# Post-Hoc Feature Selection Layer for Neural Networks Interpretability

João Kenji Suwa
*Institute of Informatics*
*UFRGS*
Porto Alegre, Brazil
joao.suwa10@gmail.com

João Pedro Silveira e Silva
*Institute of Informatics*
*UFRGS*
Porto Alegre, Brazil
jpedross1999@gmail.com

Bruno Iochins Grisci
*Institute of Informatics*
*UFRGS*
Porto Alegre, Brazil
bigrisci@inf.ufrgs.br

***Abstract*—The interpretability of complex neural networks remains a critical challenge, especially for models already deployed in high-stakes domains. To address this, we adapted the Feature Selection Layer (FSL) to optimize feature weights post-training, a variant we term post-hoc FSL. Our approach reframes the FSL as a lightweight, trainable module that integrates with already frozen pre-trained models on tabular datasets to highlight the features the original model considers most important. This post-hoc FSL learns feature relevance by fine-tuning its weights based on the original model's learned outputs. Crucially, this process is non-invasive, operating without altering the original model's architecture or its learned parameters. We conducted our experiments using both statistical and visual metrics, including accuracy, F1 score, recall, precision, weighted t-SNE and silhouette score, and also analyzed the stability of the post-hoc FSL on high-dimensional synthetic and real-world tabular datasets. We compared the post-hoc FSL feature weighting method using these metrics against the original embedded FSL and other post-hoc interpretability methods, such as Integrated Gradients, Noise Tunnel, DeepLIFT, Gradient SHAP, and Feature Ablation. Experimental results demonstrate that the post-hoc FSL feature weighting method successfully identified relevant features across the different datasets, maintaining the predictive power of the original neural network while enhancing its interpretability.**

***Index Terms*—feature selection, deep learning interpretability, post hoc feature weighing**

## I. Introduction

Over the past decade, the availability of large-scale data has established Deep Neural Networks (DNNs) as state-of-the-art tools for knowledge extraction in domains such as computer vision and healthcare [1], [2]. However, due to their lack of interpretability, their "black box" nature poses serious challenges for transparency, ethical accountability, and bias mitigation [3], [4]. Addressing the trade-off between accuracy and interpretability remains a critical challenge in the safe and reliable deployment of deep learning systems. To address this interpretability problem, feature weighting methods assign relevance scores to input features, providing insights into feature behavior in DNN models [5]. Among these techniques, the Feature Selection Layer (FSL), proposed by [6], is an embedded feature attribution method designed to learn feature relevance based on the model's internal dynamics during the training phase. However, its embedded nature requires joint training from scratch, limiting its use for interpreting pre-trained DNN models. To overcome this limitation, we adapted the FSL to optimize feature weights after the initial training phase, a variant we term post-hoc FSL. Post-hoc FSL preserves the efficacy of the original method while enabling feature relevance analysis on pre-trained networks. This approach enhances interpretability and potentially improves predictive performance. Furthermore, by keeping the backbone network frozen, this method improves computational efficiency, as the optimization of FSL weights converges rapidly compared to training from scratch.

This research was funded by Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS) [24/2551-0000725-3; 25/2551-0002116-2]. This work was supported by the Serrapilheira Institute (grant number Serra – R-2501-51351). This work was supported by PIBIC CNPq-UFRGS. This study was financed in part by the Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES) – Finance Code 001.

## II. Related Work

### A. Feature Selection

Feature selection aims to identify a compact subset of relevant features to improve model performance and reduce complexity [7]. Most approaches rely on feature weighting to assign relevance scores and are commonly categorized by their interaction with the learning algorithm [8], [9]. Filter methods are model-agnostic and computationally efficient but often ignore feature interactions. Wrapper and embedded methods exploit a specific predictive model, achieving strong performance at the cost of higher computational expense and limited generalization across architectures [7]. Hybrid methods combine multiple strategies to balance efficiency and accuracy [10], while ensemble methods aggregate multiple selection runs to improve robustness and stability.

### B. Feature Selection Layer

The Feature Selection Layer (FSL), introduced by [6], is an embedded feature weighting method for neural networks. It consists of a simple layer between the neural network's input and its first hidden layer, where each neuron maintains a one-to-one correspondence with an input feature. Weights, initialized as $1/n$, where $n$ is the total number of features, are jointly trained with the network and reflect feature relevance during learning. A modified L1 regularization term is added

to ensure the interaction between features, which is calculated as:

$$r(W^{\text{FSL}}) = \lambda \cdot \left| \sum_{t=1}^{n} W_t^{\text{FSL}} - 1 \right| \quad (1)$$

where $W^{\text{FSL}}$ represents the weights from the FSL, $n$ is the number of features and $\lambda$ is a value between 0 and 1 that determines the strength of the regularization. FSL then adds two hyperparameters to the network: the FSL activation function and the $\lambda$ term for regularization.

### C. Post-hoc Interpretability Methods

Post-hoc feature attribution methods are essential for interpreting "black box" models such as neural networks. Gradient-based approaches, including Integrated Gradients [11], DeepLIFT [12], and Gradient SHAP [13], estimate feature relevance relative to a reference input. Integrated Gradients and Gradient SHAP compute attributions along paths from the baseline to the input, while DeepLIFT propagates relevance based on differences from reference activations. Other methods include Feature Ablation, a perturbation-based technique, and Noise Tunnel, which improves the robustness of gradient-based attributions by smoothing [14]. In our experiments, we apply Noise Tunnel with Integrated Gradients and compare the proposed post-hoc FSL method against these approaches using the metrics described in Section IV-B.

## III. Proposed Method

We propose an adaptation of FSL designed to enhance the interpretability of pre-trained neural networks, which we term post-hoc FSL. Our method adapts the FSL [6], reformulating it as a trainable module applicable to any compatible pre-trained neural network. The core objective is to identify the most relevant features for a model's predictions without altering its original learned parameters.

### A. Architecture and Mechanism

Architecturally, the proposed post-hoc FSL remains faithful to the original FSL. It is a simple, trainable layer with a one-to-one mapping between neurons and input features, placed between the input and a pre-trained network (Figure 1). Each feature $\vec{x_i}$ is scaled by a trainable weight $\vec{w_i}$, producing a weighted vector $\vec{wx}$ passed to the subsequent layers of the network. Unlike the original FSL, our method leverages the pre-learned internal weights of the network to highlight which features are considered most relevant, without modifying the model's internal parameters. After the analysis, the post-hoc FSL yields a set of feature weights that quantifies feature importance, while also being capable of improving predictive performance by amplifying relevant features and suppressing noise. Importantly, the post-hoc FSL can be removed after the analysis, allowing the pre-trained model to revert to its original state. Despite its distinct training paradigm, we retain the inherent benefits of the FSL architecture, which is computationally efficient and performs competitively in dimensionality reduction and accuracy when compared to AFS [15], Random Forest, ReliefF, and Mutual Information [6].

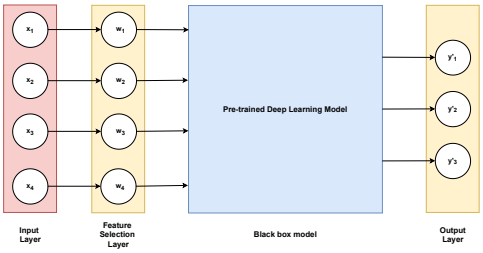



Fig. 1: **Representation of post-hoc Feature Selection Layer.** In this example, four input features are individually multiplied by their activated weights within the FSL before being passed to the pre-trained model. During fine-tuning, only the weights within the FSL are updated, while all parameters of the pre-trained model remain frozen.

### B. Post-Hoc Training Procedure

Our feature weighing approach operates via a distinct post-hoc training procedure. Given a pre-trained model $f_{\text{pre}}$ with its internal parameters frozen, a vector $\vec{x}$ that represents a single input, a vector $\vec{w}$ which represents the weights from the FSL, an activation function $a$, a loss function $\ell$, and a learning rate $\eta$, the training process is as follows:

**Step 1:** Our feature weighting layer is integrated between the input layer and the hidden layers of the pre-trained neural network (Equation 2).

**Step 2:** Data is fed through the combined model (FSL and the frozen network), and the final output is used to compute the loss function of the whole model (Equation 3).

**Step 3:** During backpropagation the parameters of the pre-trained network remain unchanged, while gradients are computed only for the FSL weights.

**Step 4:** Steps 2 and 3 are repeated over several epochs. This allows the weights of the post-hoc FSL to converge, with their final magnitudes representing the feature importance.

$$\vec{x}' = a(\vec{w}^{(i)}) \odot \vec{x} \quad (2)$$

$$\vec{y} = f_{\text{pre}}(\vec{x}') \quad (3)$$

### C. Layer Configuration: Regularization, Activation, and Initialization

We adopt three design choices to regulate the Feature Selection Layer (FSL). First, we apply the standard $L1$ regularization to promote sparsity by shrinking weights of irrelevant features toward zero. Second, we use ReLU activation to ensure non-negative weights, simplifying interpretation: a weight of zero indicates irrelevance, and positive values represent features' importance. Third, we initialize weights to $1.0$ instead of $1/n$, providing a stable starting point in which all features are ingested by the pre-trained model initially using unchanged values and preventing initializations with weights close to zero.

## IV. Experiments

This section details the experimental setup used to evaluate the proposed post-hoc FSL method. We describe the datasets

(Section IV-A), the evaluation metrics (Section IV-B), and the specific implementation details (Section IV-C).

### A. Datasets

To evaluate the proposed feature selection layer, we employed synthetic datasets with predefined relevant and noisy features (Section IV-A1) and real-world datasets, including email spam classification [16] and high-dimensional microarray data with limited samples [17] (Section IV-A2). Using diverse datasets allows us to evaluate post-hoc FSL's effectiveness in reducing dimensionality and improving neural network interpretability by identifying relevant features and suppressing noise. All datasets consist of complete numerical features with no missing values and are fully described in Table I.

TABLE I: **Datasets used for the experiments**

| Name | Feature type | Labels | Features | Informative | Folds | Samples | Test Split |
|---|---|---|---|---|---|---|---|
| XOR | Binary | 2 | 50 | 2 | 11 | 500 | 20% |
| SynthA | Numerical | 2 | 100 | 30 | 11 | 3000 | 10% |
| Spam | Numerical | 2 | 3000 | Unknown | 11 | 5172 | 20% |
| Liver-22405 | Numerical | 2 | 22284 | Unknown | 15 | 48 | 10% |
| Breast-45827 | Numerical | 6 | 54676 | Unknown | 15 | 151 | 10% |

*1) Synthetic Datasets:* For this analysis, we used the XOR and SynthA synthetic datasets, with controlled proportions of relevant and irrelevant features. SynthA was generated by first defining an $n$-dimensional hypercube, where $n$ is the number of informative features. Then, Gaussian clusters of data points are created at each vertex of the hypercube, with an equal number of clusters assigned to each class. The informative features are the ones that are coordinates for the created points in the hypercube [7]. Synthetic data provides ground truth for feature relevance, enabling direct evaluation of post-hoc FSL's selection accuracy (Section IV-B), based on the weights assigned to each feature.

*2) Real-world datasets:* To evaluate the post-hoc FSL on real-world high-dimensionality and low-sample-size challenges, we selected datasets from two domains. First, we utilized microarray datasets from the CuMiDa repository [17], a database of cancer datasets selected from over $30,000$ GEO experiments, following bias-reduction guidelines by [18]. We used Liver-GSE22405 (binary classification) and Breast-GSE45827 (six-class classification). Additionally, we evaluated a spam email dataset [16], represented by word frequencies in the email corpus.

### B. Evaluation Metrics

To evaluate the post-hoc FSL, we employed statistical, visual, and stability metrics (Sections IV-B1, IV-B2, and IV-B3) to assess its effectiveness.

*1) Performance Metrics:* We assess model predictive performance using accuracy, precision, recall, and F1 score. Given a synthetic dataset with known ground truth and a subset of top-ranked features, we evaluate feature selection using two metrics [7] ranging from $0.0$ to $1.0$: the percentage of informative features selected, **PIFS** (Equation 4), and the percentage of selected features that are informative, **PSFI** (Equation 5).

$$PIFS = \frac{|S_{\text{selected}} \cap S_{\text{informative}}|}{|S_{\text{informative}}|} \tag{4}$$

$$PSFI = \frac{|S_{\text{selected}} \cap S_{\text{informative}}|}{|S_{\text{selected}}|} \tag{5}$$

*2) Visual Metrics:* For visual analysis, we use a variant of t-SNE (t-Distributed Stochastic Neighbor Embedding) [19], a technique that projects high-dimensional data into a low-dimensional space (typically two dimensions) based on pairwise distances between points. To visualize the discriminative power of post-hoc FSL features, we use weighted t-SNE [20]. This algorithm scales Euclidean distances by feature importance, ensuring that highly relevant dimensions contribute more significantly to the final spatial arrangement. Higher values indicate well-separated and dense clusters, while values near $0$ suggest boundary points and values close to $-1$ indicates that the points are closer to some other cluster than the one they are assigned. The primary objective of using weighted t-SNE is to provide a global visual inspection tool to compare how different feature scoring techniques impact data clustering.

*3) Stability Metrics:* To evaluate the stability of the post-hoc FSL feature weighting method, we adopted a stability analysis protocol based on data perturbation, which assesses how the method's output varies under small changes in the training data [21]. We employed $k$-fold stratified cross-validation, where the dataset is partitioned into $k$ folds and the feature weighting method is applied $k$ times, each time using $k$–1 folds for training and one fold for validation. This process produces $k$ sets of feature weights, which are then analyzed to assess stability by measuring their consistency across folds using three metrics proposed by [7]. The Jaccard Index measures the overlap between two sets with the top-$n$ selected features, capturing the subset stability. The Pearson correlation evaluates the similarity of feature weights. Lastly, the Spearman rank correlation quantifies the agreement between feature rankings.

### C. Experiments setup

In our experimental evaluation, we trained 10 models, being two per dataset: (1) a baseline model without embedded feature weighting and (2) a model with an embedded Feature Selection Layer (FSL). The baseline models were designed to be interpreted by the proposed post-hoc FSL and compared post-hoc methods (Section II-C). All models employed different architectures and were trained using the Adam optimizer (learning rate $0.001$) with a weighted cross-entropy loss. To assess the effectiveness of the post-hoc FSL, we compared it with state-of-the-art post-hoc methods and the original FSL using the statistical, visual, and stability metrics described in Section IV-B, computed as the average across cross-validation folds. Additionally, we extended the evaluation to TabPFN, a pre-trained transformer-based neural network [22].

## V. Results and Discussion

### A. Results for Synthetic Datasets

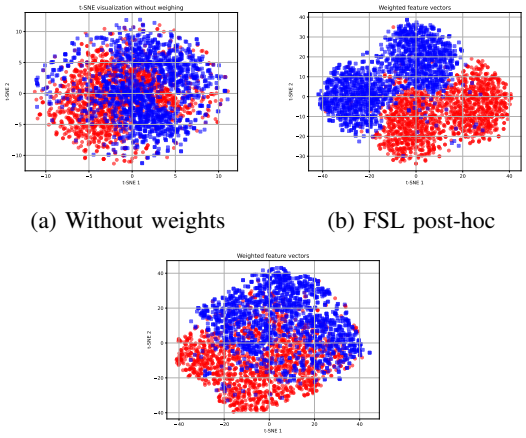
When comparing the predictive performance for the XOR dataset, the baseline model achieved near-optimal predictive performance, 0.999 across all metrics, while both FSL and post-hoc FSL reached perfect scores of 1.0. All models consistently identified the two relevant features across all folds, achieving perfect PIFS, PSFI, and Jaccard similarity scores of 1.0 when evaluated on the top-2 features. This consistency is further supported by the Pearson correlation metric, which yielded values near 1.0. Despite this, the Spearman rank correlation remained near 0.0 across all methods, indicating that while relevant features were prioritized over irrelevant ones, their relative rankings can vary. When comparing the silhouette scores and weighted t-SNE visualizations, all models achieved statistically similar results, while improving cluster separability when compared to the standard t-SNE ($p < 0.01$).

For the SynthA dataset, both FSL variants surpassed the baseline across all predictive metrics (F1 score, accuracy, precision and recall). Post-hoc FSL showed moderate improvement, yielding scores of approximately 0.916, whereas FSL consistently delivered the highest performance, with scores approaching 0.954. Furthermore, both FSL variants and the alternative post-hoc methods yielded statistically comparable results in terms of cluster separability (with all methods significantly outperforming the baseline, $p < 0.01$). Despite this statistical parity, numerical variations were observed in the average silhouette scores. Post-hoc FSL recorded the lowest average silhouette score (0.115), contrasting with higher averages for FSL (0.191) and the other post-hoc approaches, which ranged from a minimum of 0.181 (Feature Ablation) to a maximum of 0.235 (Integrated Gradients). Regardless of these constraints, the FSL post-hoc reached PFSI and PIFS scores of 0.906 on the top-30 features, demonstrating its ability to identify informative features, albeit in this specific case, less effectively than the FSL, which reached perfect scores of 1.0 on both metrics. In comparison, the other post-hoc approaches achieved scores in the range of 0.96–0.99. These results influenced post-hoc FSL stability scores (Table II), which also had an inferior performance when compared to the other methods. We hypothesize that the synthetic dataset's structure may have hindered post-hoc FSL's feature identification. Nonetheless, its performance on real-world, high-dimensional datasets remained strong.

To complement our experiments, we used TabPFN [22], a pre-trained Transformer-based model for supervised classification on small tabular datasets that operates without hyperparameter tuning and achieves competitive performance against state-of-the-art classification methods. Due to source code compatibility, comparisons were made using the original Kernel SHAP implementation. Both post-hoc techniques successfully identified the most relevant features, yielding identical PSFI and PIFS scores of 0.966 when evaluating the top-30 ranked features, which ideally should include all, and only, the informative features. However, post-hoc FSL visualization outperformed both weighted t-SNE using Kernel SHAP weights and standard t-SNE (Figure 2), achieving a higher silhouette score of 0.316 compared to 0.186 and 0.070, respectively.

TABLE II: **Results for stability metrics using the SynthA dataset.** Bold indicates the highest average. Jaccard was calculated using the top-30 weighted features.

| Model | Jaccard | Spearman | Pearson |
|---|---|---|---|
| Integrated Gradients | $0.855 \pm 0.028$ | $0.758 \pm 0.038$ | $0.969 \pm 0.009$ |
| Noise Tunnel | $0.954 \pm 0.031$ | $0.752 \pm 0.038$ | $0.974 \pm 0.007$ |
| Deep Lift | $0.976 \pm 0.034$ | $0.760 \pm 0.037$ | $0.971 \pm 0.008$ |
| Gradient SHAP | $0.939 \pm 0.041$ | $0.750 \pm 0.041$ | $0.975 \pm 0.006$ |
| Feature Ablation | $0.956 \pm 0.050$ | $0.759 \pm 0.037$ | $0.980 \pm 0.005$ |
| FSL | $\mathbf{1.000 \pm 0.000}$ | $\mathbf{0.805 \pm 0.030}$ | $\mathbf{0.980 \pm 0.003}$ |
| Post-hoc FSL | $0.759 \pm 0.052$ | $0.759 \pm 0.030$ | $0.770 \pm 0.027$ |




(a) Without weights (b) FSL post-hoc

(c) Kernel SHAP

Fig. 2: **Weighted t-SNE** of the post-hoc FSL and Kernel SHAP methods using the TabPFN and the SynthA dataset.

### B. Results for Real-world Datasets

We evaluated the post-hoc FSL on real-world datasets (Section IV-A2), including spam email detection and high-dimensional, low-sample-size microarray analysis. The results for the spam dataset are presented in Table III, IV and Figure 3. As shown in Table III, the baseline and post-hoc FSL models achieved the highest performance metrics. The baseline model recorded the best average scores across accuracy, F1 score, recall, and precision. While the baseline model demonstrated the top performance, post-hoc FSL obtained comparable results and outperformed FSL. In terms of visual separability metrics, the FSL post-hoc yielded the best results, achieving a higher average silhouette coefficient than both the FSL and the other post-hoc methods (Table IV and Figure 3). However, in terms of stability, both FSL variants underperformed compared to the other post-hoc approaches. The Jaccard index, computed using the top-100 weighted features, ranged from 0.382 to 0.486 across all methods, indicating relatively consistent feature selection across runs. In contrast, Spearman rank correlation

was substantially lower for FSL and post-hoc FSL (0.157 and 0.312), while the remaining post-hoc methods achieved averages between 0.553 and 0.741. Pearson correlation was generally higher across all approaches (0.614–0.818), likely reflecting the large number of consistently irrelevant features with near-zero weights. Although our proposed FSL achieved comparable Jaccard and Pearson scores, it exhibited lower overall stability than other post-hoc methods. Nevertheless, it consistently assigned meaningful weights to a subset of highly relevant features, suggesting that while the global ranking may vary, the selected features remain informative.

TABLE III: **Prediction performance for real-world datasets.** Bold indicates the highest average; best statistically significant results ($p < 0.01$) from Kruskal-Wallis and Dunn's tests are marked in orange.

(a) Spam dataset

| Model | F1 Score | Accuracy | Precision | Recall |
|---|---|---|---|---|
| Baseline | $\mathbf{0.976 \pm 0.005}$ | $\mathbf{0.976 \pm 0.005}$ | $\mathbf{0.977 \pm 0.005}$ | $\mathbf{0.976 \pm 0.005}$ |
| FSL | $0.968 \pm 0.004$ | $0.968 \pm 0.004$ | $0.969 \pm 0.004$ | $0.968 \pm 0.004$ |
| FSL post-hoc | $0.973 \pm 0.003$ | $0.973 \pm 0.003$ | $0.974 \pm 0.003$ | $0.973 \pm 0.003$ |

(b) Liver dataset

| Model | F1 Score | Accuracy | Precision | Recall |
|---|---|---|---|---|
| Baseline | $0.800 \pm 0.076$ | $0.800 \pm 0.076$ | $0.858 \pm 0.079$ | $0.800 \pm 0.076$ |
| FSL | $\mathbf{0.840 \pm 0.083}$ | $\mathbf{0.840 \pm 0.083}$ | $\mathbf{0.893 \pm 0.055}$ | $\mathbf{0.840 \pm 0.083}$ |
| Post-hoc FSL | $0.827 \pm 0.070$ | $0.827 \pm 0.070$ | $0.884 \pm 0.047$ | $0.827 \pm 0.070$ |

(c) Breast dataset

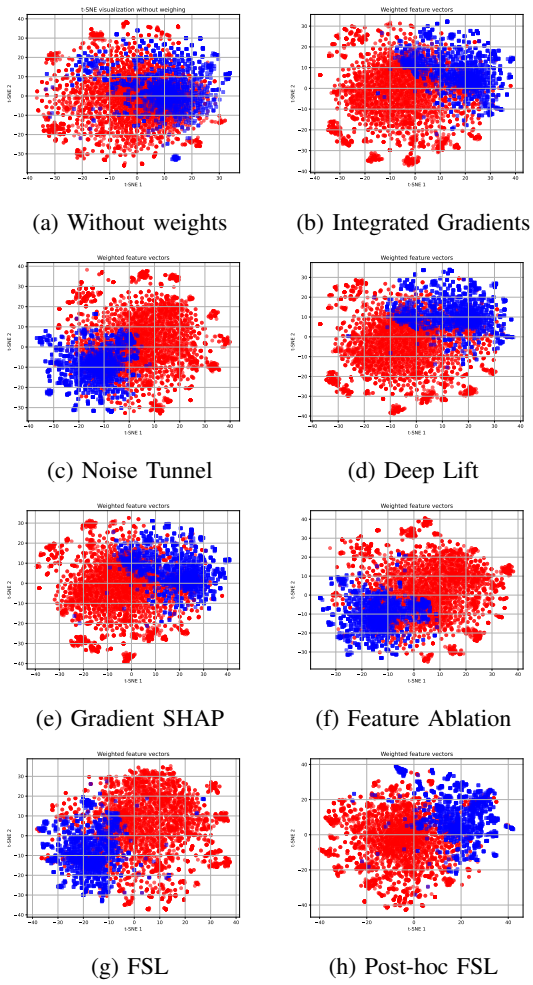
| Model | F1 Score | Accuracy | Precision | Recall |
|---|---|---|---|---|
| Baseline | $0.942 \pm 0.024$ | $0.945 \pm 0.021$ | $0.955 \pm 0.024$ | $0.945 \pm 0.021$ |
| FSL | $0.955 \pm 0.033$ | $0.958 \pm 0.030$ | $0.964 \pm 0.030$ | $0.958 \pm 0.030$ |
| FSL post-hoc | $\mathbf{0.970 \pm 0.033}$ | $\mathbf{0.970 \pm 0.032}$ | $\mathbf{0.978 \pm 0.024}$ | $\mathbf{0.970 \pm 0.032}$ |

TABLE IV: **Silhouette score for real-world dataset.** Bold indicates the highest average; best statistically significant results ($p < 0.01$) from Kruskal-Wallis and Dunn's tests are marked in orange.

| Model | Spam | Liver | Breast |
|---|---|---|---|
| None | $0.090 \pm 0.000$ | $0.066 \pm 0.000$ | $0.192 \pm 0.000$ |
| Integrated Gradients | $0.202 \pm 0.031$ | $0.170 \pm 0.031$ | $0.451 \pm 0.036$ |
| Noise Tunnel | $0.189 \pm 0.034$ | $0.178 \pm 0.044$ | $0.447 \pm 0.045$ |
| Deep Lift | $0.202 \pm 0.025$ | $0.172 \pm 0.049$ | $0.449 \pm 0.030$ |
| Gradient SHAP | $0.209 \pm 0.022$ | $0.199 \pm 0.042$ | $0.450 \pm 0.040$ |
| Feature Ablation | $0.196 \pm 0.029$ | $0.181 \pm 0.030$ | $0.450 \pm 0.033$ |
| FSL | $0.234 \pm 0.013$ | $0.574 \pm 0.141$ | $\mathbf{0.647 \pm 0.021}$ |
| Post-hoc FSL | $\mathbf{0.235 \pm 0.069}$ | $\mathbf{0.666 \pm 0.012}$ | $0.511 \pm 0.078$ |

TABLE V: **Results for stability metrics using the Liver dataset.** In bold is the highest average of each metric. Jaccard was calculated using the top-100 weighted features.

| Model | Jaccard | Spearman | Pearson |
|---|---|---|---|
| Integrated Gradients | $0.263 \pm 0.078$ | $\mathbf{0.649 \pm 0.091}$ | $0.739 \pm 0.081$ |
| Noise Tunnel | $0.226 \pm 0.102$ | $0.647 \pm 0.092$ | $0.728 \pm 0.088$ |
| Deep Lift | $0.263 \pm 0.077$ | $\mathbf{0.649 \pm 0.091}$ | $0.739 \pm 0.081$ |
| Gradient SHAP | $0.122 \pm 0.047$ | $0.599 \pm 0.098$ | $0.666 \pm 0.084$ |
| Feature Ablation | $0.261 \pm 0.077$ | $\mathbf{0.649 \pm 0.092}$ | $0.739 \pm 0.081$ |
| FSL | $0.174 \pm 0.064$ | $0.017 \pm 0.016$ | $0.454 \pm 0.180$ |
| Post-hoc FSL | $\mathbf{0.472 \pm 0.086}$ | $0.090 \pm 0.014$ | $\mathbf{0.832 \pm 0.058}$ |



(a) Without weights (b) Integrated Gradients (c) Noise Tunnel (d) Deep Lift (e) Gradient SHAP (f) Feature Ablation (g) FSL (h) Post-hoc FSL

Fig. 3: **Weighted t-SNE** of all feature weighing metrics using the spam dataset.

On the Liver and Breast datasets, post-hoc FSL achieved predictive performance comparable to the baseline and original FSL (Table III), while consistently producing more meaningful feature attributions. Both FSL variants outperformed state-of-the-art post-hoc methods in silhouette scores (Table IV), with post-hoc FSL peaking on the Liver dataset (0.666) and original FSL on the Breast dataset (0.647), demonstrating their ability to create well-defined class clusters. Regarding stability (Table V), post-hoc FSL yielded the highest Jaccard and Pearson scores on the Liver dataset, although both FSL variants showed lower Spearman scores. Conversely, both methods exhibited higher instability across all metrics on the Breast dataset, likely because the FSL architecture struggles to capture local, class-specific relationships across its six classes.

### C. *Feature Erasure*

We further evaluated feature relevance using Feature Erasure, which measures performance impact by zeroing out the top-$n$ features ranked by post-hoc weights and later performing inference on the baseline model. We progressively erased

the top-1 to 50 features for the XOR and SynthA datasets, and top-100 to 4000 features for the Spam dataset. After masking the top-2 for XOR, top-30 for SynthA and top-2000 for Spam the baseline model's F1-Score and accuracy significantly degraded. All methods, including post-hoc FSL, exhibited similar performance degradation, demonstrating their effectiveness in selecting relevant features.

## VI. Conclusion

This paper explores the development of a variation of the Feature Selection Layer (FSL) as a feature weighing method to improve the interpretability of pre-trained Deep Neural Networks. Our results demonstrate that the proposed method effectively isolates relevant features in high-dimensional datasets, matching or outperforming state-of-the-art post-hoc feature weighting methods. Unlike traditional methods, Post-hoc FSL can be integrated directly into the model architecture for subsequent inferences. By driving irrelevant feature weights toward zero, the method systematically excludes non-contributory inputs, thereby increasing reliability. Experimental results further indicate that coupling Post-hoc FSL typically yields a positive impact on the model's predictive performance.

### A. Limitations and future work

While effective for binary classification, our proposed method underperforms in multi-class settings, as it fails to capture local and class-specific feature relationships. Furthermore, our proposed method is specifically designed for tabular data, in which each feature has a fixed meaning within the dataset; however, this assumption does not hold for images, where the meaning of each pixel can vary across samples. Future work should focus on adapting the method to overcome these limitations and extend its interpretability to such domains.

### *Reproducibility Statement*

**Paper source code** is available at https://github.com/joaosuwa/FSL_post-hoc with all necessary dependencies, seeds and steps to reproduce each of the experiments presented in this paper. The original code of the weighted t-SNE [20] can be found at https://github.com/sbcblab/weighted_tSNE. **CuMiDa** datasets [17] can be found at https://sbcb.inf.ufrgs.br/cumida and **Spam** dataset [16] can be found at https://www.kaggle.com/datasets/balaka18/email-spam-classification-dataset-csv. Both **XOR** and **SynthA** dataset are available at the same repository of the paper source code. **TabPFN source code** is available at https://github.com/PriorLabs/TabPFN and **Kernel SHAP documentation** is available at https://shap.readthedocs.io/en/latest/.

## References


[1] M. Badawy, N. Ramadan, and H. A. Hefny, "Healthcare predictive analytics using machine learning and deep learning techniques: a survey," *Journal of Electrical Systems and Information Technology*, vol. 10, no. 1, p. 40, 2023.

[2] W. Khan, A. Daud, K. Khan, S. Muhammad, and R. Haq, "Exploring the frontiers of deep learning and natural language processing: A comprehensive overview of key challenges and emerging trends," *Natural Language Processing Journal*, vol. 4, p. 100026, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2949719123000237

[3] N. Y. Murad, M. H. Hasan, M. H. Azam, N. Yousuf, and J. S. Yalli, "Unraveling the black box: A review of explainable deep learning healthcare techniques," *IEEE Access*, vol. 12, pp. 66 556–66 568, 2024.

[4] S. J. Prince, *Understanding Deep Learning*. The MIT Press, 2023. [Online]. Available: http://udlbook.com

[5] C. Molnar, G. Casalicchio, and B. Bischl, "Interpretable machine learning–a brief history, state-of-the-art and challenges," in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2020, pp. 417–431.

[6] J. Figueroa Barraza, E. López Droguett, and M. R. Martins, "Towards interpretable deep learning: a feature selection framework for prognostics and health management using deep neural networks," *Sensors*, vol. 21, no. 17, p. 5888, 2021.

[7] M. C. Barbieri, B. I. Grisci, and M. Dorn, "Analysis and comparison of feature selection methods towards performance and stability," *Expert Systems with Applications*, vol. 249, p. 123667, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417424005335

[8] M. A. Tahir, A. Bouridane, and F. Kurugollu, "Simultaneous feature selection and feature weighting using hybrid tabu search/k-nearest neighbor classifier," *Pattern Recognition Letters*, vol. 28, no. 4, pp. 438–446, 2007.

[9] J. Miao and L. Niu, "A survey on feature selection," *Procedia computer science*, vol. 91, pp. 919–926, 2016.

[10] J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed, "Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 13, no. 5, pp. 971–989, 2015.

[11] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.

[12] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *International conference on machine learning*. PMlR, 2017, pp. 3145–3153.

[13] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[14] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson, "Captum: A unified and generic model interpretability library for pytorch," 2020. [Online]. Available: https://arxiv.org/abs/2009.07896

[15] N. Gui, D. Ge, and Z. Hu, "Afs: An attention-based mechanism for supervised feature selection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 3705–3713.

[16] Balaka, "Email spam classification dataset (csv)," https://www.kaggle.com/datasets/balaka18/email-spam-classification-dataset-csv, 2020, accessed: 2025-08-20.

[17] B. Feltes, E. B. Chandelier, B. I. Grisci, and M. Dorn, "Cumida: An extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research," *Journal of Computational Biology*, vol. 26, no. 4, pp. 376–386, 2019.

[18] B. I. Grisci, B. C. Feltes, J. de Faria Poloni, P. H. Narloch, and M. Dorn, "The use of gene expression datasets in feature selection research: 20 years of inherent bias?" *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 14, no. 2, p. e1523, 2024.

[19] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[20] B. I. Grisci, M. Inostroza-Ponta, and M. Dorn, "Assessing feature scorer results on high-dimensional datasets with t-sne," *Neurocomputing*, p. 130561, 2025.

[21] W. Awada, T. M. Khoshgoftaar, D. Dittman, R. Wald, and A. Napolitano, "A review of the stability of feature selection techniques for bioinformatics data," in *2012 IEEE 13th International conference on information reuse & integration (IRI)*. IEEE, 2012, pp. 356–363.

[22] N. Hollmann, S. Müller, K. Eggensperger, and F. Hutter, "Tabpfn: A transformer that solves small tabular classification problems in a second," *arXiv preprint arXiv:2207.01848*, 2022.