InterpreTabNet: Enhancing Interpretability of Tabular Data Using Deep Generative Models and Large Language Models

Jacob Yoke Hong Si[1,*] **Michael Cooper**[1,2,†] **Wendy Yusi Cheng**[1,†] **Rahul G. Krishnan**[1,2]
Department of Computer Science, University of Toronto[1]
Vector Institute[2]
{jacobyhsi, coopermj, wendyc, rahulgk}@cs.toronto.edu

Abstract

Tabular data are omnipresent in various sectors of industries. Neural networks for tabular data such as TabNet have been proposed to make predictions while leveraging the attention mechanism for interpretability. We find that the inferred attention masks on high-dimensional data are often dense, hindering interpretability. To remedy this, we propose the InterpreTabNet, a variant of the TabNet model that models the attention mechanism as a latent variable sampled from a Gumbel-Softmax distribution. This enables us to regularize the model to learn distinct concepts in the attention masks via a KL Divergence regularizer. It prevents overlapping feature selection which maximizes the model's efficacy and improves interpretability. To automate the interpretation of the features from our model, we employ GPT-4 and use prompt engineering to map from the learned feature mask onto natural language text describing the learned signal. Through comprehensive experiments on real-world datasets, we demonstrate that our InterpreTabNet Model outperforms previous methods for learning from tabular data while attaining competitive accuracy and interpretability.

1 Introduction

The primary objective of our paper is to enhance the interpretability and accuracy of machine learning models applied to tabular data. We introduce InterpreTabNet, a neural network architecture that extracts actionable insights from tabular data while maintaining high accuracy in classification tasks.

The utility of this endeavor is underscored by the multifaceted applications of tabular data across diverse industries such as healthcare [Clore and Strack, 2014] and finance [Moro and Cortez, 2012], where the translation of data into intelligible insights is paramount. In complex, data-driven environments, the utility of a machine learning model is markedly amplified when it combines predictive accuracy with interpretability. This amalgamation facilitates informed, strategic decision-making, emphasizing the necessity for models that are as explanatory as they are accurate for practitioners.

Despite commendable advancements made by existing models such as TabNet [Arik and Pfister, 2020], there remains a discernible gap in achieving a harmonious integration of accuracy and interpretability. TabNet's ability to generate learnable mask for soft salient feature selection is limited as its interpretation is ambiguous. The considerable overlap between multiple masks makes it challenging for a user to discern the salient features used by the model for reasoning at each decision step. Other interpretability metrics such as attention weights [Vaswani et al., 2017] and

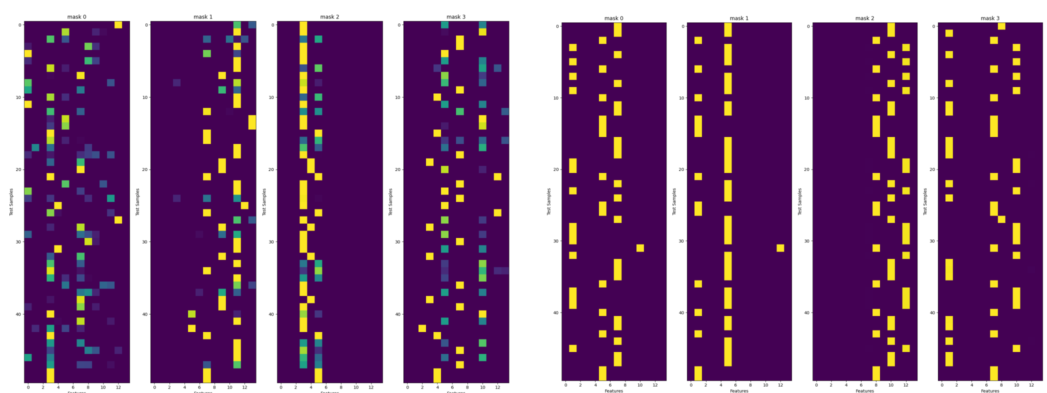
[*] Correspondence to: Jacob Yoke Hong Si.
[†] Equal contribution.

SHAP values [Lundberg and Lee, 2017] have been criticized for their inconsistency in providing meaningful insights [Roberts et al., 2022] and computational intensity for complex datasets [Jain and Wallace, 2019].

The premise of our work is that we can map the predictive signal onto a modified variant of the TabNet neural architecture, enabling us to sparsify the identity of the predictive signal; then, using tools such as large language models (LLMs) [OpenAI, 2023] we can perform a post-hoc interpretation of the source of the learned signal.

Our work makes the following contributions:

1. TabNet's Sparsity Regularizer from [Grandvalet and Bengio, 2004] promotes sparsity in the form of entropy therefore, it reuses certain features without promoting diverse feature usage. We devise a regularization scheme that maximizes diversity between masks in the TabNet architecture. Empirically, under our regularization scheme, the model learns to concentrate attention around fewer features, reducing the challenges implicit in interpreting the "soft" salient feature masks generated by TabNet. Furthermore, our method suffers from only a modest tradeoff between accuracy and interpretability: we find that our approach performs comparably to (and in most cases, better than) TabNet on a broad suite of benchmark evaluation tasks.
2. Our regularization scheme relies on maximizing the KL divergence [Kullback and Leibler, 1951] between the distributions from which each TabNet attention mask is implicitly sampled. Whereas the original TabNet formulation does not explicitly characterize these distributions, we leverage tools from variational inference to model the attention weights within TabNet as samples drawn from a Gumbel-Softmax distribution. By reformulating the attention weights within TabNet as a latent variable model, we can directly control properties of the mask distributions (such as the KL divergence) using regularized gradient-based optimization.
3. Because our method simplifies the learned importance masks generated under TabNet, one potential concern is that our method is unable to capture the rich interdependencies between features that are needed to interpret model predictions in complex settings. We show that leveraging rich linguistic priors in interpretation by means of a large language model largely ameliorates these concerns. We demonstrate how language models can relate the learned feature masks to a world model underlying the LLM [Hao et al., 2023] to form detailed hypotheses about what is being learned at each step of the TabNet decision-making pipeline.



(a) Original TabNet Model (b) InterpreTabNet ($r_M^* = 9$)

Figure 1: X/Y-axis labels denote the features and test samples for each respective mask. **Left** (a): Learned masks associated with TabNet. Observe how, for each example, there exist overlaps in the attention learned for each mask with no clear salience. This makes the masks challenging to interpret, as there is no obvious way to reconcile attention that is distributed across multiple masks in this manner. **Right** (b): Learned masks associated with InterpreTabNet. Observe how for each example, there is minimal overlap in the attention learned across different masks with high salience. This mutual exclusivity of attention across masks makes for easier visual interpretation of the learned signal that InterpreTabNet leverages in its predictions. More details can be found in Section 4.1.

2 Related Works

Learning from Tabular Data. Early work on deep learning architecture for tabular data, such as TabNet, uses a sequential attention mechanism for tabular data analysis [Arik and Pfister, 2020]. Their prominent strength is the capability to outperform other neural networks and decision trees on tabular datasets while yielding some level of interpretability for feature selections. However, TabNet's self-attention transformers' inability to capture diversifying latent variables can lead to suboptimal feature selection. To address this limitation, diversity-promoting regularizers and latent models attempt to solve this problem [Xie et al., 2017] [Xie et al., 2016]. Subsequent work on tabular data includes Net-DNF [Katzir et al., 2020], SubTab [Ucar et al., 2021], and TabTransformer [Huang et al., 2020]. Net-DNF [Katzir et al., 2020] introduces an inductive bias that aligns model structures with disjunctive normal form (DNF) and emphasizes localized decisions. SubTab [Ucar et al., 2021] transforms tabular data into a multi-view representation learning task, enhancing latent representation. Furthermore, TabTransformer [Huang et al., 2020] is a deep tabular data modelling architecture built upon self-attention-based Transformers.

Latent Variable Models. Latent variable models like VAEs [Kingma and Welling, 2022] and their variations have demonstrated attractive abilities to model complex distributions and produce latent values. DirVAE has more interpretable latent values with no collapsing issues [Joo et al., 2019], while the cVAE [Kristiadi, 2016] models latent variables and observed data, both on random variables, which gain control of the data generation process on the VAE. Additionally, the cVAE also generates diverse but realistic output representations using stochastic inference [Sohn et al., 2015]. Transformer-based cVAE demonstrates its excellent representation learning capability and controllability [Fang et al., 2021]. We draw inspiration from these VAE extensions and incorporate the cVAE into TabNet's architecture to capture and reconstruct discrete data.

Recent works in approximate inference for categorical data include Categorical Reparameterization with Gumbel-Softmax [Jang et al., 2016]. In our paper, we leverage the Gumbel-Softmax distribution as a key component of our methodology to strike a balance between interpretability and performance.

Model Interpretability. Methods from interpretability aim to surface information about *why* a machine learning model is making certain predictions to user. Broadly, there are two families of methods in model interpretability. *Intrinsic interpretability* refers to the scenario in which the user can directly leverage the parameters learned by the model to understand the rationale underlying the predictions. Linear models [Gauss, 1877], decision trees, Transformers (by means of their learned attention weights), and TabNet [Arik and Pfister, 2020], are all, to varying degrees, intrinsincally interpretable methods. In contrast, methods from *post-hoc interpretability* tackle the scenario in which the model may be black-box: these methods instead attempt to approximate the decision-making process underlying the model, which is then surfaced to the user. Methods like SHAP Lundberg and Lee [2017], LIME Ribeiro et al. [2016], and Grad-CAM [Selvaraju et al., 2017] are methods for post-hoc interpretability. The central tradeoff between intrinsic and post-hoc interpretability is this: while an intrinsically interpretable model is (definitionally) faithful to its underlying decision rule, it may be necessary to make simplifying assumptions in the design of the model. Conversely, while post-hoc interpretability methods can interpret models of arbitrary complexity, the interpretable decision rule surfaced by such procedures is only an approximate one Du et al. [2019]. Our approach draws upon insights from both classes of methods: we leverage tools from variational inference to improve upon the intrinsic interpretability of TabNet, and we employ a large language model to provide a richer contextual interpretation of the learned features post-hoc.

3 The InterpreTabNet Model

Let $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathcal{X} \times \mathcal{Y}$ represent the covariates and outcome that we want to model, respectively. As we are operating in the tabular data regime, assume that $X \in \mathbb{R}^{N \times D}$, where each $d \in [1, ..., D]$ corresponds to a single discrete feature in the data. Then, each $x^{(i)}, y^{(i)}$ represents D-vector and label corresponding to a particular example.[3] Let $P(\cdot|\cdot)$ denote true probability density functions, and $Q(\cdot|\cdot)$ denote variational approximations of those densities.

[3]Unless otherwise stated, our notation uses uppercase letters to refer to distribution-level quantities, such as the distribution over the covariates, and lowercase letters to refer to specific samples drawn from those distributions.

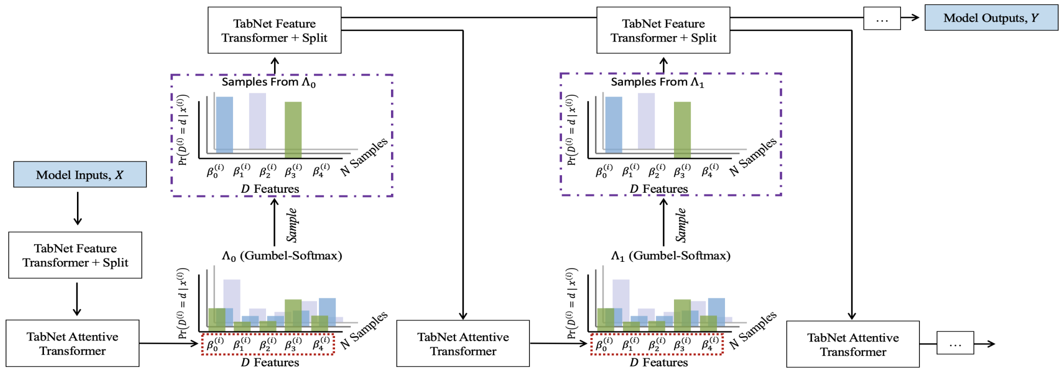


Figure 2: The InterpreTabNet architecture presents a variational formulation of the TabNet encoder. In our formulation, the weights of the attention masks produced by the TabNet encoder at each step k are treated as the parameters, $\beta_0^{(i)}, ..., \beta_{D-1}^{(i)}$, of a Gumbel-Softmax distribution, Λ_k, unique to each instance (shown by the red dotted rectangle). This distribution is then sampled to produce a single feature that is highlighted for each feature at each step (purple dot-dashed rectangle). This figure shows $k = 2$ steps of the encoder architecture, over $D = 5$ features, for $N = 3$ samples.

3.1 High-Level Approach

The TabNet encoder architecture models the prediction process, $P(y \mid x)$, as a nonlinear combination of the covariates, x, and a sequence of k *learned attention masks*. Feature importance mask m_k depicts the feature selected at the k-th decision step. Each mask, m_k, is learned by applying the TabNet Transformer in the encoder to the covariates and previous attention mask at each step of a multi-step decision process. Since the nonlinear combination is modeled using a multi-layer perceptron [Haykin, 1994], inference within TabNet's encoder can be expressed as:

$$\Pr(y \mid x) = f_\psi^{(\text{MLP})}\left(\sum_{k=0}^{K-1} f_\psi^{(\text{TabNet\_Transformer})}(x, m_k)\right), \tag{1}$$

where $m_k = \varnothing$ if $k = 0$, and where ψ is a general-purpose variable used to denote the parameters associated with a given MLP or TabNet Transformer sub-model. Our goal is to construct a version of this model wherein each mask is treated as a latent variable in a deep generative model. Then we can learn the model via amortized variational inference by inferring m_k using some parametric distribution Q that admits backpropagation by means of the reparameterization trick. While the attention masks are represented in a continuous space $m_k^{(i)} \in \mathbb{R}^{D\times N}$, the prominent features within these masks exhibit distinct patterns that we treat in a discrete-like manner for our modeling purposes thus, we will sample these masks from a discrete latent distribution. By specifying the form of the distribution from which the masks are sampled, we can directly adjust the properties of this latent variable by regularizing the loss function. Specifically, as our objective is to promote sparsity among the masks, we will then aim to maximize the KL divergence between learned masks.

In the following sections, we will introduce the Gumbel-Softmax distribution [Jang et al., 2016], and how we leverage it in our variational formulation of the TabNet encoder. The Gumbel-Softmax distribution offers a continuous relaxation of discrete variables, such as categorical variables, and facilitates the reparameterization of categorical latent factors by approximating samples from a categorical distribution, making it possible to compute gradients during training smoothly.

3.2 Mask Sampling Process

The mask sampling process for our model is the following, where Λ represents a Gumbel-Softmax distribution.

$$\begin{aligned} P(m_k|X) &\sim \Lambda_k(\text{TabNet\_Transformer}(X)), && k = 0 \\ P(m_k|\hat{Y}_k, X) &\sim \Lambda_k(\text{TabNet\_Transformer}(\hat{Y}_k, X)), && \forall k \in [1, \ldots, K-1]. \end{aligned}$$

TabNet's model does not leverage its feature importance masks to make predictions. Instead, it acts as a deterministic system by producing its masks directly via its attentive transformer. On the other hand, since TabNet outputs a feature mask from the first training iteration onwards, we can utilize these masks in the subsequent iterations as latent variables. These latent variables serve as a rich source of embedded knowledge, allowing the model to improve its generalizations by acting as a stochastic process. Furthermore, sampling this latent variable from the Gumbel-Softmax distribution will act as a crucial component in improving interpretability (details explored in Section 3.4).

Let us represent the collection of all k masks, $[m_0, ..., m_{k-1}]$ as a single latent variable, $z \in \mathbb{R}^{D\times k}$, drawn from a Gumbel-Softmax distribution (note that TabNet employs ReLU to construct the overall decision embedding from the samples however, the Gumbel-Softmax does not require this attribute). Drawing samples z from a categorical distribution with class probabilities π is as follows.

$$z = \text{one\_hot}\left(\arg\max_i(\beta_i + \log \pi_i)\right)$$

where $\beta_0, ..., \beta_{D-1}$ are i.i.d samples drawn from the TabNet_Transformer$(\cdot)$ Gumbel(0,1) output.

As observed, the mask sampling process is characterized as a latent variable problem. This necessitates the implementation of inference techniques for effective learning.

3.3 Generating Predictions with the Conditional Variational Autoencoder

TabNet's innate encoder-decoder architecture enables us to integrate a cVAE to generate predictions. To reiterate, Y represents the predicted outcome, z represents the concatenation of all the m_k masks sampled from a Gumbel-Softmax distribution, and X represents the data. Using the aforementioned problem setup, TabNet's encoder is now conditioned on two variables, Y and X: $Q(z|Y, X)$. Similarly, TabNet's decoder is also conditioned on two variables, z and X: $P(Y|z, X)$. Therefore, the cVAE's objective is to model the outcome, $P(Y|X)$ as follows: $\int P(Y|X, z)P(z|X)dz$. To do so, we need to infer $P(z)$ through $P(z|Y)$ using $Q(z|Y)$. This gives us a variational lower bound objective of the following which we will maximize. The full derivation can be found in Appendix 5.1.

$$\log P(Y|X) - D_{KL}[Q(z|Y,X)||P(z|Y,X)] = E[\log P(Y|z,X)] - D_{KL}[Q(z|Y,X)||P(z|X)] \quad (2)$$

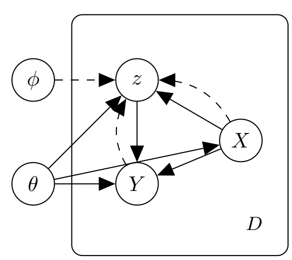


Figure 3: Graphical model of InterpreTabNet with D i.i.d samples. Solid lines denote the generative model $p_\theta(Y|z, X)p_\theta(z|X)$, dashed lines denote the variational approximation $q_\phi(z|X, Y)$ to the intractable posterior $p_\theta(z|X, Y)$. The variational parameters ϕ are learned jointly with the generative model parameters θ.

3.4 Sparsity-Promoting Regularization

In the previous sections, we have modeled the original TabNet as a stochastic cVAE. This provides us with a setup to leverage the Gumbel-Softmax distributed masks to promote sparsity. The idea is to encourage diversity between masks for a more even selection of features and better interpretability. Hence, we can incorporate a KL Divergence Sparsity Regularizer (r_M) in the model architecture. With the KL Divergence, we aim to maximize the difference between the distribution of masks that are subsequent to one another. This would reduce the number of selected features, ensuring that the features selected are independent between masks. Additionally, with a sparser feature selection, the model can focus on fewer high-salience features. Therefore, the ELBO of the InterpreTabNet model is as follows with r_M as a tunable regularizer weight.

$$E[\log P(Y|z,X)] - \sum_{i} D_{KL}\Big(\left(Q\left(z_i|Y,X\right)\right) \Big|\Big| \left(P\left(z_i|X\right)\right)\Big) + r_M \cdot \sum_{i \neq j} D_{KL}\Big(\left(Q\left(z_i|Y,X\right)\right) \Big|\Big| \left(Q\left(z_j|Y,X\right)\right)\Big) \quad (3)$$

3.5 Sparsity Regularizer (r_M) Algorithm

To assess the level of interpretability a feature mask provides, we divide it into two set criteria.

1. Number of selected features.
2. "Salience" of each feature.

Within a feature mask, we would like to swiftly identify the features that are of the highest salience which contributes to its prediction. Thus, our aim is to minimize the number of selected features, and only select those of the highest salience, while maintaining a competitive accuracy. This would yield the most interpretable mask.

We propose an adaptive algorithm to optimize our KL Divergence Sparsity Regularizer, r_M, for better interpretability of the feature masks. Our method involves iterative training and evaluation of the InterpreTabNet model with varying values of r_M within a pre-defined range. Simultaneously, the model's feature importance masks are analyzed to validate that they meet a set criterion. Upon fulfilling the criterion a specified number of times, the algorithm terminates. For increased efficiency, the algorithm also employs a recursive search to narrow down the value range around the current best r_M, thereby reducing computational overhead. The end result is the optimal r_M value corresponding to the most interpretable feature mask and highest classification accuracy, improving the overall efficacy of our model. The algorithm can be found in Appendix 5.2.

4 Experiments and Discussions

We evaluate the performance of the InterpreTabNet on real-world classification tasks both quantitatively and qualitatively.

Datasets The model performance is evaluated on real-world tabular datasets from UCI Machine Learning Repository[Kelly et al., 2023] and OpenML[Vanschoren et al., 2013]. These datasets were selected since they were used to evaluate the existing methods (baselines). Additionally, they have varied size and nature, with both categorical and continuous features to ensure a holistic evaluation of our methodology across multiple domains and scenarios. The training/validation/testing proportion of the datasets for each split are 80/10/10% apart from the Higgs dataset. Due to the inherently large Higgs dataset, we follow TabNet's method of data split with 500k training samples, 100k validation samples and 100k testing samples. Details of the datasets can be found in Appendix 5.4.

Baselines We compare our model against five other ML methods for tabular classification. This includes the Original TabNet, transformer-based tabular model, TabTransformer [Huang et al., 2020], tree-based boosting methods, XGBoost [Chen and Guestrin, 2016], LightGBM [Ke et al., 2017] and multi-layer perceptrons [Huang et al., 2020]. For each ML model, we utilize the recommended hyperparameters mentioned by the authors of their respective papers. Furthermore, we also conduct a grid search within the range of the recommended hyperparameters to optimize the models, selecting the best-performing hyperparameter configuration.

4.1 Results

The performance of our method relative to baselines for tabular learning is shown in Table 1.

In the following section, we will be exploring the Adult Census Income dataset [Becker and Kohavi, 1996] to evaluate InterpreTabNet against Original TabNet.

Figure 1 highlights the learned masks associated with InterpreTabNet using a sparsity regularizer value of $r_M = 9$ (right) compared to those of TabNet (left). The rows of each mask represent individual data samples, while the columns represent discrete features in the tabular data. Colors that are brighter indicate features of higher salience. As observed in Figure 1, feature masks of the

Table 1: Test Accuracy Scores with Optimal Mask Regularizer Values (r_M) across Different Models and Datasets. Our InterpreTabNet achieved substantial improvements in interpretability across all the datasets and remains competitive in terms of accuracy in most datasets.

Model / Dataset	Adult Census	Forest Cover	Poker Hand	Mushroom	Blastchar	Diabetes	Higgs
InterpreTabNet	**87.42**	**94.75**	99.50	96.62	72.96	55.37	53.08
Original TabNet	85.55	94.18	99.00	99.94	76.22	56.91	52.94
TabTransformer	85.09	82.55	**99.81**	**100.00**	73.17	44.45	51.97
XGBoost	86.60	92.30	75.57	99.69	77.29	**61.44**	**72.70**
LightGBM	86.20	86.38	78.47	**100.00**	**77.86**	60.87	72.62
MLP	79.76	84.89	99.70	99.82	75.16	53.99	63.17

Original TabNet are more difficult to interpret since each mask may highlight multiple features for a given data sample. On the contrary, our InterpreTabNet model highlights mutually exclusive features of high importance that are more easily interpretable.

In an ablation study on how varying r_M values affect our masks (found in Appendix 5.7), we notice that at low r_M values, test accuracy is high but feature selection diversity is poor and interpretability is difficult since almost all features are selected in the decision-making process. On the other hand, at high r_M values, the masks are sparse which are easily interpretable but at a cost of accuracy.

In terms of computational efficiency, our model necessitates an additional computation through the Gumbel-Softmax reparameterization and also requires conditioning on the mask from the previous time step when compared to TabNet. Nonetheless, this extra step incurs a minimal cost, leading to a mere several-minute increase in training time. Furthermore, likewise to TabNet, our model maintains greater computational efficiency compared to other baseline models without necessitating an extensive search for fine-grained hyperparameters.

4.2 Feature Mask Aggregate Interpretability Analysis using Prompts with LLMs

To perform aggregate analysis on the feature masks, we leverage LLMs such as GPT-4 and design a prompt to generate an informative explanation of the feature masks. The prompt is constructed in the order of complexity from the simplest prompt to a complex prompt with more information. Initially, a basic prompt of "Conduct aggregate analysis on the description of the following feature mask. (followed by the mask description)". However, as predicted, the output clearly lacks a lot of information such the ability to analyze the masks by its salient features and a precise output mapping.

In order to generate a precise output mapping, instructions are provided to GPT-4 that the extracted salient features should be formatted into a dictionary where each mask corresponds to an individual analysis followed by an aggregate analysis of all masks. Furthermore, a statement to ensure that no other natural language generation is produced by GPT-4 is added in order to maintain a consistent output map. Last but not least, GPT-4 is provided with in-context examples to enable prompt tuning through few-shot learning. This is conducted via 3-fold cross-validation where dataset D1 and D2 is used as part of prompt for tuning on D3; D2 and D3 as part of prompt for tuning on D1 and so on. Only a 3-fold CV is conducted since increasing the subsets will decrease GPT-4's performance as it is unable to process extremely long sequences of texts.

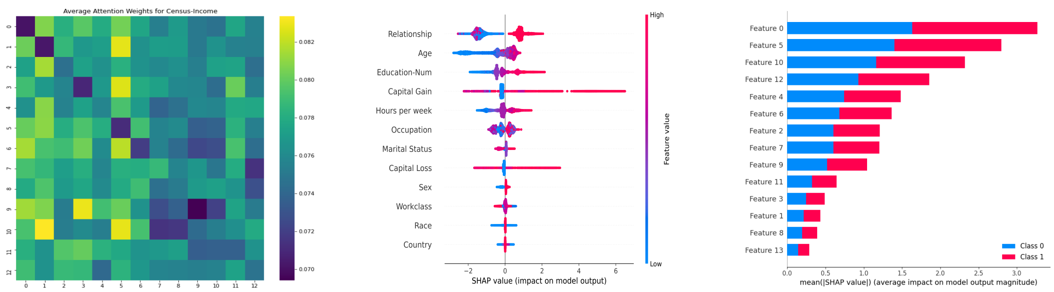
Overall, this improves the generalization of GPT-4 when extracting salient features from new datasets. The structure of the designed prompt can be found in Table 2. The full prompts and outputs can be found in Appendix 5.8 and 5.9 respectively.

4.3 Evaluating the Interpretability of InterpreTabNet against other baselines

Figure 4 illustrates the complex pathways of model interpretation inherent in prominent machine learning architectures like TabTransformer, XGBoost, and LightGBM. These models, while powerful, necessitate nuanced or additional tools to render interpretative insights. TabTransformer relies on attention weights, and both XGBoost and LightGBM are augmented with SHAP values derived from external SHAP packages to achieve interpretability.

Table 2: Prompt Structure Design

Section	Description
Dataset Description	The Adult Census Income dataset is considered...
Mask Description	At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 1, 5, and 7 which are workclass, marital status, and relationship...
In-Context Example 1	The Poker Hand dataset is considered...
In-Context Output 1	Output: {"Mask 0": "Initially, the rank of card 2 is recognized...}
In-Context Example 2	The Forest Cover Type dataset is considered...
In-Context Output 2	Output: {"Mask 0": "The initial feature selection identifies...}
GPT-4 Output	*{"Mask 0": "This suggests that early in the feature selection process, the model finds that work-related factors...}*



(a) TabTransformer Att. Wgts. (b) XGBoost SHAP (c) LightGBM SHAP

Figure 4: (a), (b), and (c) depict the TabTransformer Attention Weights, XGBoost SHAP analysis, and LightGBM SHAP analysis for the Adult Census Income Dataset, respectively.

Building upon the foundation laid by the original TabNet, our InterpreTabNet elevates the standard of intrinsic interpretability. While TabNet is celebrated for its feature selection masks, which offer real-time insights into feature contributions, the InterpreTabNet refines and enhances this feature, offering an advanced level of clarity and detail in feature importance interpretation.

The enhanced masks of the InterpreTabNet illuminate feature contributions with heightened clarity, offering practitioners and stakeholders an unprecedented level of understanding. This improvement facilitates more informed decision-making, bridging gaps in knowledge and enabling a closer alignment between model predictions and actionable insights. The immediacy and clarity of these enhanced masks combined with the feature analysis capabilities of GPT-4 underscore the model's adaptability and effectiveness in scenarios demanding rapid, yet deeply insightful, interpretations.

5 Conclusion

We propose an interpretable variant of the TabNet neural network that is as expressive in learning the distributions of tabular data while enabling an enhanced level of interpretability. This model is designed by blending a Gumbel-Softmax distribution with a KL divergence sparsity regularizer between the attention-based feature masks to create a sparse and semantically meaningful decomposition of the predictive signal. Relative to TabNet, our model outputs more interpretable feature masks while maintaining its competitive accuracy across all baselines for most datasets. The salient features from our masks are channelled into GPT-4 via a carefully engineered prompt that outputs an analysis of the features. For practitioners, the InterpreTabNet stands as a practical toolkit for understanding where predictive signal from tabular data comes from. It bridges the often challenging gap between intricate machine learning outputs and real-world decision-making, ensuring that insights are not just extracted but are also intuitively understood and readily actionable.

References

Mushroom. UCI Machine Learning Repository, 1987. DOI: https://doi.org/10.24432/C5959T.

S. O. Arik and T. Pfister. Tabnet: Attentive interpretable tabular learning, Feb 2020. URL `https://arxiv.org/abs/1908.07442v4`.

B. Becker and R. Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

BlastChar. Telco customer churn, Feb 2018. URL `https://www.kaggle.com/datasets/blastchar/telco-customer-churn`.

R. Cattral and F. Oppacher. Poker Hand. UCI Machine Learning Repository, 2007. DOI: https://doi.org/10.24432/C5KW38.

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

C. K. D. J. Clore, John and B. Strack. Diabetes 130-US hospitals for years 1999-2008. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5230J.

M. Du, N. Liu, and X. Hu. Techniques for interpretable machine learning. *Communications of the ACM*, 63(1):68–77, 2019.

D. Dua and C. Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

L. Fang, T. Zeng, C. Liu, L. Bo, W. Dong, and C. Chen. Transformer-based conditional variational autoencoder for controllable story generation, Jul 2021. URL `https://arxiv.org/abs/2101.00828v2`.

C. F. Gauss. *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*, volume 7. FA Perthes, 1877.

Y. Grandvalet and Y. Bengio. Semi-supervised learning by entropy minimization. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004. URL `https://proceedings.neurips.cc/paper_files/paper/2004/file/96f2b50b5d3613adf9c27049b2a888c7-Paper.pdf`.

S. Hao, Y. Gu, H. Ma, J. J. Hong, Z. Wang, D. Z. Wang, and Z. Hu. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*, 2023.

S. Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.

X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*, 2020.

S. Jain and B. C. Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax, 2016. URL `https://arxiv.org/pdf/1611.01144.pdf`.

W. Joo, W. Lee, S. Park, and I.-C. Moon. Dirichlet variational autoencoder, Jan 2019. URL `https://arxiv.org/abs/1901.02739v1`.

L. Katzir, G. Elidan, and R. El-Yaniv. Net-dnf: Effective deep modeling of tabular data. In *International conference on learning representations*, 2020.

G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf`.

M. Kelly, R. Longjohn, and K. Nottingham. The UCI machine learning repository, 2023. URL https://archive.ics.uci.edu.

D. P. Kingma and M. Welling. Auto-encoding variational bayes, 2022.

A. Kristiadi. Conditional variational autoencoder: Intuition and implementation, 2016. URL https://agustinus.kristia.de/techblog/2016/12/17/conditional-vae/.

S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

R. P. Moro, S. and P. Cortez. Bank Marketing. UCI Machine Learning Repository, 2012. DOI: https://doi.org/10.24432/C5K306.

OpenAI. Gpt-4 technical report, 2023.

M. T. Ribeiro, S. Singh, and C. Guestrin. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016. URL http://arxiv.org/abs/1602.04938.

C. V. Roberts, E. Elahi, and A. Chandrashekar. On the bias-variance characteristics of lime and shap in high sparsity movie recommendation explanation tasks. *arXiv preprint arXiv:2206.04784*, 2022.

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models, Jan 2015. URL https://papers.nips.cc/paper/2015/hash/8d55a249e6baa5c06772297520da2051-Abstract.html.

T. Ucar, E. Hajiramezanali, and L. Edwards. Subtab: Subsetting features of tabular data for self-supervised representation learning. *Advances in Neural Information Processing Systems*, 34: 18853–18865, 2021.

J. Vanschoren, J. N. van Rijn, B. Bischl, and L. Torgo. Openml: Networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2013. doi: 10.1145/2641190.2641198. URL http://doi.acm.org/10.1145/2641190.2641198.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

D. Whiteson. HIGGS. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5V312.

P. Xie, J. Zhu, and E. Xing. Diversity-promoting bayesian learning of latent variable models. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 59–68, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/xiea16.html.

P. Xie, J. Zhu, and E. P. Xing. Diversity-promoting bayesian learning of latent variable models, Nov 2017. URL https://arxiv.org/abs/1711.08770.

Appendix

5.1 Proof: cVAE Evidence Lower Bound

$$
\begin{aligned}
D_{KL}[Q(z|Y,X)||P(z|Y,X)] &= \sum_z Q(z|Y,X)\log\frac{Q(z|Y,X)}{P(z|Y,X)} \\
&= E[\log\frac{Q(z|Y,X)}{P(z|Y,X)}] \\
&= E[\log Q(z|Y,X) - \log P(z|Y,X)] \\
&\text{using Bayes' rule} \\
&= E[\log Q(z|Y,X) - \log\frac{P(z,Y,X)}{P(Y,X)}] \\
&= E[\log Q(z|Y,X) - \log\frac{P(Y|z,X)P(z,X)}{P(Y,X)}] \\
&= E[\log Q(z|Y,X) - \log\frac{P(Y|z,X)P(z|X)P(X)}{P(Y,X)}] \\
&= E[\log Q(z|Y,X) - \log\frac{P(Y|z,X)P(z|X)P(X)}{P(Y|X)P(X)}] \\
&= E[\log Q(z|Y,X) - \log\frac{P(Y|z,X)P(z|X)}{P(Y|X)}] \\
&= E[\log Q(z|Y,X) - (\log P(Y|z,X) + \log P(z|X) - \log P(Y|X))] \\
&= E[\log Q(z|Y,X) - \log P(Y|z,X) - \log P(z|X) + \log P(Y|X)] \\
&= E[\log Q(z|Y,X) - \log P(Y|z,X) - \log P(z|X)] + \log P(Y|X) \\
D_{KL}[Q(z|Y,X)||P(z|Y,X)] - \log P(Y|X) &= E[\log Q(z|Y,X) - \log P(Y|z,X) - \log P(z|X)] \\
&\text{rearranging the sign to rewrite RHS as another KL Divergence} \\
\log P(Y|X) - D_{KL}[Q(z|Y,X)||P(z|Y,X)] &= E[\log P(Y|z,X) - (\log Q(z|Y,X) - \log P(z|X))] \\
\log P(Y|X) - D_{KL}[Q(z|Y,X)||P(z|Y,X)] &= E[\log P(Y|z,X)] - E[\log Q(z|Y,X) - \log P(z|X)] \\
\log P(Y|X) - D_{KL}[Q(z|Y,X)||P(z|Y,X)] &= E[\log P(Y|z,X)] - E[\log\frac{Q(z|Y,X)}{P(z|X)}] \\
\log P(Y|X) - D_{KL}[Q(z|Y,X)||P(z|Y,X)] &= E[\log P(Y|z,X)] - D_{KL}[Q(z|Y,X)||P(z|X)] \\
&\text{which is the cVAE objective function}
\end{aligned}
$$

5.2 Algorithm: KL Divergence Sparsity Regularizer r_M

Algorithm 1 Our proposed algorithm for interpretability optimization. Good default settings for the tested machine learning problems are $\alpha = 0$, $\beta = [0, 10000000]$, $\delta = [0.20, 0.25]$, $\gamma = [2, 3]$ $\epsilon = [3, 5]$. For β, δ and γ, it would depend on the nature of the dataset. More samples require higher parameter values.

Require: α: Starting range (start)
Require: β: Ending range (end)
Require: δ: Percentage of feature importance captured by one feature in each feature mask (col_threshold_val)
Require: γ: Number of columns that satisfies δ in each feature mask (col_threshold)
Require: ι: Number of complete-feature masks that passes the algorithm's feature selection criteria (all_mask_pass)
Require: ϵ: Threshold for the number of complete-feature masks that passes the algorithm's feature selection criteria (all_mask_pass_thresh)
Require: ζ: Step size computed using a logarithmic scale at high levels (step_size)
Require: θ: Dictionary storing r_M-accuracy pairs (reg_m_acc_dict)
Require: λ: Flag for recursion (is_recursive)
Ensure: Optimal regularization parameter r_M^*

Initialize θ if θ is None.
Initialize ι if ι is None.
if $\iota = \epsilon$ **then**
    $r_M^* = \arg\max(\theta)$
    return r_M^*
end if
while $\alpha \leq \beta$ and $\iota < \epsilon$ **do**
    Train TabNet, Compute Accuracy and Generate Masks ▷ Inner loop evaluating each feature mask here.
    if Criteria for updating θ and ι are met **then**
        Update θ, ι
    end if
    if λ **then**
        $\alpha = \alpha + \zeta$
    else if $\alpha = 0$ **then**
        $\alpha = 10$
    else
        $\alpha * = 10$
    end if
end while
if r_M^* is Not None & Length of $\theta = 1$ **then**
    Recurse with updated boundaries.
else
    $r_M^* = \arg\max(\theta)$
    return r_M^*
end if

5.3 Reproducibility

Code Release The code for InterpreTabNet and files to reproduce the experiments are available on GitHub at `https://github.com/jacobyokehongsi/InterpreTabNet`. The code will be released once the paper is finalized, and the release aims to ensure the research's reproducibility.

Availability of Datasets The datasets used in this paper are all freely accessible on OpenML. `OpenML.org` and `UCI Machine Learning Repository`. Download links and additional statistical details about the datasets can be found in Appendix 5.4 of the paper.

5.4 Additional Dataset Information

We evaluated our model on 7 datasets. These datasets contain 4 binary classification tasks and 3 multi-class classification tasks. We provided statistical details in Table 3, and download links in Table 4. In each of our datasets, we applied label encoding to the categorical features to transform textual values into numerical representations. Additionally, we introduced a distinct token to handle missing data within these categorical columns. This uniform preprocessing approach was applied consistently across all datasets, ensuring compatibility and reliability for subsequent machine learning analyses.

Table 3: Datasets used for evaluation

Dataset	**Task**	**# Features**	**# Categorical**	**# Instances**	**# Classes**	**# NaNs**
Adult Census Income	Binary	14	8	32,560	2	0
Forest Cover Type	Multi-Class	54	44	581,012	7	0
Poker Hand	Multi-Class	10	10	1,025,010	10	0
Mushroom	Binary	22	22	8,124	2	0
Blastchar	Binary	20	17	7,043	2	0
Diabetes	Multi-Class	49	39	101,766	3	0
Higgs	Binary	28	0	11,000,000	2	0

Table 4: Dataset Links

Dataset Name	**Dataset Link**
Adult Census Income	`https://archive.ics.uci.edu/dataset/2/adult`
Forest Cover Type	`https://archive.ics.uci.edu/dataset/31/covertype`
Poker Hand	`https://archive.ics.uci.edu/dataset/158/poker+hand`
Mushroom	`https://archive.ics.uci.edu/dataset/73/mushroom`
Blastchar	`https://www.kaggle.com/datasets/blastchar/telco-customer-churn`
Diabetes	`https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008`
Higgs	`https://archive.ics.uci.edu/dataset/280/higgs`

5.5 Hyperparameters Search Space

We provided hyperparameter search spaces for all models in Table 5. For TabTransformer, we used the same hyperparameter space mentioned in their paper Huang et al. [2020]. XGboost and LightGBM were designed from scratch and used common hyperparameter choices with suggestions from the official documentation Chen and Guestrin [2016] Ke et al. [2017]. For MLP, we followed the exact hyperparameter search space as Huang et al. [2020].

Table 5: Hyperparameter spaces for all models

Model	**Hyperparameter Space**
InterpreTabNet	$N_d = N_a$ (output dimension): [16, 32, 128], N_{steps}: [3, 4, 5], γ: [1.0, 1.2, 1.5, 2.0], λ: [0.001, 0.01, 0.1, 0.3], Learning Rate: [0.005, 0.01, 0.02, 0.025], r_M: range from [0, 1,000,000,000,000]
Original TabNet	$N_d = N_a$ (output dimension): [16, 32, 128], N_{steps}: [3, 4, 5], γ: [1.0, 1.2, 1.5, 2.0], λ: [0.001, 0.01, 0.1, 0.3], Learning Rate: [0.005, 0.01, 0.02, 0.025], r_M: range from [0, 1,000,000,000,000]
TabTransformer	Hidden Dimension: [32, 54, 128, 256], Number of Layers: [1, 2, 3, 6, 12], Number of Attention Heads: [2, 4, 8], MLP First Hidden Layer: $x = m \times l, m \in \mathbb{Z} \mid 1 \leq m \leq 8$, where l is the input size, MLP Second Hidden Layer: $x = m \times l, m \in \mathbb{Z} \mid 1 \leq m \leq 3$, where l is the input size
XGBoost	learning_rate: [0.01, 0.1, 0.2], max_depth: [3, 4, 5, 6], n_estimators: [50, 100, 200], subsample: [0.8, 0.9], colsample_bytree: [0.8, 0.9], min_child_weight: [1, 2, 3]
LightGBM	num_leaves: [20, 30, 40], learning_rate: [0.05, 0.1, 0.2], n_estimators: [100, 200], subsample: [0.8, 0.9], colsample_bytree: [0.8, 0.9]
MLP	First Hidden Layer: $x = m \times l, m \in \mathbb{Z} \mid 1 \leq m \leq 8$, where l is the input size, Second Hidden Layer: $x = m \times l, m \in \mathbb{Z} \mid 1 \leq m \leq 3$, where l is the input size

5.6 Results from other datasets (Accuracies & Masks)

Forest Cover Type [Dua and Graff, 2017]

Model	**Test Accuracy (%)**
XGBoost	92.30
LightGBM	86.38
TabTransformer	82.55
MLP	94.27
Original TabNet	94.18
InterpreTabNet ($r^*_M = 900$)	**94.75**

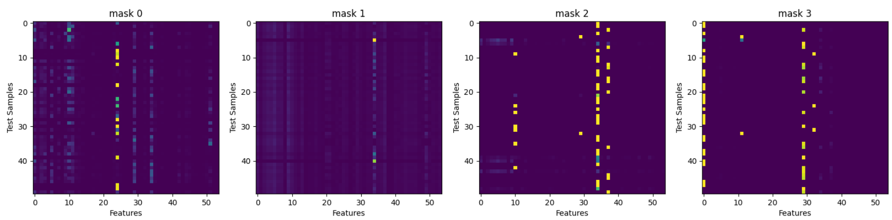


Figure 5: InterpreTabNet with $r^*_M = 900$ (Best Performing Model) for Forest Cover Type Dataset

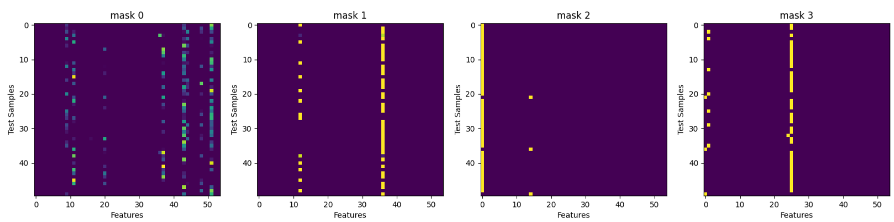


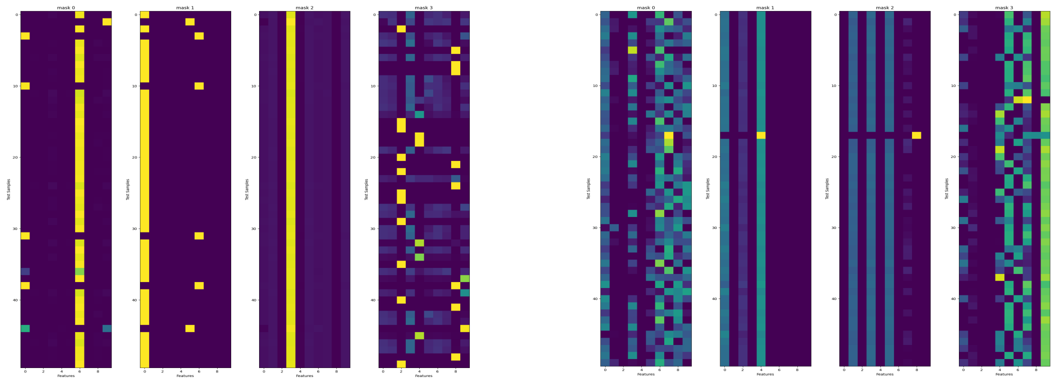
Figure 6: Original TabNet Model for Forest Cover Type Dataset

Poker Hand [Cattral and Oppacher, 2007]

Model	**Test Accuracy (%)**
XGBoost	75.41
LightGBM	78.47
TabTransformer	99.81
MLP	99.98
Original TabNet	99.00
InterpreTabNet ($r^*_M = 1000$)	**99.50**

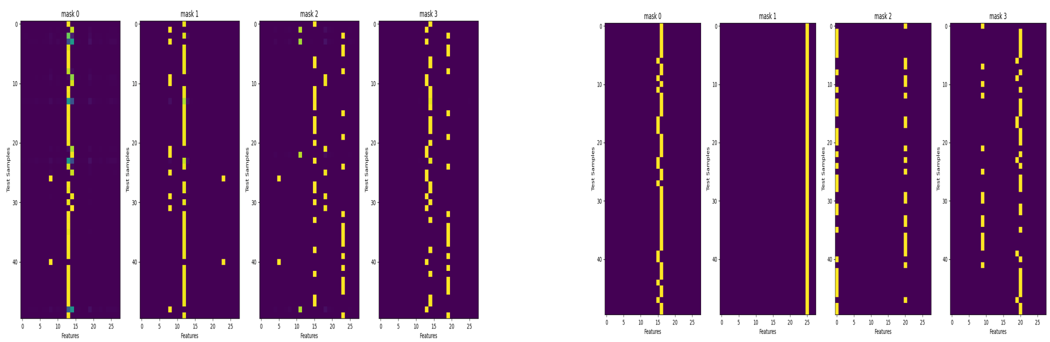
Higgs [Whiteson, 2014]

Model	**Test Accuracy (%)**
XGBoost	72.91
LightGBM	72.62
TabTransformer	51.97
MLP	68.67
Original TabNet	52.94
InterpreTabNet ($r^*_M = 10000$)	**53.08**



(a) InterpreTabNet (b) Original TabNet Model

Figure 7: (a) Best performing model using InterpreTabNet, $r_M^* = 1000$, with an accuracy of 99.13% on the Poker Hand Dataset. (b) The baseline performance using the Original TabNet model, attaining an accuracy of 99.23%.



(a) InterpreTabNet (b) Original TabNet Model

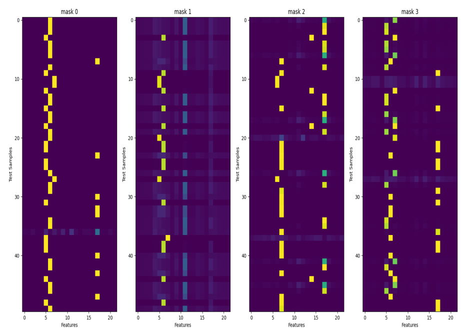
Figure 8: (a) Best performing model using InterpreTabNet, $r_M^* = 10000$, with an accuracy of 53.08% on the Higgs Dataset. (b) The baseline performance using the Original TabNet model, attaining an accuracy of 60.22%.

Mushroom [mus, 1987]

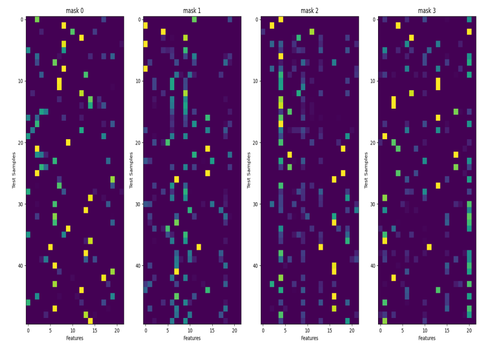
Model	**Test Accuracy (%)**
XGBoost	100.00
LightGBM	100.00
TabTransformer	100.00
MLP	100.00
Original TabNet	99.94
InterpreTabNet ($r_M^* = 10,000,000,000,000$)	**96.62**

Blastchar [BlastChar, 2018]

Model	**Test Accuracy (%)**
XGBoost	77.29
LightGBM	77.86
TabTransformer	73.17
MLP	73.67
Original TabNet	76.22
InterpreTabNet ($r_M^* = 10,000,000,000,000$)	**72.96**

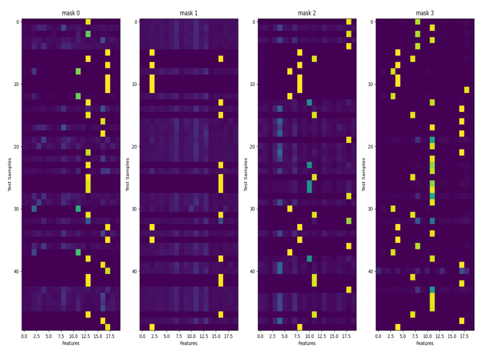


(a) InterpreTabNet

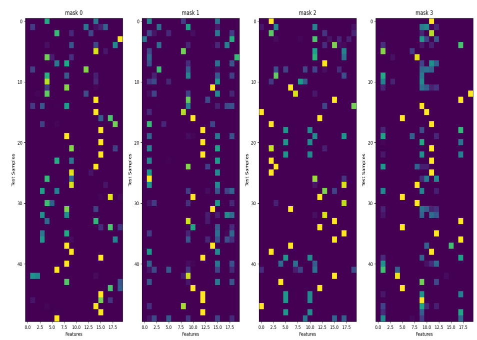


(b) Original TabNet Model

Figure 9: (a) Best performing model using InterpreTabNet, $r_M^* = 1,000,000,000,0000$, with an accuracy of 96.62% on the Mushroom Dataset. (b) The baseline performance using the Original TabNet model, attaining an accuracy of 99.94%.



(a) InterpreTabNet



(b) Original TabNet Model

Figure 10: (a) Best performing model using InterpreTabNet, $r_M^* = 10,000,000,000,000$, with an accuracy of 72.96% on the Blastchar Dataset. (b) The baseline performance using the Original TabNet model, attaining an accuracy of 76.22%.

Diabetes [Clore and Strack, 2014]

Model	**Test Accuracy (%)**
XGBoost	61.44
LightGBM	60.87
TabTransformer	44.45
MLP	57.19
Original TabNet	56.91
InterpreTabNet ($r_M^* = 100,000,000$)	**55.37**

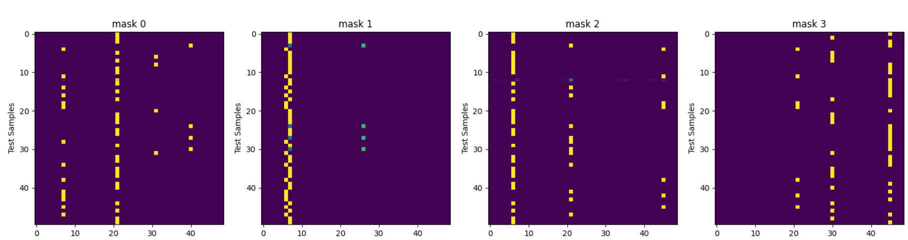


Figure 11: InterpreTabNet with $r_M^* = 100,000,000$ (Best Performing Model) for Diabetes Dataset

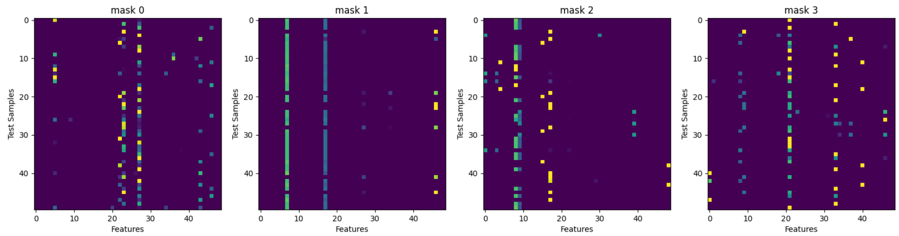
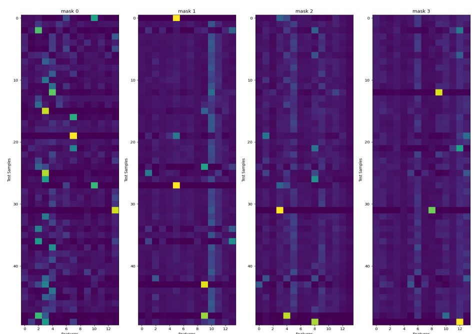


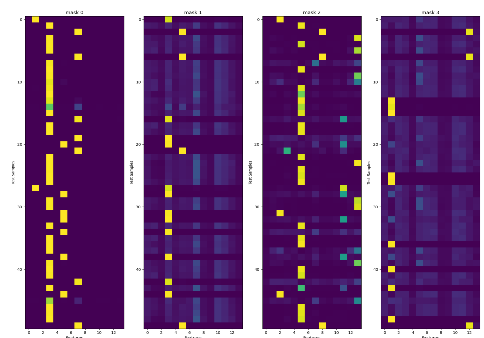
Figure 12: Original TabNet Model for Diabetes Dataset

5.7 Ablation Study on InterpreTabNet's Interpretability for varying r_M values

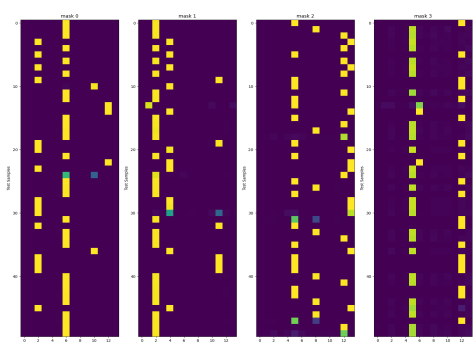
Adult Census Income Mask Figures



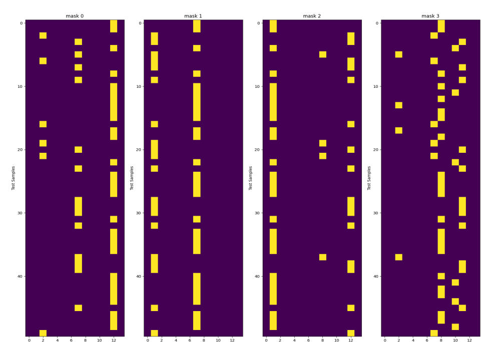
(a) InterpreTabNet ($r_M = 0$)



(b) InterpreTabNet ($r_M = 1$)



(c) InterpreTabNet ($r_M = 100$)



(d) InterpreTabNet ($r_M^* = 1000$)

Figure 13: As the r_M value increases, both feature mask sparsity and feature importance increases, improving the interpretability of the masks.

5.8 Prompts for all datasets

Adult Census Income

Conduct aggregate analysis on the description of the following feature mask. Please output ONLY a dictionary and no other natural language generation when generating the sentence as shown in the in-context example below. Please use single-word classification that encapsulates the meaning of the features if possible.

The Adult Census Income dataset is considered which predicts whether an individual's income exceeds $50,000 per year. There are 50 test samples and 14 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 1, 5 and 7 which are workclass, marital status and relationship. At the 1st step of feature selection, we observe mask 0 with the main features highlighted as 1 and 5 which are workclass and marital status. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 8, 10 and 12 which are race, capital gain and hours-per-week. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 1, 7 and 10 which are workclass, relationship and capital-gain.

Here are in-context examples for few-shot learning.

The Poker Hand dataset is considered, which is a classification of poker hands into 10 distinct types based on attributes like rank of individual cards, suits, and card combinations. There are 50 test samples and 10 features.

At the 0th step of feature selection, we observe mask 0 with the main feature highlighted as 3 which is the rank of the card 2. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 7 which is the rank of card 3. At the 2nd step of feature selection, we observe mask 2 with the main feature highlighted as 6 which is the suit of card 3. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 3, 6, 7 and 10 which are the rank of cards 1 and 3 as well as the suits of cards 3 and 5.

Output: {"Mask 0": "Initially, the rank of card 2 is recognized as a significant predictor in the classification of poker hands. The rank of individual cards seems crucial at this stage.", "Mask 1": "At this step, the rank of card 3 is highlighted, further emphasizing the importance of individual card ranks in determining poker hand types.", "Mask 2": "Here, besides the rank of card 3, the suit of the same card is considered, suggesting that both rank and suit play roles in classifying poker hands.", "Mask 3": "This step combines multiple features, both ranks (cards 1 and 3) and suits (cards 3 and 5), indicating that a combination of these factors provides a broader understanding of poker hand classifications.", "Aggregate": "The rank of cards, specifically card 2 and card 3, appears consistently across different masks, indicating their primary role in poker hand classification. Suits, particularly of card 3, also have importance but are introduced later in the feature selection. The combined importance of both ranks and suits in the final step suggests that while individual card ranks are pivotal in initial classifications, understanding the relationship between card ranks and their respective suits provides a more comprehensive insight into the poker hand types. The consistent presence of card 3's attributes (both rank and suit) underscores its pivotal role in determining poker hand categories."}

The Forest Cover Type dataset is considered which is a classification of pixels into 7 forest cover types based on attributes such as elevation, aspect, slope, and more. There are 50 test samples and 54 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 24, 29 and 34 which are different soil types. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 34 which is soil type. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 10, 34 and 37 which are wilderness area and soil types. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 0 and 29 which are elevation, and soil type.

Output: {"Mask 0": "The initial feature selection identifies that various soil types are significant predictors for the forest cover types. It appears that the characteristics of the soil play a vital role in determining the type of forest cover.", "Mask 1": "At this step, the model emphasizes even more on one particular soil type (34), suggesting it as a primary driver for the classification. This specific soil type may have unique characteristics making it a distinctive feature for certain forest cover types.", "Mask 2": "Here, besides the emphasis on soil type 34, another soil type (37) and wilderness area are considered. This suggests that both the nature of the terrain (wilderness) and the characteristics of the soil together contribute to the forest cover type.", "Mask 3": "This step introduces elevation as a significant predictor while still retaining a soil type. Elevation, combined with soil characteristics, might influence specific flora and subsequently determine forest cover types.", "Aggregate": "Soil type 34 is the most consistent feature across all the masks, appearing in every step, emphasizing its importance in the forest cover type classification. Another soil type, 29, is also consistent, appearing in both mask 0 and 3. The presence of different features across the steps, like wilderness area (mask 2) and elevation (mask 3), indicates that the landscape's nature and the physical height of the terrain are also important but might offer secondary or nuanced insights into forest cover types. The dominant predictor seems to be soil types, particularly soil type 34, suggesting that specific soil characteristics play a crucial role in determining forest cover types. However, aspects like elevation and the nature of the terrain (wilderness area) also contribute to the classification and shouldn't be overlooked."}

Forest Cover Type

Conduct aggregate analysis on the description of the following feature mask. Please output ONLY a dictionary and no other natural language generation when generating the sentence as shown in the in-context example below. Please use single-word classification that encapsulates the meaning of the features if possible.

The Forest Cover Type dataset is considered which is a classification of pixels into 7 forest cover types based on attributes such as elevation, aspect, slope, and more. There are 50 test samples and 54 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 24, 29 and 34 which are different soil types. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 34 which is soil type. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 10, 34 and 37 which are wilderness area and soil types. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 0 and 29 which are elevation, and soil type.

Here are in-context examples for few-shot learning.

The Adult Census Income dataset is considered which predicts whether an individual's income exceeds $50,000 per year. There are 50 test samples and 14 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 1, 5 and 7 which are workclass, marital status and relationship. At the 1st step of feature selection, we observe mask 0 with the main features highlighted as 1 and 5 which are workclass and marital status. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 8, 10 and 12 which are race, capital gain and hours-per-week. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 1, 7 and 10 which are workclass, relationship and capital-gain.

Output: {"Mask 0": "This suggests that early in the feature selection process, the model finds that work-related factors (like the type of work) and personal relationships (both marital status and other relationships) are significant predictors for income exceeding $50,000.", "Mask 1": "Here, the model emphasizes even more on work nature and marital status, suggesting these two as the primary drivers for the prediction. Relationship as a feature has been excluded in this step, which might indicate a possible correlation with marital status.", "Mask 2": "At this step, the model seems to be looking beyond just personal and work nature to more financial and demographic aspects. Capital gain, which is directly related to financial health, is considered along with hours-per-week, potentially pointing to a relationship between longer working hours and higher income. The inclusion of race might indicate underlying societal or economic factors at play.", "Mask 3": "This combination brings back features from both the initial and second masks. Workclass and relationship are revisited, which further underscores their importance. Capital gain remains consistent from the second step, emphasizing its importance as a predictor for income.", "Aggregate": "The features that consistently appear across multiple masks are: workclass (masks 0, 1, and 3), relationship (masks 0 and 3), and capital gain (masks 2 and 3). This highlights that these features play a pivotal role in predicting whether an individual's income exceeds $50,000 per year. Marital status is seen in the initial stages but disappears later, while hours-per-week and race come into play at the second step. This could indicate that while some features provide foundational predictive power, others may provide more nuanced insight or refinement to the model's predictions. Work-related aspects, personal relationships, and financial metrics seem to be the main predictors of an individual's income level in this dataset. It would be crucial to further analyze these features' exact impacts and understand any potential correlations among them for a more comprehensive understanding."}

The Poker Hand dataset is considered, which is a classification of poker hands into 10 distinct types based on attributes like rank of individual cards, suits, and card combinations. There are 50 test samples and 10 features.

At the 0th step of feature selection, we observe mask 0 with the main feature highlighted as 3 which is the rank of the card 2. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 7 which is the rank of card 3. At the 2nd step of feature selection, we observe mask 2 with the main feature highlighted as 6 which is the suit of card 3. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 3, 6, 7 and 10 which are the rank of cards 1 and 3 as well as the suits of cards 3 and 5.

Output: {"Mask 0": "Initially, the rank of card 2 is recognized as a significant predictor in the classification of poker hands. The rank of individual cards seems crucial at this stage.", "Mask 1": "At this step, the rank of card 3 is highlighted, further emphasizing the importance of individual card ranks in determining poker hand types.", "Mask 2": "Here, besides the rank of card 3, the suit of the same card is considered, suggesting that both rank and suit play roles in classifying poker hands.", "Mask 3": "This step combines multiple features, both ranks (cards 1 and 3) and suits (cards 3 and 5), indicating that a combination of these factors provides a broader understanding of poker hand classifications.", "Aggregate": "The rank of cards, specifically card 2 and card 3, appears consistently across different masks, indicating their primary role in poker hand classification. Suits, particularly of card 3, also have importance but are introduced later in the feature selection. The combined importance of both ranks and suits in the final step suggests that while individual card ranks are pivotal in initial classifications, understanding the relationship between card ranks and their respective suits provides a more comprehensive insight into the poker hand types. The consistent presence of card 3's attributes (both rank and suit) underscores its pivotal role in determining poker hand categories."}

Poker Hand Prompt

Conduct aggregate analysis on the description of the following feature mask. Please output ONLY a dictionary and no other natural language generation when generating the sentence as shown in the in-context example below. Please use single-word classification that encapsulates the meaning of the features if possible.

The Poker Hand dataset is considered, which is a classification of poker hands into 10 distinct types based on attributes like rank of individual cards, suits, and card combinations. There are 50 test samples and 10 features.

At the 0th step of feature selection, we observe mask 0 with the main feature highlighted as 3 which is the rank of the card 2. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 7 which is the rank of card 3. At the 2nd step of feature selection, we observe mask 2 with the main feature highlighted as 6 which is the suit of card 3. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 3, 6, 7 and 10 which are the rank of cards 1 and 3 as well as the suits of cards 3 and 5.

Here are in-context examples for few-shot learning.

The Adult Census Income dataset is considered which predicts whether an individual's income exceeds $50,000 per year. There are 50 test samples and 14 features. At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 1, 5 and 7 which are workclass, marital status and relationship. At the 1st step of feature selection, we observe mask 0 with the main features highlighted as 1 and 5 which are workclass and marital status. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 8, 10 and 12 which are race, capital gain and hours-per-week. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 1, 7 and 10 which are workclass, relationship and capital-gain.

Output: {"Mask 0": "This suggests that early in the feature selection process, the model finds that work-related factors (like the type of work) and personal relationships (both marital status and other relationships) are significant predictors for income exceeding $50,000.", "Mask 1": "Here, the model emphasizes even more on work nature and marital status, suggesting these two as the primary drivers for the prediction. Relationship as a feature has been excluded in this step, which might indicate a possible correlation with marital status.", "Mask 2": "At this step, the model seems to be looking beyond just personal and work nature to more financial and demographic aspects. Capital gain, which is directly related to financial health, is considered along with hours-per-week, potentially pointing to a relationship between longer working hours and higher income. The inclusion of race might indicate underlying societal or economic factors at play.", "Mask 3": "This combination brings back features from both the initial and second masks. Workclass and relationship are revisited, which further underscores their importance. Capital gain remains consistent from the second step, emphasizing its importance as a predictor for income.", "Aggregate": "The features that consistently appear across multiple masks are: workclass (masks 0, 1, and 3), relationship (masks 0 and 3), and capital gain (masks 2 and 3). This highlights that these features play a pivotal role in predicting whether an individual's income exceeds $50,000 per year. Marital status is seen in the initial stages but disappears later, while hours-per-week and race come into play at the second step. This could indicate that while some features provide foundational predictive power, others may provide more nuanced insight or refinement to the model's predictions. Work-related aspects, personal relationships, and financial metrics seem to be the main predictors of an individual's income level in this dataset. It would be crucial to further analyze these features' exact impacts and understand any potential correlations among them for a more comprehensive understanding."}

The Forest Cover Type dataset is considered which is a classification of pixels into 7 forest cover types based on attributes such as elevation, aspect, slope, and more. There are 50 test samples and 54 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 24, 29 and 34 which are different soil types. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 34 which is soil type. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 10, 34 and 37 which are wilderness area and soil types. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 0 and 29 which are elevation, and soil type.

Output: {"Mask 0": "The initial feature selection identifies that various soil types are significant predictors for the forest cover types. It appears that the characteristics of the soil play a vital role in determining the type of forest cover.", "Mask 1": "At this step, the model emphasizes even more on one particular soil type (34), suggesting it as a primary driver for the classification. This specific soil type may have unique characteristics making it a distinctive feature for certain forest cover types.", "Mask 2": "Here, besides the emphasis on soil type 34, another soil type (37) and wilderness area are considered. This suggests that both the nature of the terrain (wilderness) and the characteristics of the soil together contribute to the forest cover type.", "Mask 3": "This step introduces elevation as a significant predictor while still retaining a soil type. Elevation, combined with soil characteristics, might influence specific flora and subsequently determine forest cover types.", "Aggregate": "Soil type 34 is the most consistent feature across all the masks, appearing in every step, emphasizing its importance in the forest cover type classification. Another soil type, 29, is also consistent, appearing in both mask 0 and 3. The presence of different features across the steps, like wilderness area (mask 2) and elevation (mask 3), indicates that the landscape's nature and the physical height of the terrain are also important but might offer secondary or nuanced insights into forest cover types. The dominant predictor seems to be soil types, particularly soil type 34, suggesting that specific soil characteristics play a crucial role in determining forest cover types. However, aspects like elevation and the nature of the terrain (wilderness area) also contribute to the classification and shouldn't be overlooked."}

Mushroom Prompt

Conduct aggregate analysis on the description of the following feature masks. Start off with an analysis of the individual masks, followed by an aggregate analysis of all masks combined. Please format the output into a dictionary as shown in the in-context examples. The output should only contain the formatted output, no other natural language generation is required.

The Mushroom dataset is considered, which is a classification of mushrooms into edible or poisonous categories based on attributes like cap shape, gill color, stalk length, and other morphological characteristics. There are 50 test samples and 22 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 5, 6 and 17 which are bruises, odor and veil-type. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 5, 6, 11 which are bruises, odor and stalk-shape. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 7, 14, 17 which are gill-attachment, stalk-surface-below-ring and veil-type. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 5, 6, 7 and 17 which are bruises, odor, gill-attachment and veil-type.

Here are in-context examples for few-shot learning.

The Adult Census Income dataset is considered which predicts whether an individual's income exceeds $50,000 per year. There are 50 test samples and 14 features. At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 1, 5 and 7 which are workclass, marital status and relationship. At the 1st step of feature selection, we observe mask 0 with the main features highlighted as 1 and 5 which are workclass and marital status. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 8, 10 and 12 which are race, capital gain and hours-per-week. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 1, 7 and 10 which are workclass, relationship and capital-gain.

Output: {"Mask 0": "This suggests that early in the feature selection process, the model finds that work-related factors (like the type of work) and personal relationships (both marital status and other relationships) are significant predictors for income exceeding $50,000.", "Mask 1": "Here, the model emphasizes even more on work nature and marital status, suggesting these two as the primary drivers for the prediction. Relationship as a feature has been excluded in this step, which might indicate a possible correlation with marital status.", "Mask 2": "At this step, the model seems to be looking beyond just personal and work nature to more financial and demographic aspects. Capital gain, which is directly related to financial health, is considered along with hours-per-week, potentially pointing to a relationship between longer working hours and higher income. The inclusion of race might indicate underlying societal or economic factors at play.", "Mask 3": "This combination brings back features from both the initial and second masks. Workclass and relationship are revisited, which further underscores their importance. Capital gain remains consistent from the second step, emphasizing its importance as a predictor for income.", "Aggregate": "The features that consistently appear across multiple masks are: workclass (masks 0, 1, and 3), relationship (masks 0 and 3), and capital gain (masks 2 and 3). This highlights that these features play a pivotal role in predicting whether an individual's income exceeds $50,000 per year. Marital status is seen in the initial stages but disappears later, while hours-per-week and race come into play at the second step. This could indicate that while some features provide foundational predictive power, others may provide more nuanced insight or refinement to the model's predictions. Work-related aspects, personal relationships, and financial metrics seem to be the main predictors of an individual's income level in this dataset. It would be crucial to further analyze these features' exact impacts and understand any potential correlations among them for a more comprehensive understanding."}

The Forest Cover Type dataset is considered which is a classification of pixels into 7 forest cover types based on attributes such as elevation, aspect, slope, and more. There are 50 test samples and 54 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 24, 29 and 34 which are different soil types. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 34 which is soil type. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 10, 34 and 37 which are wilderness area and soil types. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 0 and 29 which are elevation, and soil type.

Output: {"Mask 0": "The initial feature selection identifies that various soil types are significant predictors for the forest cover types. It appears that the characteristics of the soil play a vital role in determining the type of forest cover.", "Mask 1": "At this step, the model emphasizes even more on one particular soil type (34), suggesting it as a primary driver for the classification. This specific soil type may have unique characteristics making it a distinctive feature for certain forest cover types.", "Mask 2": "Here, besides the emphasis on soil type 34, another soil type (37) and wilderness area are considered. This suggests that both the nature of the terrain (wilderness) and the characteristics of the soil together contribute to the forest cover type.", "Mask 3": "This step introduces elevation as a significant predictor while still retaining a soil type. Elevation, combined with soil characteristics, might influence specific flora and subsequently determine forest cover types.", "Aggregate": "Soil type 34 is the most consistent feature across all the masks, appearing in every step, emphasizing its importance in the forest cover type classification. Another soil type, 29, is also consistent, appearing in both mask 0 and 3. The presence of different features across the steps, like wilderness area (mask 2) and elevation (mask 3), indicates that the landscape's nature and the physical height of the terrain are also important but might offer secondary or nuanced insights into forest cover types. The dominant predictor seems to be soil types, particularly soil type 34, suggesting that specific soil characteristics play a crucial role in determining forest cover types. However, aspects like elevation and the nature of the terrain (wilderness area) also contribute to the classification and shouldn't be overlooked."}

Blastchar Prompt

Conduct aggregate analysis on the description of the following feature masks. Start off with an analysis of the individual masks, followed by an aggregate analysis of all masks combined. Please format the output into a dictionary as shown in the in-context examples. The output should only contain the formatted output, no other natural language generation is required.

The BlastChar Telco Customer Churn dataset is considered, which is a classification of customers into retained or churned categories based on attributes like gender, seniority, tenure, service subscriptions, contract type, billing methods, and charges, among others. There are 50 test samples and 21 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 13, 16, and 17 which are StreamingTV, PaperlessBilling and PaymentMethod. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 2 and 16 which are SeniorCitizen and PaperlessBilling. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 8, 11, and 18 which are InternetService, DeviceProtection, and MonthlyCharges. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 3, 11, and 17 which are Partner, DeviceProtection, and PaymentMethod.

Here are in-context examples for few-shot learning.

The Adult Census Income dataset is considered which predicts whether an individual's income exceeds $50,000 per year. There are 50 test samples and 14 features. At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 1, 5 and 7 which are workclass, marital status and relationship. At the 1st step of feature selection, we observe mask 0 with the main features highlighted as 1 and 5 which are workclass and marital status. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 8, 10 and 12 which are race, capital gain and hours-per-week. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 1, 7 and 10 which are workclass, relationship and capital-gain.

Output: {"Mask 0": "This suggests that early in the feature selection process, the model finds that work-related factors (like the type of work) and personal relationships (both marital status and other relationships) are significant predictors for income exceeding $50,000.", "Mask 1": "Here, the model emphasizes even more on work nature and marital status, suggesting these two as the primary drivers for the prediction. Relationship as a feature has been excluded in this step, which might indicate a possible correlation with marital status.", "Mask 2": "At this step, the model seems to be looking beyond just personal and work nature to more financial and demographic aspects. Capital gain, which is directly related to financial health, is considered along with hours-per-week, potentially pointing to a relationship between longer working hours and higher income. The inclusion of race might indicate underlying societal or economic factors at play.", "Mask 3": "This combination brings back features from both the initial and second masks. Workclass and relationship are revisited, which further underscores their importance. Capital gain remains consistent from the second step, emphasizing its importance as a predictor for income.", "Aggregate": "The features that consistently appear across multiple masks are: workclass (masks 0, 1, and 3), relationship (masks 0 and 3), and capital gain (masks 2 and 3). This highlights that these features play a pivotal role in predicting whether an individual's income exceeds $50,000 per year. Marital status is seen in the initial stages but disappears later, while hours-per-week and race come into play at the second step. This could indicate that while some features provide foundational predictive power, others may provide more nuanced insight or refinement to the model's predictions. Work-related aspects, personal relationships, and financial metrics seem to be the main predictors of an individual's income level in this dataset. It would be crucial to further analyze these features' exact impacts and understand any potential correlations among them for a more comprehensive understanding."}

The Forest Cover Type dataset is considered which is a classification of pixels into 7 forest cover types based on attributes such as elevation, aspect, slope, and more. There are 50 test samples and 54 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 24, 29 and 34 which are different soil types. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 34 which is soil type. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 10, 34 and 37 which are wilderness area and soil types. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 0 and 29 which are elevation, and soil type.

Output: {"Mask 0": "The initial feature selection identifies that various soil types are significant predictors for the forest cover types. It appears that the characteristics of the soil play a vital role in determining the type of forest cover.", "Mask 1": "At this step, the model emphasizes even more on one particular soil type (34), suggesting it as a primary driver for the classification. This specific soil type may have unique characteristics making it a distinctive feature for certain forest cover types.", "Mask 2": "Here, besides the emphasis on soil type 34, another soil type (37) and wilderness area are considered. This suggests that both the nature of the terrain (wilderness) and the characteristics of the soil together contribute to the forest cover type.", "Mask 3": "This step introduces elevation as a significant predictor while still retaining a soil type. Elevation, combined with soil characteristics, might influence specific flora and subsequently determine forest cover types.", "Aggregate": "Soil type 34 is the most consistent feature across all the masks, appearing in every step, emphasizing its importance in the forest cover type classification. Another soil type, 29, is also consistent, appearing in both mask 0 and 3. The presence of different features across the steps, like wilderness area (mask 2) and elevation (mask 3), indicates that the landscape's nature and the physical height of the terrain are also important but might offer secondary or nuanced insights into forest cover types. The dominant predictor seems to be soil types, particularly soil type 34, suggesting that specific soil characteristics play a crucial role in determining forest cover types. However, aspects like elevation and the nature of the terrain (wilderness area) also contribute to the classification and shouldn't be overlooked."}

Diabetes Prompt

Conduct aggregate analysis on the description of the following feature masks. Start off with an analysis of the individual masks, followed by an aggregate analysis of all masks combined. Please format the output into a dictionary as shown in the in-context examples. The output should only contain the formatted output, no other natural language generation is required.

The Diabetes 130-US hospitals for years 1999-2008 dataset is considered, which is a classification of patient encounters into readmitted or not readmitted categories based on attributes like the number of laboratory tests performed, the number of medications prescribed, diagnoses, and other clinical and administrative data. There are 50 test samples and 50 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 7, 21, 31, and 40 which are discharge disposition id, number diagnoses, glyburide, and citoglipton. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 6 and 7 which are admission type id and discharge disposition id. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 6, 21, and 45 which are admission type id, number diagnoses, and metformin. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 21, 30, 45 which are number diagnoses, glipizide, and metformin.

Here are in-context examples for few-shot learning.

The Adult Census Income dataset is considered which predicts whether an individual's income exceeds $50,000 per year. There are 50 test samples and 14 features. At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 1, 5 and 7 which are workclass, marital status and relationship. At the 1st step of feature selection, we observe mask 0 with the main features highlighted as 1 and 5 which are workclass and marital status. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 8, 10 and 12 which are race, capital gain and hours-per-week. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 1, 7 and 10 which are workclass, relationship and capital-gain.

Output: {"Mask 0": "This suggests that early in the feature selection process, the model finds that work-related factors (like the type of work) and personal relationships (both marital status and other relationships) are significant predictors for income exceeding $50,000.", "Mask 1": "Here, the model emphasizes even more on work nature and marital status, suggesting these two as the primary drivers for the prediction. Relationship as a feature has been excluded in this step, which might indicate a possible correlation with marital status.", "Mask 2": "At this step, the model seems to be looking beyond just personal and work nature to more financial and demographic aspects. Capital gain, which is directly related to financial health, is considered along with hours-per-week, potentially pointing to a relationship between longer working hours and higher income. The inclusion of race might indicate underlying societal or economic factors at play.", "Mask 3": "This combination brings back features from both the initial and second masks. Workclass and relationship are revisited, which further underscores their importance. Capital gain remains consistent from the second step, emphasizing its importance as a predictor for income.", "Aggregate": "The features that consistently appear across multiple masks are: workclass (masks 0, 1, and 3), relationship (masks 0 and 3), and capital gain (masks 2 and 3). This highlights that these features play a pivotal role in predicting whether an individual's income exceeds $50,000 per year. Marital status is seen in the initial stages but disappears later, while hours-per-week and race come into play at the second step. This could indicate that while some features provide foundational predictive power, others may provide more nuanced insight or refinement to the model's predictions. Work-related aspects, personal relationships, and financial metrics seem to be the main predictors of an individual's income level in this dataset. It would be crucial to further analyze these features' exact impacts and understand any potential correlations among them for a more comprehensive understanding."}

The Forest Cover Type dataset is considered which is a classification of pixels into 7 forest cover types based on attributes such as elevation, aspect, slope, and more. There are 50 test samples and 54 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 24, 29 and 34 which are different soil types. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 34 which is soil type. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 10, 34 and 37 which are wilderness area and soil types. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 0 and 29 which are elevation, and soil type.

Output: {"Mask 0": "The initial feature selection identifies that various soil types are significant predictors for the forest cover types. It appears that the characteristics of the soil play a vital role in determining the type of forest cover.", "Mask 1": "At this step, the model emphasizes even more on one particular soil type (34), suggesting it as a primary driver for the classification. This specific soil type may have unique characteristics making it a distinctive feature for certain forest cover types.", "Mask 2": "Here, besides the emphasis on soil type 34, another soil type (37) and wilderness area are considered. This suggests that both the nature of the terrain (wilderness) and the characteristics of the soil together contribute to the forest cover type.", "Mask 3": "This step introduces elevation as a significant predictor while still retaining a soil type. Elevation, combined with soil characteristics, might influence specific flora and subsequently determine forest cover types.", "Aggregate": "Soil type 34 is the most consistent feature across all the masks, appearing in every step, emphasizing its importance in the forest cover type classification. Another soil type, 29, is also consistent, appearing in both mask 0 and 3. The presence of different features across the steps, like wilderness area (mask 2) and elevation (mask 3), indicates that the landscape's nature and the physical height of the terrain are also important but might offer secondary or nuanced insights into forest cover types. The dominant predictor seems to be soil types, particularly soil type 34, suggesting that specific soil characteristics play a crucial role in determining forest cover types. However, aspects like elevation and the nature of the terrain (wilderness area) also contribute to the classification and shouldn't be overlooked."}

Higgs Prompt

Conduct aggregate analysis on the description of the following feature mask. Please output ONLY a dictionary and no other natural language generation when generating the sentence as shown in the in-context example below. Please use single-word classification that encapsulates the meaning of the features if possible.

The Higgs dataset is considered, which is a classification of events into signal and background processes based on attributes like jet kinematics, lepton momentum, and other high-energy physics variables. There are 50 test samples and 28 features. At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 22 and 25 which are the invariant mass of the lepton, the missing energy, and the leading jet, as well as the four leading jets. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 6 which is the Azimuthal angle for the second leading jet. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 15 and 25 which are the transverse momentum for the second leading jet and the invariant mass of the lepton, the missing energy, and the leading jet.. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 8 and 16 which are the B-tagging score for the second leading jet and the pseudorapidity of the lepton.

Here are in-context examples for few-shot learning.

The Adult Census Income dataset is considered which predicts whether an individual's income exceeds $50,000 per year. There are 50 test samples and 14 features. At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 1, 5 and 7 which are workclass, marital status and relationship. At the 1st step of feature selection, we observe mask 0 with the main features highlighted as 1 and 5 which are workclass and marital status. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 8, 10 and 12 which are race, capital gain and hours-per-week. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 1, 7 and 10 which are workclass, relationship and capital-gain.

Output: {"Mask 0": "This suggests that early in the feature selection process, the model finds that work-related factors (like the type of work) and personal relationships (both marital status and other relationships) are significant predictors for income exceeding $50,000.", "Mask 1": "Here, the model emphasizes even more on work nature and marital status, suggesting these two as the primary drivers for the prediction. Relationship as a feature has been excluded in this step, which might indicate a possible correlation with marital status.", "Mask 2": "At this step, the model seems to be looking beyond just personal and work nature to more financial and demographic aspects. Capital gain, which is directly related to financial health, is considered along with hours-per-week, potentially pointing to a relationship between longer working hours and higher income. The inclusion of race might indicate underlying societal or economic factors at play.", "Mask 3": "This combination brings back features from both the initial and second masks. Workclass and relationship are revisited, which further underscores their importance. Capital gain remains consistent from the second step, emphasizing its importance as a predictor for income.", "Aggregate": "The features that consistently appear across multiple masks are: workclass (masks 0, 1, and 3), relationship (masks 0 and 3), and capital gain (masks 2 and 3). This highlights that these features play a pivotal role in predicting whether an individual's income exceeds $50,000 per year. Marital status is seen in the initial stages but disappears later, while hours-per-week and race come into play at the second step. This could indicate that while some features provide foundational predictive power, others may provide more nuanced insight or refinement to the model's predictions. Work-related aspects, personal relationships, and financial metrics seem to be the main predictors of an individual's income level in this dataset. It would be crucial to further analyze these features' exact impacts and understand any potential correlations among them for a more comprehensive understanding."}

The Forest Cover Type dataset is considered which is a classification of pixels into 7 forest cover types based on attributes such as elevation, aspect, slope, and more. There are 50 test samples and 54 features.

At the 0th step of feature selection, we observe mask 0 with the main features highlighted as 24, 29 and 34 which are different soil types. At the 1st step of feature selection, we observe mask 1 with the main feature highlighted as 34 which is soil type. At the 2nd step of feature selection, we observe mask 2 with the main features highlighted as 10, 34 and 37 which are wilderness area and soil types. Lastly, at the 3rd step of feature selection, we observe mask 3 with the main features highlighted as 0 and 29 which are elevation, and soil type.

Output: {"Mask 0": "The initial feature selection identifies that various soil types are significant predictors for the forest cover types. It appears that the characteristics of the soil play a vital role in determining the type of forest cover.", "Mask 1": "At this step, the model emphasizes even more on one particular soil type (34), suggesting it as a primary driver for the classification. This specific soil type may have unique characteristics making it a distinctive feature for certain forest cover types.", "Mask 2": "Here, besides the emphasis on soil type 34, another soil type (37) and wilderness area are considered. This suggests that both the nature of the terrain (wilderness) and the characteristics of the soil together contribute to the forest cover type.", "Mask 3": "This step introduces elevation as a significant predictor while still retaining a soil type. Elevation, combined with soil characteristics, might influence specific flora and subsequently determine forest cover types.", "Aggregate": "Soil type 34 is the most consistent feature across all the masks, appearing in every step, emphasizing its importance in the forest cover type classification. Another soil type, 29, is also consistent, appearing in both mask 0 and 3. The presence of different features across the steps, like wilderness area (mask 2) and elevation (mask 3), indicates that the landscape's nature and the physical height of the terrain are also important but might offer secondary or nuanced insights into forest cover types. The dominant predictor seems to be soil types, particularly soil type 34, suggesting that specific soil characteristics play a crucial role in determining forest cover types. However, aspects like elevation and the nature of the terrain (wilderness area) also contribute to the classification and shouldn't be overlooked."}

5.9 Output text from GPT-4 on all datasets for mask interpretation

Dataset	Output from GPT-4
Adult Census Income	{"Mask 0": "This suggests that early in the feature selection process, the model finds that work-related factors (like the type of work) and personal relationships (both marital status and other relationships) are significant predictors for income exceeding $50,000.", "Mask 1": "Here, the model emphasizes even more on work nature and marital status, suggesting these two as the primary drivers for the prediction. Relationship as a feature has been excluded in this step, which might indicate a possible correlation with marital status.", "Mask 2": "At this step, the model seems to be looking beyond just personal and work nature to more financial and demographic aspects. Capital gain, which is directly related to financial health, is considered along with hours-per-week, potentially pointing to a relationship between longer working hours and higher income. The inclusion of race might indicate underlying societal or economic factors at play.", "Mask 3": "This combination brings back features from both the initial and second masks. Workclass and relationship are revisited, which further underscores their importance. Capital gain remains consistent from the second step, emphasizing its importance as a predictor for income.", "Aggregate": "The features that consistently appear across multiple masks are: workclass (masks 0, 1, and 3), relationship (masks 0 and 3), and capital gain (masks 2 and 3). This highlights that these features play a pivotal role in predicting whether an individual's income exceeds $50,000 per year. Marital status is seen in the initial stages but disappears later, while hours-per-week and race come into play at the second step. This could indicate that while some features provide foundational predictive power, others may provide more nuanced insight or refinement to the model's predictions. Work-related aspects, personal relationships, and financial metrics seem to be the main predictors of an individual's income level in this dataset. It would be crucial to further analyze these features' exact impacts and understand any potential correlations among them for a more comprehensive understanding."}

Forest Cover Type	{"Mask 0": "The initial feature selection identifies that various soil types are significant predictors for the forest cover types. It appears that the characteristics of the soil play a vital role in determining the type of forest cover.", "Mask 1": "At this step, the model emphasizes even more on one particular soil type (34), suggesting it as a primary driver for the classification. This specific soil type may have unique characteristics making it a distinctive feature for certain forest cover types.", "Mask 2": "Here, besides the emphasis on soil type 34, another soil type (37) and wilderness area are considered. This suggests that both the nature of the terrain (wilderness) and the characteristics of the soil together contribute to the forest cover type.", "Mask 3": "This step introduces elevation as a significant predictor while still retaining a soil type. Elevation, combined with soil characteristics, might influence specific flora and subsequently determine forest cover types.", "Aggregate": "Soil type 34 is the most consistent feature across all the masks, appearing in every step, emphasizing its importance in the forest cover type classification. Another soil type, 29, is also consistent, appearing in both mask 0 and 3. The presence of different features across the steps, like wilderness area (mask 2) and elevation (mask 3), indicates that the landscape's nature and the physical height of the terrain are also important but might offer secondary or nuanced insights into forest cover types. The dominant predictor seems to be soil types, particularly soil type 34, suggesting that specific soil characteristics play a crucial role in determining forest cover types. However, aspects like elevation and the nature of the terrain (wilderness area) also contribute to the classification and shouldn't be overlooked."}
Poker Hand	{"Mask 0": "Initially, the rank of card 2 is recognized as a significant predictor in the classification of poker hands. The rank of individual cards seems crucial at this stage.", "Mask 1": "At this step, the rank of card 3 is highlighted, further emphasizing the importance of individual card ranks in determining poker hand types.", "Mask 2": "Here, besides the rank of card 3, the suit of the same card is considered, suggesting that both rank and suit play roles in classifying poker hands.", "Mask 3": "This step combines multiple features, both ranks (cards 1 and 3) and suits (cards 3 and 5), indicating that a combination of these factors provides a broader understanding of poker hand classifications.", "Aggregate": "The rank of cards, specifically card 2 and card 3, appears consistently across different masks, indicating their primary role in poker hand classification. Suits, particularly of card 3, also have importance but are introduced later in the feature selection. The combined importance of both ranks and suits in the final step suggests that while individual card ranks are pivotal in initial classifications, understanding the relationship between card ranks and their respective suits provides a more comprehensive insight into the poker hand types. The consistent presence of card 3's attributes (both rank and suit) underscores its pivotal role in determining poker hand categories."}

Mushroom	{"Mask 0": "At the outset of the feature selection process, the model identifies physical attributes such as bruises, odor, and veil-type as significant indicators for classifying mushrooms into edible or poisonous categories. This suggests that visible physical changes and inherent characteristics play an essential role in the initial classification stages.", "Mask 1": "In this step, stalk-shape is introduced while retaining bruises and odor from the previous mask. Here, the model seems to emphasize the physical appearance (bruises and stalk-shape) in conjunction with inherent characteristics (odor) for the classification. Veil-type, although significant in the initial step, is omitted here, hinting at its reduced relevance or a possible correlation with other selected features.", "Mask 2": "The model shifts its attention here, emphasizing gill-attachment and stalk-surface-below-ring while still retaining the veil-type. This shift indicates the significance of features directly related to the mushroom's growth structure and its underlying layers. The absence of previously highlighted features suggests a refined approach to feature importance at this step.", "Mask 3": "This combination merges the significant features from the earlier masks, highlighting bruises, odor, gill-attachment, and veil-type. The recurrence of these features underscores their consistent importance in classifying mushrooms. The introduction of gill-attachment here, combined with its presence in the previous mask, further emphasizes its role.", "Aggregate": "The features that consistently appear across multiple masks are: bruises (masks 0, 1, and 3), odor (masks 0, 1, and 3), and veil-type (masks 0, 2, and 3). This suggests that these attributes play a crucial role in distinguishing between edible and poisonous mushrooms in this dataset. Gill-attachment, although appearing in the latter stages (masks 2 and 3), indicates its significance in the classification process. The dataset appears to highlight the importance of physical attributes (like bruises and stalk-shape), inherent characteristics (like odor), and growth-related features (like gill-attachment and stalk-surface-below-ring) in determining the edibility of mushrooms. Further investigation into these features' interactions and exact impacts would provide deeper insights into the classification process."}

Blastchar	{ "Mask 0": "The initial feature selection suggests that aspects related to customer service preferences such as StreamingTV and billing and payment methods like PaperlessBilling and PaymentMethod are significant predictors for customer churn. These features might be pointing towards a customer's comfort and adaptability with the service's digital offerings.", "Mask 1": "In this step, demographic aspects come into play with the inclusion of SeniorCitizen, suggesting that age or seniority might have a bearing on customer churn. The continued emphasis on PaperlessBilling also underlines its importance in this prediction.", "Mask 2": "At this juncture, the model considers the type of service subscribed by the customer, as seen from the inclusion of InternetService and DeviceProtection. MonthlyCharges also comes into the picture, indicating that the amount being charged could play a role in a customer's decision to stay or leave the service. It's possible that customers assess the value they receive from these services against what they are being charged.", "Mask 3": "In this combination, there's a mix of service-related (DeviceProtection), demographic (Partner), and payment (PaymentMethod) features. The appearance of Partner might indicate the influence of family decisions or joint subscriptions on customer churn.", "Aggregate": "PaperlessBilling (masks 0 and 1) and DeviceProtection (masks 2 and 3) consistently appear in multiple steps, highlighting their significance in predicting customer churn. PaymentMethod also appears in masks 0 and 3, suggesting the importance of billing methods in this context. On the other hand, unique features like SeniorCitizen and InternetService give insight into the potential impact of demographics and service type respectively. In essence, customer service preferences, billing methods, and charges, along with certain demographic elements, are key in predicting customer churn in this dataset. For a comprehensive understanding, a deeper dive into the interaction of these features and their respective weights in the prediction model would be beneficial." }

Diabetes	{ "Mask 0": "In the initial step, the model identifies discharge disposition id, number of diagnoses, and specific medications (glyburide, and citoglipton) as key features. This suggests a combination of administrative data (discharge process), clinical data (diagnoses), and prescribed medications play a significant role in predicting patient readmission.", "Mask 1": "Here, the model focuses on the patient's admission type and discharge disposition, emphasizing the importance of the hospital admission and discharge process in determining the likelihood of readmission. The absence of clinical and medication features in this step indicates a potential higher influence of administrative processes at this stage of feature selection.", "Mask 2": "The model reintroduces clinical data (number of diagnoses) and adds a new medication (metformin), while retaining admission type id. This combination underscores a nuanced interplay between the nature of hospital admission, the complexity of patient's health conditions, and specific treatments in predicting readmissions.", "Mask 3": "This step focuses on the number of diagnoses and two medications (glipizide and metformin). The consistency of number of diagnoses across masks highlights its pivotal role, while the introduction of glipizide and retention of metformin point towards an evolving emphasis on medication in conjunction with clinical data.", "Aggregate": "Number diagnoses (masks 0, 2, and 3) and discharge disposition id (masks 0 and 1) are recurrent, underscoring their persistent importance in predicting readmissions. Admission type id also appears consistently (masks 1 and 2), linking the administrative process tightly with readmission predictions. The changing set of medications across the masks (glyburide, citoglipton, metformin, glipizide) points to the dynamic nature of treatment's role in the prediction, warranting a closer look into specific medication combinations and their interactions with clinical and administrative data. The aggregate analysis reveals a complex interplay between administrative processes, clinical data, and medication in predicting patient readmissions, suggesting a multifaceted approach for model refinement and interpretation." }

Higgs	{ "Mask 0": "The initial feature selection emphasizes on variables related to kinematics and energy in the Higgs dataset, particularly invariant mass of the lepton, missing energy, and jets, implying these measurements play a fundamental role in distinguishing signal from background processes.", "Mask 1": "At this step, the model focuses on the Azimuthal angle for the second leading jet, suggesting that the orientation or direction of this jet may have a unique significance in the classification process.", "Mask 2": "Here, the model considers the transverse momentum of the second leading jet alongside the invariant mass of the lepton, missing energy, and leading jet. This suggests an interplay between momentum and energy variables as critical determinants in the classification.", "Mask 3": "In this step, the B-tagging score for the second leading jet and the pseudorapidity of the lepton are highlighted. B-tagging, used to identify the presence of a b-quark, combined with the pseudorapidity, indicates the model's attempt to understand particle behavior and properties for classification.", "Aggregate": "Over the feature selection steps, emphasis is consistently placed on high-energy physics variables, particularly those related to jets and leptons. The invariant mass of the lepton and missing energy are repeated features, appearing in both masks 0 and 2, showcasing their importance in the classification process. Momentum and angular measurements, like the Azimuthal angle and transverse momentum, also play a pivotal role. As the steps progress, there's an evident shift from energy-related features to more particle-specific characteristics, such as B-tagging. Overall, the classification in the Higgs dataset relies heavily on a combination of energy measurements, momentum, and particle properties."}