FSL-Net for Feature Importance

Jan Aguiló Plana

17/12/2025

1 Pre-trained FSL-Net Inference and Qualitative Analysis

In the first phase, we evaluate the behavior of the pre-trained Feature Shift Localization Network (FSL-Net) through inference-only experiments. FSL-Net is designed to detect feature-level distributional shifts between a clean reference dataset and a potentially shifted query dataset, without requiring labels or fine-tuning. We apply the model to the MNIST dataset by treating each pixel as an individual feature. Specifically, we construct a reference set composed of 1,000 images of digit 3 and a query set composed of 1,000 images of digit 8, with each image flattened into a 784-dimensional feature vector. The pre-trained FSL-Net model is then used to estimate, for each pixel, an independent probability of being distributionally shifted between the two datasets.

Figure 1 shows the resulting shift probability map produced by FSL-Net. High probability values are concentrated along regions where the digit shapes differ, while background pixels remain inactive. This indicates that FSL-Net successfully localizes distributional changes induced by the class shift.

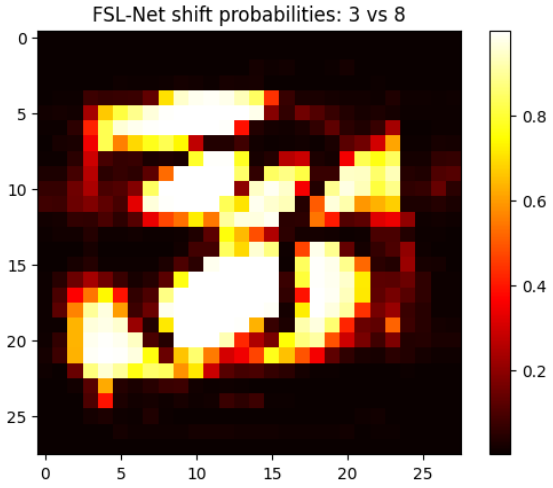

Figure 1: FSL-Net shift probability map for MNIST digit comparison (3 vs 8)

2 Feature Importance via FSL-Net Pre-trained

We study whether feature-level distributional shifts can be exploited as a proxy for feature importance in supervised digit classification. We consider the MNIST dataset, where each image is represented as a 784-dimensional vector of normalized pixel intensities, and use three standard classifiers: Logistic Regression, a Multilayer Perceptron (MLP), and XGBoost.

Since FSL-Net estimates for a given pair of datasets the probability that each feature exhibits a distributional shift and MNIST is a multi-class problem, we adopt a one-vs-all strategy. For each digit class $c \in \{0, \ldots, 9\}$, FSL-Net is applied between a reference set containing all training samples of class c and a query set containing all remaining samples. This yields ten class-conditional shift vectors $P_c \in [0, 1]^{784}$.

The class-conditional vectors are aggregated by taking their mean across classes, resulting in a single global shift-based importance vector. This vector is normalized by its maximum value to emphasize relative differences between features. Importantly, this score reflects distributional differences across classes rather than supervised discriminative importance.

Using this aggregated importance vector, we evaluate two approaches. First, percentile-based feature selection is performed by retaining the top 20%, 40%, 60%, and 80% of features according to their importance scores. Second, feature weighting is applied by scaling each input feature by its corresponding importance value while retaining all dimensions. For each configuration, the same classifiers are retrained using an identical train test split and hyperparameters, and performance is evaluated using balanced accuracy and macro-averaged ROC-AUC in a one-vs-rest setting.

2.1 Results

Table 1 reports the classification performance for each feature selection and weighting strategy, including the number of retained features, balanced accuracy, and ROC-AUC for each model. The results indicate that only XGBoost consistently maintains baseline-level performance under feature selection, even when retaining as few as 20% of the features, highlighting its robustness to aggressive dimensionality reduction. In contrast, Logistic Regression suffers a deep degradation in balanced accuracy across all feature selection and weighting strategies, suggesting that linearity relies on a broader set of complementary features not captured by shift-based importance. The MLP exhibits intermediate behavior, with noticeable performance drops under strong feature selection (Top 20%) but more moderate degradation at higher percentiles. Feature weighting similarly harms Logistic Regression and provides no clear benefit for the other models. While both MLP and XGBoost are non-linear, only the tree-based model consistently preserves performance under feature selection, indicating that internal feature selection plays a more critical role than non-linearity alone.

Scenario	**# Features**	**Model**	**Balanced Acc.**	**ROC-AUC**
Baseline	784	Logistic Regression	0.904687	0.987975
		MLP	0.974449	0.999398
		XGBoost	0.975461	0.999578
Top 20%	157	Logistic Regression	0.665054	0.942627
		MLP	0.868159	0.986084
		XGBoost	0.966892	0.999210
Top 40%	314	Logistic Regression	0.695404	0.958372
		MLP	0.910816	0.992916
		XGBoost	0.974769	0.999574
Top 60%	470	Logistic Regression	0.708343	0.961003
		MLP	0.921197	0.994250
		XGBoost	0.975249	0.999591
Top 80%	627	Logistic Regression	0.708339	0.961212
		MLP	0.923694	0.994238
		XGBoost	0.975107	0.999587
Weighted	784	Logistic Regression	0.689701	0.949915
		MLP	0.908046	0.991969
		XGBoost	0.975461	0.999578

Table 1: Phase 1 results: classification performance using feature selection (percentiles) and weighting based on FSL-Net distributional shifts.

3 Feature Importance via FSL-Net Fine-Tuned

In this phase, we investigate whether FSL-Net can be fine-tuned to encode a supervised notion of feature importance derived from downstream classifiers. Rather than relying solely on unsupervised distributional shifts, the goal is to align the output of FSL-Net with model-specific discriminative relevance.

For each baseline classifier (Logistic Regression, MLP, and XGBoost), a global feature-importance vector of dimension 784 is computed from the trained model parameters using standard model-specific heuristics. These vectors serve as pseudo-supervised targets, acknowledging that no ground-truth feature importance annotations are available for MNIST.

Starting from the same pre-trained FSL-Net backbone, three separate fine-tuned models are obtained, one per classifier. Fine-tuning is performed by freezing the majority of the network parameters and optimizing the final layers so that, given a pair of datasets, the predicted shift vector matches the corresponding classifier-derived importance vector. As in Phase 1, a one-vs-all strategy is used to generate meaningful distributional contrasts: for each digit class, the reference set consists of samples from that class, while the query set contains all remaining samples. The same global importance target is used across all one-vs-all pairs.

After fine-tuning, each FSL-Net variant is used to recompute a global importance vector by aggregating its outputs across the ten one-vs-all comparisons. The resulting vectors are normalized and evaluated using the same feature selection and feature weighting strategies as in Phase 1. For each configuration, the classifiers are retrained from scratch using identical splits and hyperparameters, and performance is measured using balanced accuracy and macro-averaged ROC-AUC. This allows a direct comparison between unsupervised shift-based importance and supervised-aligned importance learned by FSL-Net.

3.1 Results

Table 2 reports the classification performance obtained when feature selection and weighting are driven by the fine-tuned variants of FSL-Net, including the baseline performance without feature selection. Overall, supervised alignment does not improve downstream performance compared to the frozen model. Logistic Regression continues to exhibit severe degradation under feature selection and weighting, with balanced accuracy remaining substantially below the baseline across all configurations. The MLP shows gradual recovery as more features are retained, but consistently underperforms its baseline, particularly under aggressive feature selection. XGBoost remains the most robust model, preserving near-baseline performance even when retaining a small subset of features, although no systematic gains over the frozen FSL-Net are observed. These results indicate that, in the MNIST setting, unsupervised shift-based importance captures discriminative structure more effectively than supervised alignment.

Scenario	**# Features**	**Model**	**Balanced Acc.**	**ROC-AUC**
Baseline	784	Logistic Regression	0.904687	0.987975
		MLP	0.974449	0.999398
		XGBoost	0.975461	0.999578
Top 20%	157	Logistic Regression	0.668843	0.946395
		MLP	0.755506	0.962687
		XGBoost	0.936323	0.997223
Top 40%	314	Logistic Regression	0.695547	0.959115
		MLP	0.853422	0.983719
		XGBoost	0.958289	0.998945
Top 60%	470	Logistic Regression	0.708360	0.961006
		MLP	0.888425	0.989223
		XGBoost	0.967992	0.999367
Top 80%	627	Logistic Regression	0.708342	0.961175
		MLP	0.916109	0.993167
		XGBoost	0.973047	0.999484
Weighted	784	Logistic Regression	0.688010	0.951639
		MLP	0.903615	0.991263
		XGBoost	0.975461	0.999578

Table 2: Phase 2 results: classification performance using feature selection and weighting based on fine-tuned FSL-Net importance, including baseline performance without feature selection.