SPARSE FEATURE ROUTING FOR TABULAR LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

The landscape of high-performance tabular learning is defined by a difficult compromise between the opaque ensembles of gradient-boosted trees and deep models that rely on elaborate pre-training to adapt ill-suited, monolithic backbones. We argue this compromise stems from a fundamental architectural mismatch. We propose a more principled path forward with a decomposed architecture that performs instance-wise selection over independent feature experts. Our model, the Sparse Feature Routing Network (SFR Net), assigns a small expert to each feature and uses a sparse router to dynamically compose expert results into an instance-specific representation, while a low-rank module captures higher-order interactions. This design yields native instance-level attributions and remains computationally efficient. A comprehensive empirical study validates these advantages. Across diverse benchmarks, SFR Net consistently outperforms strong specialized baselines, including Transformer-based models. Furthermore, it remains highly competitive with powerful self-supervised learning methods, despite being trained end-to-end without the pre-training step. Our ablation studies rigorously quantify the contribution of each architectural module, proving that the performance gains stem from the principled decomposition and dynamic routing, not brute-force capacity. These results position Sparse Feature Routing as a transparent, efficient, and powerful foundation for deep tabular learning.

1 INTRODUCTION

Tabular data are among the most widespread data modalities, arising in domains such as healthcare, finance, and the social sciences. Despite their ubiquity, learning effective representations from tabular datasets remains a fundamental challenge. Unlike image, text, or audio inputs, tabular records consist of heterogeneous features of varying types and scales, often with weak or irregular dependencies. These characteristics limit the transfer of inductive biases that have powered deep learning breakthroughs in unstructured modalities. As a result, classical approaches such as gradient-boosted decision trees remain dominant in practice, while neural networks have historically struggled to consistently outperform them.

Recent research has begun to narrow this gap by improving training strategies, designing novel architectures, and exploring self-supervised objectives tailored to tabular domains. Advances in normalization, regularization, and feature encoding have enhanced the robustness of deep models, while architectural innovations such as attention mechanisms and modular processing blocks have sought to capture complex feature interactions. In parallel, representation learning methods have shown that self-supervision can extract informative embeddings without labels, improving downstream classification and regression tasks. Yet, despite this progress, a core difficulty persists: most existing models still treat tabular inputs monolithically, processing all features through shared encoders. This design blurs the role of individual columns, complicates interpretability, and can reduce robustness in the presence of irrelevant or noisy features.

In this work, we propose the **Sparse Feature Routing Network (SFR Net)**, a new architecture designed to address these limitations. Our approach decomposes each feature into a specialized expert network, ensuring that heterogeneous columns are modeled according to their individual distributions. These experts are then dynamically composed by a lightweight router that assigns sparse, entropy-regularized attention weights, selecting only the most informative features for each instance. This design provides three key advantages: (i) principled handling of heterogeneity through feature-wise specialization, (ii) efficiency and robustness via sparse routing that scales with feature count

and resists irrelevant inputs, and (iii) native interpretability, as the router's weights yield direct, transparent per-instance attributions without the need for post-hoc analysis.

Beyond raw performance, the novelty of SFR Net lies in introducing sparse, feature-level routing as an inductive principle for tabular learning. While previous approaches have relied on either dense transformations of all features or indirect mechanisms such as mask prediction, our framework directly encodes the idea that each column should contribute selectively and independently to the decision process. This perspective not only improves predictive accuracy but also aligns with how practitioners naturally interpret tabular data—by analyzing the marginal and conditional relevance of individual features. In doing so, SFR Net provides a step toward architectures that are not only competitive with established baselines but also inherently interpretable and resilient under distributional shifts.

Our contributions can be summarized as follows:

- **Feature-wise specialization with sparse routing.** We introduce a feature-expert decomposition with an entropy-regularized router that performs instance-conditioned, sparse selection, providing a principled inductive bias for heterogeneous tabular data.
- **Competitive accuracy without pre-training.** SFR Net achieves performance competitive with or superior to recent deep learning methods across a range of tabular benchmarks, while avoiding complex self-supervised pre-training and maintaining a simple, efficient architecture.
- **Native interpretability and efficiency.** The router's sparse weights yield direct, explicit per-instance attributions from the model itself, removing the need for post-hoc methods. The architecture scales linearly with feature count, allowing effective training even on CPUs.

2 Related Work

2.1 Sparse Feature-Expert Routing

Conditional computation has long been studied as a way to scale neural models while improving efficiency and interpretability. The Mixture-of-Experts (MoE) paradigm (Jacobs et al., 1991; Shazeer et al., 2017; Lepikhin et al., 2021) introduced routing inputs to one of a set of large, interchangeable subnetworks. Concurrently, developments in differentiable gating, such as `sparsemax` and `entmax` (Martins & Astudillo, 2016; Peters et al., 2019), provided tools to enforce controllable top-k sparsity in model components. Our work leverages these principles for differentiable selection but reframes the target of sparsity. Instead of routing inputs to generic experts, we introduce *feature-expert routing*, where each input feature is assigned its own dedicated expert network. Sparsity is thus applied not to a pool of undifferentiated experts, but directly to the semantic columns of the data itself. In parallel, low-rank interaction modules like factorization machines (Rendle, 2010) and cross networks (Wang et al., 2021) efficiently capture higher-order dependencies, but do so within monolithic backbones that obscure per-feature contributions. Our method departs from these traditions by creating a novel architecture where a router enforces instance-dependent sparsity over feature-specific experts, and a rank-controlled mixer captures their higher-order interactions under an explicit budget. This design directly models feature-level sparsity and low-rank interactions, a combination of properties not jointly addressed by prior approaches.

2.2 Representation Learning for Tabular Data

Within the tabular domain, deep learning models have sought to overcome the challenges of heterogeneous and unstructured feature sets. Transformer-based models such as `TabTransformer` (Huang et al., 2020) and `FT-Transformer` (Gorishniy et al., 2021) rely on dense, all-to-all self-attention, assuming that broad feature interactions are the dominant signal source, which may not be efficient or optimal. At the other end of the spectrum, Generalized Additive Models like `NAMs` (Agarwal et al., 2021) enforce a strict feature-wise decomposition, enhancing interpretability but limiting interaction modeling to purely additive forms. `TabNet` (Arik & Pfister, 2021) represents a critical predecessor, introducing instance-specific feature selection via sequential masking. However, its design employs a multi-step process where the feature selection (via an attentive trans-

former) and the feature transformation are tightly intertwined at each step. In contrast, our approach decouples this process: a single, global routing operation selects a sparse subset of feature-experts which are then processed in parallel before their outputs are composed. This architectural choice avoids sequential dependencies and provides a more direct, interpretable view of which features are selected for a given instance. Other approaches, like `TabCaps` (Chen et al., 2023), import part-whole reasoning at the cost of opaque latent capsules. On the self-supervised front, various pre-training objectives (Yoon et al., 2020; Bahri et al., 2022; Ucar et al., 2021; Thimonier et al., 2024) have been proposed to learn transferable representations, but often require costly pre-training and rely on monolithic encoders that blur per-feature contributions. Our architecture, in contrast, directly encodes a strong, feature-centric inductive bias without this overhead, achieving a unique balance of performance, efficiency, and transparency.

3 METHOD

Our proposed **Sparse Feature Routing Network (SFR Net)** is designed to model tabular data by directly addressing the core challenge of feature heterogeneity through a modular and interpretable architecture, as illustrated in Figure 1. The model comprises three principal components: (1) a set of specialized **expert networks**, one for each input feature; (2) an instance-wise **sparse feature router** that dynamically selects the most relevant experts; and (3) a **low-rank interaction head** that efficiently captures higher-order dependencies among the selected features before making a final prediction.

3.1 FEATURE-WISE EXPERT NETWORKS

To effectively handle the diverse types and distributions inherent in tabular data, SFR Net eschews a monolithic encoder. Instead, for an input instance with F features, $\mathbf{x} = [x_1, x_2, \ldots, x_F]$, each feature x_j is processed by its own dedicated expert network E_j. This "one-expert-per-feature" principle allows the model to learn specialized transformations tailored to the semantics of each column, producing a high-dimensional feature representation $\mathbf{h}_j \in \mathbb{R}^D$.

Numeric Experts For a scalar numerical feature x_j, the corresponding expert E_j^{num} is a small Multi-Layer Perceptron (MLP) that maps the scalar input to the D-dimensional representation space: $\mathbf{h}_j = \text{MLP}_{\text{num}}(x_j)$.

Categorical Experts For a categorical feature x_j with cardinality C_j, the expert E_j^{cat} first projects it into a dense embedding space using an embedding layer Emb_j to obtain a vector $\mathbf{e}_j \in \mathbb{R}^{D_{\text{emb}}}$, which is then transformed by an MLP: $\mathbf{h}_j = \text{MLP}_{\text{cat}}(\text{Emb}_j(x_j))$. For robustness to out-of-distribution data, the embedding layer for each categorical expert reserves a dedicated index to represent unknown categories encountered during inference.

After processing all F features, we obtain a set of expert representations $\{\mathbf{h}_1, \ldots, \mathbf{h}_F\}$, which are then conceptually stacked to form a representation matrix $\mathbf{H} \in \mathbb{R}^{F \times D}$ for the input instance.

3.2 INSTANCE-WISE SPARSE FEATURE ROUTER

Rather than naively combining all feature representations, SFR Net employs a lightweight routing mechanism to perform instance-specific feature selection. The router learns to assign an attention weight α_j to each expert representation $\mathbf{h}_j$, effectively determining the importance of each feature for a given input.

A shared scoring network—a simple MLP with a Tanh activation—computes a scalar score s_j for each feature representation. These scores are subsequently normalized into a probability distribution $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_F]$ over the features using the softmax function:

$$\alpha_j = \frac{\exp(s_j)}{\sum_{k=1}^{F} \exp(s_k)}. \tag{1}$$

To encourage the model to select a small subset of highly informative features, thereby inducing sparsity and improving interpretability, we introduce an entropy regularization term to the train-

ing objective. Minimizing the entropy of the attention distribution, $H(\boldsymbol{\alpha}) = -\sum_{j=1}^{F} \alpha_j \log(\alpha_j)$, encourages $\boldsymbol{\alpha}$ to become "peaky," concentrating its mass on a few features.

3.3 Low-Rank Interaction and Prediction Head

The sparse weights $\boldsymbol{\alpha}$ gate two parallel pathways. While the first captures additive effects, the second is designed to explicitly model higher-order relationships efficiently through a **low-rank interaction module**.

First-Order Representation The first-order representation $\mathbf{r}^{(1)} \in \mathbb{R}^D$ is computed as the attention-weighted sum of the expert outputs, capturing the additive effects of the selected features:

$$\mathbf{r}^{(1)} = \sum_{j=1}^{F} \alpha_j \mathbf{h}_j. \tag{2}$$

Higher-Order Interaction Each expert representation $\mathbf{h}_j$ is projected into two separate low-dimensional "key" and "value" spaces using shared projection matrices $\mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{D \times K}$, where $K \ll D$ is the interaction rank. The interaction representation $\mathbf{r}^{(2)} \in \mathbb{R}^K$ is then computed as the element-wise product of the attention-weighted keys and values:

$$\mathbf{r}^{(2)} = \sum_{j=1}^{F} \alpha_j (\mathbf{k}_j \odot \mathbf{v}_j), \quad \text{where} \quad \mathbf{k}_j = \mathbf{h}_j^T \mathbf{W}_K \text{ and } \mathbf{v}_j = \mathbf{h}_j^T \mathbf{W}_V. \tag{3}$$

This formulation efficiently captures second-order interactions between the routed features under an explicit rank budget K. The final, enriched instance representation $\mathbf{r}_{\text{final}}$ is the concatenation of the first-order and higher-order representations:

$$\mathbf{r}_{\text{final}} = [\mathbf{r}^{(1)}; \mathbf{r}^{(2)}]. \tag{4}$$

This combined vector is passed to a final `Prediction Head` (a standard MLP) that maps $\mathbf{r}_{\text{final}}$ to the output logits for the given task.

3.4 Training Objective

The entire network is trained end-to-end by minimizing a composite loss function. This objective combines the standard task-specific loss (e.g., Binary Cross-Entropy, $\mathcal{L}_{\text{task}}$) with the entropy regularization term from the router, balanced by a hyperparameter λ:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}}(\hat{y}, y) + \lambda H(\boldsymbol{\alpha}). \tag{5}$$

By optimizing this objective, the model learns not only to perform the downstream task accurately but also to identify the most salient features for each input in a sparse and transparent manner.

4 Experiments

We conduct a series of experiments to evaluate the performance of our proposed Sparse Feature Routing Network (SFR Net). Our evaluation is designed to answer two key questions: (1) How does SFR Net's architecture perform against other specialized deep tabular models when trained from scratch? (2) How does SFR Net, trained end-to-end, compare to powerful backbone models that have been enhanced with computationally expensive Self-Supervised Learning (SSL) pre-training?

4.1 Experimental Setup

Datasets Following previous work, we conduct experiments on four public benchmark datasets with heterogeneous features: two for binary classification (**AD**, **JA**) and two for regression (**HE**, **CA**). We use Accuracy (↑) for classification and Root Mean Squared Error (RMSE, ↓) for regression as performance metrics. For all experiments, we report the mean and standard deviation over multiple runs with different random seeds to ensure the robustness and reliability of our findings.

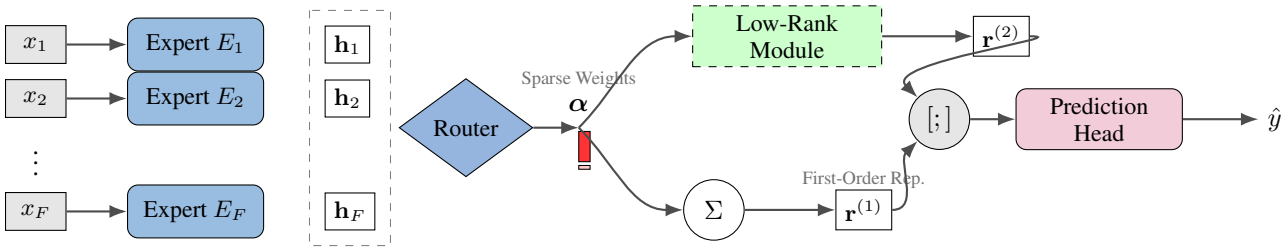

Figure 1: The **Sparse Feature Routing Network (SFR Net)** Architecture. An input instance $\mathbf{x}$ is processed by a set of parallel, feature-wise expert networks (E_j) to produce specialized representations ($\mathbf{h}_j$). A central `Router` then computes instance-specific, sparse attention weights ($\boldsymbol{\alpha}$), dynamically selecting a small subset of the most relevant features. These weights simultaneously gate two parallel pathways: (i) a first-order representation ($\mathbf{r}^{(1)}$) is formed by an attention-weighted sum (Σ), capturing the additive effects of the selected features; and (ii) a higher-order representation ($\mathbf{r}^{(2)}$) is produced by a `Low-Rank Mixer` that efficiently models interactions only among the same selected features. Finally, the two representations are concatenated (the $[\,;\,]$ node) and passed to a `Prediction Head` to produce the output $\hat{y}$.

Baselines To assess the effectiveness of our approach, we compare SFR Net against three strong groups of baselines. (1) **Deep Tabular Architectures:** This group includes a standard `MLP` as a fundamental baseline, as well as specialized SOTA models designed for tabular data: `DCNv2` and `AutoInt` focus on explicit feature interactions, while `FT-Transformer` adapts the powerful attention mechanism for tabular inputs. (2) **Self-Supervised Learning Methods:** To provide a rigorous comparison against representation learning approaches, we include several state-of-the-art SSL methods applied to a powerful `ResNet` backbone, establishing a very strong performance ceiling. (3) **Gradient Boosted Decision Trees (GBDTs):** We include `XGBoost` and `CatBoost` as they are often considered the *de facto* state-of-the-art in practice and represent a crucial point of reference for any new tabular model.

4.2 Results and Discussion

Comparison with Deep Tabular Architectures As displayed in Table 1, SFR Net's performance when trained from scratch establishes it as a state-of-the-art neural architecture for tabular data. **SFR Net is the top-performing neural network across all four datasets**, often by a significant margin. Interestingly, despite their design advantages for tabular data, recent attention-based models such as `FT-Transformer` and `TabNet` did not surpass our proposed architecture. This outcome suggests that SFR Net's architectural inductive bias—combining dedicated feature experts with an explicit, sparse routing mechanism that performs instance-wise feature selection—is a more effective approach for these tabular tasks than the dense, all-to-all attention of Transformers or the sequential attention of TabNet. When compared to the GBDT models for reference, SFR Net significantly narrows the performance gap that often exists between deep learning models and tree-based ensembles, even matching the performance of XGBoost on the HE dataset. This establishes SFR Net as a powerful standalone architecture.

Comparison with Self-Supervised Learning Methods As displayed in Table 2, we position SFR Net against a more challenging set of baselines: powerful ResNet backbones enhanced with various SSL pre-training techniques. The goal is to assess whether SFR Net's built-in architectural priors can compete with the learned representations from SSL. The results are highly compelling. Despite being trained end-to-end without a separate, computationally expensive pre-training stage, **SFR Net remains highly competitive and even achieves the best performance on the AD dataset**, outperforming all SSL-enhanced ResNet models. Furthermore, it ranks as the second-best model on both the JA and CA datasets. This demonstrates that SFR Net provides an effective and significantly more efficient alternative to the pre-train-then-finetune paradigm. The strong performance suggests that encoding explicit structural priors about feature-wise specialization and instance-wise sparsity directly into the model can be as, or more, powerful than learning these priors implicitly through unsupervised pre-training tasks.

Table 1: Comparison of SFR Net with specialized deep tabular models. **Best** and second-best results among neural networks are highlighted. SFR Net demonstrates superior performance over strong baselines like FT-Transformer, which adapts the powerful attention mechanism for tabular data. State-of-the-art GBDT models are included for reference.

Model	AD ↑	HE ↑	JA ↑	CA ↓
Deep Tabular Architectures				
MLP	0.827±1e-3	0.353±1e-3	0.672±1e-3	0.511±3e-3
DCNv2	0.829±4e-3	0.347±3e-3	0.662±3e-3	0.504±4e-3
AutoInt	0.823±1e-3	0.338±3e-3	0.653±6e-3	0.501±3e-3
FT-Trans	0.821±7e-3	0.363±2e-3	0.677±2e-3	0.473±5e-3
SFR Net (ours)	**0.868**±1e-3	**0.375**±2e-3	**0.720**±4e-3	**0.456**±1e-3
Gradient Boosted Decision Trees (for reference)				
XGBoost	0.872±5e-4	0.375±1.2e-3	0.721±1e-3	0.433±2e-3
CatBoost	0.873±1e-3	0.381±1e-3	0.721±1e-3	0.430±7e-4

Table 2: SFR Net compared against strong Self-Supervised Learning (SSL) methods applied to a ResNet backbone. Despite being trained end-to-end without a separate pre-training stage, SFR Net remains highly competitive. It achieves the top rank on the AD dataset and is the second-best performing model on JA and CA, demonstrating that its architecture provides an effective alternative to representation learning via SSL.

Model	AD ↑	HE ↑	JA ↑	CA ↓
Self-Supervised Learning Networks (on ResNet)				
+PTaRL	0.862±5e-3	0.383±2e-3	**0.723**±5e-3	0.498±1e-3
+VIME	0.851±1e-3	0.372±2e-3	0.699±3e-3	0.505±1e-2
+BinRecon	0.828±9e-3	0.327±1e-2	0.699±3e-3	0.471±1e-2
+SubTab	0.823±3e-3	0.365±3e-3	0.702±1e-3	0.487±2e-2
+T-JEPA	0.865±3e-3	**0.401**±2e-3	0.718±3e-3	**0.441**±8e-2
SFR Net (ours)	**0.868**±1e-3	0.375±2e-3	0.720±4e-3	0.456±1e-3

4.3 Ablation Studies

To dissect the architectural drivers of SFR Net's performance, we conduct a crucial ablation study on the Adult dataset. As shown in Table 3, we start with a strong, overparameterized monolithic MLP and incrementally introduce SFR Net's core components. This analysis reveals not only that our final model is nearly 2× more parameter-efficient, but also that it achieves its superior performance in significantly fewer training steps.

Table 3: Quantifying the impact of SFR Net's components on the Adult dataset. Each row builds upon the previous one, with performance gains shown as percent enhancement over the prior step. Error reduction is calculated for AUC (1-AUC) and LogLoss.

Model Variant	#Params	LogLoss Reduction ↓	AUC Error Reduction ↑
(a) Monolithic MLP (Baseline)	27.4k	(Baseline Performance 0.31 LogLoss, 0.908 AUC)	
(b) Decomposed	13.9k	4.4%	5.4%
(c) Decomposed + Routing	13.9k	1.3%	1.0%
(d) **SFR Net (Sparse Routing)**	13.9k	0.3%	1.4%

The Foundational Leap from Monolithic Design. The first and most impactful step is moving from a monolithic MLP (a) to a decomposed, feature-wise expert architecture (b). This foundational shift, even when combined with a simple averaging of expert outputs, is remarkably effective. As shown in Table 3, this change alone slashes the test log-loss by 4.4% and cuts the classification error (1-AUC) by a massive 5.4%—all while using half the parameters of the monolithic baseline. This

result strongly validates our core hypothesis: specializing model components to individual features provides a superior and more efficient inductive bias for tabular data.

The Multiplier Effect of Learned Routing and Rapid Convergence. Building on this strong decomposed base, we replace average pooling with our learned, instance-wise router. Even with dense attention (c), the impact is immediate, further reducing the remaining classification error by another 1.1%. This demonstrates the power of dynamic, context-aware feature selection. Crucially, we also observed that this performance gain is achieved with remarkable efficiency. The decomposed architectures (b, c, and d) consistently reached optimal validation loss in significantly fewer training epochs than the monolithic MLP, which required a longer training schedule and still converged to a suboptimal solution. This highlights that SFR Net's architecture not only performs better but also learns more efficiently, a critical advantage in practical applications.

The Final Polish from Sparsity. Finally, applying our full model with sparsity-inducing regularization (d) provides the final performance gain. While numerically the smallest step, its role is critical: it acts as a powerful regularizer, forcing the model to commit to a minimal subset of high-signal features. Cumulatively, our architectural choices provide a total test log-loss reduction of nearly 6% and a total classification error reduction of 7.8% compared to a overparameterized baseline, demonstrating the value of each component in the SFR Net design.

5 Conclusion

In this work, we introduced the Sparse Feature Routing Network (SFR Net), a novel paradigm for deep tabular learning that challenges the prevailing use of monolithic encoders. SFR Net is built on the principle of feature-wise specialization, employing dedicated experts for each feature that are dynamically selected by an instance-wise, sparse router. Our experiments demonstrated that this architectural prior is highly effective, with SFR Net consistently outperforming specialized deep tabular baselines, including FT-Transformer and TabNet. Furthermore, we showed that SFR Net, trained end-to-end, remains highly competitive with powerful ResNet backbones enhanced by SSL pre-training, suggesting that encoding strong, problem-aligned inductive biases directly into the architecture is a more efficient and equally powerful alternative to representation learning via SSL. While our analysis focused on datasets with a moderate number of features, a promising direction for future work is scaling this sparse, modular approach to high-dimensional data.

References

Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xinyi Zhang, Ben Lengerich, Rich Caruana, and Geoffrey Hinton. Neural additive models: Interpretable machine learning with neural nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 4699–4711, 2021.

Sercan O Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 6679–6687, 2021.

Dara Bahri, Heinrich Jiang, Ishan Gupta, and Luke Metz. SCARF: Self-Supervised Contrastive Learning using Random Feature Corruption. In *International Conference on Learning Representations (ICLR)*, 2022.

Yih-Wenn Chen, Hsin-Yuan Lin, and Yuan-Hao Huang. TABCAPS: A Novel CapsNet for Tabular Data Classification. In *Proceedings of the 2023 International Conference on Signal Processing and Machine Learning*, pp. 1–6, 2023.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 18932–18943, 2021.

Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. TabTransformer: Tabular Data Modeling Using Contextual Embeddings, 2020.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Cui. GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. In *International Conference on Learning Representations (ICLR)*, 2021.

André FT Martins and Ramón F Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International conference on machine learning (ICML)*, pp. 1614–1623. PMLR, 2016.

Ben Peters, Vlad Niculae, and André FT Martins. Sparse sequence-to-sequence models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1504–1519, 2019.

Steffen Rendle. Factorization machines. In *2010 IEEE International Conference on Data Mining (ICDM)*, pp. 995–1000. IEEE, 2010.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

Théo Thimonier, Nika Zabërgja, Andrea Zancristo, Corentin Dubuis, and Philippe Cudré-Mauroux. T-JEPA: A Tabular Joint-Embedding Predictive Architecture for Self-Supervised Learning, 2024. Preprint, referenced in text as 2025 submission.

Tarik Ucar, Yassine Ouali, Lambert T. T. Le, Trung-Hieu Hoang, Davaadorj Battulga, Hae-Yong Kim, Se-Yoon Oh, and E-G-Youn. SubTab: Subsetting Features of Tabular Data for Self-Supervised Representation Learning. In *2021 IEEE International Conference on Big Data (Big Data)*, pp. 1370–1376, 2021.

Ruoxi Wang, Ronna Fu, Jyun-Cheng Sun, and Min Li. Dcn v2: Improved deep & cross network and practical lessons for web-scale learning to rank. In *Proceedings of the 2021 ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3754–3762, 2021.

Jinsung Yoon, Yao Zhang, James Jordon, and Mihaela Van der Schaar. Vime: Extending the success of self-and semi-supervised learning to tabular domain. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 11358–11369, 2020.