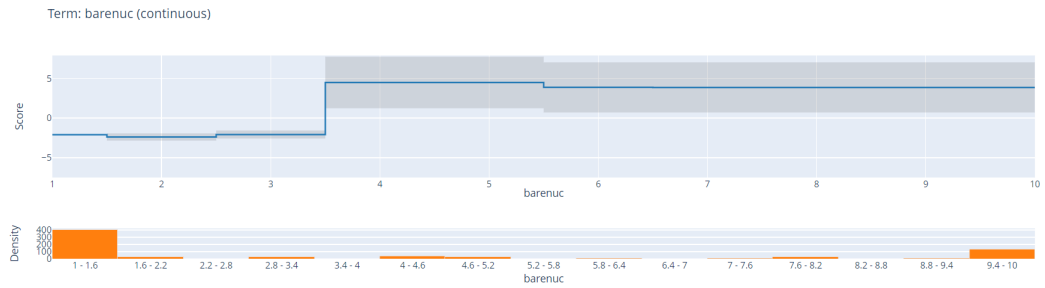


Figure 4.12 Individual Feature Graph for *barenuc* in Cluster C2. The Score graph reveals high uncertainty for the unselected data points. Additionally, the Density graph illustrates the distribution of *barenuc* values across the entire dataset.

4.2.2.3 Analyzing Cluster C3

We conclude our examination with Cluster C3. A closer inspection of the EBM's Feature Importance graph reveals that the four primary features for this cluster are obviously different from those of other clusters. Notably, the first and third features are combinations of two different attributes. The most important features for C3 are *sizeun* and *blandchr*, *blandchr*, *marg_adh* and *blandchr*, and *sizeun*, respectively (see Figure 4.13).

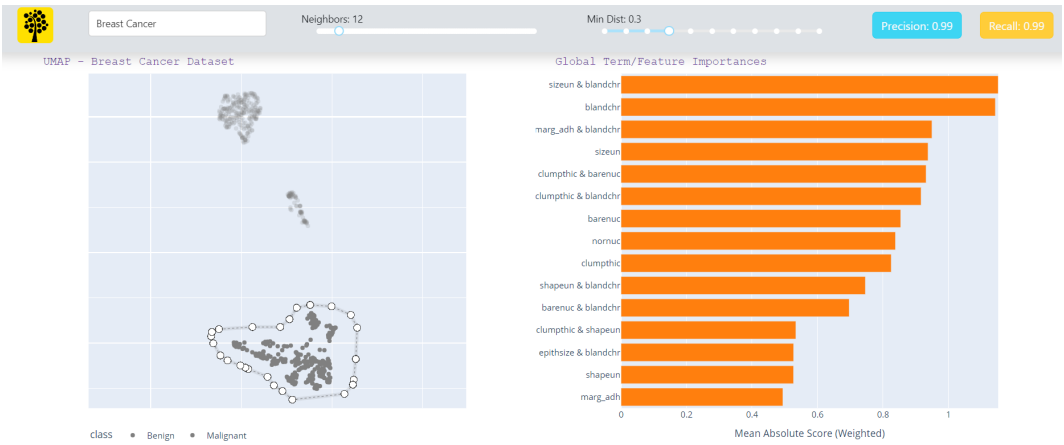


Figure 4.13 Selecting C3 retrains the EBM model to highlight the differences between C3 and the other clusters. The most important features are *sizeun*, *blandchr*, *marg_adh*, and *sizeun*

To investigate more and gain a deeper understanding of Cluster C3, we analyze each feature individually, as we did with the previous clusters, C1 and C2. We begin with the most important feature, the combination of *sizeun* and *blandchr* and illustrates their interaction on the model's prediction. Upon selecting this combinatorial feature, the graph shows that the selected data points exhibit very low values for one feature and high values for the other. This pattern suggests a strong likelihood that the selected data points are Benign cases, similar to our hypothesis for C2.

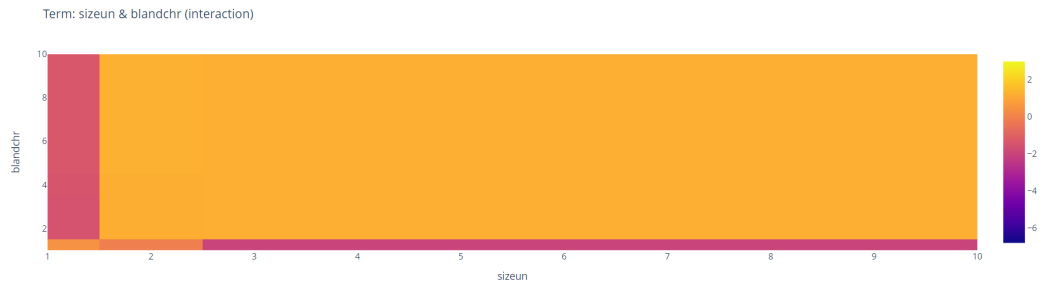


Figure 4.14 Pairwise Feature Graph for the combination of *sizeun* and *blandchr* in Cluster C3. The Score graph illustrates the combined effect of these two features on the selected data points, with their intersection representing our selection.

Following the same methodology, we proceed to investigate and explore the remaining three most important features. Their respective outcomes are shown in Figure 4.15, Figure 4.16, and Figure 4.17. Upon thorough examination, we notice significant similarities among these features, which supports our hypothesis regarding the possible class label of this selection. Interestingly, this cluster displays

a more mixed range of feature values compared to C2, which predominantly consisted of points with low values.

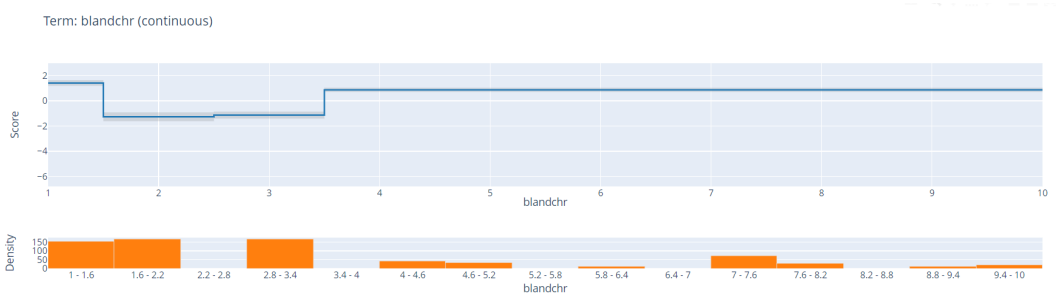


Figure 4.15 Individual Feature Graph for *blandchr* in Cluster C3. The Score graph displays an almost linear trend with low uncertainty for both selected and unselected data points. Additionally, the Density graph indicates a nearly uniform distribution of *blandchr* values across the entire dataset.

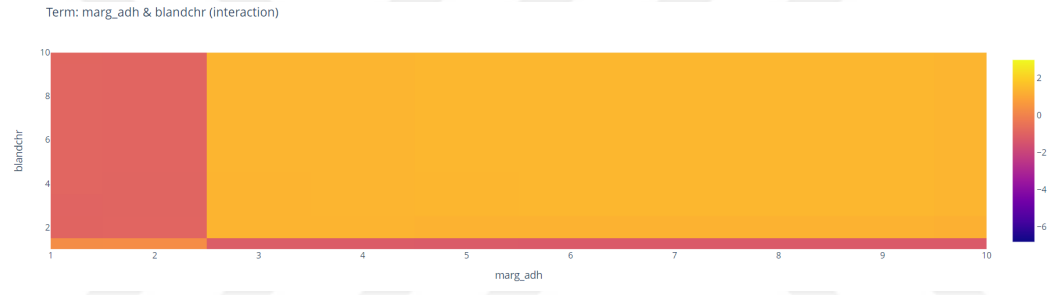


Figure 4.16 Pairwise Feature Graph for the combination of *marg_adh* and *blandchr* in Cluster C3. This combinatorial graph shows the intersection of both features, highlighting low feature values corresponding to our selection.

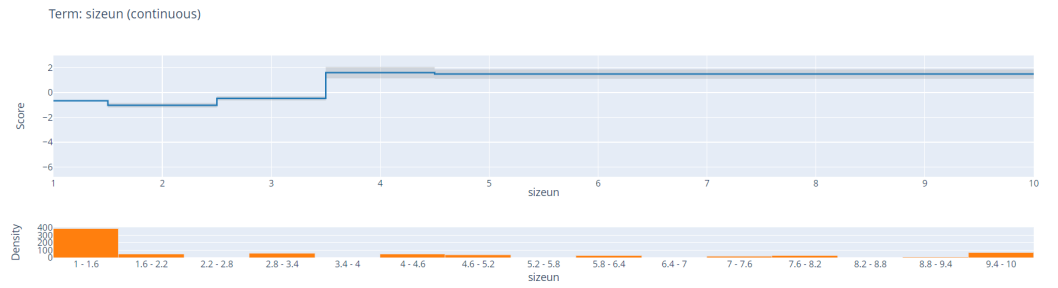


Figure 4.17 Individual Feature Graph for *sizeun* in Cluster C3. The Score graph exhibits a similar pattern to other Individual Feature Graphs for clusters C2 and C3. The Density graph illustrates the distribution of *sizeun* across the dataset.

4.2.2.4 Insights and Summary of Results

Through a detailed exploration of the three provided and generated clusters (C1, C2, and C3) by the UMAP algorithm, we have uncovered several notable patterns

and insights. Our investigations revealed that the top four features in the EBM's Feature Importance Graph are remarkably similar across clusters. Specifically, in C2, which we hypothesize to be a pure cluster with all data points belonging to a single class label, there is no combination of features in the top ranks of the EBM's Feature Importance Graph. The EBM's Feature Importance graphs have significantly enhanced our understanding of the data, allowing us to hypothesize about the class labels of the data points within each cluster.

Cluster C1 shows a strong tendency towards the *Malignant* class label, with high feature values (closer to 10) reinforcing this hypothesis. Further analysis of prominent features such as *barenuc* and *sizeun* supports this indication, suggesting that majority of the instances represented by C1 in the dataset have a higher risk of being *Malignant*.

Clusters C2 and C3 tell a different story. Very low feature values (tending towards 10), especially for attributes like *clumpthic* and *shapeun* in C2, strongly suggest that the relevant class label is *Benign*. This is further supported by the constant linear trend in the *sizeun* feature graph, indicating uniformity and homogeneity in C2. Although C3 also shows similar insights with low feature values, it lacks the homogeneity seen in C2, implying that the majority of its data points are likely *Benign*. Despite being separate clusters, C2 and C3 share similar attributes regarding their potential class labels.

The EBM's Individual Feature graphs have provided a deep perspective into each cluster and the reason behind their formation, enabling precise and thorough analysis of groups of data points. The trends, patterns, and graphs examined in each cluster have informed our hypotheses about the dataset. While C1 exhibits features indicative of *Malignancy*, clusters C2 and C3 demonstrate *Benign* tendencies. These insights not only improve our understanding and interpretation of the dataset but also suggest potential directions for further medical investigations in this field.

4.2.2.5 Validating Our Hypotheses Against the Ground Truth

Through To confirm and validate the hypotheses and conclusions derived from our dataset investigation, we reran our application to visualize the UMAP clustering with class labels displayed, as shown in Figure 4.18. We compared our findings against the established ground truth of the Breast Cancer dataset.

As illustrated in Figure 4.18, most data points in cluster C1 are labeled as *Malignant*. This observation corroborates our initial hypothesis regarding this cluster.

Next, we focused on cluster C2. Our initial hypothesis suggested that this cluster represents a homogeneous class label, specifically *Benign*. By comparing our results with Figure 4.18, we confirmed that our assumption about C2 was accurate.

Finally, we examined cluster C3. Comparing the results of our investigation with the ground truth revealed that the majority of data points within this cluster are labeled as *Benign*, aligning with our findings using DimVis.

Overall, these comparisons validate our initial hypotheses and demonstrate the reliability of our analysis.

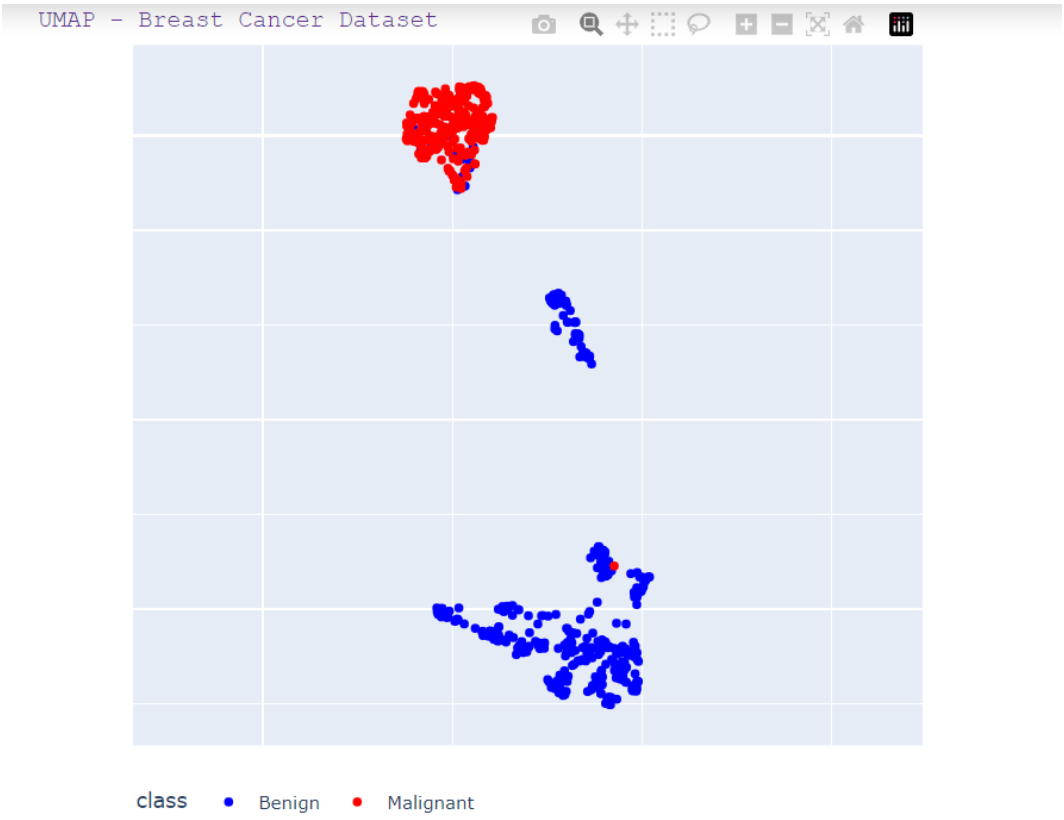


Figure 4.18 This figure displays a projection of the breast cancer dataset with hyperparameters set to 12 for Number of Neighbors and 0.3 for Minimum Distance. It is used to validate our findings, showing the same three clusters as before, now color-coded with the ground-truth labels.

5
APPLICATION 2: PIMA INDIAN DIABETES

5.1 DimVis in Practice: The Second Case Study

To demonstrate the versatility and effectiveness of our tool, we present another use case that explores its application within a different context. This case study highlights DimVis's capability to aid users, such as medical professionals and researchers, in understanding and interpreting data projections. In this instance, we focus on the Pima Indian Diabetes dataset to illustrate how DimVis can be employed to investigate the clustering of data points and derive meaningful insights. In the following section, we will detail the process and outcomes of the second use case, Pima Indian Diabetes, showing how DimVis fulfills its intended goals.

5.2 Use Case: Understanding a Projection

Sara, a medical doctor aspiring to specialize in diabetes diagnosis, is determined to identify individuals at risk for the condition well before symptoms appear. Recently, she discovered UMAP, a widely-used dimensionality reduction (DR) algorithm. Although she is excited about UMAP's potential, Sara harbors some doubts about its reliability. Therefore, she prefers using an interactive visualization tool like DimVis to assess how UMAP forms data clusters and understand the reasons behind their formation.

Sara plans to explore the Pima Indian Diabetes dataset, which she downloaded from the Kaggle repository [23]. This dataset includes 768 female patients of Pima Indian heritage, aged between 21 and 81, and classifies them into two categories: positive or negative for diabetes. The dataset contains 268 patients diagnosed with diabetes and 500 non-diabetic subjects. Each record in the dataset includes eight features: the number of pregnancies (*Pregnan*), age (*Age*), plasma glucose concentration (*Glucose*), diastolic blood pressure (*BloodPr*), skin thickness (*SkinThick*), insulin level (*Insulin*), body mass index (*BMI*), and a diabetes pedigree function (*DPF*) that measures genetic risk for diabetes.

5.2.1 Overall Accuracy

Sara selects her dataset from the dropdown menu in DimVis and begins examining the UMAP-generated results. Upon applying her selection, she observes four distinct and well-separated clusters generated by the UMAP algorithm, as depicted in Figure 5.1. However, she quickly notices that these clusters are not homogeneous with respect to their class labels; each cluster contains a mix of both *diabetic* and *non-diabetic cases*.

Understanding that DimVis excels at in-depth projection analysis using the EBM's Feature Importance Graph, Sara reviews the performance metrics of the supervised EBM model. The precision and recall values are 0.81 and 0.82, respectively (see Figure 5.1(c)). These relatively high EBM results, when compared to the projection shown in Figure 5.1(d), indicate that further investigation is warranted.

DimVis allows Sara to upload and utilize the dataset pertinent to her professional field, thus fulfilling Goal 3. By checking the validation metrics, she gains deeper insights into the dataset, further demonstrating the achievement of Goal 3.

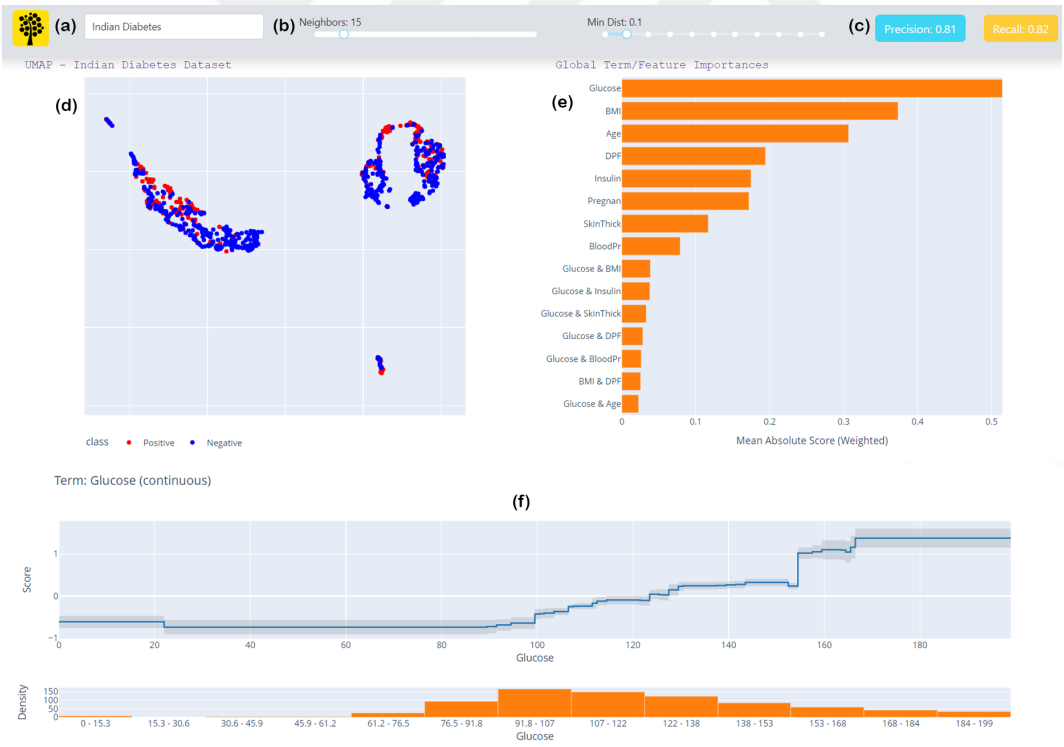


Figure 5.1 Illustrates the Pima Indian Diabetes dataset before any selection. (a) Displays the dropdown menu for selecting the dataset. (b) Illustrates the hyperparameter settings. (c) Shows the Precision and Recall values for the EBM classifier trained on the entire dataset. (d) Provides an overview of UMAP clusters, including data clusters and their corresponding class labels. (e) Demonstrates the features sorted by the EBM's importance ranking. (f) Shows the selected features dynamically presented to enhance user understanding and interpretation.

5.2.1.1 Adjusting Hyperparameters

Sara understands that the Neighbors and Min Dist hyperparameters are crucial in controlling the density of the projections (see Figure 5.1(b)). The Neighbors parameter determines how many data points are used to reflect the data density in a specific region, while Min Dist dictates how close data points must be in the projected space to be considered neighbors.

Eager to explore further, Sara adjusts these hyperparameters to observe how changes affect the clusters. She sets Neighbors to a smaller value, like 10, and increases Min Dist to 0.4. Although these adjustments result in different numbers and shapes of clusters, they still contain a mix of both class labels, which she finds confusing. After these unsatisfactory attempts, Sara decides to reset the hyperparameters and focus on interpreting each cluster separately using the EBM's Feature Importance Graph. Adjusting the hyperparameters to explore various visualizations to find a satisfactory one is a fundamental aspect of Goal 3, which has been achieved in this step.

5.2.1.2 Understanding Clusters

Sara Uses DimVis to validate the projections and identify and interpret similar groups of data points at the same time. Despite the clusters being mixed (because of carrying different class labels inside the same cluster or overlapping different class labels) and thus, initially challenging to interpret, Sara gains insights into the formation of data clusters by selecting them individually. She decides to evaluate each cluster one by one to understand how the features contribute to their formation and the reason behind this clustering. Before proceeding, she notes the feature rankings for the entire dataset, with *Glucose*, *BMI*, *Age*, and *DPF* being the top four (cf. Figure 5.1(e)). The most influential feature, which is *Glucose*, significantly impacts the differentiation between positive and negative diabetes cases as shown in Figure 5.1(e). In addition, low glucose values are negatively associated with diabetes, while high values are prevalent among diabetic patients.

Sara begins the exploration of clusters by drawing a polyline using the lasso tool around them. By doing so she observes that the tool dynamically updates the Feature Importance Graph (EBM plot) based on her selections. The ability to select different groups of data points to explore and interpret various visualizations is a key aspect of Goal 1 and this means that this goal has been successfully achieved.

5.2.1.3 Starting the Analysis with Selecting the Clusters to Gain More Insight

When Sara draws her first selection around cluster C1 for this dataset, the Feature Importance Ranking Graph shows a significant change which is displayed in Figure 5.2. The top four important features now are *Glucose*, *Number of Pregnancies*, *Skin Thickness*, and *Insulin*, which is totally different from the total feature ranking.

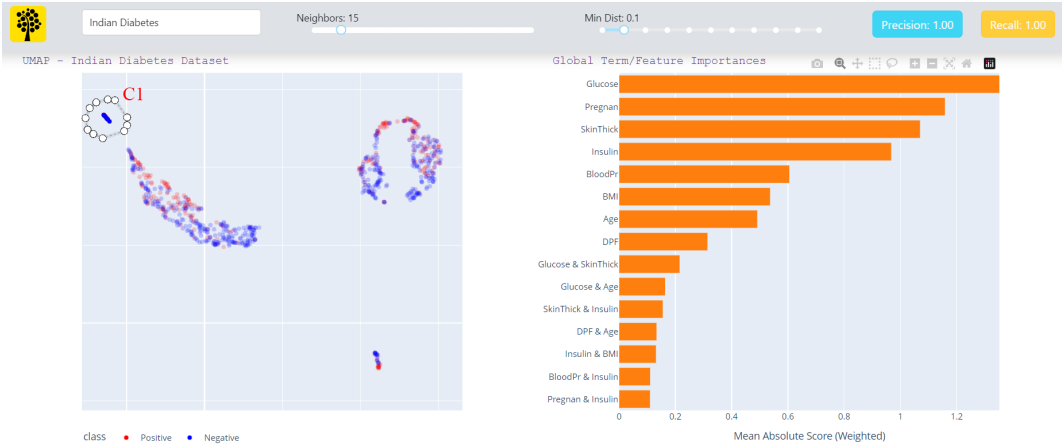


Figure 5.2 As illustrated in the figure, C1 is selected as the first cluster and it is depicted that the Feature Importance Ranking Graph has changed significantly. The first four important features in the EBM graph are highly important rather than the rest of them.

As Sara notices the significant change in the EBM graph with her selection, her curiosity grows. She decides to select all four clusters, examine the Feature Importance Graph, and review the feature rankings for each cluster. Based on these rankings, she plans to investigate the most important features and see if any feature holds the top rank in multiple clusters simultaneously.

Having said this, Sara navigates to the next cluster and selects C2, as shown in Figure 5.3. The EBM is re-trained and recalculated for the new selection, resulting in a new ranking: *Insulin*, *Skin Thickness*, *Skin Thickness and Insulin*, and *Insulin and DPF*.

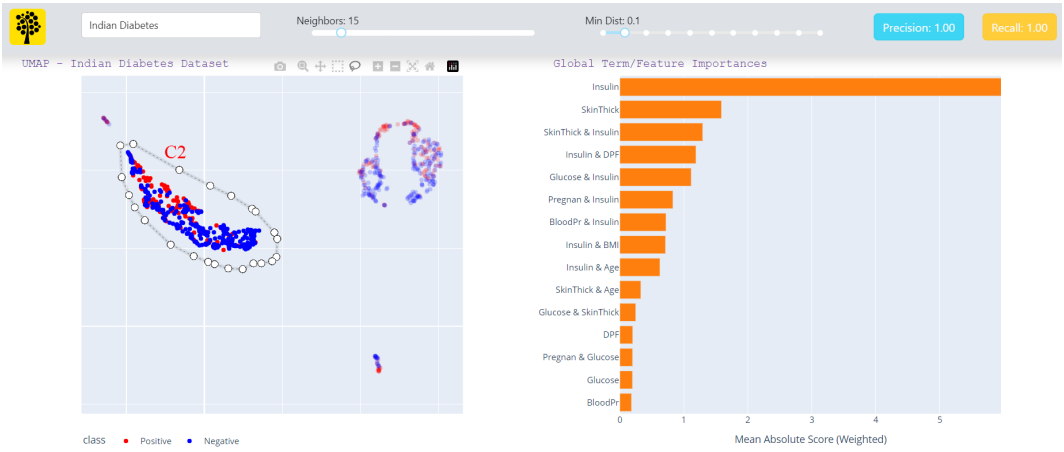


Figure 5.3 As displayed in the figure, C2 is selected cluster. Again, the feature importance graph has changed remarkably and the ranking is different. For this selection, *Insulin*, *Skin Thickness*, *Skin Thickness and Insulin* and *Insulin and DPF* are the most important four features.

This surprises her, and she becomes more eager to continue. She then selects cluster C3, which shows a completely different Feature Ranking: *Insulin*, *Skin Thickness*, *Blood Pressure*, and *Age* (Figure 5.4). Noticing that *Insulin* remains the most important feature, Sara wonders if it will be the top feature for the last cluster, C4 which is the most interesting cluster because of its shape.

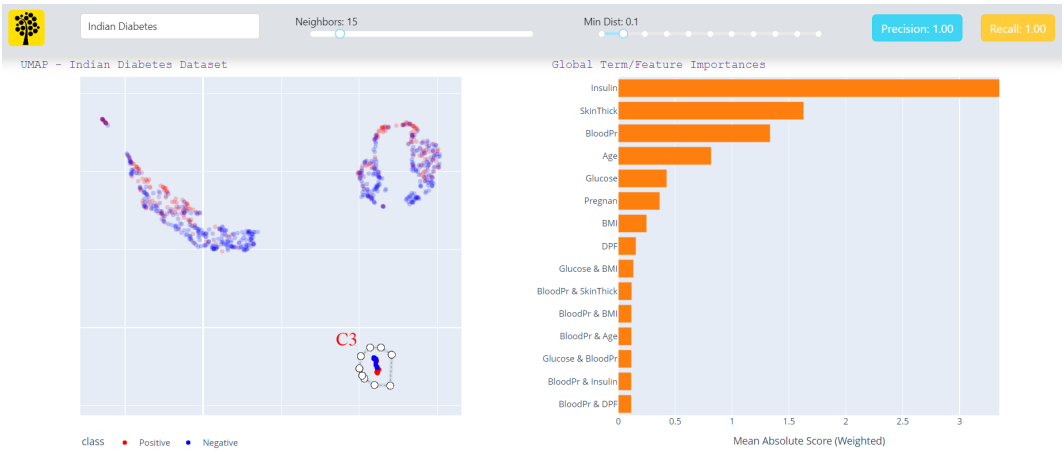


Figure 5.4 Depicted the figure, C3 is the selection. For this new selection, the feature importance graph has changed again notably and the most four important features are: *Insulin*, *Skin Thickness*, *Blood Pressure*, and *Age*, respectively. The is a huge gap between *Insulin* and other features for this selection also.

She selects cluster C4 and sees a new Feature Importance Graph, where *Insulin* is again the most important feature (Figure 5.5). Additionally, there is a significant gap between the importance of the first and second features. Recognizing that *Insulin*

consistently appears as the most important feature in clusters C2, C3, and C4, Sara decides to explore *Insulin* in depth.

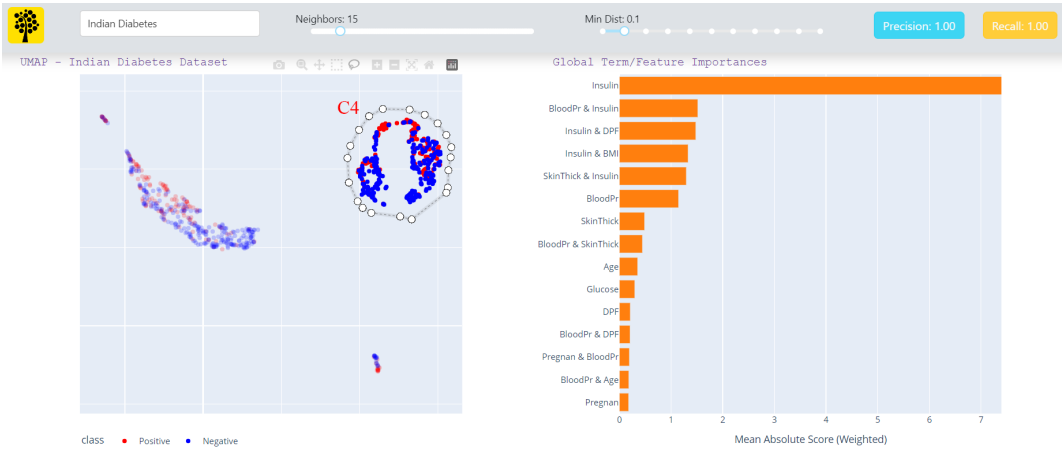


Figure 5.5 Cluster C4 is selected in this figure. The feature rankings has been changed once more and the most important feature is *Insulin* again which is remarkably distinguished among the rest of the features in the EBM plot.

Before delving into the Individual Feature Graph for *Insulin*, the shape of cluster C4 catches her attention. It appears as two distinct clusters divided by a thin line into subclusters C5 and C6. She selects cluster C5 and sees a significant change in the Feature Importance Graph (Figure 5.6). Interestingly, *Insulin* is again the top feature, followed by *Skin Thickness*, with a large gap in importance between these two and the rest.

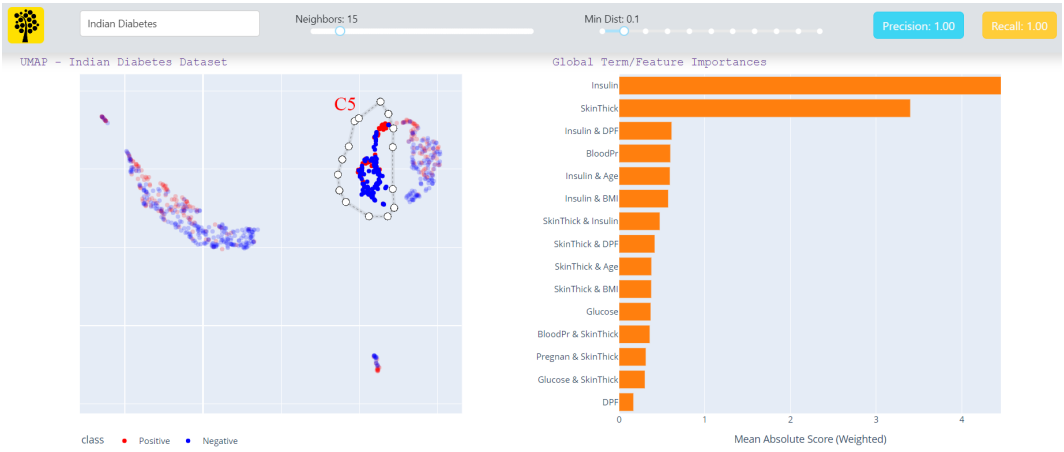


Figure 5.6 The selected cluster C5, is subcluster of a another cluster named C4. This subcluster is the left half of the cluster C4. These clusters are examined separately to find the reason behind their division into to separate clusters. On the other hand, for this selection, the most two important features have extremely different importance values from the rest of the feature in the provided feature ranking list.

Sara then investigates cluster C6 separately to see what is the reason that these instances have shaped cluster C4 and also inside this cluster they have been divided into two subclusters. The EBM plot changes as shown in Figure 5.7, showing different rankings, but the top two features remain the same as in C5, although in reverse order: for C5, *Insulin* is first and *Skin Thickness* second, while for C6, *Skin Thickness* is first and *Insulin* second. This reversal in feature ranking explains why C4 split into two subclusters.

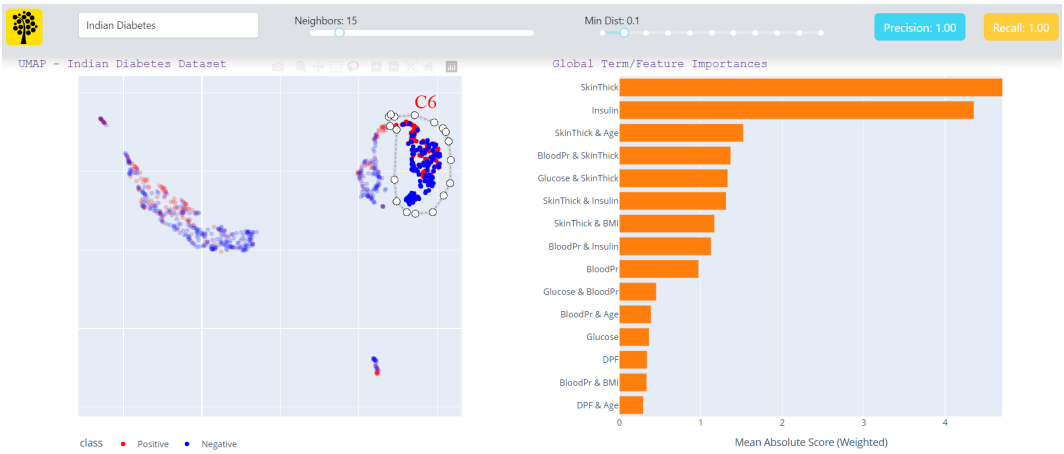


Figure 5.7 The selected cluster C6 is the right part of the larger cluster C4. In the previous figure, Sara explored the other subcluster, and with this selection, she investigates the C6 subcluster. For this selection, the first two important features are the same as in the other subcluster; however, their order is reversed.

With these insights, Sara decides to start her in-depth investigation with *Insulin*. She begins with C6 and clicks on *Insulin* to examine its Individual Feature Graph for this cluster, as shown in Figure 5.8. It shows very low *Insulin* values for the selected parts (values below zero correspond to the selection) and a linear graph for the rest of the instances in the dataset. This gives her the idea that the majority of the datapoints inside this cluster carry very low values (almost zero) for the *Insulin* feature. In addition, the uncertainty of the model is very high for unselected parts as the gray background shows a wide interval. Furthermore, the Density graph shows a very high population for *Insulin* values around our selection.

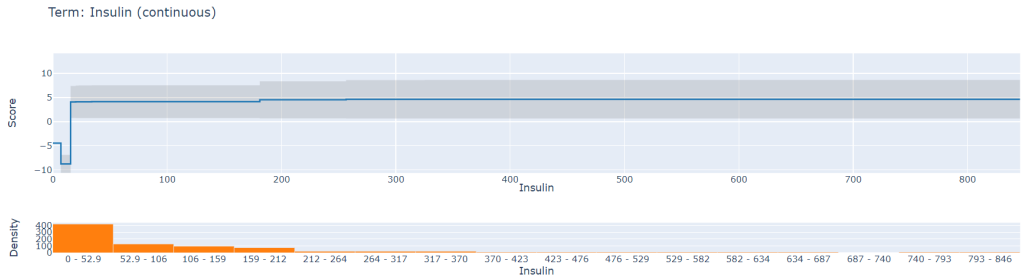


Figure 5.8 As illustrated in the figure, *Insulin* Individual Feature Graph for C6 is displayed. It shows very low values for the selection and also a linear trend for the rest of the datapoints.

In order to be able to make a good and effective comparison between C5 and C6, Sara wants to check the second important feature, *Skin Thickness*. So, she clicks on this feature in the EBM plot. As shown in Figure 5.9. The first part of this figure (score values above zero) shows the unselected portions of the dataset. This part demonstrates very high uncertainty, which is one of the important aspects that catches Sara's attention. Another observation she makes is the linear trend of this feature for her selection. Although satisfied with this investigation, Sara realizes that to fully understand the cluster formation, she must also examine these two features in the other subcluster. Before doing so, she checks the Density graph and notices that the distribution of this feature is concentrated in the first half of the graph. This is another crucial insight she gains from her in-depth investigation.

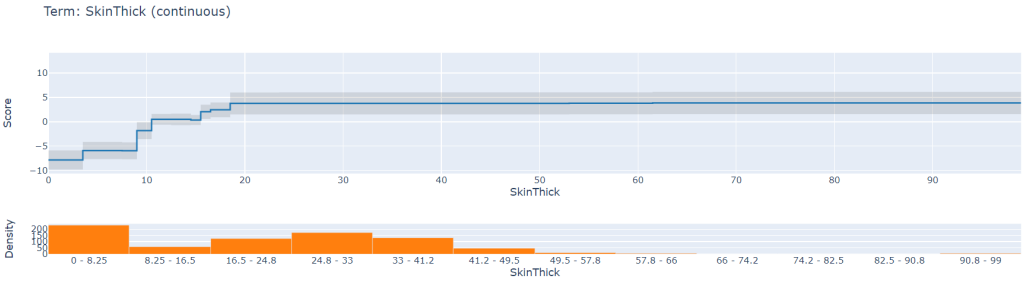


Figure 5.9 The Individual Feature Graph of *Skin Thickness* for C6 is shown in this figure. It has an increasing trend at first which is followed by a linear trend.

After exploring the two most important features of C6, Sara reselects the subcluster C5 to investigate further and understand the reasons behind the division and formation of these subclusters. Upon selecting C5 again, she clicks on the *Insulin's* Individual Feature Graph as displayed in Figure 5.10. Although the graph shows slight differences, it is very similar overall. In this Individual Feature Graph, the selected data points have very low *Insulin* values, just like the previous selection. For the rest of the dataset, *Insulin* follows a linear trend, and the model shows high

uncertainty. Sara realizes that since *Insulin* displays the same trend for both selected and unselected parts of the dataset in both subclusters, it is likely the feature that has brought both subclusters together, forming the larger cluster C4.

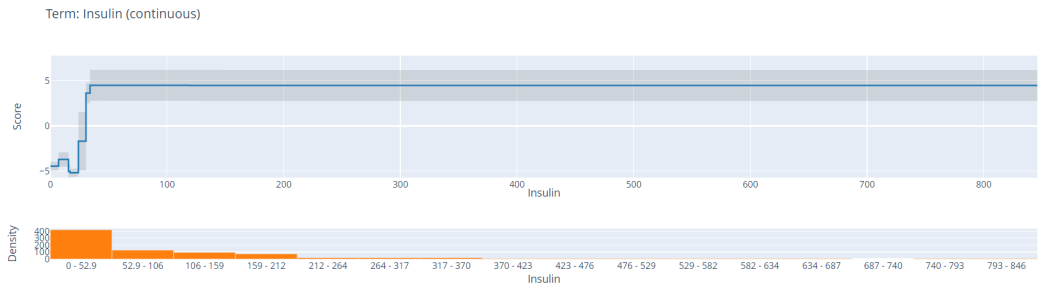


Figure 5.10 The Individual Feature Graph of *Insulin* for C5 is displayed in this figure. It shows very low values for the selection and also a linear trend for the rest of the datapoints.

Happy and satisfied with her findings, Sara clicks on the *Skin Thickness* feature to explore its Individual Feature Graph, hoping to find the reason behind the division of these two subclusters. The relevant graph is illustrated in Figure 5.11. She observes an almost exact opposite trend for *Skin Thickness* compared to her previous selection (C6). Sara concludes that this feature is the reason why the two subclusters were created within the larger cluster C4.

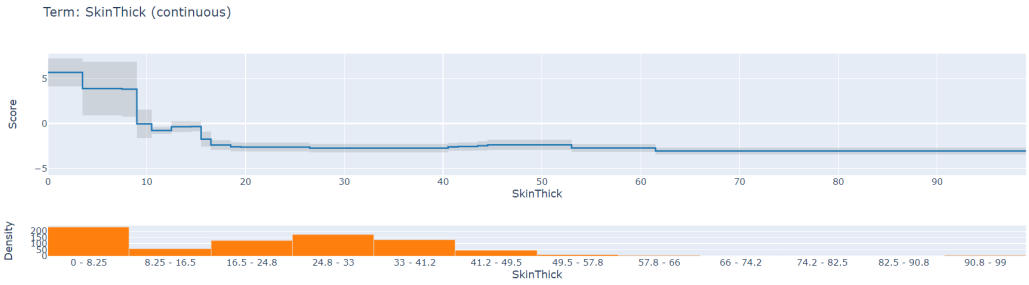


Figure 5.11 The Individual Feature Graph of *Skin Thickness* for C5 is displayed in this figure. This is the second important feature for C5 selection and has wide intervals of uncertainty for selected data points.

5.2.1.4 Deep Dive into Insulin Feature

After this satisfactory investigation, Sara decides to continue exploring *Insulin*, knowing it is the most important feature for the C2, C3, and C4 clusters. She initiates an in-depth analysis of *Insulin* by selecting C4 again and clicking on the *Insulin* feature. The Individual Feature Graph appears at the bottom of the page and is illustrated in Figure 5.12. She notices an increasing trend at the beginning of the graph: the first half, representing the selected points, carries negative values,

while the second half, representing the unselected parts of the dataset, carries positive values. The selected data points have *Insulin* levels up to ≈20 in the blood, corresponding to a score value up to 0. In contrast, the effect remains constant for the rest of the data points (unselected ones) due to the straight line.

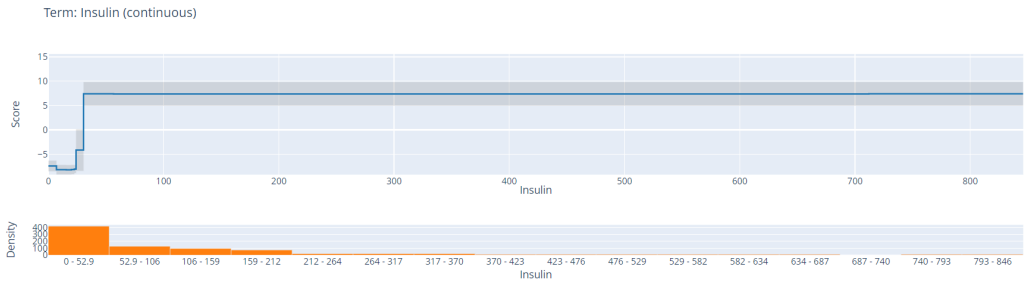


Figure 5.12 The Individual Feature Graph of *Insulin* for C4 is illustrated in this figure. It displays the *Insulin* trend for this selection. The value of *Insulin* is so low for the selected datapoints and then it demonstrates a linear trend for the rest of the dataset. This graphs shows very similar trend to the *Insulin* graph of C5 and C6 selection which is another way that supports Sara's hypothesis about the feature which gathered those two subclusters and shaped a bigger one.

After investigating the *Insulin* feature's Individual Feature Graph for C4, Sara moves to C3 and reselects that cluster again. The Feature Importance Graph changes and then, she clicks on the *Insulin*. She sees a totally different graph for *Insulin* this time as shown in Figure 5.13. For this selection, the *Insulin* shows an extremely low value (almost zero). In contrast, for the unselected data points, the trend of *Insulin* is linear. As the number of selected data points in C3 is very low and the majority of those data points belong to the non-diabetic class label, a very low value for *Insulin* feature and a narrow selection window for its value are to be expected.

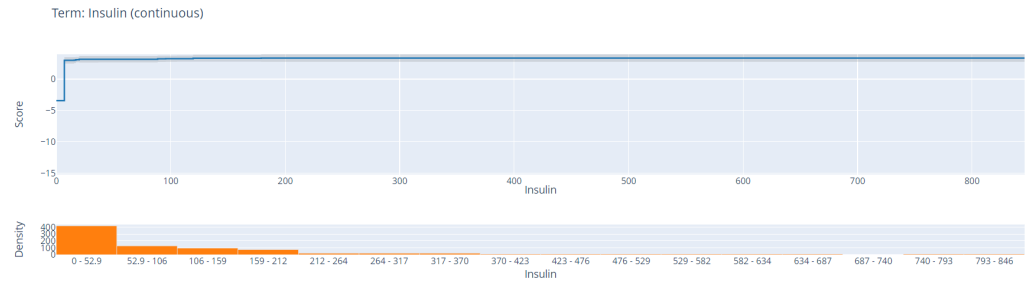


Figure 5.13 The Individual Feature Graph of *Insulin* for C3 is illustrated in this figure. It displays the Insulin trend for this selection. The selected cluster is so small and that is why the area that corresponds to the selection is so small.

Another different plot has been provided by EBM to Sara for her new selection

which is C2. Her selection is illustrated in the middle of the figure and the beginning and the end of the plot demonstrate the unselected datapoints of the dataset. Furthermore, the model exhibits high uncertainty across the entire dataset, affecting both selected and unselected data points. This discovery is particularly surprising and noteworthy for Sara.

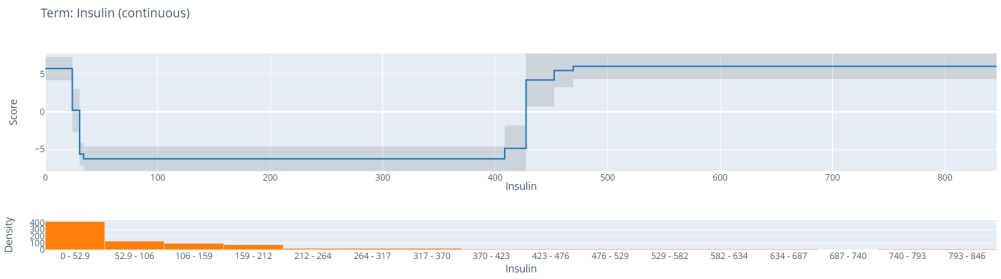


Figure 5.14 The Individual Feature Graph of *Insulin* for C2 is illustrated in this figure. It shows a decreasing trend at first, then a linear trend, after that it displays an increasing and again a linear trend. The selection C2 corresponds to Insulin values ≈20 up to ≈420. Any section out of this interval belongs to the rest of the dataset.

Finally, she returns to C1 and selects that cluster once more. In the newly ranked feature graph, she clicks on *Insulin* and examines its Individual Feature Graph. The results for *Insulin* are very different and intriguing for this selection as shown in Figure 5.15. Sara observes that the selected parts of the dataset and the unselected parts appear to be exact opposites of each other. Another noteworthy observation is that the uncertainty levels for both selected and unselected data points are quite high and consistent. This indicates that the model lacks certainty across all the data points in the dataset.

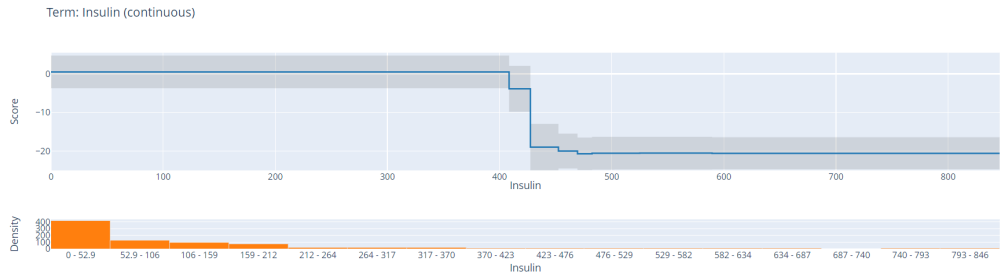


Figure 5.15 The Individual Feature Graph of *Insulin* for C1 is illustrated in this figure. The selection covers from *Insulin* values ≈430 up to the end and the model shows a high uncertainty for the whole dataset prediction.

5.2.1.5 Concluding Observations of Sara

Sara has finished her investigation now and she is satisfied with her exploration and the insight she has gained. She feels gratified having discovered informative patterns and interpretations that validate her initial thoughts about *Insulin's* significance in the dataset. Encouraged by these findings, she is eager to explore new diabetes-related datasets and plans to extend her analysis using this tool to further her research and become an expert in this field.

6
APPLICATION 3: HYPERSPECTRAL IMAGES

6.1 HyEx in Action

To showcase the applicability and effectiveness of our tool in a completely different domain, we present a use case involving the Indian Pines hyperspectral images dataset. Hyperspectral images are known for their high dimensionality, with each pixel containing a vast number of spectral features. In hyperspectral image processing, techniques such as Principal Component Analysis (PCA) and Minimum Noise Fraction (MNF) are commonly used as pre-processing steps to reduce dimensionality, and edge-preserving filters like bilateral filters are applied to enhance feature extraction. These traditional approaches, however, often fall short in providing clear and interpretable insights into the formation of data clusters.

This example demonstrates HyEx's ability to assist users, such as data analysts, in evaluating the formation of data clusters in a supervised setting within the hyperspectral imaging domain. With 16 class labels provided, this use case highlights HyEx's capability to handle complex, labeled datasets effectively, addressing the limitations of conventional methods. By employing advanced dimensionality reduction and feature importance techniques, HyEx provides a more nuanced and interpretable analysis of hyperspectral data.

In the following section, we will detail where and how these goals have been achieved, illustrating how HyEx fills the existing gaps in hyperspectral image analysis and offers significant improvements over traditional methods.

6.2 Use Case: Analyzing Hyperspectral Image Data with Ground-Truth Labels

For this use case, we select the Indian Pines hyperspectral images dataset from the Hyperspectral Remote Sensing Scenes repository [24] to demonstrate how HyEx assists users in exploring and analyzing data clusters based on various dimensions.

Unlike the previous example, this dataset includes ground-truth labels, allowing us to evaluate HyEx's effectiveness in a supervised setting.

The Indian Pines dataset consists of hyperspectral images labeled with 16 different classes. These classes are *Alfalfa* (46 samples), *Corn-notill* (1428 samples), *Corn-mintill* (830 samples), *Corn* (237 samples), *Grass-pasture* (483 samples), *Grass-trees* (730 samples), *Grass-pasture-mowed* (28 samples), *Hay-windrowed* (478 samples), *Oats* (20 samples), *Soybean-notill* (972 samples), *Soybean-mintill* (2455 samples), *Soybean-clean* (593 samples), *Wheat* (205 samples), *Woods* (1265 samples), *Buildings-Grass-Trees-Drives* (386 samples), and *Stone-Steel-Towers* (93 samples).

Initially, we will analyze the data clusters using the provided ground-truth labels to validate our findings. This approach will enable us to observe how HyEx can effectively explain the formation of these clusters and highlight the most important features (Figure 6.1). By using this dataset, we achieve Goal 3, demonstrating HyEx's capability to select and analyze an appropriate dataset from various options, thus providing valuable insights into complex hyperspectral data.

6.2.1 Data Prepration

As the Indian Pines dataset is so complicated (shown in Figure 6.1), to effectively prepare the Indian Pines hyperspectral dataset for clustering analysis, we applied four different data reduction techniques in Matlab before utilizing UMAP. The primary goal of these techniques is to reduce the dimensionality of the dataset while retaining its essential features and characteristics. Given the high dimensionality and complexity of hyperspectral images, a straightforward dimensionality reduction may not be sufficient to capture all relevant information. Therefore, to enhance the interpretability of the data, we created different combinations of dimensionality reduction and feature reinforcement methods. Specifically, we used MNF and PCA for dimensionality reduction and coupled them with Bilateral filtering for feature reinforcement.

By employing these methods, we aimed to tackle the inherent noise and redundancy in the hyperspectral images, which often obscure meaningful patterns and structures. MNF and PCA are effective in isolating the most significant components of the data, but when combined with Bilateral filtering, they can better preserve spatial details and edge information, which are crucial for accurate interpretation and analysis.

We began by individually applying MNF and PCA to reduce the number of

dimensions, making the dataset more manageable and highlighting the primary variance within the data. Following this, we integrated Bilateral filtering to reinforce these reduced-dimensional datasets. The Bilateral filter plays a key role in smoothing while preserving important edges, thereby maintaining the integrity of the spatial information.

This multi-faceted approach allowed us to systematically explore which preprocessing combination provided the most clarity and interpretability. By comparing the results from MNF alone, PCA alone, MNF with Bilateral filtering, and PCA with Bilateral filtering, we could identify the optimal method that balances dimensionality reduction with feature preservation.

Here, we describe each technique in detail, discuss their specific application to our dataset, and present the results for comparison. This approach allowed us to evaluate which preprocessing combination was more interpretable and effective for subsequent analysis. Through this comprehensive evaluation, we were able to identify the preprocessing steps that most effectively prepared the Indian Pines hyperspectral dataset for meaningful clustering analysis using UMAP.

In the subsequent sections, we will delve deeper into the specifics of each technique, providing a detailed explanation of their application and demonstrating how each preprocessing method impacted the dataset. This analysis enabled us to select the most appropriate preprocessing strategy, thereby facilitating a more accurate and insightful clustering analysis.

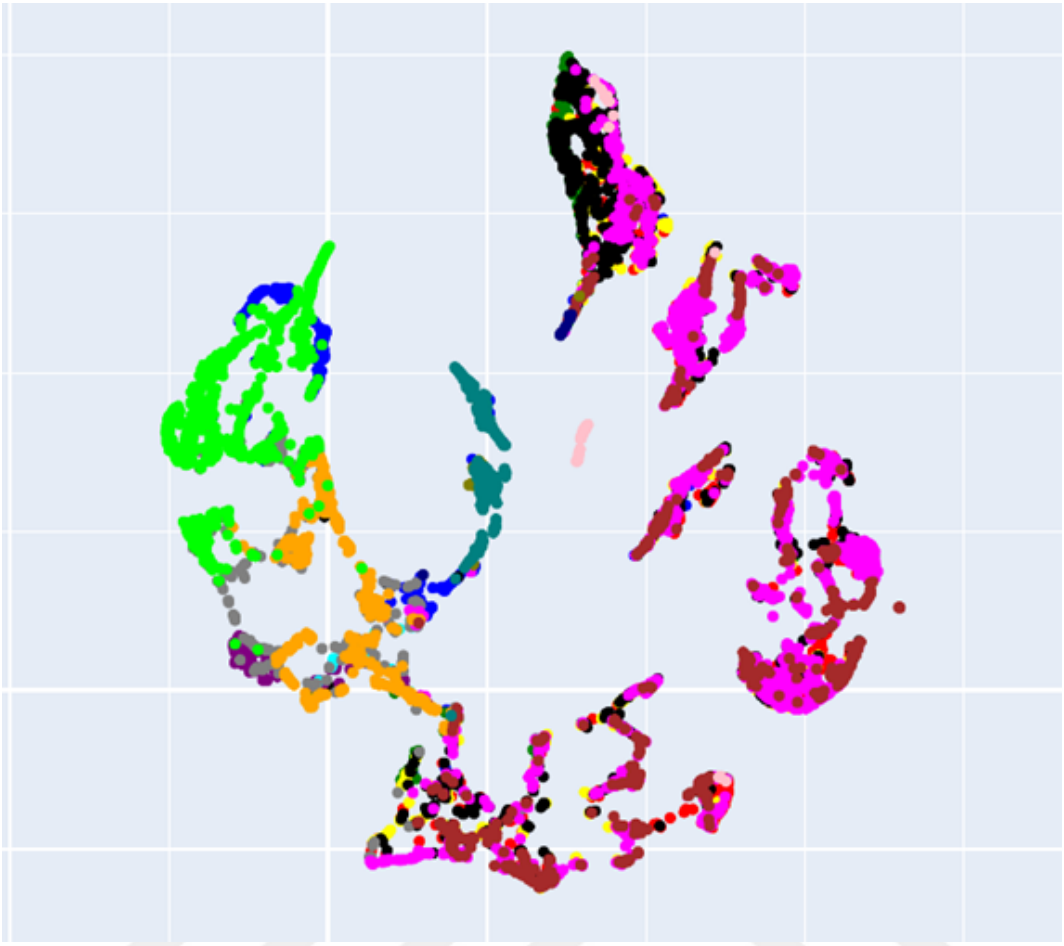


Figure 6.1 Displays the Indian Pines dataset before any feature reduction. This initial visualization highlights the complexity and overlapping different class labels of the data, underscoring the need for effective dimensionality reduction techniques to simplify the analysis and improve the clarity of the resulting clusters. The original dataset includes hyperspectral images with ground-truth labels for 16 different classes, which will be crucial in validating the effectiveness of the reduction techniques and subsequent clustering.

6.2.1.1 Minimum Noise Fraction (MNF)

Minimum Noise Fraction (MNF) is a data reduction method designed to improve the signal-to-noise ratio by transforming the data into a new coordinate system. This technique is particularly beneficial for hyperspectral images, as it reduces noise and emphasizes meaningful spectral information. MNF works by separating the noise from the signal and discarding the noise-dominated components. This results in cleaner data that highlights the significant spectral variations. Figure 6.2 displays the result of this technique on our dataset.

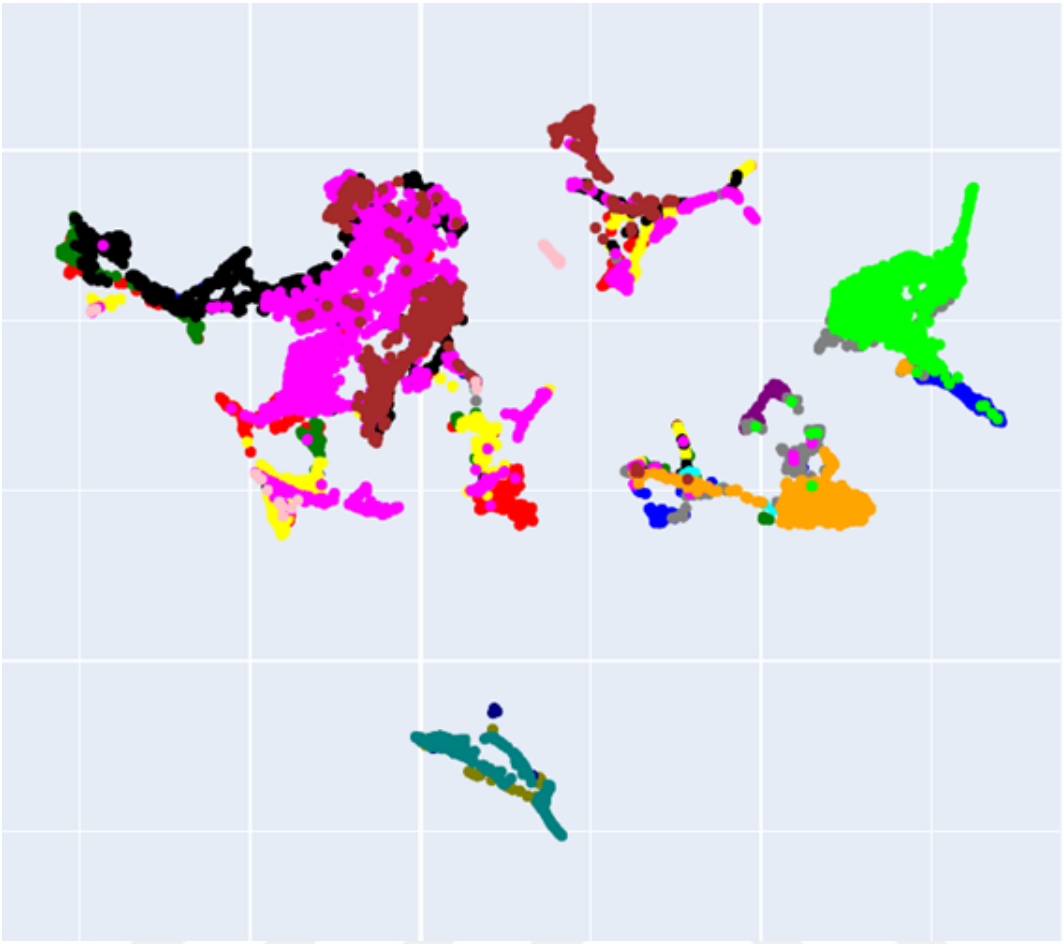


Figure 6.2 Displays the result of applying MNF to the Indian Pines dataset. The transformed data shows a reduction in noise, making the underlying patterns more discernible.

6.2.1.2 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a widely used dimensionality reduction technique that transforms the original high-dimensional data into a lower-dimensional space. This transformation is achieved by projecting the data onto a set of orthogonal axes (principal components) that capture the maximum variance in the data. PCA simplifies complex datasets and highlights the most significant patterns, making it easier to analyze and interpret the data. Figure 6.3 illustrates the results of PCA on our dataset.

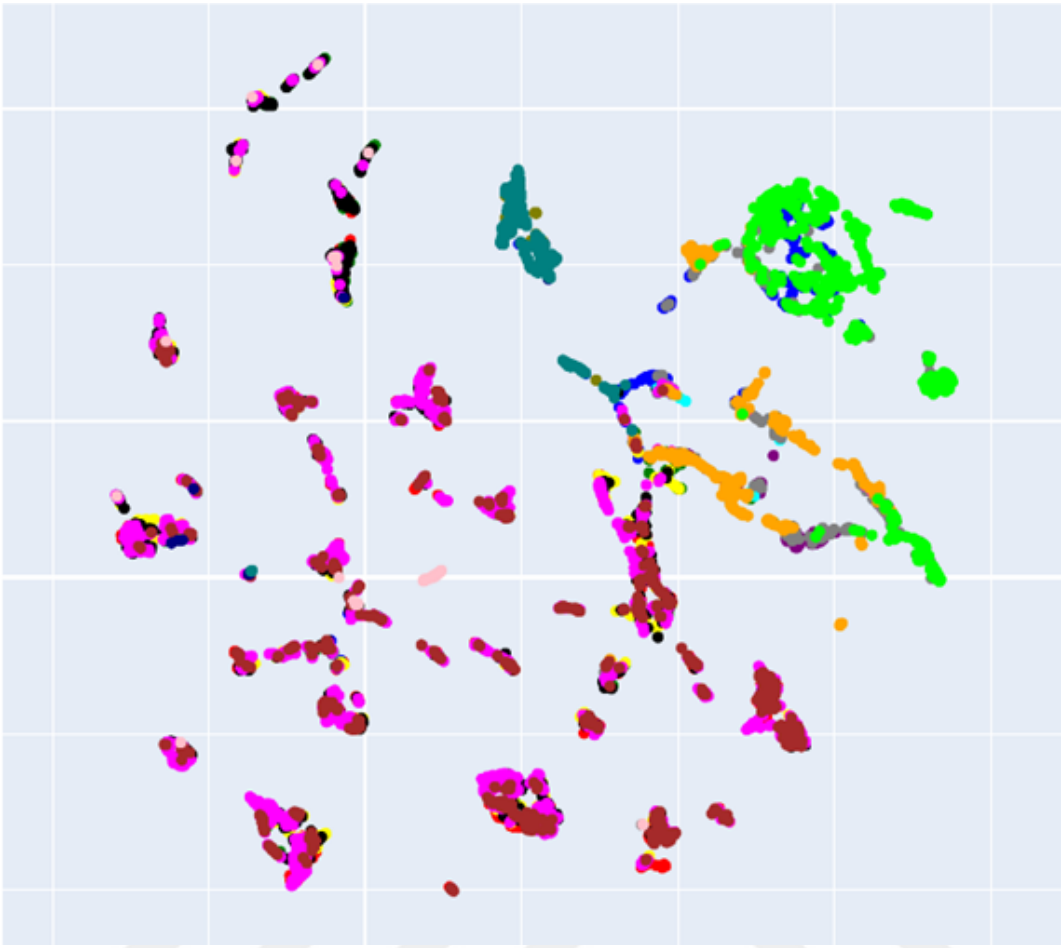


Figure 6.3 Shows the result of applying PCA to the Indian Pines dataset. The first few principal components capture most of the variance, providing a simplified yet informative representation of the data.

6.2.1.3 MNF Coupled with Bilateral (MNF+Bilateral)

MNF with Bilateral combines the advantages of MNF with bilateral filtering, a non-linear, edge-preserving, and noise-reducing smoothing filter. This technique not only reduces noise and enhances signal but also preserves important spatial features and edges in the data. MNF coupled with Bilateral is particularly effective for hyperspectral images where maintaining spatial integrity is crucial. In Figure 6.4, you can see the results of MNF coupled with Bilateral on our dataset.

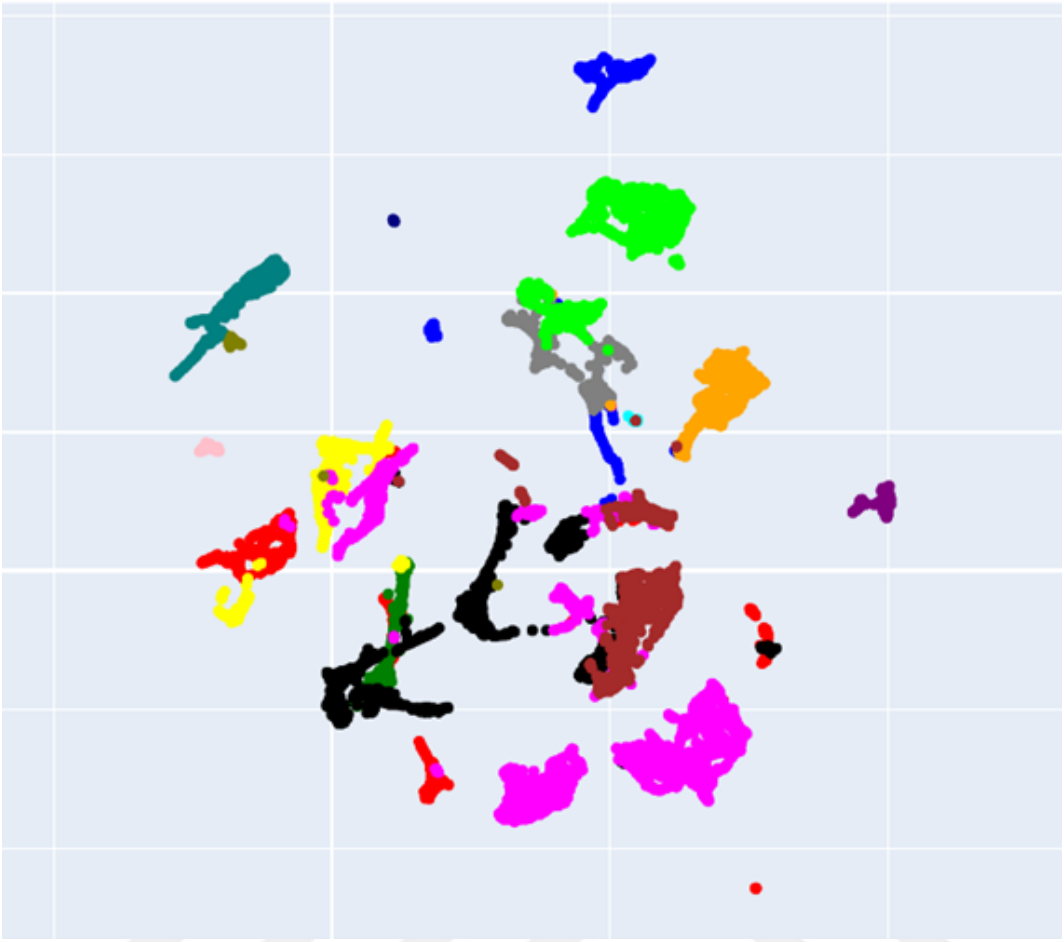


Figure 6.4 Presents the result of applying MNF coupled with Bilateral to the Indian Pines dataset. The data exhibits enhanced signal clarity with well-preserved edges, making it suitable for subsequent clustering analysis.

6.2.1.4 PCA Coupled with Bilateral (PCA+Bilateral)

Similar to MNF coupled Bilateral, PCA coupled with Bilateral combines PCA with bilateral filtering. By applying PCA to reduce dimensionality and bilateral filtering to preserve edges and reduce noise, this technique aims to retain essential spectral and spatial characteristics of the hyperspectral data. You can see the results in Figure 6.5.

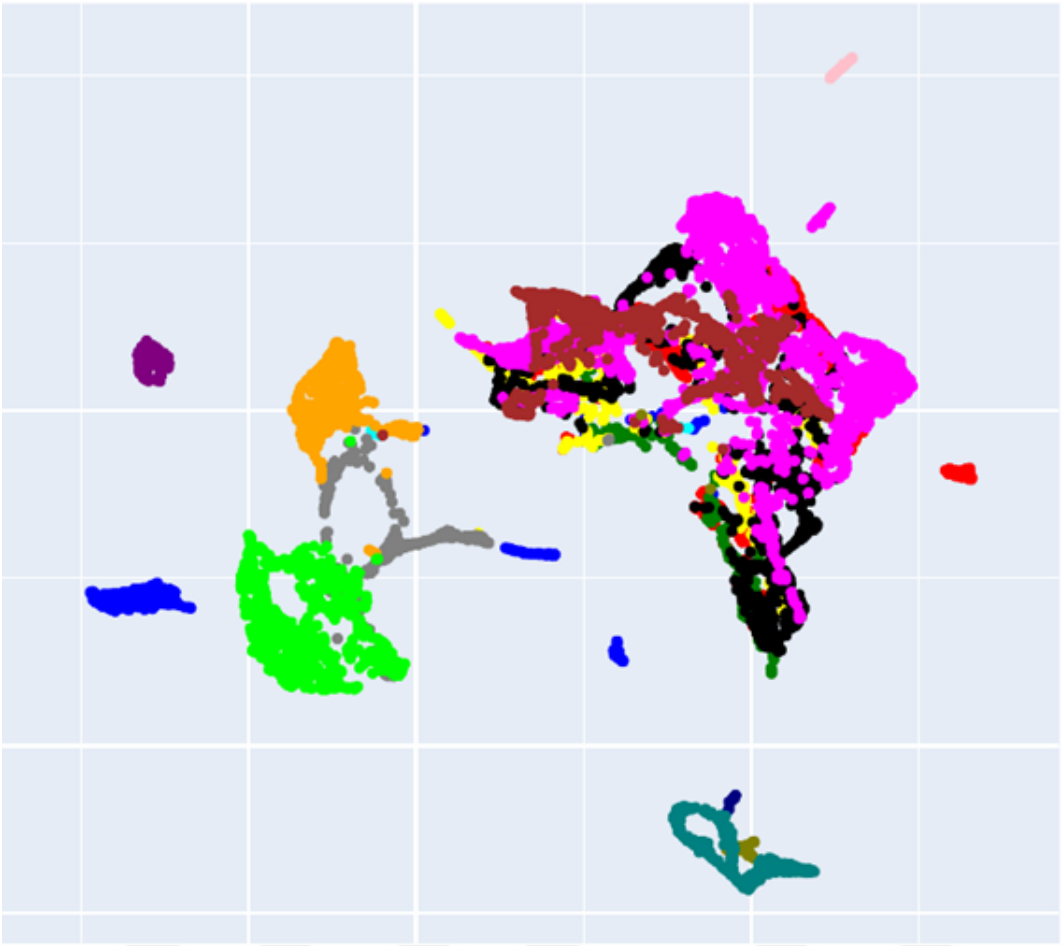


Figure 6.5 Illustrates the result of applying PCA coupled with Bilateral to the Indian Pines dataset. The data retains significant spectral information while minimizing noise, providing a balanced and clear representation for analysis.

6.2.1.5 Comparison and Selection of the Best Technique

Each data reduction technique was evaluated based on its ability to prepare the Indian Pines dataset for effective clustering with UMAP. The results of each technique are depicted in Figure 6.2 through Figure 6.5. Upon comparison, MNF coupled with Bilateral was found to provide the most distinct and well-separated clusters. This method effectively reduced noise while preserving important spatial features, resulting in clearer and more interpretable clusters. You can see the side-by-side comparison of all techniques in Figure 6.6.

By assessing the outcomes from each data reduction technique, we concluded that MNF coupled with Bilateral offers the best results for our dataset. Its ability to enhance signal clarity and maintain spatial information makes it the most suitable choice for preparing the Indian Pines hyperspectral data for clustering analysis with UMAP. As shown in Figure 6.6, MNF coupled Bilateral significantly improved the visualization and interpretation of the data clusters, affirming its effectiveness in our

analysis.

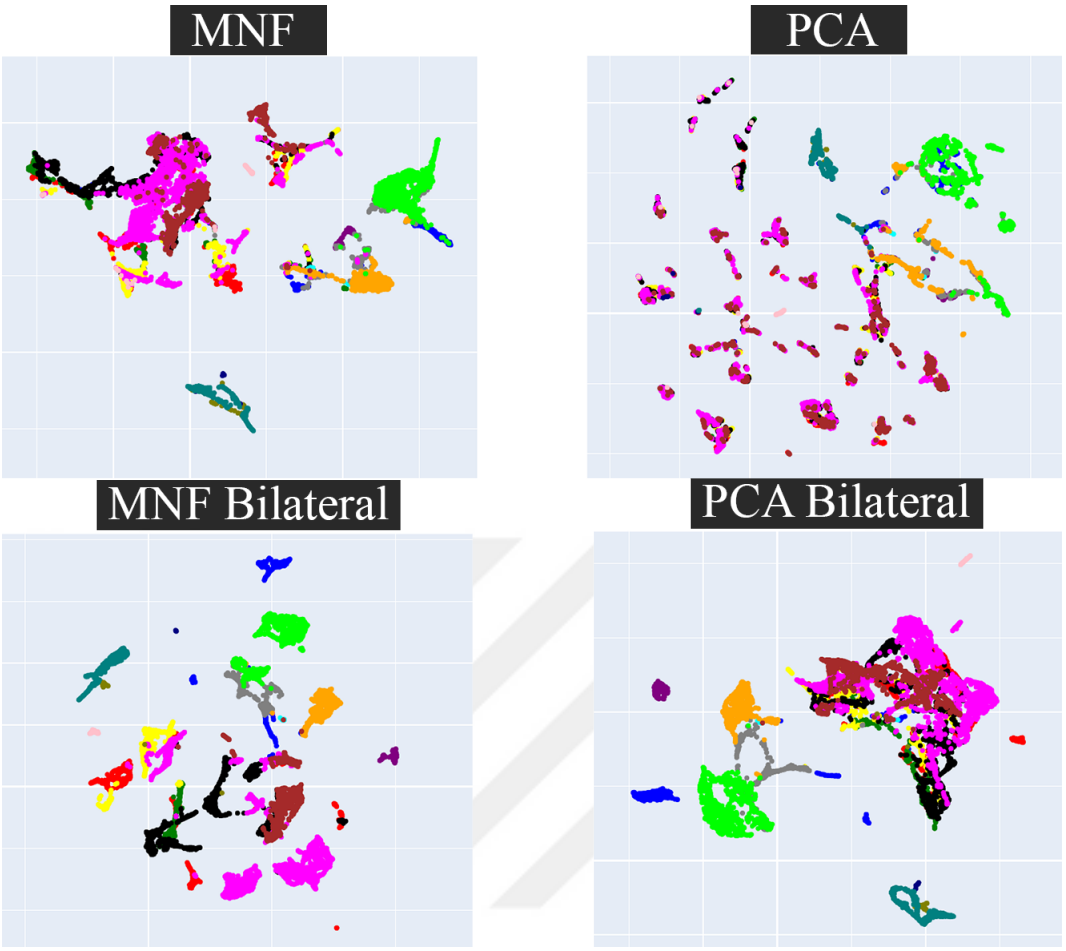


Figure 6.6 Comparison of UMAP results applied to the Indian Pines dataset after each feature reduction technique has been applied.

6.2.2 Investigating Cluster Formations

As shown in Figure 6.7, the application of UMAP on the Indian Pines dataset, after employing the MNF coupled with Bilateral feature reduction technique, reveals multiple clusters. Among these, the most distinct or interesting clusters which could be effective on dataset interpretation were selected for further investigation and labeled C1 through C6.

These selected clusters (C1 to C6) represent the best separations we could achieve, though some overlaps still exist. Overlapping clusters were intentionally excluded from our study when possible to minimize confusion and enhance the clarity of our analysis. The separation and shape of these clusters highlight the effectiveness of UMAP hyperparameters in delineating the data into meaningful groups.

By focusing on these clusters, we aim to understand the reasons behind the

clustering and the factors that shape these clusters. Despite some overlap, the distinctness of these clusters provides a robust basis for further exploration.

Examining these chosen clusters allows us to gain deeper insights into the underlying structure of the Indian Pines dataset. To achieve this, we will start with C1 and move forward to C6, highlighting the important factors affecting the formation of these clusters.

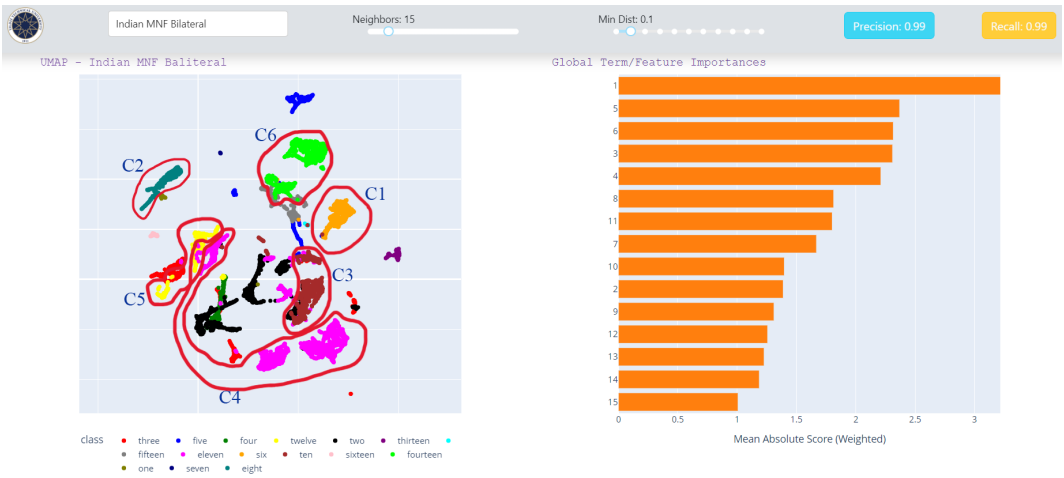


Figure 6.7 Depicts the six data clusters generated by UMAP (left) and a Feature Importance Graph for the most important features as identified by the EBM model trained on the entire dataset (right). By selecting each of the clusters using the lasso tool, we can observe the changes in EBM's Feature Importance Graph.

6.2.2.1 Analyzing Cluster C1

We begin our analysis by selecting cluster C1. The EBM plot re-trains and updates immediately, highlighting the differences between the selected data points and the rest of the dataset. This selection significantly changes the ranking of the features (see Figure 6.8). The top two important features are: *5th* band and *8th* band of the hyperspectral image. The ability to select different clusters and retrain the EBM model to generate various visualizations of ranked features aligns with the objectives of Goal 1.

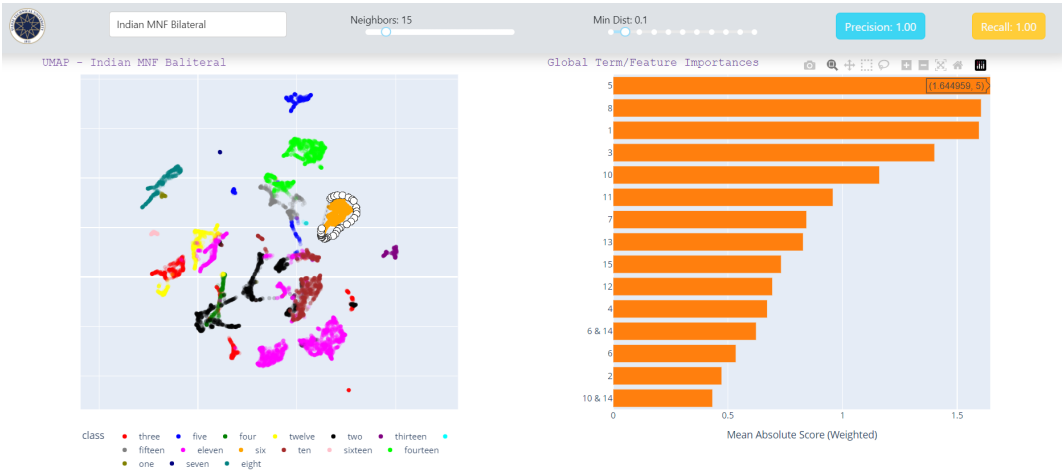


Figure 6.8 Selecting C1 retrains the EBM model to highlight the differences between the datapoints of cluster C1 and the rest of the dataset. The top features identified are *band 5* and *band 8*.

Upon selecting C1, we turn our attention to the validation metrics. The EBM proxy model performs exceptionally well in explaining the data points within this cluster, achieving perfect scores of 1 (or 100%) for both precision and recall. These optimal performance metrics underscore the model's robustness and reliability.

To delve deeper, we begin with the two most important features for this cluster: *band 5* and *band 8* of the hyperspectral image. Clicking on the bar representing *band 5* in the Feature Importance Graph reveals a Score graph and a Density graph below, allowing us to analyze this feature in isolation (shown in Figure 6.9).

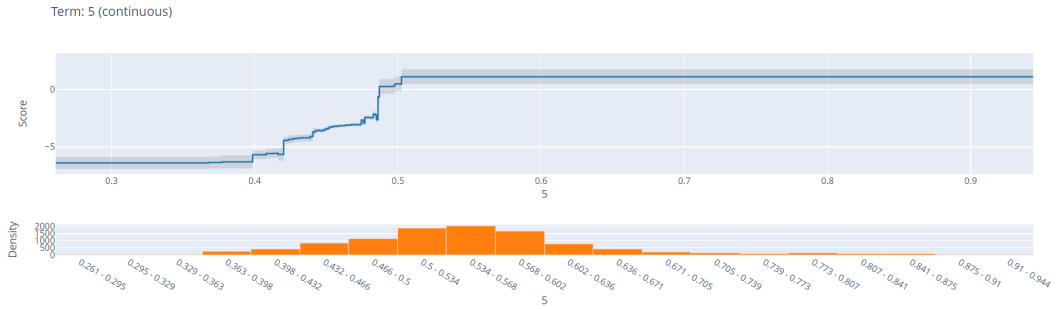


Figure 6.9 Individual Feature Graph for *5th* band of the hyperspectral image dataset in our first cluster which is C1. Showing very intense effect of this feature on the selected datapoints. Additionally, a Density graph is also displayed in this figure which shows a higher population for *5th* band between 0.4 to 0.6 value range.

The Score graph illustrates our selection with score values below 0. The closer these values are to -5, the more intense their effect, while values closer to 0 indicate a lesser effect. Hovering over band 5 in the EBM plot shows an intensity effect

value of 1.644959 (see Figure 6.8). According to EBM criteria, values between 1 and 2 suggest looking for light gray shapes (the locations and coordinates of these shapes are provided in the ground truth image shown in Figure 6.10) in the specific hyperspectral image band. For an in-depth examination, we print the relevant band, which is the *5th* band of the hyperspectral image, using Matlab (see Figure 6.11) to compare it with our ground truth image.

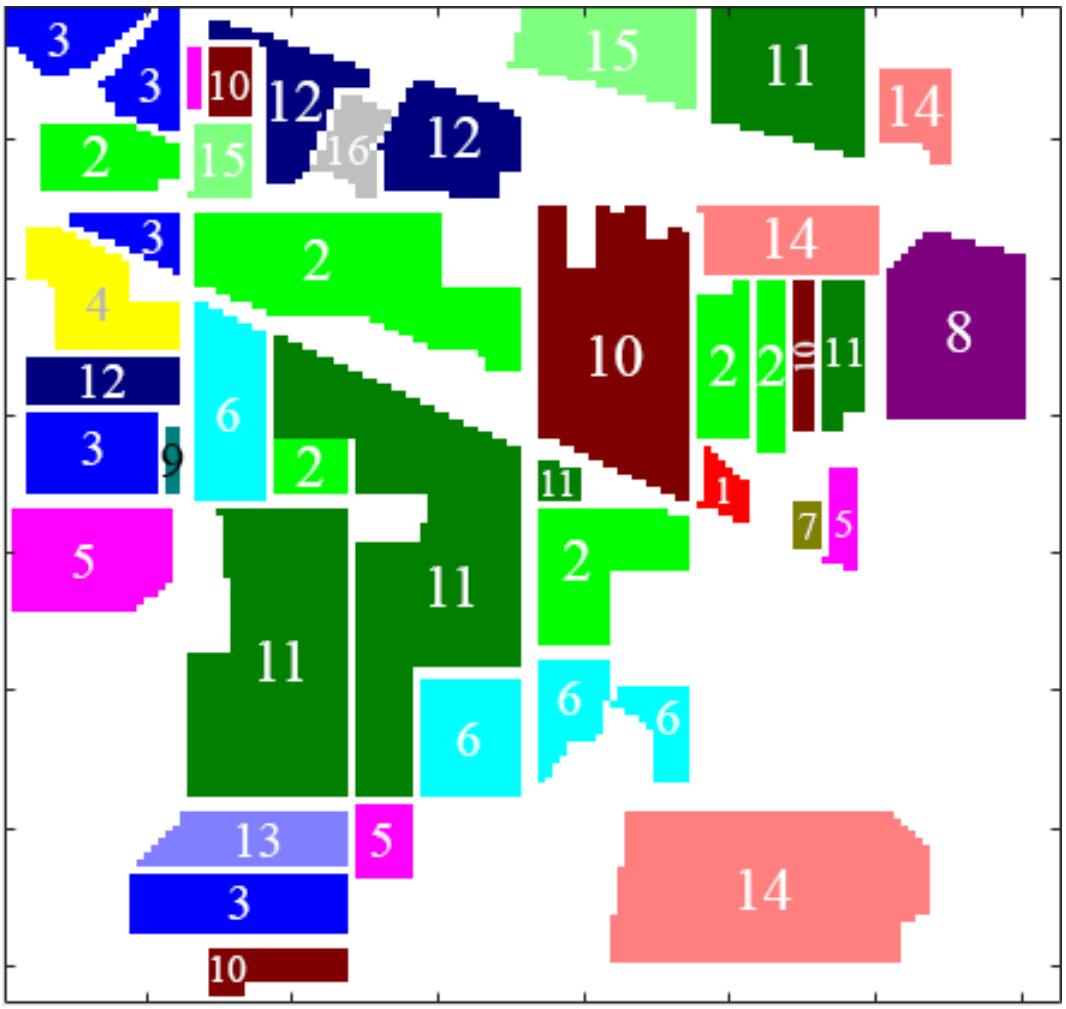


Figure 6.10 Ground truth of the Indian Pines dataset with class labels. Each color represents a different class, as indicated by the numbers on the image. This visual representation provides the actual class distribution and locations of the data points, serving as a reference for comparing and validating the clustering results.



Figure 6.11 *5th* band of the Indian Pines hyperspectral dataset. This grayscale image represents one spectral band of the hyperspectral data, highlighting the variations in reflectance across the different areas of the scene. This band is used for detailed analysis and comparison with the ground truth to identify and interpret specific features within the dataset for the selected cluster C1.

We then compare this with our ground truth image, which includes class labels. Upon close inspection, we observe that moderate gray shapes in the specific locations correspond to the ground truth image. You can see these areas highlighted with red rectangles and compare the results with the ground truth image provided in Figure 6.12. It's important to note that while other shapes may also be visible in this image band, the effect of this feature for our selection is particularly significant as its value is close to -5 in score graph and exhibit intense effects. This does not negate the presence of other shapes in the band; rather, it highlights the importance of the values for our selected points, which can be discerned based on their intensity color.

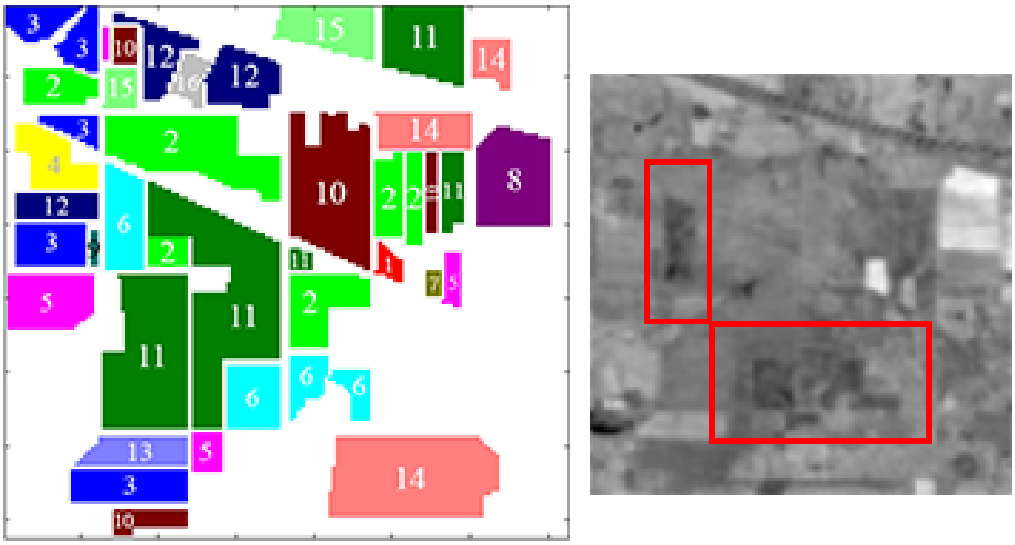


Figure 6.12 Comparison of the *5th* band of the Indian Pines hyperspectral dataset with the ground truth image. The red rectangles highlight the specific areas in the *5th* band that correspond to the shapes that cover class label *6* seen in the ground truth image. This comparison is used to validate the intensity effect values for our selected points, which are particularly significant due to their low values and intense effects. Other shapes may be visible in this band, but the highlighted areas emphasize the importance of our selected points based on their intensity color.

Next, we repeat the process for *band 8*. Analyzing the Score graph for *band 8* reveals similar patterns (see Figure 6.13), with values below 0 indicating our selection. We print the corresponding *band 8* image using Matlab and observe the intensity effect value in the EBM plot. For *band 8*, this value is lower, guiding us to look for lighter gray shapes in the hyperspectral image as shown in Figure 6.14.

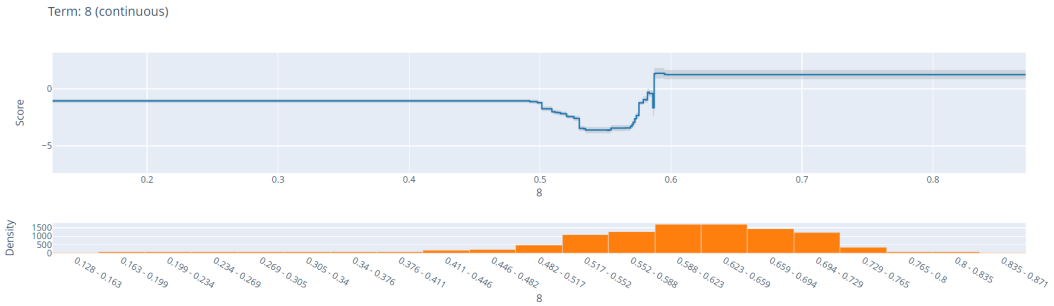


Figure 6.13 Individual Feature Graph for *8th* band of the hyperspectral image dataset in our first cluster which is C1. Showing moderate effect of this feature on the selected datapoints. Additionally, a Density graph is also displayed in this figure which shows a higher population for *5th* band between 0.5 to 0.7 value range.

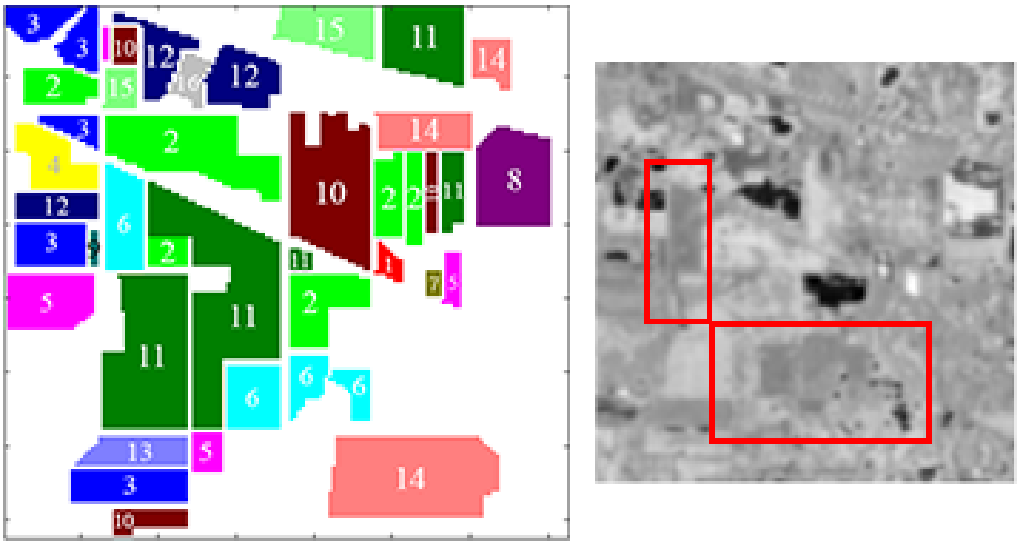


Figure 6.14 Comparison of the *8th* band of the Indian Pines hyperspectral dataset with the ground truth image. The red rectangles highlight the specific areas in the *8th* band that correspond to the shapes that cover class label *6* seen in the ground truth image. This comparison is used to validate the effect values for our selected points. Other shapes may be visible in this band, but the highlighted areas emphasize the importance of our selected points based on their intensity color.

By comparing these observations with our ground truth image, we gain further insights. Despite the potential presence of other shapes, the gray shapes aligned with our selection criteria provide valuable information about the data points in cluster C1. Additionally, the numeric values of the specific image bands highlighted in our selections also aid in a more detailed interpretation of the dataset.

This thorough investigation satisfies our goal of understanding the factors shaping these clusters and highlights the utility of our approach in analyzing hyperspectral image data. After this, we will move on to the next cluster.

6.2.2.2 Continuing Investigation with Cluster C2

We continue our investigation by selecting cluster C2. The EBM plot re-trains and updates immediately, highlighting the differences between the selected data points and the rest of the dataset. This selection changes the ranking of the features again (see Figure 6.15). The top two important features for this cluster are the *2nd* band and the 1st band of the hyperspectral image. As mentioned before, the ability to select different clusters and retrain the EBM model to generate various visualizations of ranked features aligns with the objectives of Goal 1.

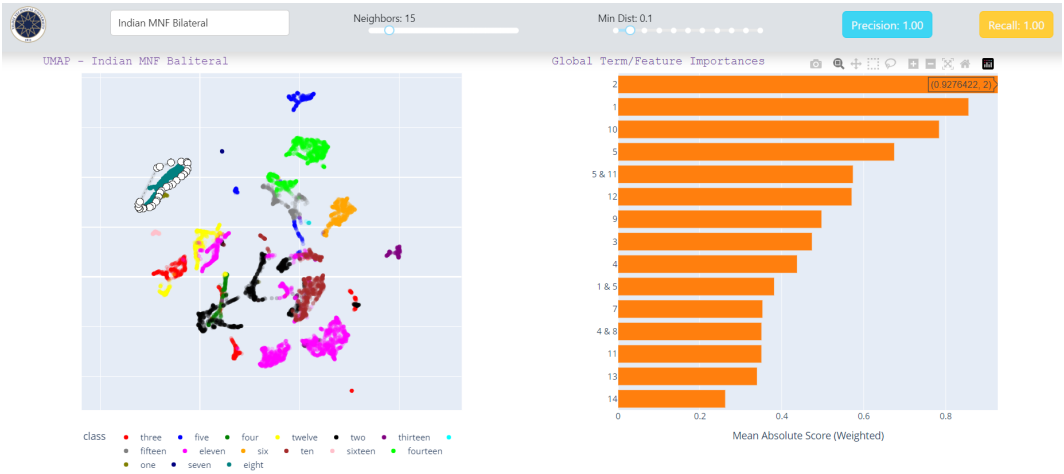


Figure 6.15 Selecting C2 retrains the EBM model to highlight the differences between the datapoints of cluster C2 and the rest of the dataset. The top features identified are band *2* and band *1* respectively.

To conduct a deeper investigation, we begin with band *2* of the dataset, which is the most important feature for this selection. We click on it and examine the Score and Density graphs. By exploring these graphs, we observe that the scores are very close to 0, indicating that these data points do not have a very intense effect. Additionally, we see that the majority of the values are between 0.3 and 0.5 (shown in Figure 6.16).

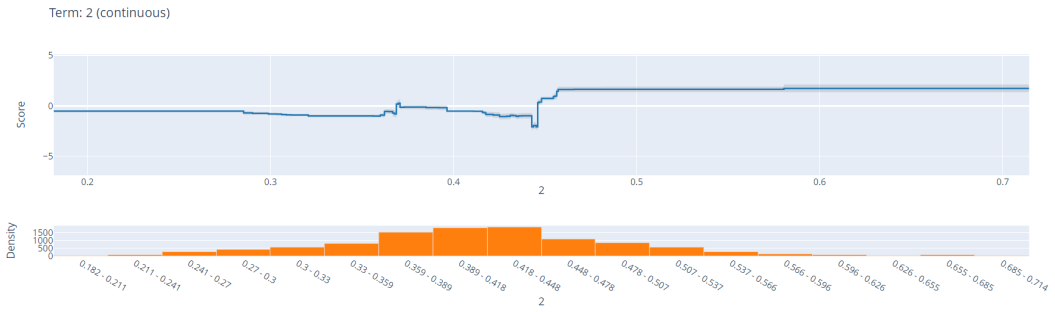


Figure 6.16 Individual Feature Graph for band *2* of the hyperspectral image dataset in our second cluster which is C2. Showing moderate effect of this feature on the selected datapoints. Additionally, a Density graph is also displayed in this figure which shows a higher population for *5th* band between 0.3 to 0.5 value range.

To compare this band with the ground truth image for our selection, which belongs to class label *8*, we print out the image band *2* using Matlab. The results are displayed in Figure 6.17. As band *2* for the selection has a moderate EBM value of 1.644959, we need to look for light gray shapes in the relevant image band.

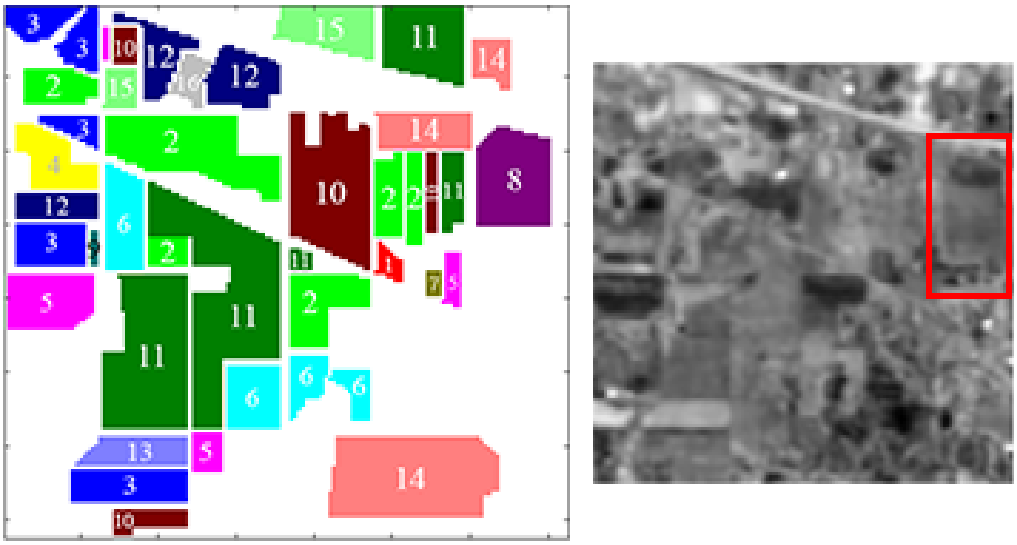


Figure 6.17 Comparison of the *2nd* band of the Indian Pines hyperspectral dataset with the ground truth image. The red rectangles highlight the specific areas in the *2nd* band that correspond to the shapes that cover class label *8* seen in the ground truth image. This comparison is used to validate the effect values for our selected points. Other shapes may be visible in this band, but the highlighted areas emphasize the importance of our selected points based on their intensity color.

By comparing these observations with our ground truth image, we gain further insights. Despite the presence of other shapes, the gray shape highlighted with a red rectangle accurately corresponds to class label *8*, matching the position shown in the ground truth. Additionally, the numeric values of the specific image bands highlighted in our selections further aid in a detailed interpretation of the dataset.

This thorough investigation meets our goal of understanding the factors shaping these clusters. For further analysis, we proceed to investigate the second most important feature for cluster C2, which is band *1* of the image dataset.

When examining the Individual Feature graph for band *1* in cluster C2, we see that we need to look for a similar light gray shape, as its value is smaller than that of the first feature. The Individual Feature Graph and the comparison of the ground truth with band *1* for the selected data points are shown in Figure 6.18 and Figure 6.19, respectively.

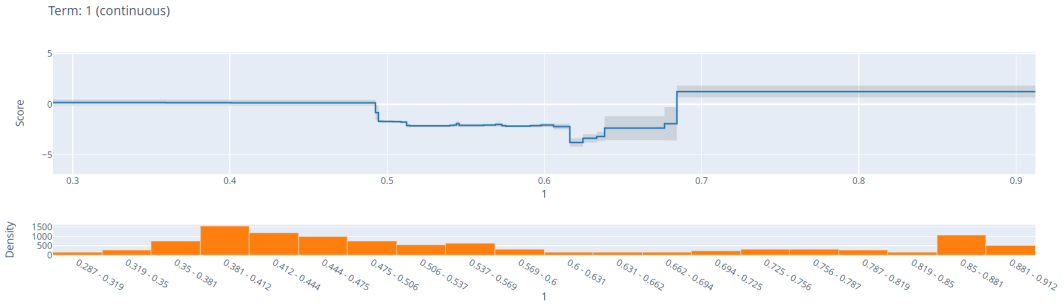


Figure 6.18 Individual Feature Graph for *1st* band of the hyperspectral image dataset in our first cluster which is C2. Showing very close score values to 0 which depicts that these leads do not have very intense effect but carry important factors for the selection between the values 0.5 and 0.65. Also, the Density graph shows almost equally distributed values among the dataset for this band.

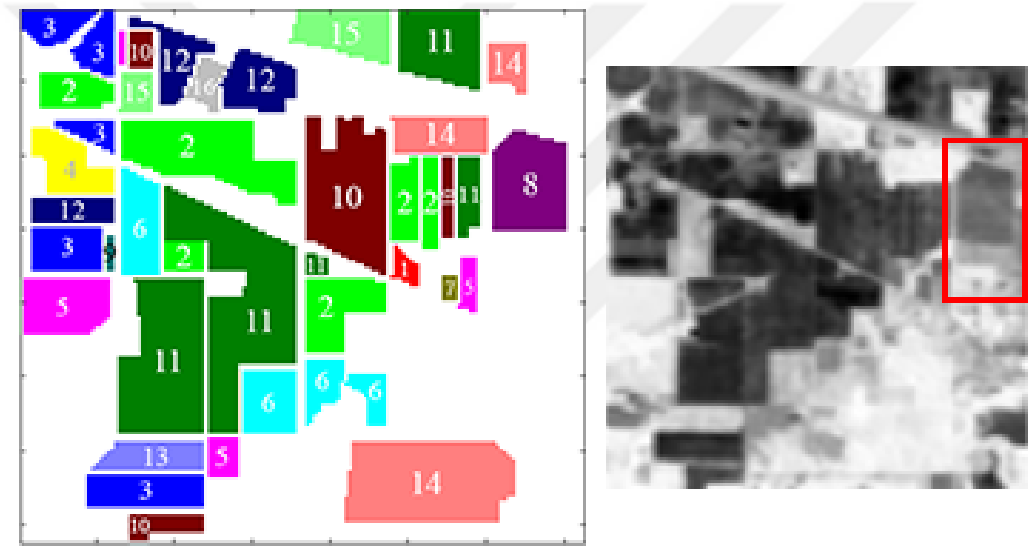


Figure 6.19 Comparison of the band *1* of the Indian Pines hyperspectral dataset with the ground truth image. The red rectangle highlights the specific areas related to the feature values for class label *8* in the relevant selection and displays the comparison between the ground truth and the specified area.

6.2.2.3 Advancing the Investigation with Cluster C3

We continue our investigation by moving to the next selection, cluster C3. As shown in Figure 6.20, features *1* and *11* are the two most important bands of the Indian Pines dataset for class label *10*. We observe a significant gap between the EBM rank values of the first and second features. This discrepancy intrigues us, prompting a deeper investigation into the first feature.

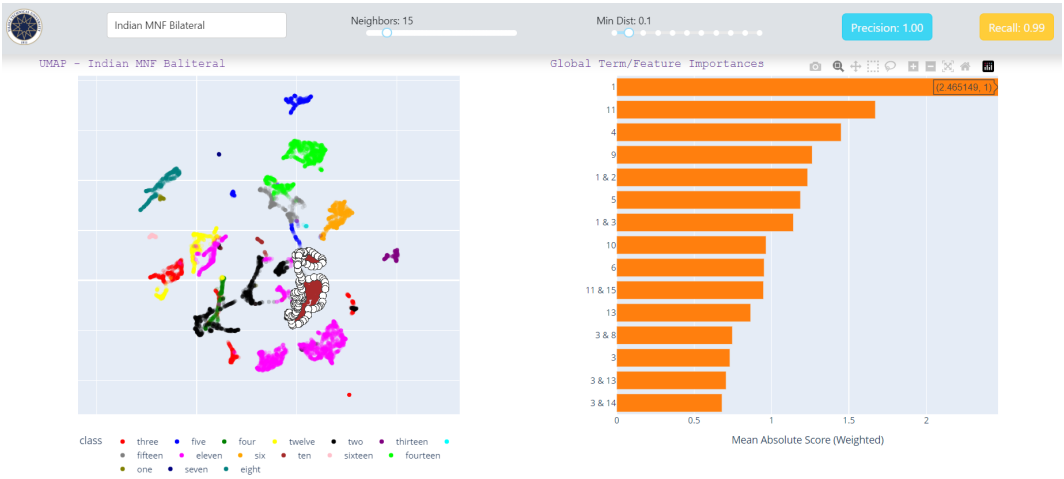


Figure 6.20 Selecting C3 retrains the EBM model to highlight the differences between the datapoints of cluster C3 and the rest of the dataset. The top features identified are band *1* and band *11* respectively. There is a huge gap between the first two important features which demosntrates the importance difference between these two important features.

We click on the first feature and in the provided Individual Feature graph, we notice a trend that initially decreases and then increases. The interval representing our selection ranges from 0.36 to 0.66 (see Figure 6.21). Another notable point is that these values have a very high EBM value (2.465149), indicating that we should look for either dark gray or very light gray (almost white) shapes in the *1st* band of the Indian Pines dataset.

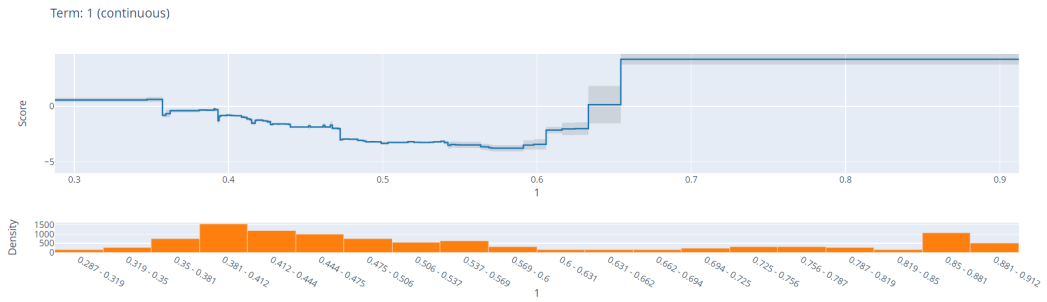


Figure 6.21 Individual Feature Graph for band *1* of the hyperspectral image dataset in C3. Showing almost intense effect of this feature on the selected datapoints. Additionally, the Density graph shows almost equally distributed values among the dataset for this band.

Based on these details, we compare the ground truth with the 1st band, as shown in Figure 6.22. The ground truth validates our findings, with the relevant areas highlighted within red rectangles.

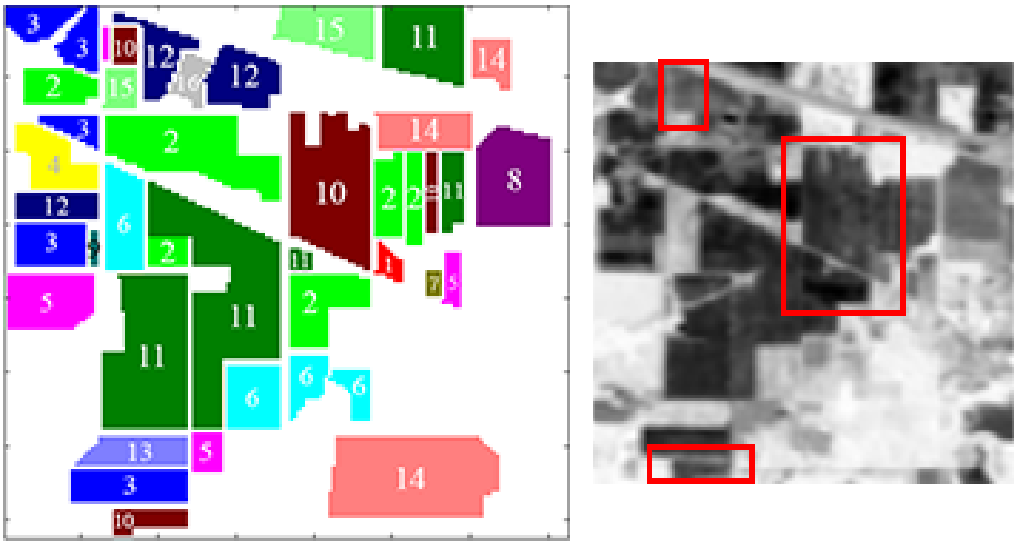


Figure 6.22 Comparison of the *1st* band of the Indian Pines hyperspectral dataset with the ground truth image. The red rectangles highlight the specific areas in the *1st* band that correspond to the shapes that cover class label *10* seen in the ground truth image. This comparison is used to validates our investigation.

After exploring the first feature, we move on to the second feature, which is band *11* of the hyperspectral image in the Indian Pines dataset. After clicking on this feature, its Individual Feature graph appears at the bottom of the page, as shown in Figure 6.23. The Score graph shows a very intense effect for our selection group, with score values nearly reaching -5. However, the Density graph indicates a low distribution for data points with score values close to -5, while the distribution is dense around 0.7 to 0.8. These feature values correspond to score values between -2 to 2. This means that a very small portion of the shapes in band *11* will be very dark gray or white, and the rest will be light gray. This example demonstrates that investigating both the Score graph and the Density graph within the same plot can be very useful in some cases.

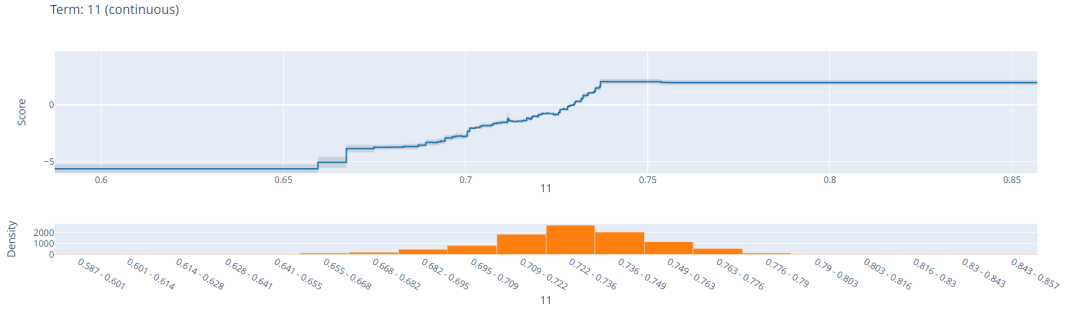


Figure 6.23 Individual Feature Graph for band *11* of the hyperspectral image dataset in C3. Showing very intense effect of this feature on some part of the selected datapoints. Additionally, the Density graph shows the distribution almost equally for selected and unselected datapoints.

The comparison of the ground truth and band *11* of the Indian Pines dataset image validates our findings and is illustrated in Figure 6.24 for reference.

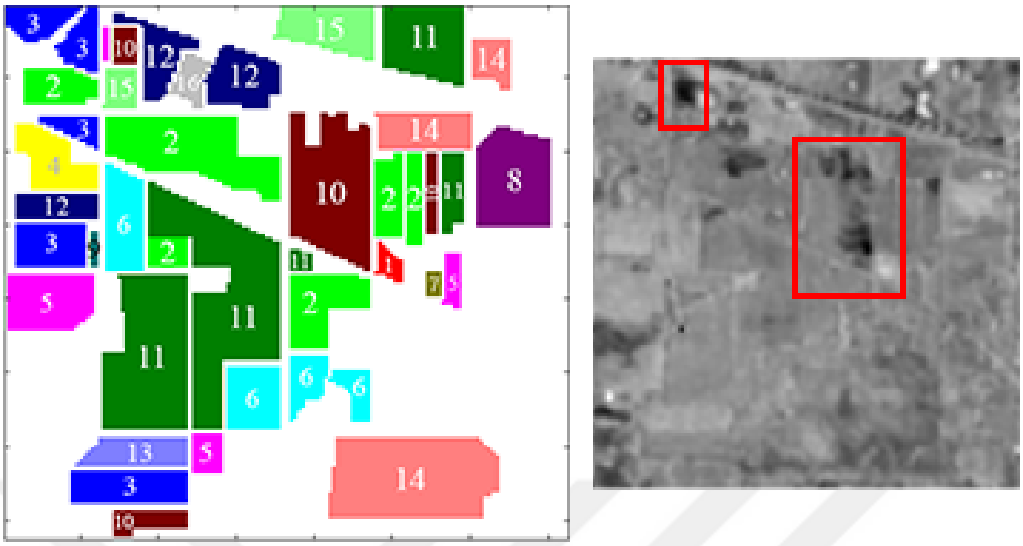


Figure 6.24 Comparison of the *11th* band of the Indian Pines hyperspectral dataset with the ground truth image. The red rectangles highlight the specific areas in the *1st* band that correspond to the shapes that cover class label *10* seen in the ground truth image. This comparison is used to validate our hypothesis.

6.2.2.4 Continuing the Investigation with Cluster C4

Another interesting and worth investigating cluster is cluster C4. This cluster is widely spread throughout the dataset and overlaps with many data points from other class labels, not just class label *11*. We attempted to select only data points belonging to class *11* but were unable to completely separate them, resulting in the inclusion of a few data points from other classes. Additionally, due to the widespread nature of this class, we were unable to select all data points belonging to class *11*. Consequently, this limitation introduces a bit of noise. Please see Figure 6.25.

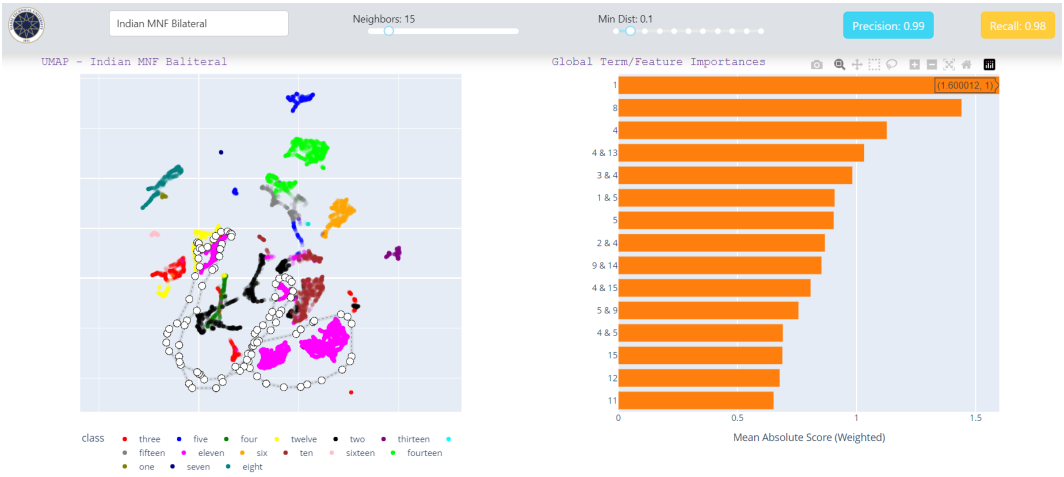


Figure 6.25 Selecting C4 retrains the EBM model to highlight the differences between the datapoints of cluster C4 and the rest of the dataset. The top features identified are band *1* and band *8* respectively.

To examine the results for this interesting cluster, we click on the first important feature, which is feature *1*. In the provided Individual Feature Graph (see Figure 6.26), we observe a linear trend that starts with very intense values (-5) and increases slightly throughout the graph. Notably, the density for feature *1* is very high at the end interval of our selection. As depicted in the Individual Feature Graph, feature *1* has a significant influence on the formation of the selected cluster.

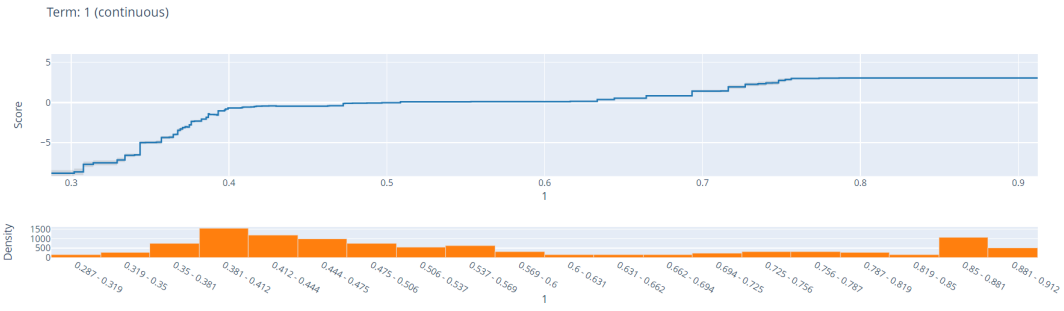


Figure 6.26 Individual Feature Graph for band *1* of the hyperspectral image dataset in C4. Showing a linear trend through the graph. Additionally, the Density graph shows a dense distribution around the end of the selection interval.

After investigating the Individual Feature Graph, it is time to examine band *1*, identify the selected class, and then compare it with the ground truth (provided in Figure 6.27). Given that the EBM value is approximately 1.6, we need to look for gray shapes in band *1*. The detected area is compared to the ground truth image, and we can see that HyEx is successful in this selection as well.

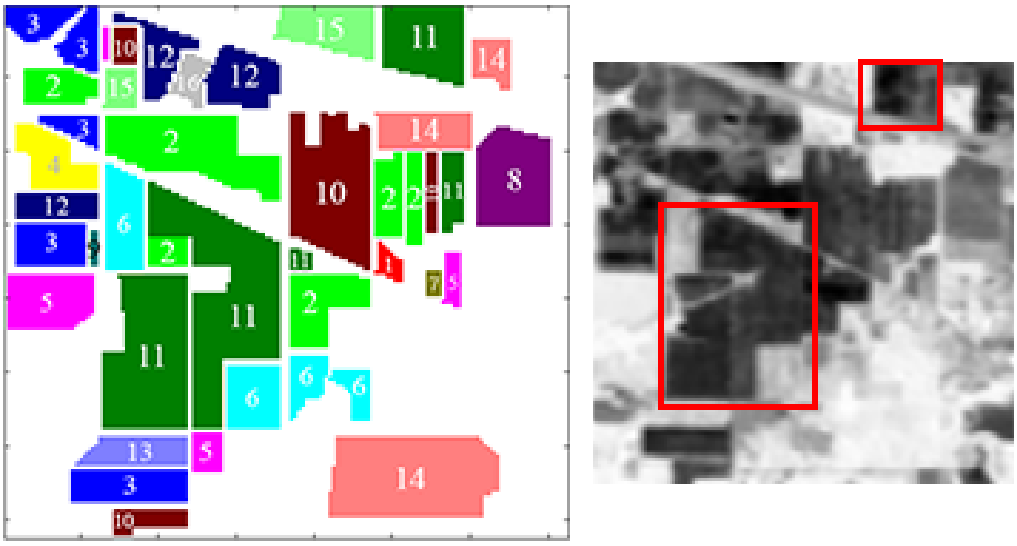


Figure 6.27 Comparison of the *1st* band of the Indian Pines hyperspectral dataset with the ground truth image. The red rectangles highlight the specific areas in the *1st* band that correspond to the shapes that cover class label *11* seen in the ground truth image. This comparison is used to validates our investigation.

As with the previous selections, we will proceed to investigate the second most important feature, which is feature *8* in this case. After clicking on this feature, its Individual Feature graph appears at the bottom of the page, as shown in Figure 6.28. The Score graph shows a very linear trend with scores close to 0, and the EBM value is also close to 1. This indicates that we need to look for very light gray shapes when examining band *8*. Additionally, the Density graph demonstrates that the majority of data points are concentrated around our selection, further confirming that this selection has been a good choice.

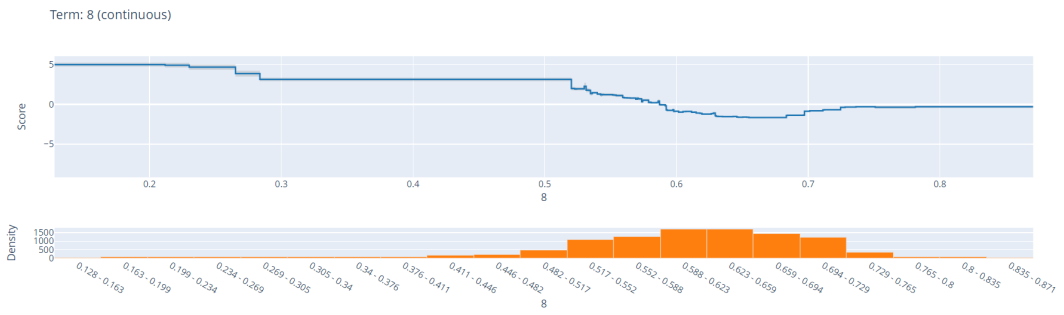


Figure 6.28 Individual Feature Graph for band *8* of the hyperspectral image dataset in C4. Showing a linear trend through the graph. Additionally, the Density graph shows a dense distribution around the selection interval.

Now, we need to explore band *8* and compare it with the ground truth illustrated in Figure 6.29. As shown in Figure 6.29, we could only recognize one part that

consists of the majority of the data points. This limitation may be due to the noise that we could not eliminate during the selection process.

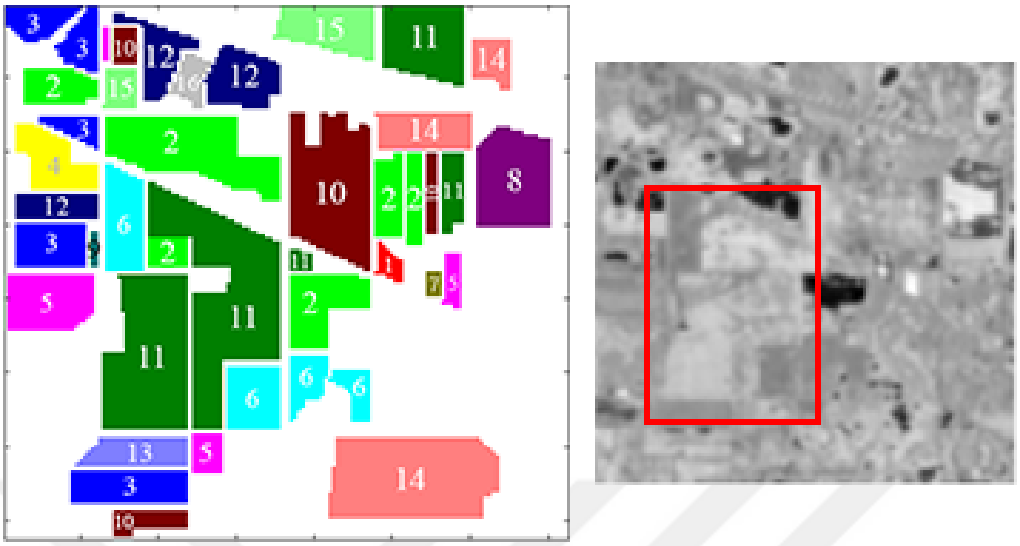


Figure 6.29 Comparison of the band *8* of the Indian Pines hyperspectral dataset with the ground truth image. The red rectangle highlights the specific areas in the *8th* band that correspond to the shapes that cover class label *11* seen in the ground truth image. This comparison is used to validates our investigation.

6.2.2.5 Continuing the Study with Cluster C5

To continue with the investigation, we select cluster C5. The EBM re-trains based on this selection compared to the rest of the dataset and provides a completely new ranking of features for this selection, as displayed in Figure 6.30. In this selection, the most important features we will be investigating are band *11* and band *1*. An important point that stands out here is that feature *1* and feature *2* are very close to each other, with only a slight difference in their ranking.

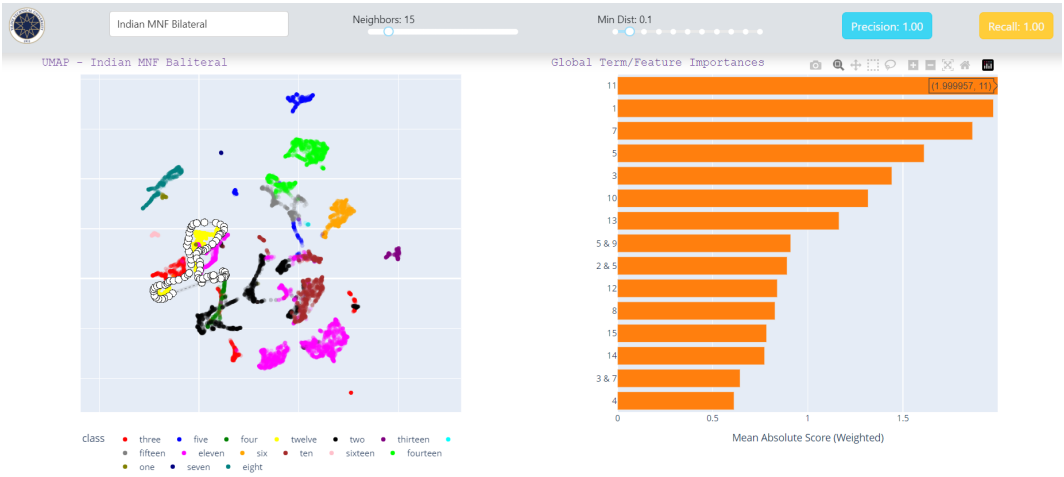


Figure 6.30 Selecting C5 retrains the EBM model to highlight the differences between the datapoints of cluster C5 and the rest of the dataset. The top features identified are band *11* and band *1* respectively.

To examine cluster C5, which consists of data points with class label *12*, we click on the first important feature, band *11* of the Indian Pines image dataset. As demonstrated in the Score graph in Figure 6.31, while our selection has a moderate effect on the formation of this cluster, the rest of the data points have a very intense effect. Additionally, the Density graph shows that the majority of the data points have values around the beginning of the selection interval and the end of the rest of the selection.

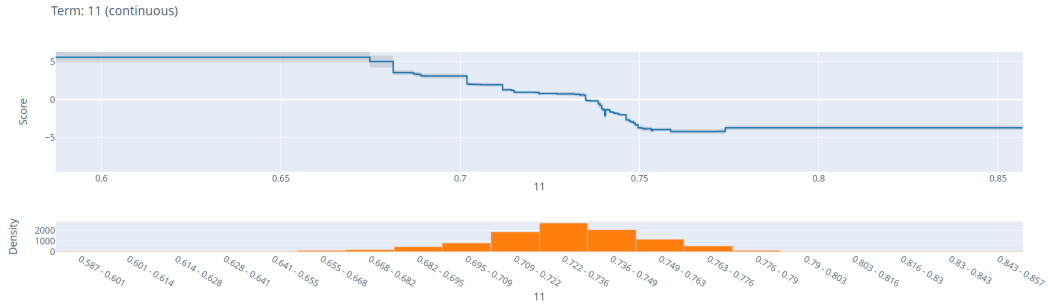


Figure 6.31 Individual Feature Graph for band *11* of cluster C5 that shows a linear trend. Additionally, the Density graph shows a dense distribution around the beginning of the selection interval and the rest of the dataset.

To explore further, we continue investigating band *11*, looking for the shapes of areas covered by class label *12*. As depicted in Figure 6.32, the red rectangles and triangle highlight the areas recognized as class label *12*, which align with the ground truth.

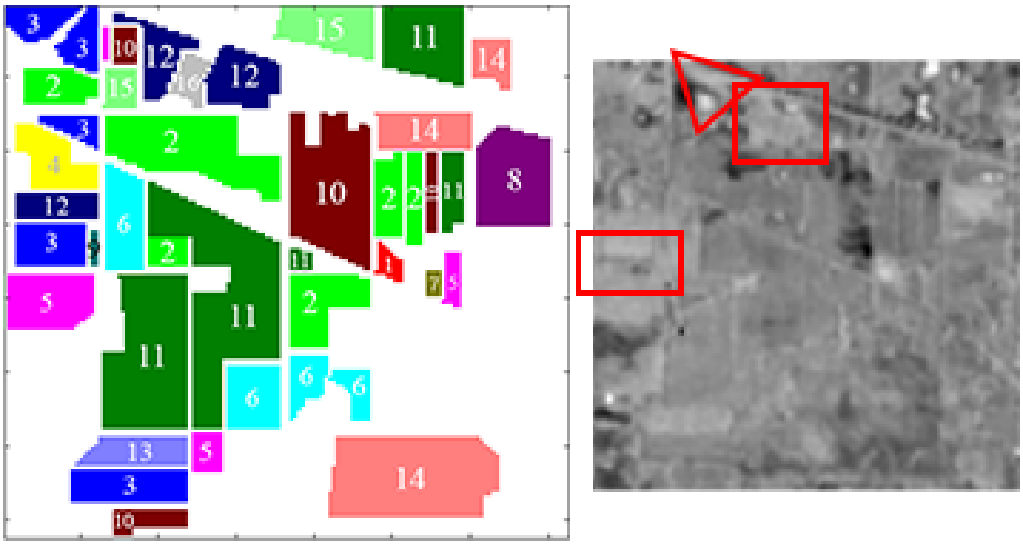


Figure 6.32 Comparison of the *11th* band of the Indian Pines hyperspectral dataset with the ground truth image. The red rectangles and the red triangle highlight the specific areas in the image band *11* that correspond to the shapes that cover class label *12* seen in the ground truth image. This comparison is used to validates our investigation.

To further interpret cluster C5, we proceed by examining feature *1*, which is the second most important feature for this cluster. As shown in Figure 6.33, there is a wide interval for our selection compared to the rest of the dataset. Although this feature is distributed almost equally across the dataset, there is noticeable density around our selection in the graph. This indicates that most of the data points have values that cluster around our selection.

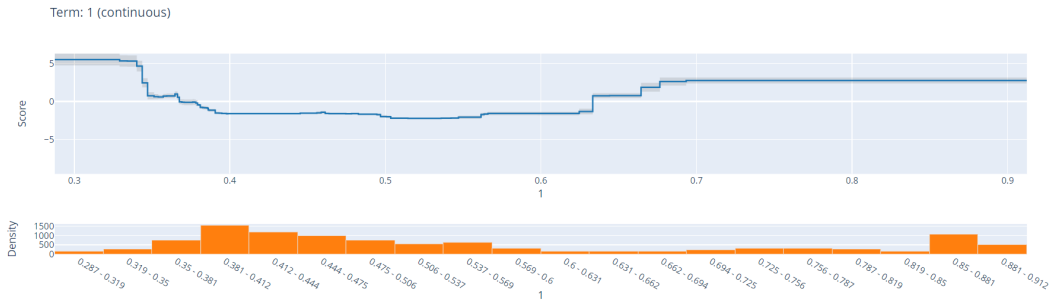


Figure 6.33 Individual Feature Graph for band *1* of cluster C5 that shows a linear trend. Additionally, the Density graph shows a dense distribution around the selection interval.

To explore further and validate our findings, we examine band *1* of the Indian Pines dataset. The shapes of class *12*, as highlighted by the ground truth image, are easily identifiable. The comparison results are provided in Figure 6.34 for reference.

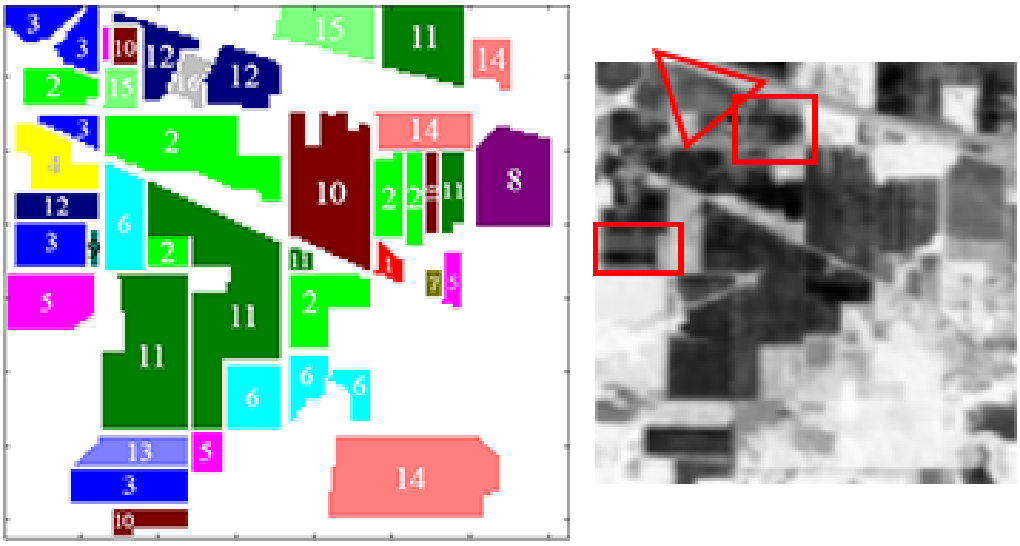


Figure 6.34 Comparison of the image band *1* of the Indian Pines hyperspectral dataset with the ground truth image. The red rectangles and the red triangle highlights the specific areas in the *1st* band that correspond to the shapes that cover class label *12* seen in the ground truth image. This comparison is used to validates our investigation.

6.2.2.6 Deepening the Analysis with Last Cluster, C6

To conclude our investigation, we finally take cluster C6 for deeper analysis. As shown in Figure 6.35, the two most important features are image band *1* and image band *6*. Notably, feature *1* is highly effective for this selection, with a significant gap between it and the rest of the important features in the EBM Feature Ranking Graph. For further analysis, we will investigate the two most important features one by one.

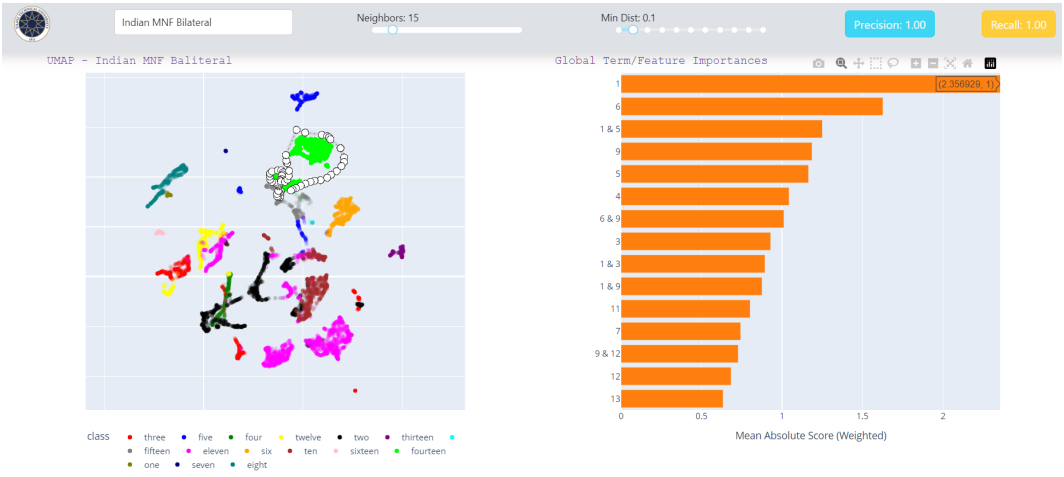


Figure 6.35 Selecting C6 retrains the EBM model to highlight the differences between the datapoints of cluster C6 and the rest of the dataset. The top features identified are band *1* and band *6* respectively.

When clicking on the feature *1*, we see a very interesting trend in the Score graph. The Score graph indicates a constant but with high uncertainty trend until the 0.725 for the image band *1* as illustrated in Figure 6.36. On the other hand, the selected datapoints follow a decreasing trend until the end of the graph. Furthermore, we can see the almost equally distribution of the datapoints in the dataset for this feature.

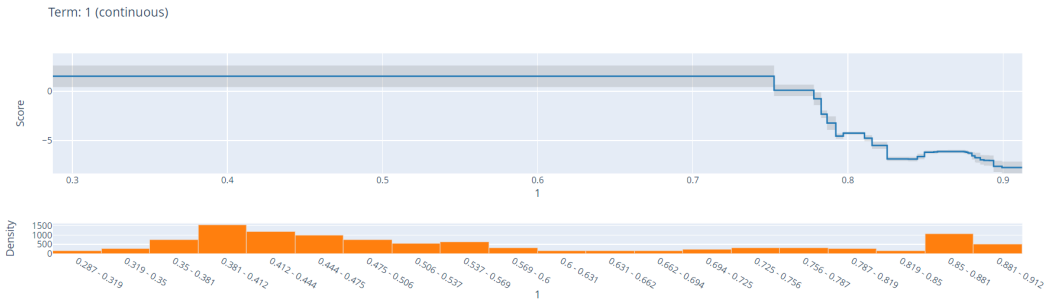


Figure 6.36 Individual Feature Graph for band *1* of cluster C6 that shows a linear trend until our selection and for the selection it demonstrates a decreasing trend until the end of the graph. Additionally, the Density graph shows an almost equally distributed datapoints in the dataset for this feature.

For further investigation, we continue with band *1*, looking for the white shapes of areas covered by class label *14*. As depicted in Figure 6.37, the red rectangles highlight the areas recognized as class label *14*, which align with the ground truth.

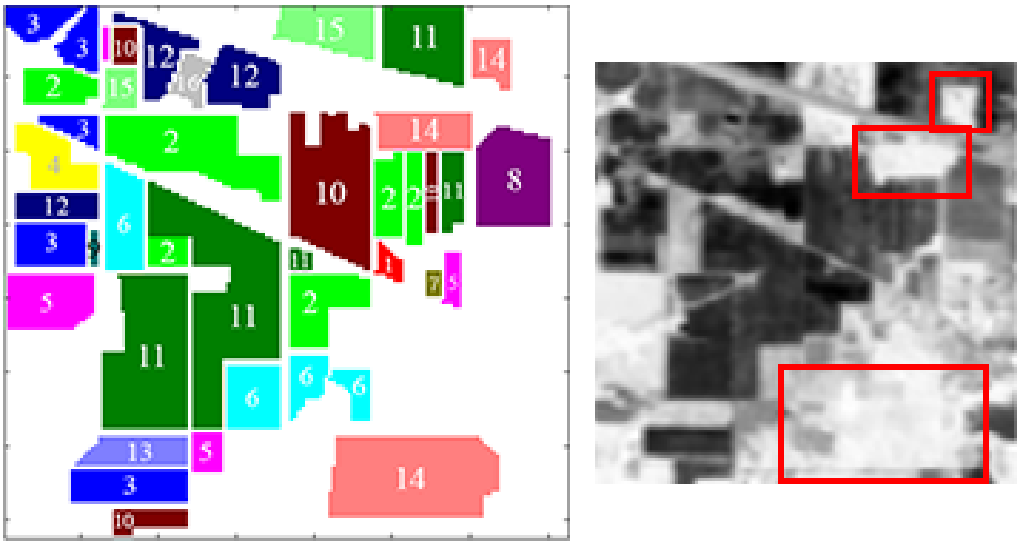


Figure 6.37 Comparison of the image band *1* of the Indian Pines hyperspectral dataset with the ground truth image. The red rectangles highlight the specific areas in the *1st* band that correspond to the shapes that cover class label *14* seen in the ground truth image. This comparison validates our investigation.

To further interpret the last cluster, we proceed by examining the second most

important feature for this cluster called image band *6*. As shown in Figure 6.38, there is a wide interval for our selection compared to the rest of the dataset. However, this interval follows a very linear trend that carries score values close to zero. In addition, the Density graph shows a dense distribution around the end of our selection interval and the beginning of the unselected datapoints interval.

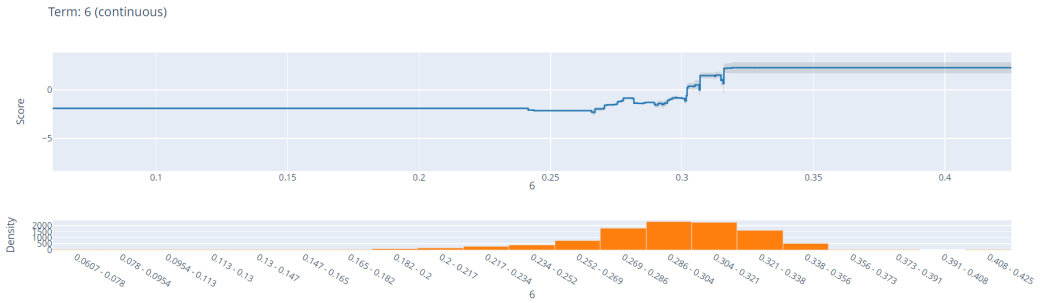


Figure 6.38 Individual Feature Graph for band *6* of cluster C6 that shows a linear trend for our selection and then an increase and finally again a linear trend for the unselected datapoints. Additionally, the Density graph shows a dense distribution around the end of our selection interval and the beginning of the unselected datapoints interval.

Finally, to finish our examination on the last cluster, we need to look for light gray shapes corresponding to the areas demonstrating class label *14* as the EBM value for this feature is between 1 and 2. As shown in Figure 6.39, the red rectangles show the area of the class label *14* that our tool has detected.

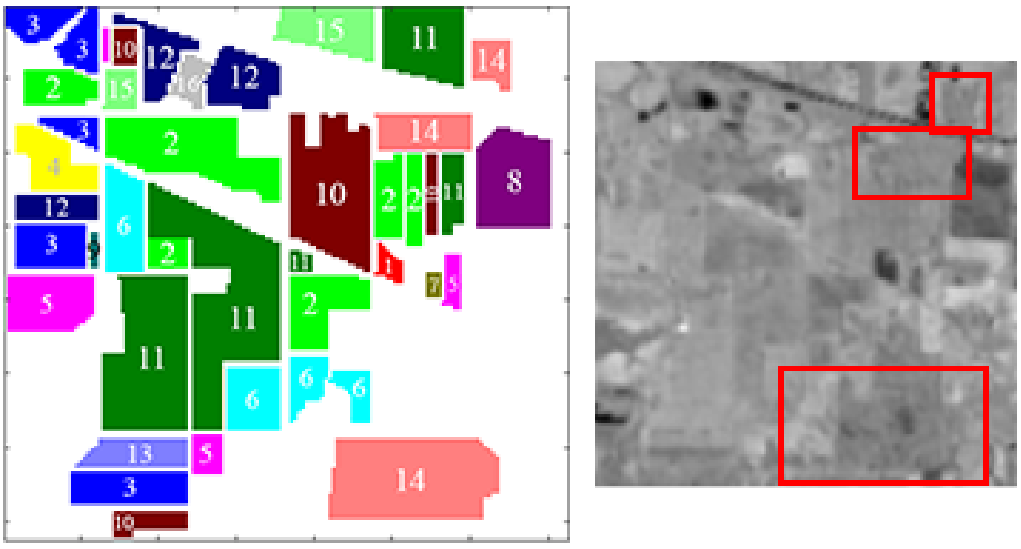


Figure 6.39 Comparison of the image band *6* of the Indian Pines hyperspectral dataset with the ground truth image. The red rectangles highlight the specific areas in the *6st* band that correspond to the shapes that cover class label *14* seen in the ground truth image. This comparison validates our investigation.

6.2.2.7 Final Thoughts

After the encouraging results gained from our thorough investigation of the Hyperspectral Image dataset, we believe that this tool is valuable across various domains. It can assist data scientists and researchers in understanding and interpreting their high-dimensional datasets more easily.

7
DISCUSSION

7.1 Development Insights and Future Directions of DimVis and HyEx

In this section, we will describe the design choices of DimVis, our tool for numeric datasets, and HyEx, a modified version of the same tool tailored for Hyperspectral Image datasets. We will also discuss the limitations and future work for these tools.

7.1.1 Design Choices

The initial goal behind developing DimVis was to create an accessible and interactive platform specifically designed for high-dimensional data exploration. Numerous design choices were implemented to ensure that the tool remained simple, user-friendly, and effective for analyzing and interpreting intricate datasets. Recognizing the potential benefits for Hyperspectral Image datasets, which are notably complex and challenging to interpret, we adapted DimVis into HyEx. While the primary design decisions have been detailed in Section 3, this section will cover additional, less critical design considerations for both DimVis and HyEx. Each part is elaborated on below:

7.1.1.1 Visual Design and Layout

The user interface was crafted to be clean and minimalistic, focusing on the core functionalities of the tool without overwhelming the user. Essential elements, such as the dataset selection dropdown, hyperparameter adjustment panel, and data visualization graphs, are prominently displayed without confusing the user. This straightforward design ensures that the primary objectives of the research are clear and unobscured by unnecessary elements. This approach was adopted to align with Goal 3.

7.1.1.2 Data Selection Interface

Flexibility: This feature allows users to seamlessly switch between various datasets from different fields, like a numeric dataset in any domain or Hyperspectral Image datasets, observing changes in real-time and interacting with the data dynamically. Datasets are uploaded in CSV format, a widely accepted standard, ensuring compatibility with most available datasets. This design choice was made to support Goal 3, enabling users to utilize DimVis and HyEx based on their need with ease.

7.1.1.3 Dynamic Hyperparameter Adjustment

UMAP Personalization: Initially, UMAP operates with default hyperparameter settings such as 15 for "Number of neighbors" and 0.1 for "Minimum Distance" hyperparameters. However, DimVis and HyEx include two sliders for adjusting these two key hyperparameters, enabling users to create up to 1,000 different visualizations for each tool. This feature allows for precise customization of UMAP projections to meet specific needs. This design choice also aligns with our Goal 3, enhancing user support.

7.1.1.4 Dynamic and Interactive UMAP Plot

Dynamic Exploration: The UMAP graph within DimVis and HyEx is designed to be interactive, enabling users to select various combinations of data points in real-time. This functionality enhances user engagement in data exploration and aids in more effective data interpretation.

This user interaction is a key innovation of our proposed tool. By allowing users to select data points that form specific patterns, or any patterns they desire using a lasso tool, we leverage UMAP's strength in visually distinguishing interesting patterns while involving users in the process. This dual approach translates user selections into actionable input for the underlying EBM model, which then provides interpretations for either the whole dataset or any specific selection. While automatic clustering could achieve similar results, it would introduce uncertainties in the clustering process and could not help users in interpreting specific clusters that need to be analyzed in isolation, potentially affecting the interpretation outcomes.

We also operate under the assumption that visually distinctive patterns will emerge in the DR output, which might not always be the case. Although the tool functions even if the selected points do not form clear patterns, the interpretation may not be as useful. However, given UMAP's capability in visually separating patterns, we believe the likelihood of identifying and selecting meaningful patterns is high,

making the interpretations valuable.

Additionally, precision and recall metrics provide insights into whether the selected pattern is significant. High values suggest that the selected pattern is distinct and interpretable. This design choice aligns with Goal 1, ensuring effective user support and engagement.

7.1.1.5 EBM Feature Ranking Graphs

Real-Time Retraining: The EBM component of our tool automatically retrains each time a user adjusts the UMAP graph, instantly updating to display a new feature importance graph. This functionality provides insights into how feature importance rankings fluctuate with different data selections, aligning with Goal 1.

Performance Testing: We evaluated the dynamic retraining capability of DimVis and HyEx by selecting eight random patterns from two datasets, recording the retraining and regeneration times in seconds. For DimVis the retraining times are like below: 07.50, 02.08, 12.89, 11.09, 15.15, 02.66, 03.38, and 20.98. The average retraining time was about 09.47 seconds. However, for HyEx the time consumed to display the results is higher as the dataset is much more complex. EBM regeneration times for HyEx are as follows: 15.71, 44,64, 16.09, 07.22, 61.41, 37.50, 57.78 and 12.72. The average retraining time for HyEx was about 31.63 seconds This choice supports Goal 3 also.

7.1.1.6 EBM Individual Feature Analysis Graph

In-Depth Exploration: The graph facilitates a detailed exploration by enabling users to assess each feature individually, separate from others. This approach improves data interpretation and allows users to understand the reasons behind visual patterns by investigating important features individually. This design choice supports Goal 2.

Final Design Considerations: The design principles for DimVis and HyEx ensure a dynamic and interactive user experience, allowing for comprehensive exploration of UMAP projections and EBM features. Ongoing user feedback will be critical in improving the tool's design and functionality.

7.1.2 Limitations and Future work

DimVis and HyEx were developed utilizing Python and Plotly Dash. The use cases and experiments were conducted on a Dell laptop equipped with an Intel(R)

Core(TM) i7-10750H processor operating at 2.60GHz, an NVIDIA GeForce GTX 1650 Ti 4GB GPU, 16GB of DDR4 RAM, and running on Windows 11 Enterprise.

One limitation of our tool is that it exclusively supports numeric features in high-dimensional datasets. While DimVis and HyEx can handle string class labels, all features and attributes used for analysis must be numeric. This constraint may reduce the tool's applicability in cases where datasets include non-numeric features.

Another limitation is the reliance on a specific algorithm, which may become outdated or be superseded by other algorithms, such as TDA [25], which has gained popularity in recent literature. However, our tool can be adapted to different dimensionality reduction techniques or alternative interpretation methods with some modifications. The design is not inherently tied to UMAP, allowing for flexibility. For example, the EBM can be replaced by any interpretable machine learning (IML) supervised classification model, and the visualizations can be substituted with views that provide both global and local (individual) feature rankings.

A further limitation is the performance on large datasets. While DimVis retraining times are relatively fast, averaging about 09.47 seconds (details are mentioned before), HyEx takes significantly longer due to the complexity of the hyperspectral image data. The EBM regeneration times for HyEx average 31.63 seconds. This performance issue may impact the tool's usability and responsiveness, particularly when dealing with very large datasets. Addressing this limitation is crucial for ensuring the tool's effectiveness and supporting Goal 3.

Future Work: While we haven't conducted a formal user study, we have informally presented the tool to two participants—a medical doctor and a computer science student with experience in visualization courses. Their feedback highlighted key improvements that could enhance the tool's utility and user experience, which we plan to address in future work, separate from the use cases discussed earlier.

Firstly, one suggestion was to implement a threshold or range option for features. This would allow users to focus on specific value ranges, leading to more targeted and effective data analysis. Another significant recommendation was to enable dynamic color changes in the EBM Feature Importance Graph upon feature selection, providing immediate visual feedback. Additionally, incorporating detailed tooltips when hovering over data points could offer valuable information and deeper insights without requiring extensive exploration. These enhancements will be considered for future updates to ensure DimVis and HyEx remain user-friendly and adaptable to evolving user needs.

A formal user study is also planned to further validate our findings, gather additional feedback, and compare our tool to similar tools and techniques. This will help us understand how the explanations generated by EBMs differ from other methods and demonstrate why EBMs are more effective in this context. Conducting a user study would be an excellent future work.

We also aim to improve the computational efficiency of the underlying model by exploring GPU-based implementations for both the DR and EBM algorithms. This enhancement could enable the tool to focus on analyzing pairs of features instead of single ones, offering more comprehensive insights.

Currently, DimVis and HyEx can only compare one data selection against the remaining projection, making direct comparisons between two subclusters somewhat inefficient task. We plan to develop a feature that allows users to select and compare two specific data subsets directly, streamlining the comparison process in another future work.

8
CONCLUSION

8.1 Summary and Insights

In this thesis, we presented DimVis and HyEx, two interactive tools designed for data visualization and interpretation using UMAP and EBM techniques. These tools aim to enhance the data analysis experience by enabling real-time interactions, adjustable hyperparameters, and in-depth individual feature exploration, especially for complex datasets.

Reflecting on our primary goals, we find that our tools effectively meet them. Goal 1 (G1), which emphasizes user engagement through data exploration, is achieved by allowing users to create various clusters, interact with visualizations, and generate EBM plots based on their selections. Goal 2 (G2), focused on individual feature assessment, is fulfilled through the dynamic feature score plots in both tools, offering users the ability to analyze each feature and its impact on the dataset and clusters separately. Goal 3 (G3), aimed at user support, is realized by the tools' flexibility in handling diverse datasets across multiple fields and their user-friendly interfaces, which streamline the data analysis process.

As data grows increasingly complex, the need for tools like DimVis and HyEx becomes more pronounced. These tools and platforms are essential in a data-driven world, making intricate datasets interpretable and accessible to users through intuitive and user-friendly interfaces.

DimVis and HyEx are designed to help users make sense of complex datasets by leveraging UMAP for dimensionality reduction and providing 2D visualizations. Additionally, these tools utilize the EBM method to highlight the importance of each data point, assess the impact of individual features on the model's predictions, and rank them accordingly. Their strong point lies in their interactive visualization capabilities and the ability to make real-time changes and see the results instantly.

REFERENCES

[1] I. T. Jolliffe, *Principal Component Analysis*. Springer, 2002.

[2] I. Borg, P. J. Groenen, *Modern Multidimensional Scaling: Theory and Applications* (Springer Series in Statistics). Springer, 2005.

[3] L. van der Maaten, G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov. Pp. 2579–2605, 2008. [Online]. Available: http://www.jmlr.org/papers/v9/vandermaaten08a.html.

[4] J. H. L. McInnes, J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," *Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018.

[5] W. DeGroat, D. Mendhe, A. Bhusari, H. Abdelhalim, S. Zeeshan, Z. Ahmed, "Intelligenes: A novel machine learning pipeline for biomarker discovery and predictive analysis using multi-genomic profiles," *Bioinformatics*, vol. 39, no. 12, 2023. doi: 10.1093/bioinformatics/btad755. [Online]. Available: https://doi.org/10.1093/bioinformatics/btad755.

[6] R.-t. Zhao, J. Wang, G.-j. Chen, Q.-w. Li, Y.-j. Yuan, "A machine learning pipeline generation approach for data analysis," in *2020 IEEE 6th International Conference on Computer and Communications (ICCC)*, Chengdu, China, 2020, pp. 1701–1706, isbn: 978-1-7281-8635-1. doi: 10.1109/ICCC51575.2020.9345123. [Online]. Available: https://ieeexplore.ieee.org/document/9345123.

[7] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, Sydney, NSW, Australia, 2015, pp. 1721–1730. doi: 10.1145/2783258.2788613. [Online]. Available: https://dl.acm.org/doi/10.1145/2783258.2788613.

[8] B. D. A. Bibal A. Clarinval, B. Frénay, "Ixvc: An interactive pipeline for explaining visual clusters in dimensionality reduction visualizations with decision trees," *Array*, vol. 11, p. 100 080, 2021. doi: 10.1016/j.array.2021.100080. [Online]. Available: https://doi.org/10.1016/j.array.2021.100080.

[9] V. D. A. Bibal, B. Frénay, "Dt-sne: T-sne discrete visualizations as decision tree structures," *Neurocomputing*, vol. 529, pp. 101–112, Apr. 2023. doi: 10.1016/j.neucom.2022.12.025. [Online]. Available: https://doi.org/10.1016/j.neucom.2022.12.025.

[10] R. M. M. A. Chatzimparmpas, A. Kerren, "T-visne: Interactive assessment and interpretation of t-sne projections," *IEEE Transactions on Visualization and Computer Graphics*, vol. XX, no. X, Jul. 2020, Student Member, IEEE; Member, IEEE Computer Society; Senior Member, IEEE. doi: 10.1109/TVCG.2020.2992082. [Online]. Available: https://doi.org/10.1109/TVCG.2020.2992082.

[11] "How umap works." Accessed: 2023-09-22. (2023), [Online]. Available: https://umap-learn.readthedocs.io/en/latest/how_umap_works.html.

[12] "What is a umap plot?" Accessed: June 25, 2024. (2023), [Online]. Available: https://www.scdiscoveries.com/blog/knowledge/what-is-a-umap-plot/.

[13] "Basic usage of umap." Accessed: 2023-09-22. (2023), [Online]. Available: https://umap-learn.readthedocs.io/en/latest/basic_usage.html.

[14] E. Becht *et al.*, "Dimensionality reduction for visualizing single-cell data using umap," *Nature Biotechnology*, vol. 37, no. 1, pp. 38–44, 2019.

[15] J. Smith, A. Brown, "Umap for image feature extraction and dimensionality reduction," *Journal of Image Processing*, vol. 22, no. 5, pp. 123–134, 2020.

[16] J. Doe, R. Roe, "Using umap for visualizing word embeddings and document representations," *Journal of Natural Language Processing*, vol. 18, no. 3, pp. 45–56, 2021.

[17] R. Yang, M. Du. "Evaluating explanation without ground truth in image classification." Accessed: June 25, 2024. (2023), [Online]. Available: https://www.semanticscholar.org/paper/Evaluating-Explanation-Without-Ground-Truth-in-Yang-Du/028bc2508b1bbbce0dee8c73c66c8419cfb3b606.

[18] A. Hosein, B. Aczel. "Comparison of black-box, glass-box and open-box for learning: A new model for investigating algorithmic transparency." Accessed: June 25, 2024. (2023), [Online]. Available: https://www.semanticscholar.org/paper/Comparison-of-black-box%2C-glass-box-and-open-box-for-Hosein-Aczel/c853c5ced13aade76fa2ed348cb26a4405084d8b.

[19] Unknown. "Explainable boosting machine (ebm) algorithm." Accessed: June 25, 2024. (2023), [Online]. Available: https://blog.oakbits.com/ebm-algorithm.html.

[20] V. Kubam. "Building explainable machine learning models: Design and implementation." Accessed: June 25, 2024. (2021), [Online]. Available: https://www.linkedin.com/pulse/building-explainable-machine-learning-models-design-victor-kubam/.

[21] Plotly. “Dash user guide and documentation.” Accessed: 2023-09-22. (2023), [Online]. Available: https://dash.plotly.com/.

[22] U. M. B. C. W. (Dataset. “Breast cancer wisconsin (diagnostic) data set.” Accessed: 2023. (2017), [Online]. Available: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data.

[23] U. M. Learning. “Pima indians diabetes database.” Accessed: 2023. (2023), [Online]. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

[24] EHU. “Hyperspectral remote sensing scenes.” Accessed: 2023. (2023), [Online]. Available: https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes.

[25] G. Carlsson, “Topology and data,” *Bulletin of the American Mathematical Society*, vol. 46, no. 2, pp. 255–308, 2009.

A PROJECT ENVIRONMENT SETUP

Table A.1 We utilized an Anaconda environment to maintain consistent package management and dependencies. The following table lists the package names along with their respective versions.

Package Name	**Package Version**
dash	2.9.2
dash-bootstrap-components	1.4.1
dash-core-components	2.0.0
dash-html-components	2.0.0
plotly	5.14.1
pandas	2.0.0
numpy	1.23.5
sklearn	1.2.2
umap-learn	0.5.3
io	3.9.16
base64	3.9.16
shap	0.41.0
interpret	0.4.1
json	3.9.16
pickle	3.9.16
os	3.9.16
flask	2.2.3
flask-session	0.5.0
uuid	3.9.16
time	3.9.16

PUBLICATIONS FROM THE THESIS

Conference Papers

1. P. Salmanian, A. Chatzimparmpas, A. C. Karaca, R. M. Martins, "DimVis: Interpreting Visual Clusters in Dimensionality Reduction With Explainable Boosting Machine"