Probe-AI: Visual Analytics-based Model Interpretation System

Mirza Mujtaba Hussain* Nagarjun Lakshmipathy†

January 24, 2023

Abstract

We present Probe-AI, a visual analytics system designed to improve the interpretability of machine learning models in high-risk environments. By utilizing post-hoc methods such as LIME and SHAP, the system analyzes and explains the strategies used by the models and verifies them with the help of domain experts. Additionally, the system employs dimensionality reduction techniques like t-SNE to group data instances into clusters that are treated differently by the model, providing a global perspective on the model's behavior. This approach allows for a more comprehensive understanding of the model's decision-making process and enables businesses to confidently adopt and trust the predictions made by the models. Overall, Probe-AI aims to bridge the gap between complex and uninterpretable models and the ability to effectively utilize them in real-world scenarios.

1 Introduction

Machine learning models have great potential to solve a wide spectrum of real-world problems, but their complex and uninterpretable nature can make it difficult for businesses to confidently adopt them, particularly in high-risk environments such as healthcare or the insurance sector. In these cases, predictive performance alone is not sufficient and it is important to be able to hold the models up to scrutiny. To address this issue, we propose a visual analytics system, Probe-AI, that aims to improve the interpretability of machine learning models by identifying and interpreting different model strategies.

Probe-AI utilizes post-hoc methods such as LIME and SHAP to analyze and explain the strategies used by the models and verifies them with the help of domain experts. Furthermore, the system employs dimensionality reduction techniques like t-SNE to group data instances into clusters that are treated differently by the model, providing a global perspective on the model's behavior. This approach allows for a more comprehensive understanding of the model's decision-making process and enables businesses to confidently adopt and trust the predictions made by the models.

Additionally, the visual analytics approach of Probe-AI allows for an interactive exploration of the model strategies and the underlying data. This enables domain experts to quickly identify patterns and gain insights that would have been difficult to discover otherwise. Furthermore, our system provides an easy way to compare different strategies and understand how they differ from one another. This can lead to better decision making and help to choose the right model.

Probe-AI allows for the validation of the strategies used by the models, rather than relying solely on their performance on a test dataset alone which is a problematic approach as test datasets may not be properly sampled, leading to biased results.Therefore it ensures that only models with the right strategies, as determined by domain experts, are used in real-world scenarios and ultimately lead to more accurate and trustworthy predictions.

*mirza.1989740@studenti.uniroma1.it
†lakshmipathy.1982435@studenti.uniroma1.it

2 Related Work

In recent years, there has been a growing interest in developing methods for interpreting and understanding complex machine learning models. There are two main approaches in the machine learning community to produce insights: either creating inherently interpretable models (e.g., GAM [1] and CORELS [2]), or explaining models post-hoc, using an external method (e.g., LIME [3] and SHAP [4]).

In the field of interpretable models, researchers have proposed methods such as decision trees, rule-based systems, and linear models, which are considered to be more transparent and interpretable compared to other models like neural networks. However, these models often have limitations in terms of their predictive performance.

On the other hand, post-hoc explanation methods aim to provide insights into the decisions made by a black-box model. These methods typically use techniques such as feature importance and feature attribution to identify the features that contributed most to the model's decisions. LIME and SHAP are two popular post-hoc explanation methods that have been widely used to provide insights into the decisions made by machine learning models.

Visualization tools have also been developed to support the understanding of machine learning models. These tools typically focus on providing explanations for a single type of model or a specific model-agnostic method. Examples include Gamut [5], which investigates the role of interactive interfaces for model interpretation with additive models, iForest [6] which enables the interpretation of predictions by Random Forest models and GANLab [7] promotes education and understanding of Generative Adversarial Networks.

Probe-AI is related to the work done by STRATEGYATLAS [8], a visual analytics system that enables understanding of complex models by identifying and interpreting different model strategies. It uses dimensionality reduction to feature contribution vectors from explanation techniques such as LIME and SHAP.The clusters obtained are interpreted based on patterns in real data by using decision tree algorithm.However, the main difference between the two systems is that Probe-AI's use of parallel coordinates visualization allows for the easy identification of relationships and patterns in the feature contribution vectors and real data , providing a more comprehensive understanding of the model's decision-making process.

3 Data

In this work, we specifically focus on tabular data, as it is commonly used in many machine learning tasks, particularly in industries such as healthcare, finance, and insurance.The primary dataset we use is the Breast Cancer Wisconsin (Diagnostic).It contains a total of 569 instances and 32 features which are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. The features include information on the radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension of the cell nuclei present in the image. The dataset also includes a binary class label indicating whether the mass is benign or malignant.

In our work, we use the LIME (Local Interpretable Model-Agnostic Explanations) algorithm to obtain feature contribution vectors for each instance in the dataset. LIME is a post-hoc explanation method that can be used to explain the predictions of any black-box classifier.

The feature contribution vectors are represented by a set of weights assigned to each feature. A positive weight indicates that the feature has a positive impact on the prediction, while a negative weight indicates that the feature has a negative impact on the prediction. A weight of zero means that the feature does not have any impact on the prediction. These weights can be used to identify which features are most important for a specific instance and how they contribute to the prediction.

In the preprocessing step, the real data used for training the model was scaled to have values between 0 and 1. This normalization step is important for ensuring that all features are on the same scale and can be effectively used by the model. Similarly, the feature contribution values obtained from LIME are also scaled to have values between -1 and 1, which allows for easy comparison and interpretation of the contributions of different features.

id	area_mean	area_se	concave points_mean	concavity_mean	concavity_se	concavity_worst	diagnosis	fractal_dimension_mean	fractal_dimension_se	fractal_dimension_worst
842302	0.3637327677624600	0.27381125816682200	0.73111332007952300	0.70313964386129300	0.13568181818181800	0.56861022364217300	1	0.6055181128896380	0.18304243881541700	0.41886396431851000
842517	0.5015906680805940	0.12565978953974400	0.34875745526839000	0.20360824742268000	0.04696969696969700	0.19297124600639000	1	0.14132266217354700	0.09111009770186420	0.22287813196904100
84300903	0.4494167550371160	0.16292178902424000	0.63568588469184900	0.46251171508903500	0.09676767676767680	0.35974440894568700	1	0.21124684077506300	0.12700551386758400	0.21343303161485000
84348301	0.1029056203605510	0.03815479325660540	0.52286282306163000	0.56560449859418900	0.14295454545454500	0.54864217252396200	1	1.0	0.28720478697677000	0.77371113734750100

Figure 1: Original Data used to train ML model

id	area_mean	area_se	concave points_mean	concavity_mean	concavity_se	concavity_worst	diagnosis	fractal_dimension_mean	fractal_dimension_se	fractal_dimension_worst
842302	0.2844927312251740	0.3052544968252210	0.1563707826484440	0.0019343889275173900	-0.02792012438949850	0.2235073551963480 0	1	-0.005825416916838500	-0.02351569592734110	0.017158254053775600
842517	0.2685918759945130 0	0.3142727619099220 0	0.0240577833642796 00	0.0027855411490409100	-0.004668525723665610	0.1580848951834970 0	1	0.019418661741740 00	-0.020685996361375 70	0.005541350377196500
84300903	0.2705478851897130 0	0.2978027774290030	0.1625873136911870 0	0.0095925312458543 30	-0.015532824399533500	0.2471792229118770 0	1	-0.007978667464253590	-0.023650307347407400	0.012995642214103400

Figure 2: Feature Importance Data obtained from LIME

Note that the feature importance values are pre-computed and need to be provided to the system.It is important that data and feature importance values share a same key to recognise data point and its corresponding feature importance vector , in our case it is ID of each case.

4 Visaulizations

In this section, we will discuss about the visualizations used in the system and about their interactive and coordinated behaviour.

4.1 Scatter plot

Scatter plots are a type of visualization that allows us to represent two or more variables in a two-dimensional space. They are useful for understanding the relationship between two variables, and can be used to identify patterns and trends in the data.

We used scatter plots to represent the output obtained from of dimensionality reduction algorithm. By using scatter plots, we are able to represent the reduced data in a way that is easy to interpret, and to identify clusters and trends in the data and feature contribution values .

We used two scatter plots, one for the real data and another for the feature importance. The scatter plot for the real data allows us to see the clusters of real data and identify the patterns in it. While as in the scatter plot for feature contrinution values , we find clusters where each cluster represents a group of similar feature contributions, indicating a similar treatment by the model and the strategy taken by the model to make predictions. By identifying these clusters, we can gain a global understanding of the model's behavior and interpret the model's decision making process.

4.2 Parallel coordinates

Parallel coordinates are a type of visualization technique used to plot multidimensional data. The technique involves plotting each data point as a polyline along parallel axes, where each axis represents a feature or dimension of the data. The advantage of using parallel coordinates is that it allows for easy comparison and identification of patterns across multiple dimensions.

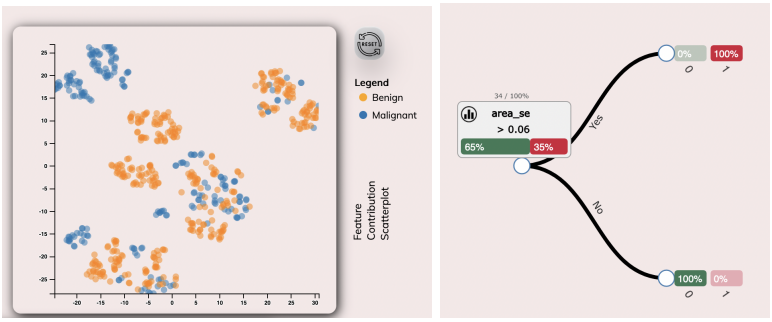

Figure 3: Scatter plot(left) and Tree visualization(right)

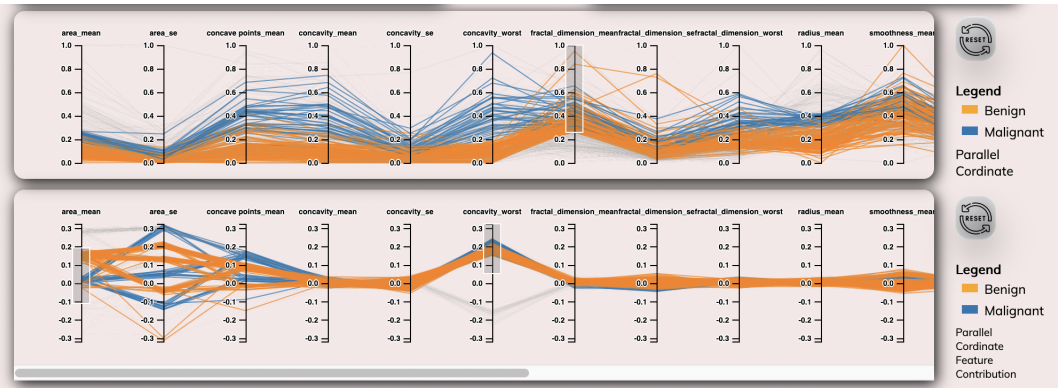

Figure 4: Parallel Plots for real and feature importance data coordinated

We used parallel coordinates to visualize both the real data and the feature importance vectors obtained from the LIME explanation technique.

For the real data, parallel coordinates allows us to see the patterns within the data by plotting each feature along a separate axis. This allows us to explore the data and identify relationships between different features.

For the feature importance vectors, parallel coordinates allows us to identify patterns in the feature importance values. By plotting each feature along a separate axis , we can see which features are most important for different instances and how they relate to one another. This allows us to understand the strategies used by the model on a high level.

Using the two parallel coordinates and keeping same scroll for them, we could easily understand the relationship between the data of a specific feature and the feature importance of that feature across all the coordinates. It also allowed to check the patterns in real data and relate that to the patterns in feature importance data.

4.3 Tree visualization

Tree visualization is a common technique used to represent hierarchical structures, such as decision trees. It allows the viewer to understand the branching structure of the tree, where each node represents a decision point and the branches represent the different outcomes of that decision. We used tree visualization to represent the rules made up by the decision tree algorithm for differentiating between the clusters.

We enable users to select two clusters in the scatter plot of feature importance vectors. After that by applying a decision tree algorithm to the selected clusters, we are able to identify the most important features that distinguish one cluster from another. We also display the accuracy with which the decision tree fits the data to give the user idea of how reliable the rules formed are.Thus, we can easily understand the strategies used by the model for predicting and how they differ between clusters.The tree visualization only appears after selecting the clusters

4.4 t-SNE

t-SNE (t-Distributed Stochastic Neighbor Embedding) is a dimensionality reduction technique that is commonly used to visualize high-dimensional data in two or three dimensions. The algorithm works by minimizing the divergence between the probability distributions of the data in the high-dimensional space and the low-dimensional space. It is particularly useful for visualizing clusters or patterns in the data that are not easily visible in the high-dimensional space.

In Probe-AI, the data from t-SNE is used to plot scatter plots. Using it we can identify clusters of similar instances in both the real data and the feature importance vectors. By setting a high value for perplexity, we can preserve the global structure of the data, which comes in handy as we want to understand the model behavior at global level and not at local level.

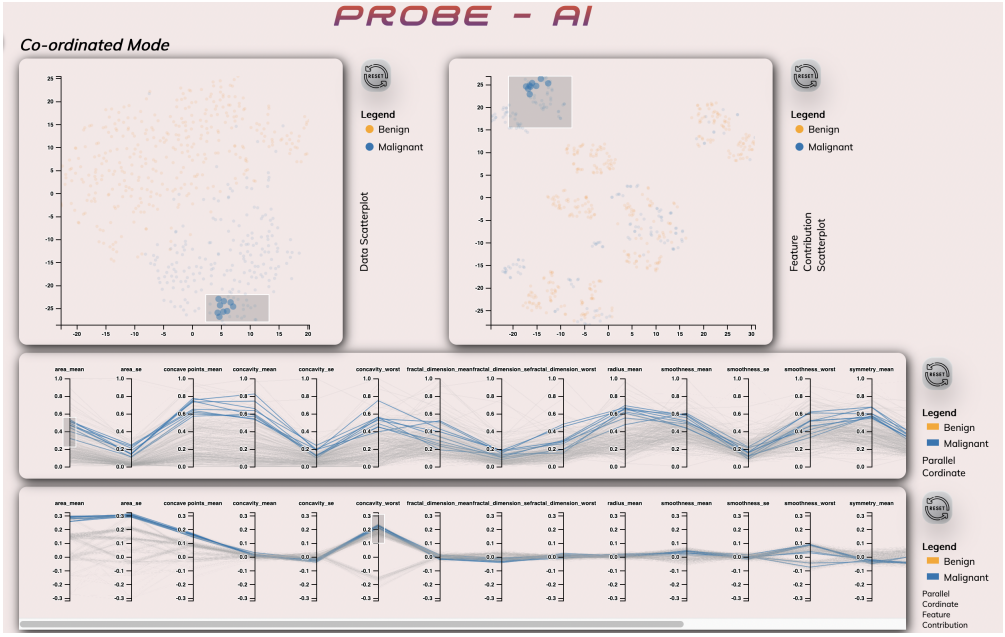

Figure 5: Coordinated view across all the visualizations

For the real data, t-SNE allows us to visualize clusters of similar instances and identify patterns in the data. This helps us to understand the relationships between different features and how they contribute to the model's predictions.

For the feature importance vectors, t-SNE helps us to find clusters that represent similar treatment by the model and helps us to understand the strategies used by the model.

Additionally, we can select a subset of features on which we want to do analysis. Every plot in Probe-AI will display the data from the new subset of features, providing a more comprehensive understanding of the model's decision-making process.

4.5 Coordination and Interaction

In Probe-AI, we have given users two modes, one is cluster mode for analysing the clusters, here the user can select two clusters using brushing in the scatter plot of feature importance vectors .The two clusters are named as red cluster and green cluster to differentiate between them easily. If there are some common points between the clusters, they are not considered while training the decision tree algorithm and not considered in either cluster.

The other mode is coordinate mode in which all of the visualizations except for the tree visualization are coordinated to work together and provide a cohesive understanding of the relationship between the data and the feature importance for each data point.

When a user selects a specific range of values of a features or set of features in the parallel co-ordinates plot, the other visualizations will automatically update to highlight the data that falls within that range. We can also use brushing on other plots at the same time and changes will be again reflected in other plot. We can see more about it in case study section. This allows the user to explore the data and the models at different levels of granularity, while still maintaining a clear understanding of the overall patterns and relationships.

The users are also given a clear button for each visualization which they can use to clear the selection/filter for a particular visualizations and undo its effect across all other visualization.Also its worth mentioning that all the visualizations are cleared when we switch from one mode to another.

All the visualizations used in the system are interactive and provide the user with information about individual points in scatter plots and parallel coordinates upon interaction.The user is given option to interact with the tree and only display the nodes he/she are interested it at the time.

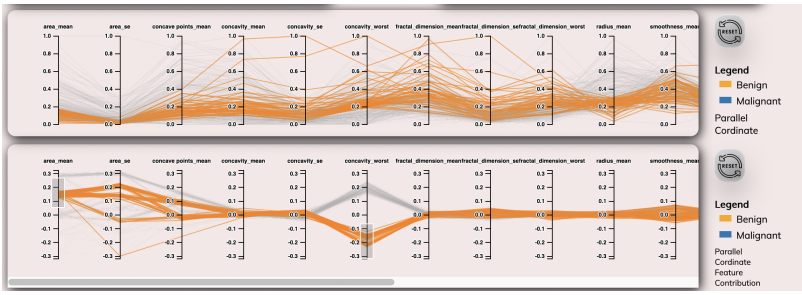

Figure 6: Concavity contradicts the prediction only when the area is much smaller

5 Case Studies

The goal of our system is to let its users like a Data Scientist, Machine Learning Engineer or a Consultant to analyse the strategies used by a Machine Learning model for prediction and validate them.Here we will discuss some scenarios that this system can be used in.

Now in order to understand the data and feature importance well, here are some general guidelines about the features:

1. radius: Larger values may indicate malignancy
2. texture: Rougher textures (lower values) may indicate malignancy
3. perimeter: Larger values may indicate malignancy
4. area: Larger values may indicate malignancy
5. smoothness: Rougher surfaces (lower values) may indicate malignancy
6. compactness: Higher values may indicate malignancy
7. concavity: Higher values may indicate malignancy
8. concave points: Higher values may indicate malignancy
9. symmetry: Lower symmetry (lower values) may indicate malignancy
10. fractal dimension: Higher values may indicate malignancy

5.1 Find relation between features and outcome

Let us suppose that a Data Science team wants to get better results but and get rid of some data that has never had a big impact on performance to cut training cost.They can use Parallel Coordinate visualization produced by feature importance vectors to find out the features that are never having impact on the final results and try to drop them from training.In our case , we found out that features such as "Fractal dimension worst" had always LIME score of close to zero.

They can even try to find features that have most impact on output and usually do not contradict with the prediction. In our case we found out the feature "Area" had huge impact when the outcome is Malignant as higher value of area almost always mean that outcome is Malignant.Also this feature rarely contradicts with the outcome when it is Malignant.

We can also find features which contradict.i.e feature importance values are mostly negative for a certain class. In our case, we found out that "concavity worst" frequently contradicts which means that even with higher value, the tumour might be benign.But in such cases area can be seen to have lower values, therefore area might be considered more important feature as compared to Concavity.

We can also find the features that the correlated features , but are not used by the model. For example, in our case, we found out radius mean can be a good feature as clearly tumours with higher radius are more likely to be Malignant as reflected by the data,but has low importance in almost all cases. This might be because of the fact that the Area and Radius are highly correlated as evident from data.

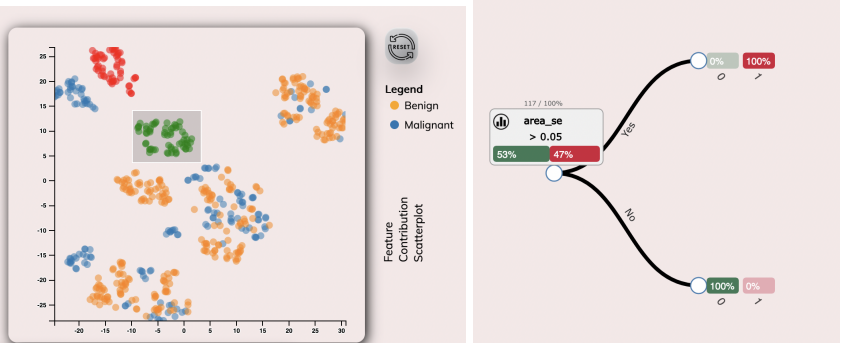

Figure 7: The red cluster contains Malignant and Green contains benign.

The team can analyse any subset of features using the drop down menu to deselect the features.In most of the cases, we see that features such as area, concave points, concavity,smoothness and texture are the features which have most impact on the outcome of the model

5.2 Identify the strategies

Now if the Data Scientist want to know if the strategies used by the model to predict are consistent, he can analyse the clusters of interest in scatter plot of importance vectors. We analysed small clusters of Malignant and Benign cases. when analysing the clusters using decision tree algorithm, we found out that the model was treating them differently based on area mean value in majority of cases.The higher area mean values were in a different cluster and were Malignant and lower area mean values were in a different cluster and were Benign.

Upon analysing another set of clusters of Malignant and Benign cases but this time with lower area mean score in feature importance, we found out that the model used texture mean value to distinguish between the clusters.In both of the cases the model has used the right strategies to distinguish between the classes.

6 Limitations

Probe-AI is a visual analytics system designed to provide a comprehensive understanding of complex machine learning models. It is primarily focused on tabular data and is only limited to classification problems. The system is dependent on the output of post-hoc explanation methods such as LIME or SHAP to make accurate assumptions about the model. There is also limitations on number of features that can be simultaneously viewed through parallel coordinates to get an overall view of data and relationships.

7 Future Work

We intend to extend Probe-AI's capabilities in the future by incorporating other post-hoc explanation methods, such as SHAP, and make it compatible with regression problems. Furthermore, we intend to improve the system by making it compatible with any datasets and models, as well as incorporating new visualization techniques such as KDE, heat maps to give more granular level information about the data.

8 Conclusion

In conclusion, Probe-AI is a visual analytics approach that enables understanding of complex machine learning models through the identification and interpretation of different model strategies. The use of parallel coordinates, scatter plots and decision trees allows for easy identification of relationships and patterns in the feature contribution vectors and real data. We used a relatively simpler data set and tested only one machine learning algorithm ,but by experimenting with

different models, we can identify how different models interpret the data and choose the most appropriate one based on right strategies,discovering real capabilities of this system.

9 Refrences

[1]Y. Lou, R. Caruana, J. Gehrke, and G. Hooker, “Accurate intelligible models with pairwise interactions,” in Proc. 19th ACM SIGKDD Conf. Knowl. Discov. Data Mining, 2013, pp. 623–631.

[2] E. Angelino, N. Larus-Stone, D. Alabi, M. Seltzer, and C. Rudin, “Learning certifiably optimal rule lists for categorical data,” arXiv preprint arXiv:1704.01701, 2017.

[3] M. T. Ribeiro, S. Singh, and C. Guestrin, ““Why should I trust you?” explaining the predictions of any classifier,” in Proc. 22nd ACM SIGKDD Conf. Knowl. Discovery Data Mining, 2016, pp. 1135–1144.

[4] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Advances Neural Inf. Process. Sys., 2017, pp. 4765–4774.

[5] F. Hohman, A. Head, R. Caruana, R. DeLine, and S. M. Drucker, “Gamut: A design probe to understand how data scientists understand machine learning models,” in Proc. CHI Conf. Human Factors Comput. Sys., 2019, pp. 1–13.

[6] X. Zhao, Y. Wu, D. L. Lee, and W. Cui, “iForest: Interpreting random forests via visual analytics,” IEEE Trans. Vis. Comput. Graphics, vol. 25, pp. 407–416, 2018.

[7] M. Kahng, N. Thorat, D. H. P. Chau, F. B. Viegas, and ´ M. Wattenberg, “GAN Lab: Understanding complex deep generative models using interactive visual experimentation,” IEEE Trans. Vis. Comput. Graphics, vol. 25, no. 1, pp. 1–11, 2018.

[8] D. Collaris and J. Van Wijk, "StrategyAtlas: Strategy Analysis for Machine Learning Interpretability," in IEEE Transactions on Visualization and Computer Graphics, doi: 10.1109/TVCG.2022.3146806.