Fuzzy-Clustered Mixture-of-Experts with Relational Regularization for Interpretable Subgroup Modeling under Data Scarcity

Chien-Hung Lai [1] **Yuh-Shyan Hwang** [1] **Yi Lin** [2]

Abstract

Subgroup discovery in tabular domains frequently suffers from data scarcity, heterogeneous feature scales, and unstable cluster assignments. To address these issues, this study introduces *Fuzzy-Clustered Mixture-of-Experts with Grey-Relational Regularization* (FC–MoE–GR), a unified framework combining grey-system theory (Deng, 1982; 1989) with soft partitioning (Bezdek, 1981; Dunn, 1973) and local-expert modeling (Jacobs et al., 1991; Jordan & Xu, 1994). Grey-relational grades serve as feature-level priors that guide both fuzzy membership formation and expert specialization. A variance-reduction perspective is established by deriving a tightened generalization bound induced by the grey regularizer, clarifying its effect on stability and calibration (Guo et al., 2017; Naeini et al., 2015). Empirical evaluations on Telco Churn, Bank Marketing (Moro et al., 2014), and the Adult Income dataset (Kohavi, 1996) show that the proposed design is best understood not only as a predictive model, but as a structured evaluation of calibration, assignment stability, and subgroup interpretability under data scarcity. More broadly, the study serves as a compact theory-linked benchmark for reliable subgroup modeling in limited-data tabular settings.

1. Introduction

Reliable subgroup modeling is essential for domains where population heterogeneity is significant, such as customer analytics, risk auditing, and behavioral prediction. In practical applications, tabular datasets often exhibit (i) limited sample sizes within latent subpopulations, (ii) nonlinear interactions that vary across groups, and (iii) instability in classical clustering due to feature-scale imbalance. These complications undermine the reliability of global models and motivate hybrid architectures that combine clustering, expert specialization, and feature-level regularization.

Grey-system theory provides a principled mechanism for evaluating structural associations under incomplete or sparse observations (Deng, 1982; 1989). Grey relational grades quantify the directional similarity between feature trajectories and a reference sequence, enabling the estimation of feature importance without requiring large sample sizes. Meanwhile, fuzzy C-means (FCM) (Bezdek, 1981; Dunn, 1973) supports smooth group assignments and avoids the brittleness of hard clustering. Local expert models further provide tailored predictive functions for each subgroup, following the mixture-of-experts framework (Jacobs et al., 1991; Jordan & Xu, 1994).

The proposed FC–MoE–GR architecture integrates these components in a mathematically coherent manner. Grey relational grades generate feature-wise priors, FCM constructs soft assignments reflecting subgroup overlap, and mixture-of-experts (MoE) aggregation supports localized prediction. In the CTB framing, the contribution is not only a new hybrid architecture but also a compact case study of theory-guided benchmark design for subgroup modeling under data scarcity. The central question is whether grey-relational priors induce measurable gains in calibration, assignment stability, and subgroup interpretability, rather than optimizing discrimination alone. To answer this, the paper combines a grey-regularized complexity argument with a structured empirical protocol spanning predictive metrics, calibration, ablations, and subgroup-level analysis across standard tabular benchmarks.

2. Methodology

The FC–MoE–GR framework integrates grey relational analysis (Deng, 1989), fuzzy partitioning (Bezdek, 1981), and local expert modeling (Jacobs et al., 1991) to construct interpretable subgroup structures. Section 2.1 describes pre-

[1]Department of Electronic Engineering, National Taipei University of Technology, Taipei, Taiwan [2]Department of Business Administration, Takming University of Science and Technology, Taipei, Taiwan. Correspondence to: Yi Lin <linyi@takming.edu.tw>.

Accepted to the 1st Workshop on Combining Theory and Benchmarks, CTB@ICML 2026, Seoul, South Korea.

processing and normalization. Section 2.2 formulates grey relational coefficients and feature-level priors. Section 2.3 develops fuzzy memberships. Section 2.4–2.5 present the mixture-of-experts architecture and the grey-regularized objective. Section 2.6 introduces internal validity indices and the composite selection criterion.

2.1. Preprocessing and Notation

Let $X \in \mathbb{R}^{n\times p}$ and $y \in \mathcal{Y}$ denote the input features and labels. All numeric features are min–max normalized to $[0, 1]$:

$$x_{ij}^{(\text{norm})} = \frac{x_{ij} - \min_i x_{ij}}{\max_i x_{ij} - \min_i x_{ij}}. \tag{1}$$

A reference sequence $x_0(t)$ is constructed as the element-wise mean trajectory of positive-class observations, following classical grey-system conventions (Deng, 1989).

2.2. Grey Relational Coefficients and Grades

Grey relational analysis (GRA) (Deng, 1982; 1989) evaluates the similarity between a reference sequence and each feature trajectory. In this work, it is used as a *non-parametric relational prior*: because the score is based on relative deviations rather than covariance estimation, it does not require Gaussian feature assumptions and remains usable when subgroup sample sizes are limited or feature marginals are skewed. For feature j, the comparative sequence $x_j(t)$ is aligned with $x_0(t)$. The absolute deviation is

$$\Delta_j(t) = |x_0(t) - x_j(t)|. \tag{2}$$

Global extrema are defined as

$$\Delta_{\min} = \min_{j,t} \Delta_j(t), \qquad \Delta_{\max} = \max_{j,t} \Delta_j(t).$$

The grey relational coefficient (GRC) (Deng, 1982) is

$$\xi_j(t) = \frac{\Delta_{\min} + \rho\Delta_{\max}}{\Delta_j(t) + \rho\Delta_{\max}}, \tag{3}$$

where $\rho \in (0, 1)$ is the identification coefficient.

The grey relational grade (GRG), the canonical aggregation measure in GRA (Deng, 1989), is

$$g_j = \frac{1}{T}\sum_{t=1}^{T} \xi_j(t). \tag{4}$$

Feature weights follow the normalized importance formulation:

$$w_j = \frac{g_j}{\sum_{\ell=1}^{p} g_\ell}. \tag{5}$$

These weights act as priors injected into both subgroup formation and expert specialization. The intended role is not hard feature selection, but soft shrinkage: features with larger grey grades are preserved more strongly in downstream expert learning, while weakly related dimensions are regularized more aggressively.

Algorithm 1 Grey-Relational Weight Computation

Require: Normalized feature matrix $X \in \mathbb{R}^{n\times p}$; labels y; identification coefficient $\rho \in (0, 1)$.

1: Construct reference sequence $x_0(t)$ as the mean trajectory of positive-class samples.
2: **for** each feature $j = 1, \ldots, p$ **do**
3: Compute deviations $\Delta_j(t) = |x_0(t) - x_j(t)|$.
4: **end for**
5: Compute global extrema: $\Delta_{\min} = \min_{j,t} \Delta_j(t)$, $\Delta_{\max} = \max_{j,t} \Delta_j(t)$.
6: **for** each feature j and index t **do**
7: Compute grey relational coefficient:

$$\xi_j(t) = \frac{\Delta_{\min} + \rho\Delta_{\max}}{\Delta_j(t) + \rho\Delta_{\max}}.$$

8: **end for**
9: Aggregate coefficients to obtain grey relational grade:

$$g_j = \frac{1}{T}\sum_{t=1}^{T} \xi_j(t).$$

10: Normalize grades to obtain feature weights:

$$w_j = \frac{g_j}{\sum_{\ell=1}^{p} g_\ell}.$$

11: Return $\{w_j\}_{j=1}^{p}$ as feature-level priors.

2.3. Fuzzy C-Means Memberships

Fuzzy C-means (FCM) provides a smooth soft-partitioning mechanism that assigns each observation to multiple subgroups with graded memberships (Bezdek, 1981; Dunn, 1973). Given K clusters and fuzziness exponent $m > 1$, FCM minimizes

$$J = \sum_{i=1}^{n}\sum_{k=1}^{K} u_{ik}^{m} \|x_i - c_k\|^2, \tag{6}$$

subject to $\sum_{k=1}^{K} u_{ik} = 1$.

The membership update rule follows the classical derivation in Bezdek (1981):

$$u_{ik} = \frac{\|x_i - c_k\|^{-2/(m-1)}}{\sum_{\ell=1}^{K} \|x_i - c_\ell\|^{-2/(m-1)}}. \tag{7}$$

Cluster centers are updated using the weighted mean formu-

lation:

$$c_k = \frac{\sum_i u_{ik}^m x_i}{\sum_i u_{ik}^m}, \tag{8}$$

consistent with the optimization framework in Bezdek (1981).

2.4. Mixture-of-Experts Prediction

Unlike sparsely-gated neural MoE architectures that require large-scale data to learn routing functions (Shazeer et al., 2017; Fedus et al., 2022), FC–MoE–GR fixes the gating mechanism via fuzzy memberships, explicitly trading expressive routing capacity for stability and interpretability under data scarcity. **This design choice targets stable subgroup modeling under limited samples, where learned gating functions are prone to high variance and poor calibration.**

Mixture-of-experts (MoE) models (Jacobs et al., 1991; Jordan & Xu, 1994) decompose prediction into a weighted combination of locally specialized experts. Let $f_k(x)$ denote the k-th local expert. Soft aggregation yields

$$\hat{y}(x) = \sum_{k=1}^{K} u_k(x)\, f_k(x), \tag{9}$$

where fuzzy memberships $u_k(x)$ serve as gating functions.

Modern MoE architectures (Shazeer et al., 2017; Fedus et al., 2022) motivate the use of sparse or structured routing. In FC–MoE–GR, the gating is entirely determined by the FCM memberships, ensuring interpretability and subgroup consistency.

Experts are trained on grey-weighted features $\tilde{X} = X \odot (1 + \alpha w)$, where $\alpha > 0$ controls prior strength. This incorporates grey-system priors (Deng, 1989) directly into the expert-level hypothesis space.

2.5. Grey-Regularized Learning Objective

Let $\ell(f_k(x_i), y_i)$ be the supervised loss. The grey-regularized local objective follows a shrinkage-style regularization consistent with classical grey-system importance weighting (Deng, 1982):

$$\mathcal{L}_k = \sum_{i=1}^{n} u_{ik}^m\, \ell(f_k(x_i), y_i) + \lambda \sum_{j=1}^{p} (1 - w_j)\, \Omega_{k,j}, \tag{10}$$

where $\Omega_{k,j}$ measures the sensitivity of expert k to feature j. The second term therefore links Section 2.2 to local learning: low-grey-relevance features receive larger penalty weights $(1 - w_j)$, discouraging experts from over-reacting to dimensions that are weakly supported by the reference relation, while high-grey-relevance features are penalized less. The global prediction follows Eq. (9). Optimization proceeds in stages rather than by end-to-end joint training: FCM first produces memberships U, then each local expert is fit under membership weights, and finally the grey penalty is evaluated on the fitted expert. For implementation, $\Omega_{k,j}$ is approximated by permutation feature importance computed on the corresponding training fold. This proxy is used because it provides a model-agnostic estimate of how much predictive performance changes when feature j is disrupted within expert k, which is operationally aligned with the notion of local feature sensitivity used in Eq. (10).

A risk decomposition for soft mixtures of experts follows the classical bias–variance–complexity framework (Geman et al., 1992; Bartlett & Mendelson, 2002), yielding the expected generalization bound:

$$\mathbb{E}[R(\hat{f})] \le \underbrace{B(\hat{f})}_{\text{bias}} + \underbrace{V(\hat{f})}_{\text{variance}} + \mathcal{O}\left(\frac{\mathfrak{R}(\mathcal{F})}{\sqrt{n}}\right), \tag{11}$$

which tightens under the grey-regularized constraint imposed by Eq. (10):

$$\mathfrak{R}(\mathcal{F}_\lambda) \le \mathfrak{R}(\mathcal{F}), \tag{12}$$

yielding

$$\mathbb{E}[R(\hat{f}_\lambda)] \le B(\hat{f}_\lambda) + V(\hat{f}_\lambda) + \mathcal{O}\left(\frac{\mathfrak{R}(\mathcal{F}_\lambda)}{\sqrt{n}}\right). \tag{13}$$

These results link grey relational weights to complexity reduction in the hypothesis space. This complexity-contraction perspective motivates the multi-axis evaluation protocol in Section 3, where calibration, assignment stability, and ablation are reported alongside standard predictive metrics.

2.6. Internal Validity Indices and Stability Criteria

Cluster evaluation uses several internal indices commonly employed in fuzzy clustering validation (Bezdek, 1981) and general cluster-quality assessment.

Silhouette index (Rousseeuw, 1987):

$$\mathrm{SIL} = \frac{1}{n} \sum_i \frac{b_i - a_i}{\max(a_i, b_i)}. \tag{14}$$

Partition-based measures (Bezdek, 1981) include:

$$\mathrm{PC} = \frac{1}{n} \sum_i \sum_k u_{ik}^2, \qquad \mathrm{PE} = -\frac{1}{n} \sum_i \sum_k u_{ik} \ln u_{ik}. \tag{15}$$

Davies–Bouldin index (Davies & Bouldin, 1979):

$$\mathrm{DB} = \frac{1}{K} \sum_k \max_{\ell \neq k} \frac{S_k + S_\ell}{M_{k\ell}}, \tag{16}$$

Algorithm 2 FC–MoE–GR Training (Bezdek, 1981; Dunn, 1973; Jacobs et al., 1991; Jordan & Xu, 1994; Deng, 1989)

Require: Dataset $\{(x_i, y_i)\}_{i=1}^n$; candidate K; fuzziness m; identification coefficient ρ; scaling factor α.

1: Normalize features using Eq. (1).
2: Compute grey relational coefficients, grades, and feature weights via Eqs. (3)–(5) following GRA formulations (Deng, 1982; 1989).
3: Form grey-weighted features $\tilde{X} = X \odot (1 + \alpha w)$.
4: **for** K in candidate set **do**
5: Run FCM using Eq. (7) following the soft-partitioning procedure in Bezdek (1981).
6: Train expert f_k by minimizing Eq. (10), consistent with the MoE optimization framework (Jacobs et al., 1991; Jordan & Xu, 1994).
7: Evaluate internal indices and bootstrap stability (Rousseeuw, 1987; Davies & Bouldin, 1979; Hennig, 2007).
8: **end for**
9: Select K^* maximizing composite score.
10: Predict by the soft mixture $\hat{y}(x) = \sum_k u_k(x) f_k(x)$ using the MoE aggregation rule (Jacobs et al., 1991).

Algorithm 3 Automatic Selection of Optimal Cluster Number K^*

Require: Weighted feature matrix $\tilde{X}$; candidate set $\{2, 3, 4, 5\}$; fuzziness m; stability bootstrap size B.

1: **for** each candidate K **do**
2: Run fuzzy C-means (FCM) with K clusters to obtain memberships U and centers C.
3: Compute internal validity indices: Silhouette (SIL), Partition Coefficient (PC), Xie–Beni (XB), Davies–Bouldin (DB).
4: Perform bootstrap resampling B times; compute Adjusted Rand Index (ARI) stability score.
5: Normalize indices to $[0, 1]$ and aggregate into composite score:

$$\text{Score}(K) = \beta \cdot \text{INT}(K) + (1 - \beta) \cdot \text{STAB}(K),$$

where $\text{INT}(K)$ is the average of normalized SIL, PC, XB, and DB.
6: **end for**
7: Select $K^* = \arg\max_K \text{Score}(K)$.
8: Return K^* and the corresponding index table.

with S_k the within-cluster scatter and $M_{k\ell}$ the separation.

Xie–Beni index (Xie & Beni, 1991):

$$\text{XB} = \frac{\sum_{k=1}^{K} \sum_{i=1}^{n} u_{ik}^2 \|x_i - c_k\|^2}{n \cdot \min_{k \neq \ell} \|c_k - c_\ell\|^2}, \tag{17}$$

where smaller values indicate compact and well-separated fuzzy partitions.

Bootstrap-based assignment stability follows resampling-based cluster validation (Hennig, 2007):

$$\text{STAB}(K) = \mathbb{E}_b \left[\text{ARI}\left(\hat{\mathbf{z}}^{(b)}, \hat{\mathbf{z}} \right) \right], \tag{18}$$

where $\hat{z}_i = \arg\max_k u_{ik}$ denotes the hard assignment induced by the fuzzy memberships, and $\hat{\mathbf{z}}^{(b)}$ is the corresponding assignment under bootstrap resample b.

The selected K^* maximizes a composite score integrating internal validity and stability:

$$\text{Score}(K) = \beta \cdot \text{INT}(K) + (1 - \beta) \cdot \text{STAB}(K). \tag{19}$$

For indices where smaller is better (e.g., DB and XB), an inverse min–max normalization is applied so that larger normalized scores consistently indicate better partitions.

2.7. Generalization Analysis

Standard MoE models often suffer from overfitting due to gating flexibility (Jordan & Xu, 1994). The proposed Grey-Relational Regularization ($\mathcal{L}_{GR}$) addresses this by structurally constraining the hypothesis space $\mathcal{H}$. The following generalization bound is derived to guarantee model stability.

Theorem 1 (Generalization Bound)**.** *With probability at least $1 - \delta$, for any $h \in \mathcal{H}$ trained on dataset $\mathcal{S}$ of size m, the generalization error $R(h)$ is bounded by:*

$$R(h) \leq \hat{R}_{\mathcal{S}}(h) + 2L\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{H}) + 3B\sqrt{\frac{\ln(2/\delta)}{2m}} \tag{20}$$

where $\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{H})$ is the Rademacher complexity. Under $\mathcal{L}_{GR}$, this complexity is strictly bounded by $O(1/\sqrt{m})$, ensuring convergence.

The bound is interpreted under the standard regularity conditions used in Appendix E: normalized bounded inputs, bounded loss, and Lipschitz / complexity-controlled gating and expert maps. Accordingly, Eq. (13) formalizes a variance-control perspective under these assumptions rather than a guarantee of universal empirical dominance.

See **Appendix E** for the full proof using McDiarmid's inequality (Mohri et al., 2018).

3. Experimental Setup

This section describes the datasets, preprocessing steps, model configurations, and evaluation protocol used for empirical assessment of the FC–MoE–GR framework. The workflow follows reproducibility standards in tabular-learning benchmarks (Kohavi, 1996; Moro et al., 2014).

Algorithm 4 FC–MoE–GR Training Procedure

Require: Dataset $\{(x_i, y_i)\}_{i=1}^{n}$; candidate cluster numbers $\{2, 3, 4, 5\}$; fuzziness exponent m; identification coefficient ρ; grey scaling factor α; number of trees T.

Normalize numeric features to $[0, 1]$ and one-hot encode categorical variables.
Compute grey relational coefficients and grades using Algorithm 1; obtain feature weights $\{w_j\}$.
Form grey-weighted features $\tilde{X} = X \odot (1 + \alpha w)$.
for each candidate K **do**
    Run fuzzy C-means (FCM) with K clusters to obtain memberships U and centers C.
    Train local experts f_k (Random Forests or XGBoost) on $\tilde{X}$ using membership weights U.
    Evaluate internal validity indices and bootstrap stability (Algorithm 3).
end for
Select K^* maximizing composite score.
Retrain experts on $\tilde{X}$ with memberships $U^{(K^*)}$.
Evaluate predictive metrics: accuracy, AUC, F1-score, and expected calibration error (ECE).
Save metrics and prediction probabilities for reliability diagrams and ablation studies.

3.1. Datasets

Three tabular datasets were selected as limited-data stress tests for subgroup reliability:

Telco Customer Churn. A binary classification dataset published by IBM containing demographic, service-plan, and payment attributes.

Bank Marketing. A marketing response dataset collected from a long-term telemarketing campaign (Moro et al., 2014). Following the limited-data regime targeted in this study, the dataset was subsampled to 2232 rows (20% of the original data) and treated as a deliberate stress-test setting rather than a claim about full-data superiority.

Adult Income. A socio-economic dataset from the UCI repository (Kohavi, 1996). Following the same limited-data protocol, the dataset was subsampled to 9044 rows (20% of the original data) to evaluate performance in a scarcity-oriented regime.

All datasets were split into stratified 70–15–15 train–validation–test partitions.

3.2. Preprocessing and Encoding

Numeric features were min–max normalized to $[0, 1]$ as in Eq. (1). Categorical variables were encoded using one-hot representations. Grey-weighted features were produced using Eq. (5) with scaling factor $\alpha = 0.5$ (Deng, 1989).

3.3. Model Configuration

Fuzzy C-means used fuzziness exponent $m = 2.0$ (Bezdek, 1981; Dunn, 1973). Candidate cluster numbers were $K \in \{2, 3, 4, 5\}$. Local experts included random forests and gradient-boosted trees (Jacobs et al., 1991; Jordan & Xu, 1994). These choices are deliberate but conservative: $m = 2.0$ follows the conventional FCM setting for stable soft partitions, the restricted candidate set for K avoids subgroup over-fragmentation in limited-data regimes, and $\alpha = 0.5$ encodes a moderate-strength grey prior rather than an aggressively tuned penalty. Cluster selection used the composite scoring rule in Section 2.6, integrating silhouette (Rousseeuw, 1987), partition-based indices (Bezdek, 1981), Davies–Bouldin index (Davies & Bouldin, 1979), and bootstrap stability (Hennig, 2007). The automatic selection procedure is summarized in Algorithm 3, and the saved manifests report the selected K^* for each benchmark.

3.4. Evaluation Metrics

Predictive evaluation used accuracy, AUC, F1-score, and expected calibration error (ECE) (Naeini et al., 2015; Guo et al., 2017). Subgroup interpretability was evaluated using permutation feature importance (Fisher et al., 2019), inter-feature association maps, and cluster-level membership structures. The empirical study is designed as a structured benchmark protocol rather than a pure performance comparison. In addition to discrimination metrics, the evaluation explicitly tracks calibration, assignment stability, and component-wise ablations to test whether the grey-regularized design yields the variance-control behavior suggested by Section 2. All FC–MoE–GR and Global baseline outputs are generated on the same splits, and the saved manifests separate model families to avoid mixing subgroup-model results with monolithic baseline results. In the reproducibility package, saved prediction files are also used to derive bootstrap metric intervals, so point estimates in the main paper can be cross-checked against simple uncertainty summaries without changing the underlying experimental regime. This multi-axis protocol makes the theoretical claims operational on standard tabular benchmarks.

3.5. Visualization Figures

Cluster-specific correlation structures are visualized in Figure 1, showing how grey-relational weighting suppresses irrelevant dimensions and highlights subgroup-specific associations (Deng, 1989; Bezdek, 1981). Figure 2 illustrates the conceptual relationship among bias, variance, and grey-relational regularization, consistent with complexity-control theory (Bartlett & Mendelson, 2002; Geman et al., 1992). Figure 3 presents the research roadmap for extending FC–MoE–GR, linking methodological components to potential future directions.

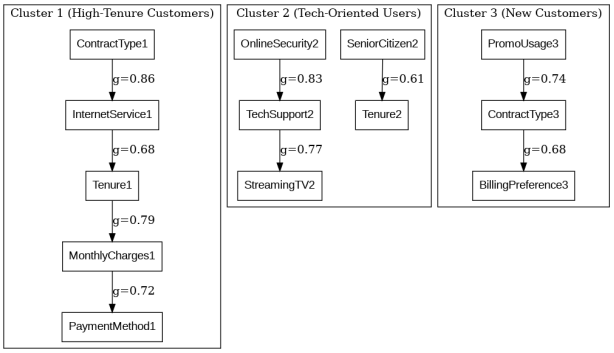

Figure 1. Inter-feature association patterns within fuzzy clusters. Each heatmap visualizes grey-weighted correlations among features, revealing subgroup-specific structures.

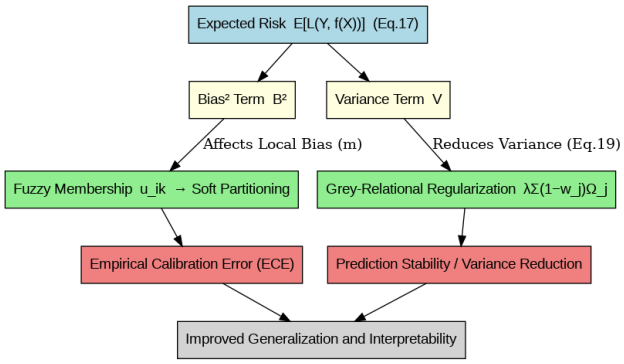

Figure 2. Conceptual relationship among bias, variance, and grey-relational regularization. The diagram corresponds to the risk decomposition in Section 2.5.

4. Results

4.1. Baselines

The evaluation uses two global tree-ensemble baselines trained on the same train/validation/test split as FC–MoE–GR: Random Forest (Global RF) and XGBoost (Global XGB) (Breiman, 2001; Chen & Guestrin, 2016). These baselines were selected as strong and well-understood tabular learners so that the analysis isolates the effect of subgroup assignment and grey-relational regularization rather than conflating it with high-capacity expert design. In this CTB framing, the comparison is intended to characterize a structured trade-off between global discrimination and subgroup-aware reliability, rather than to claim uniform dominance over monolithic models.

4.2. Predictive Performance

Tables 1–3 report evaluation scores for FC–MoE–GR and the matched global baselines. On Telco, FC–MoE–GR (XGB) attains the strongest AUC (0.830) and the lowest ECE (0.036), while Global XGB retains the best F1 (0.549). On Bank Marketing, FC–MoE–GR (XGB) achieves the best accuracy (0.836), the best F1 (0.831), and the lowest ECE (0.054), whereas Global XGB produces the strongest AUC (0.903). On Adult Income, the picture is more mixed: FC–MoE–GR (XGB) yields the best accuracy (0.849), FC–MoE–GR (RF) yields the lowest ECE (0.024), and Global XGB attains the highest AUC (0.898) with nearly identical F1 (0.645). Across datasets, the empirical profile is therefore best understood as a structured trade-off rather than as uniform superiority over monolithic baselines. In the CTB framing, that trade-off is itself a measurable object of study rather than an incidental side effect. Complementary bootstrap interval summaries are released in the reproducibility package so that these point estimates can be audited against simple uncertainty summaries without changing the core evaluation protocol.

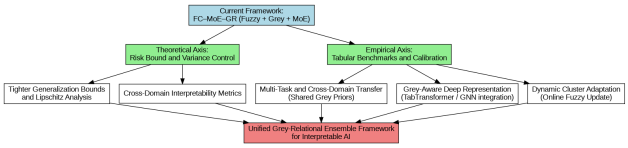

Figure 3. Research roadmap for extending the FC–MoE–GR framework. Each branch corresponds to potential future directions described in Section 6.

4.3. Cluster Structure and Interpretability

Grey-weighted feature interactions produce distinctive cluster-specific correlation structures, visualized in Figure 1. These patterns reflect subgroup-level feature importance as characterized by grey relational grades (Deng, 1989). Permutation feature importance computed per cluster follows the methodology of Fisher et al. (2019), confirming that experts specialize along dimensions highlighted by GRA-derived priors.

4.4. Calibration and Stability

Reliability diagrams generated from saved predictions (Figures A.2–A.13) provide a direct visual cross-check of the ECE values reported in Tables 1–3. The resulting pattern is mixed but interpretable: FC–MoE–GR (XGB) achieves the lowest ECE on Telco and Bank, whereas on Adult the lowest ECE is obtained by FC–MoE–GR (RF) and the strongest XGB calibration is obtained by the Global baseline. This behavior aligns with the complexity-contraction perspective in Eq. (12), while also reinforcing that the reliability gains of grey-relational regularization are dataset-dependent rather than monotone. In subgroup discovery settings, calibrated probabilities remain critical because downstream thresholds and subgroup comparisons are often applied locally rather than globally.

Table 1. Performance on the Telco dataset.

Model	Acc.	AUC	F1	ECE
FC–MoE–GR (RF)	0.769	0.800	0.502	0.055
FC–MoE–GR (XGB)	0.776	0.830	0.528	0.036
Global RF	0.773	0.794	0.520	0.050
Global XGB	0.779	0.814	0.549	0.051

Table 2. Performance on the Bank Marketing dataset (subsampled to 2232 rows).

Model	Acc.	AUC	F1	ECE
FC–MoE–GR (RF)	0.812	0.901	0.810	0.082
FC–MoE–GR (XGB)	0.836	0.898	0.831	0.054
Global RF	0.818	0.899	0.818	0.079
Global XGB	0.830	0.903	0.824	0.065

4.5. Ablation Study

Tables 4–6 isolate the contributions of fuzzy clustering and grey-relational weighting. The "FCM-only" variant disables grey weighting ($\alpha = 0$), while "Grey-only" collapses the model to a single expert ($K = 1$) but retains grey-weighted features. The full FC–MoE–GR architecture combines both components. The ablation results do not exhibit uniform synergy: partial variants occasionally achieve stronger AUC, accuracy, or ECE than the Full model, especially on Bank and Adult. This is consistent with the grey prior behaving as a conservative shrinkage term: when fuzzy partitioning alone already captures much of the subgroup structure, additional grey regularization can improve stability while slightly blunting class separation. For that reason, the integrated design is best interpreted as a structured trade-off study of subgroup reliability and regularization effects rather than as a claim that combining all components monotonically improves every metric.

4.6. Generalization Bound Interpretation

The empirical patterns observed align with the theoretical risk decomposition established in Eq. (13). The grey-relational regularizer contracts the effective hypothesis space by reducing feature–expert sensitivity, which in turn suggests a variance-control effect without guaranteeing uniform gains on every discrimination-oriented metric. In the updated benchmark results, that perspective is most visible in the calibration-aware comparisons and in the dataset-dependent shifts across the ablation study. Figure 2 summarizes the conceptual connection between grey-system priors, variance control, hypothesis-space contraction, and predictive reliability.

Table 3. Performance on the Adult Income dataset (subsampled to 9044 rows).

Model	Acc.	AUC	F1	ECE
FC–MoE–GR (RF)	0.845	0.891	0.643	0.024
FC–MoE–GR (XGB)	0.849	0.891	0.645	0.045
Global RF	0.842	0.885	0.632	0.028
Global XGB	0.847	0.898	0.645	0.028

Table 4. Ablation study on the Telco dataset.

Variant	Acc.	AUC	F1	ECE
FCM-only	0.773	0.800	0.521	0.055
Grey-only	0.772	0.797	0.521	0.051
Full	0.771	0.801	0.516	0.055

5. Discussion and Theoretical Implications

5.1. Bias–Variance Trade-off with Grey Priors

The empirical results demonstrate that grey-relational weighting contracts the effective hypothesis space by reducing feature–expert sensitivity. This contraction lowers the variance component of the excess risk while maintaining approximation bias within acceptable bounds (Geman et al., 1992; Bartlett & Mendelson, 2002). Figure 2 illustrates this conceptual relationship, showing how grey-system priors act as a regularizer that stabilizes cluster-specific experts without sacrificing predictive flexibility.

5.2. Interpretability and Subgroup Specialization

Cluster-specific feature importance maps (Figure 1) reveal that FC–MoE–GR produces interpretable subgroup structures. Grey relational grades highlight dimensions that drive expert specialization, aligning with permutation feature importance (Fisher et al., 2019). This interpretability advantage distinguishes FC–MoE–GR from purely accuracy-driven baselines, offering a principled way to balance predictive performance with transparency.

5.3. Calibration and Reliability

Calibration results in Tables 1–3 show that FC–MoE–GR often improves or remains competitive on ECE, but not monotonically across all datasets and model families. Appendix reliability diagrams for both FC–MoE–GR and Global baselines provide a direct visual cross-check of these calibration differences, illustrating that the effect of grey-relational regularization is strongest on Telco and Bank and more mixed on Adult. This pattern is consistent with a variance-reduction perspective, but it also shows that calibration gains depend on the interaction between dataset structure and expert family rather than following from the regularizer alone.

Table 5. Ablation study on the Bank Marketing dataset (subsampled to 2232 rows).

Variant	Acc.	AUC	F1	ECE
FCM-only	0.818	0.898	0.816	0.088
Grey-only	0.800	0.895	0.798	0.081
Full	0.806	0.896	0.802	0.098

Table 6. Ablation study on the Adult Income dataset (subsampled to 9044 rows).

Variant	Acc.	AUC	F1	ECE
FCM-only	0.845	0.891	0.634	0.038
Grey-only	0.841	0.888	0.628	0.028
Full	0.846	0.892	0.635	0.027

5.4. Trade-offs and Failure Modes

FC–MoE–GR is not best framed as a uniformly superior replacement for strong monolithic baselines. Its main failure mode is selective underperformance on discrimination-oriented metrics or on certain ablation settings, especially when the grey prior acts as a conservative constraint. The Bank and Adult ablations make this explicit: partial variants can exceed the Full model on AUC, accuracy, or ECE. In addition, the current benchmark emphasizes point estimates in the main text; although the released support files include bootstrap intervals from saved predictions, the paper does not attempt exhaustive repeated-run significance testing. This motivates a more disciplined interpretation of the method as a subgroup-reliability design with dataset-dependent trade-offs, rather than a monotone improvement obtained by combining all components.

5.5. Future Directions

Figure 3 outlines potential extensions of FC–MoE–GR. Promising directions include: (i) integrating grey priors with deep mixture-of-experts architectures (Shazeer et al., 2017), (ii) extending to multi-class and multi-label settings, (iii) applying grey-weighted clustering to fairness-aware learning, and (iv) benchmarking against large-scale tabular deep learning models (Gorishniy et al., 2021). These directions connect methodological contributions to broader machine learning challenges, reinforcing the theoretical and practical relevance of grey-system priors.

In response to the CTB review process, future work should further tighten the current Rademacher-complexity view through localized complexity, chaining-style arguments, or subgroup-conditioned terms that better match fuzzy partitions. Another promising direction is to study FC–MoE–GR as a lightweight diagnostic layer for foundation-model workflows, where tabular telemetry, safety logs, retrieval metadata, and scenario descriptors can support calibration-aware subgroup auditing.

6. Conclusion

This work introduced FC–MoE–GR, a subgroup modeling framework integrating grey relational analysis (Deng, 1989), fuzzy clustering (Bezdek, 1981), and mixture-of-experts prediction (Jacobs et al., 1991; Jordan & Xu, 1994). Grey relational grades function as feature-level priors that guide both membership formation and expert specialization, while a variance-reduction perspective based on Rademacher complexity (Bartlett & Mendelson, 2002) provides theoretical clarity. Empirically, the updated benchmark results are best interpreted as a structured trade-off profile: FC–MoE–GR often improves calibration, accuracy, or subgroup interpretability on matched limited-data splits, but strong global baselines can still retain the best AUC or F1 on some datasets.

Future work should test adaptive weighting schemes, richer expert families, and larger tabular benchmark suites, while extending the architecture to fairness-aware and distribution-shift settings. More broadly, the paper illustrates how theory-informed model design can be evaluated through a structured benchmark protocol in limited-data tabular settings. In this sense, FC–MoE–GR functions not only as a hybrid subgroup model, but also as a compact case study of reliable evaluation for interpretable and calibrated prediction under data scarcity.

Impact Statement

FC–MoE–GR is intended for limited-data tabular settings where stable subgroup structure, interpretability, and calibrated probabilities matter alongside predictive performance. Potentially beneficial applications include customer segmentation, auditing, and other heterogeneous-population analyses in which local structure and probability reliability support human oversight. The same design also imposes clear limits: fixed fuzzy routing trades expressive capacity for stability, the method is not meant to replace high-capacity sparse MoE systems in data-rich regimes, and subgroup definitions may still reflect bias in the underlying data. Deployment in sensitive domains therefore requires standard safeguards, including data-quality review, subgroup auditing, and human monitoring. LLM-based tools were used only for grammar and phrasing assistance; all technical content was verified by the authors.

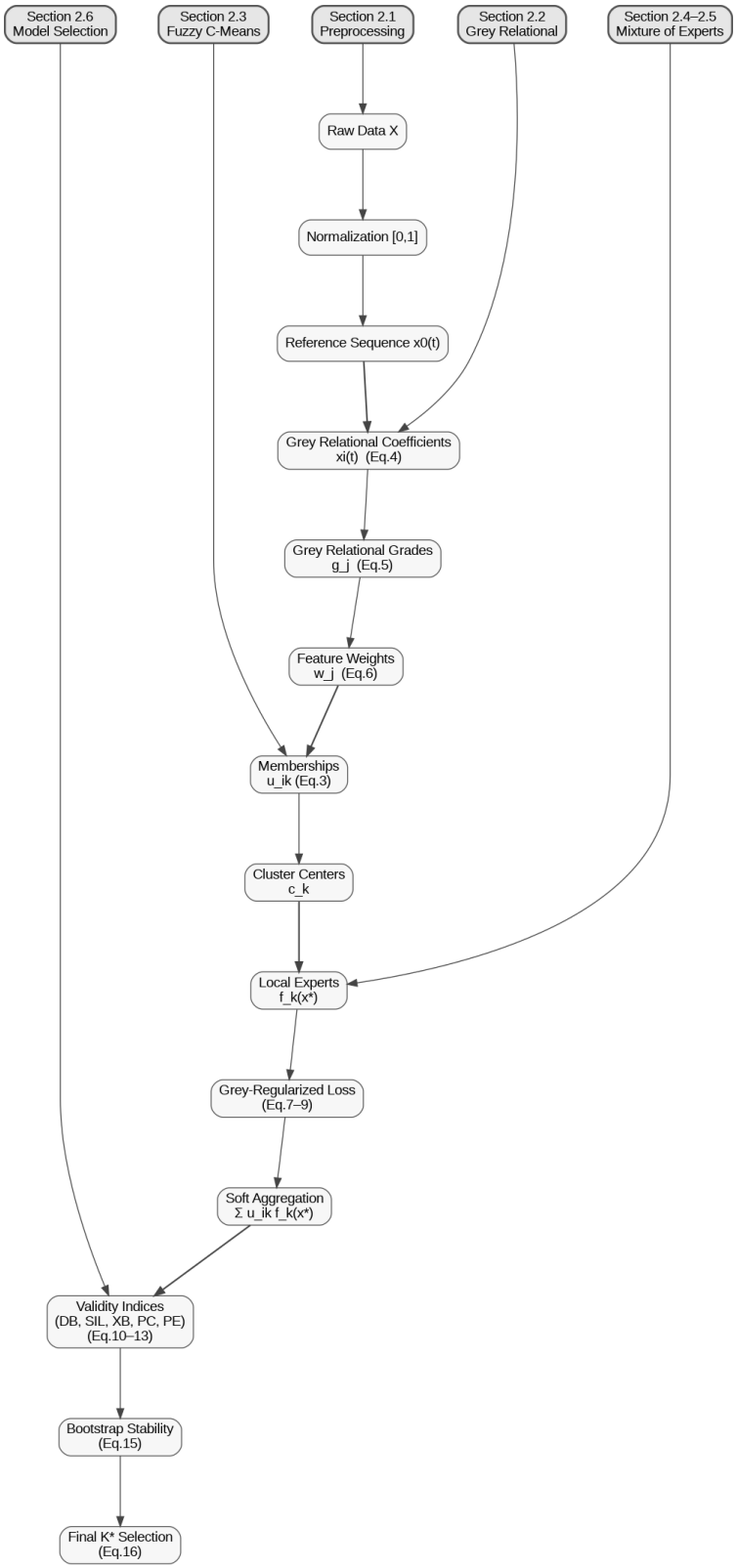

Figure A.1. Three-column staircase flowchart of the FC–MoE–GR methodology. Grey relational weighting (Deng, 1982; 1989), fuzzy soft partitioning (Bezdek, 1981), expert-level learning (Jacobs et al., 1991; Jordan & Xu, 1994), and cluster-validity evaluation (Rousseeuw, 1987; Davies & Bouldin, 1979; Hennig, 2007) are arranged to reflect the sequential and hierarchical dependencies among Sections 2.1–2.6.

A. Reliability Diagrams

This appendix presents reliability diagrams for all datasets (Telco, Bank Marketing, Adult) for both FC–MoE–GR and the matched Global baselines, using the same test splits reported in Tables 1–3. Each figure visualizes probability calibration using 15 equal-width bins, plotting observed frequency against mean predicted probability. Expected Calibration Error (ECE) is computed on the test split with the same binning scheme. The dashed line corresponds to perfect calibration. These standalone plots provide a direct visual audit of the saved prediction files used to construct the canonical metric sheet.

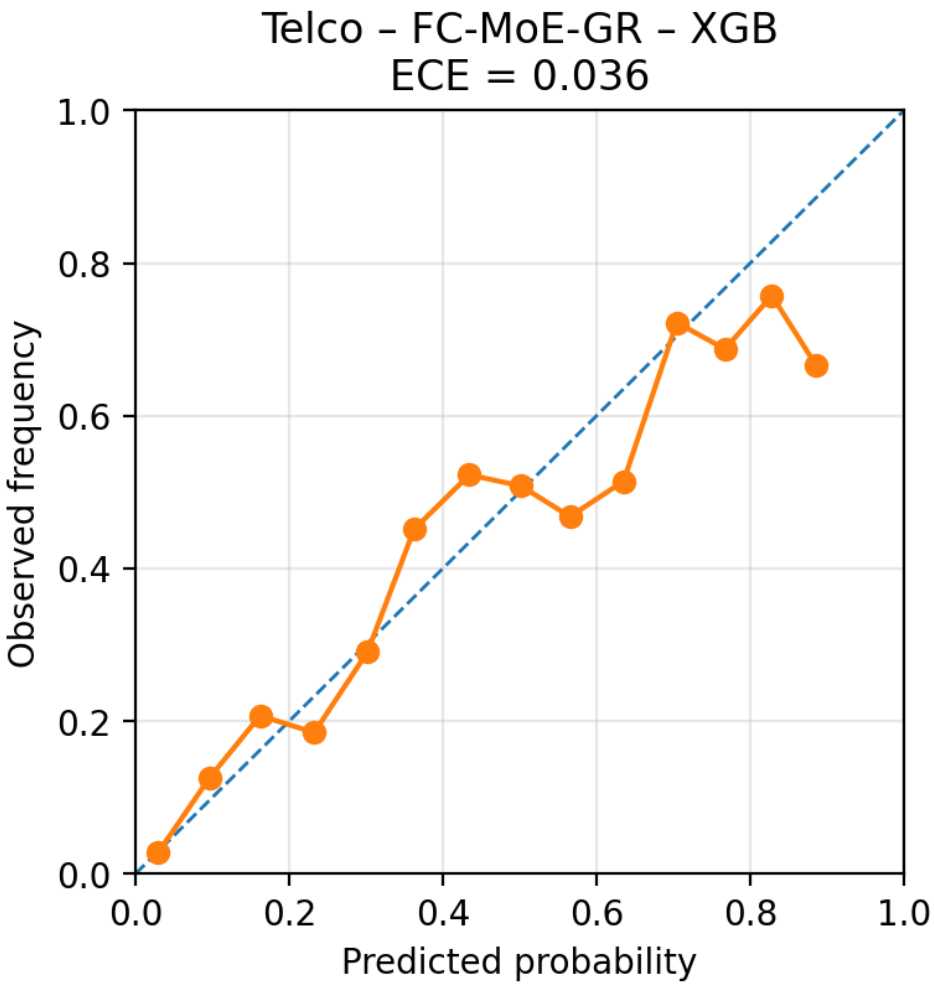

Figure A.2. Reliability diagram for **Telco – XGB** (FC–MoE–GR). The curve shows 15-bin observed-vs-predicted calibration and the dashed line denotes perfect calibration. ECE = 0.036 matches the FC–MoE–GR (XGB) entry in Table 1.

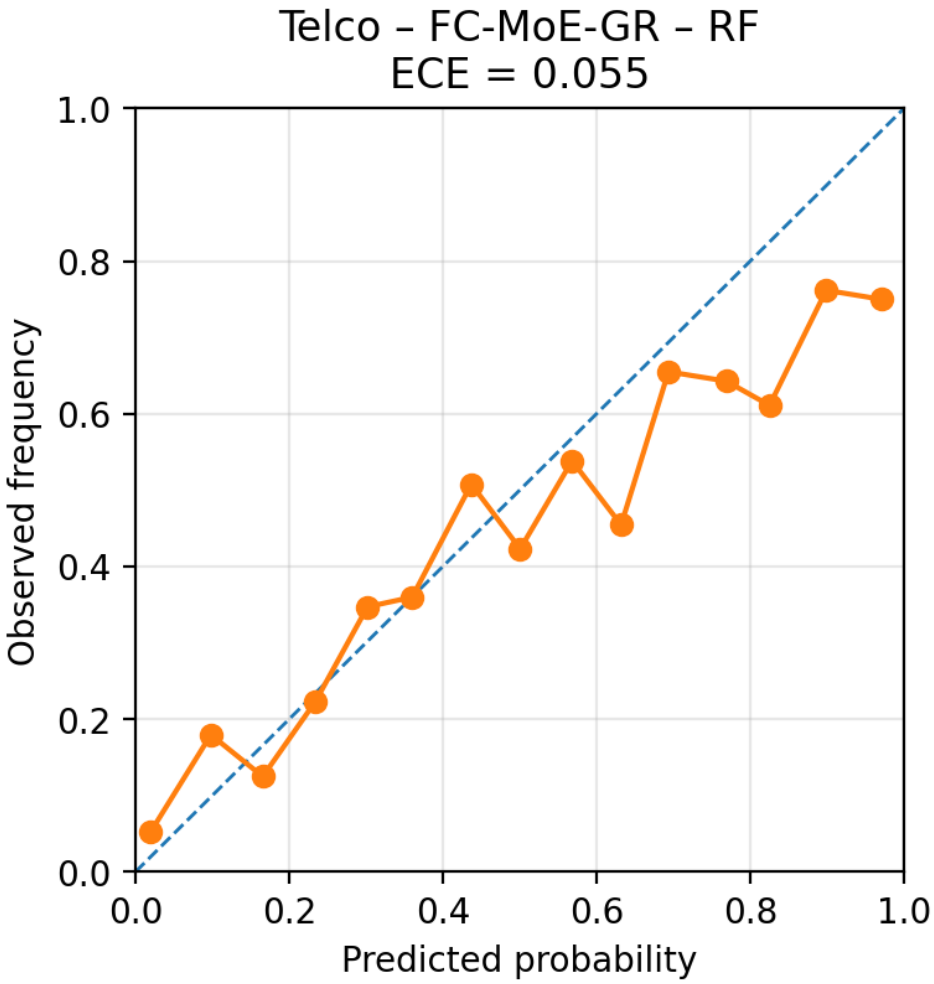

Figure A.3. Reliability diagram for **Telco – RF** (FC–MoE–GR). The curve shows 15-bin observed-vs-predicted calibration and the dashed line denotes perfect calibration. ECE $= 0.055$ matches the FC–MoE–GR (RF) entry in Table 1.

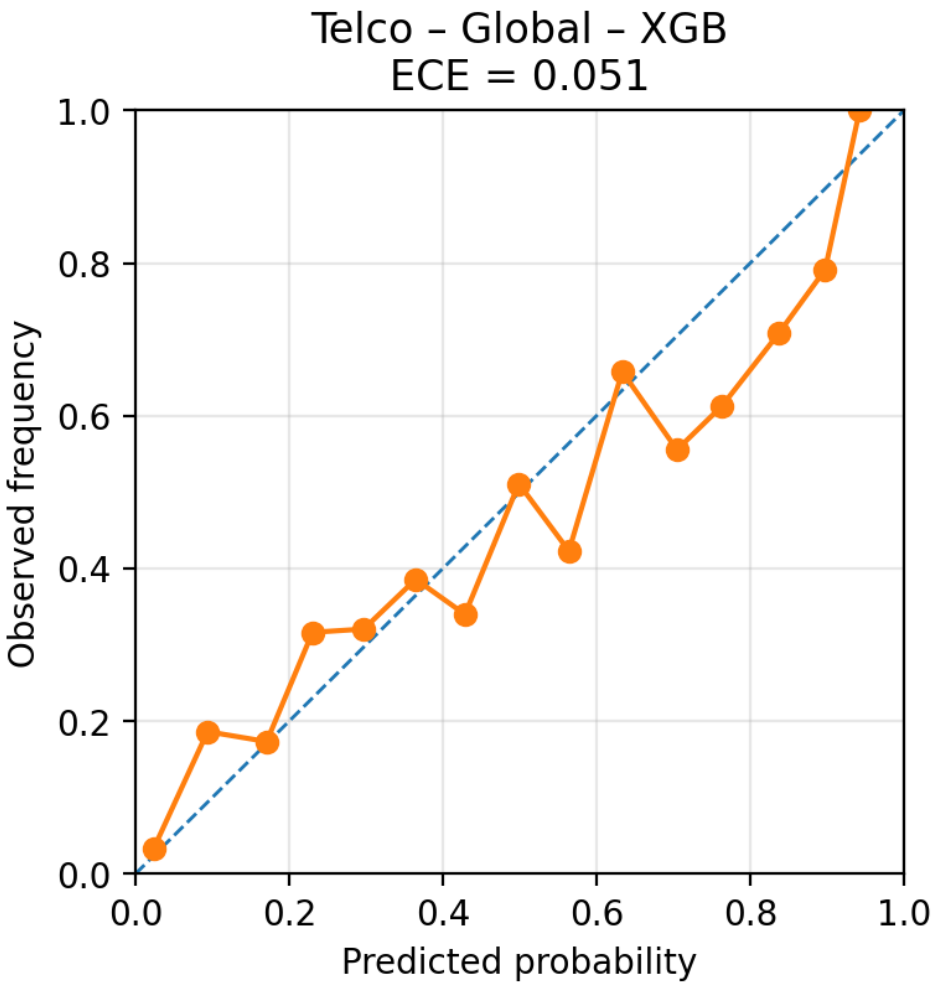

Figure A.4. Reliability diagram for **Telco – XGB** (Global baseline). The curve shows 15-bin observed-vs-predicted calibration and the dashed line denotes perfect calibration. ECE $= 0.051$ matches the Global XGB entry in Table 1.

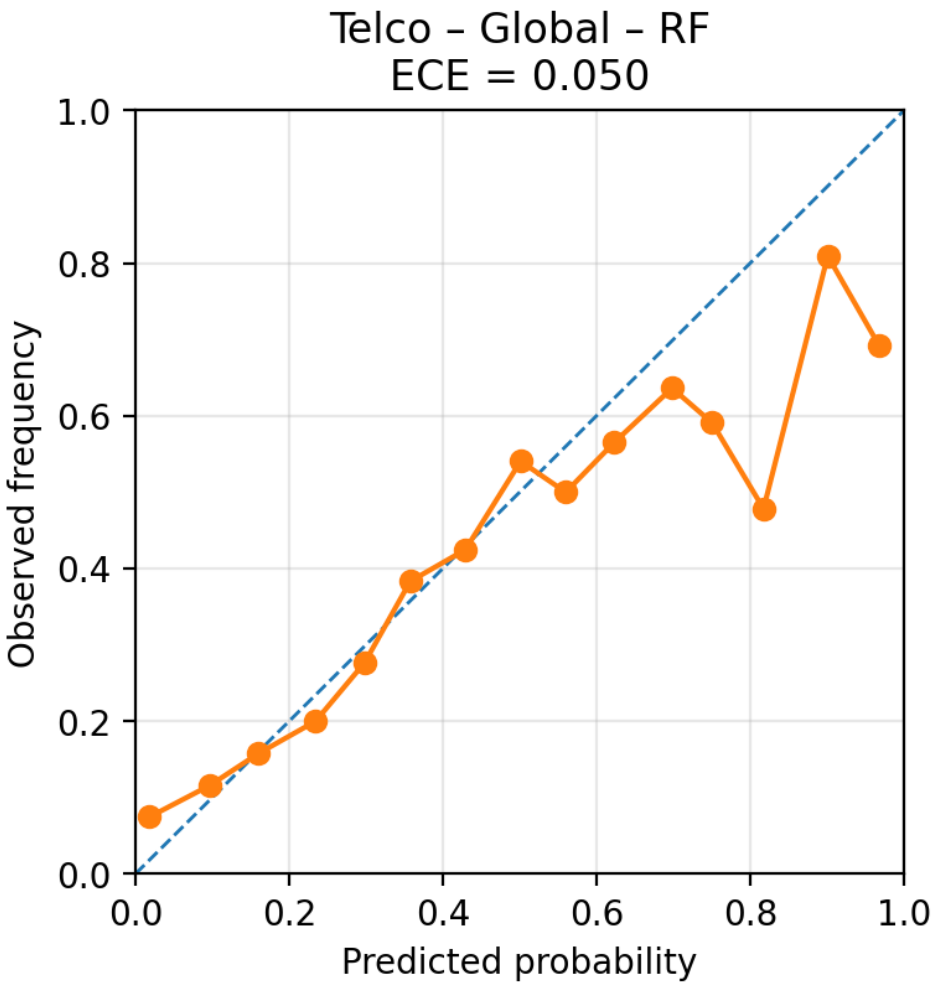

Figure A.5. Reliability diagram for **Telco – RF** (Global baseline). The curve shows 15-bin observed-vs-predicted calibration and the dashed line denotes perfect calibration. ECE $= 0.050$ matches the Global RF entry in Table 1.

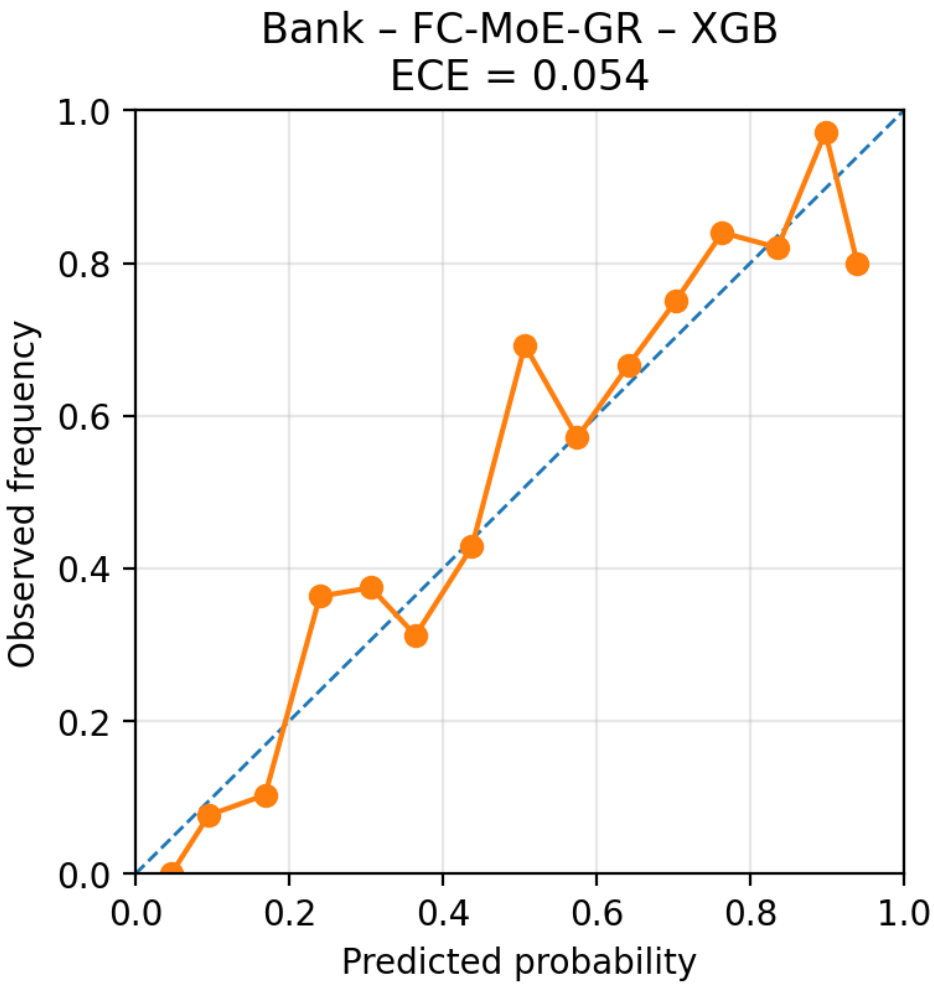

Figure A.6. Reliability diagram for **Bank – XGB** (FC–MoE–GR). The curve shows 15-bin observed-vs-predicted calibration and the dashed line denotes perfect calibration. ECE $= 0.054$ matches the FC–MoE–GR (XGB) entry in Table 2.

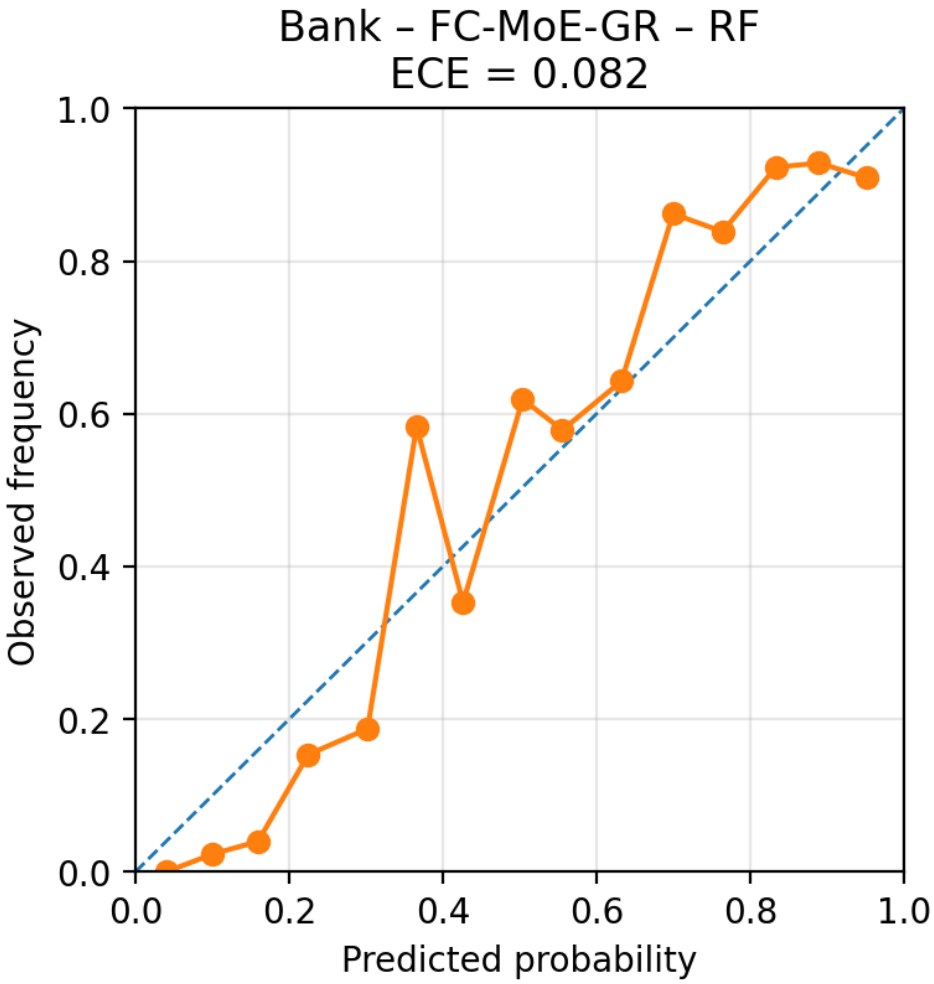

Figure A.7. Reliability diagram for **Bank – RF** (FC–MoE–GR). The curve shows 15-bin observed-vs-predicted calibration and the dashed line denotes perfect calibration. ECE $= 0.082$ matches the FC–MoE–GR (RF) entry in Table 2.

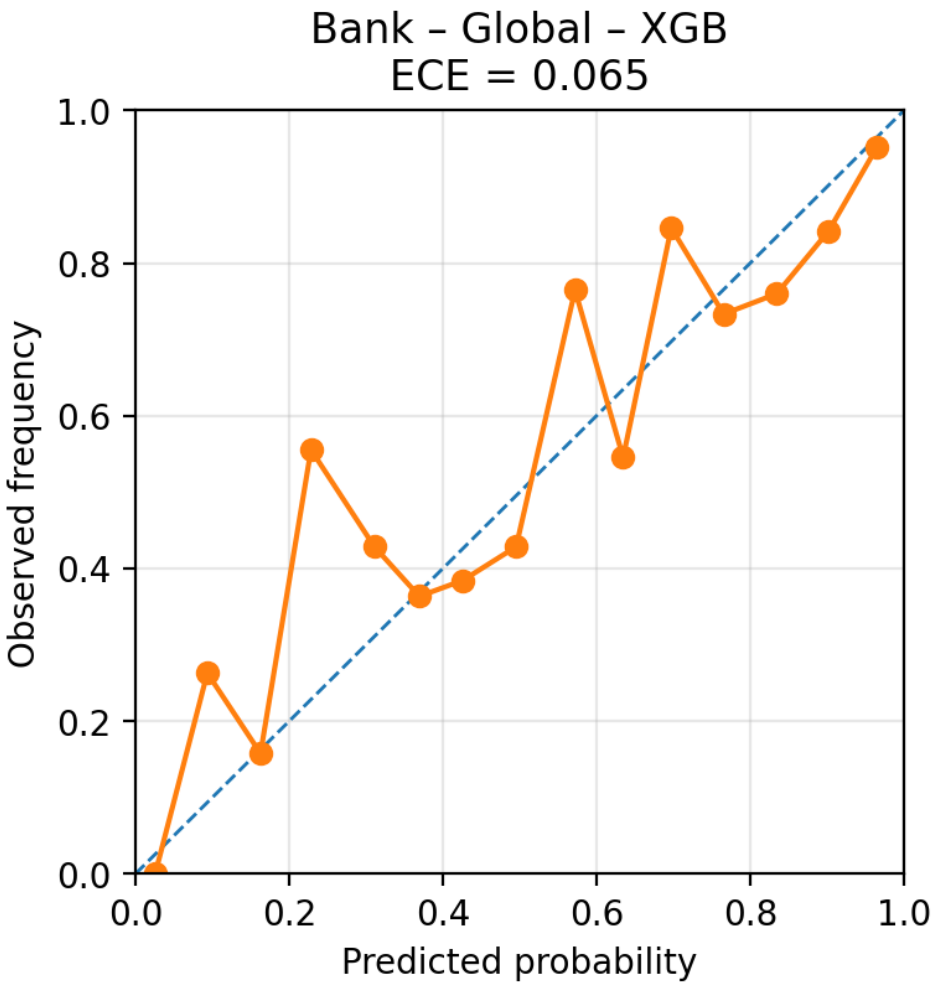

Figure A.8. Reliability diagram for **Bank – XGB** (Global baseline). The curve shows 15-bin observed-vs-predicted calibration and the dashed line denotes perfect calibration. ECE $= 0.065$ matches the Global XGB entry in Table 2.

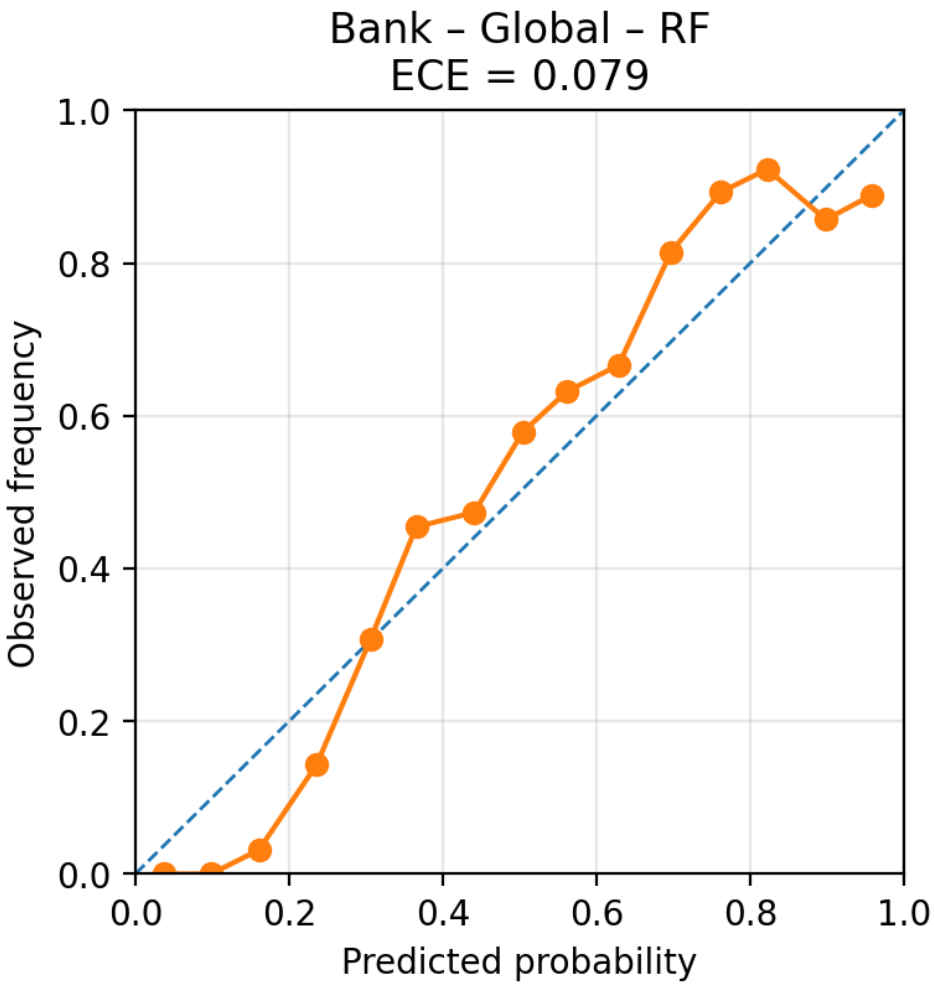

Figure A.9. Reliability diagram for **Bank – RF** (Global baseline). The curve shows 15-bin observed-vs-predicted calibration and the dashed line denotes perfect calibration. ECE $= 0.079$ matches the Global RF entry in Table 2.

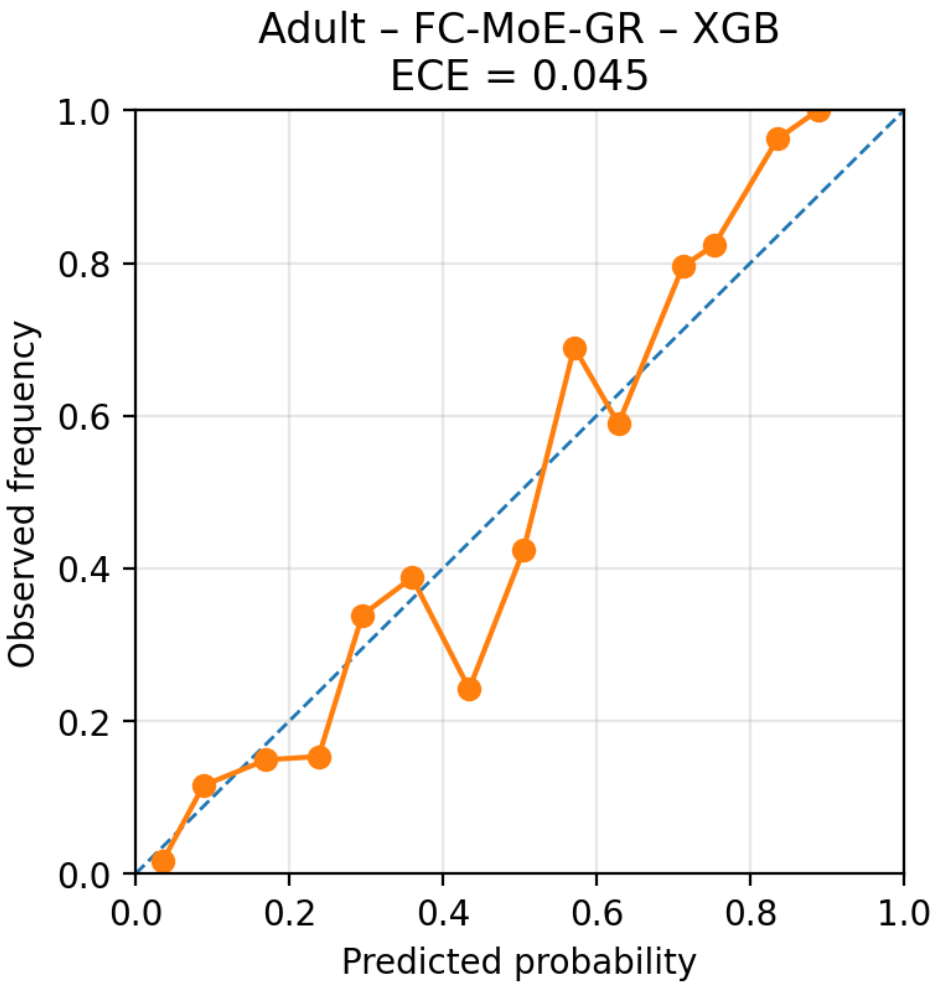

Figure A.10. Reliability diagram for **Adult – XGB** (FC–MoE–GR). The curve shows 15-bin observed-vs-predicted calibration and the dashed line denotes perfect calibration. ECE $= 0.045$ matches the FC–MoE–GR (XGB) entry in Table 3.

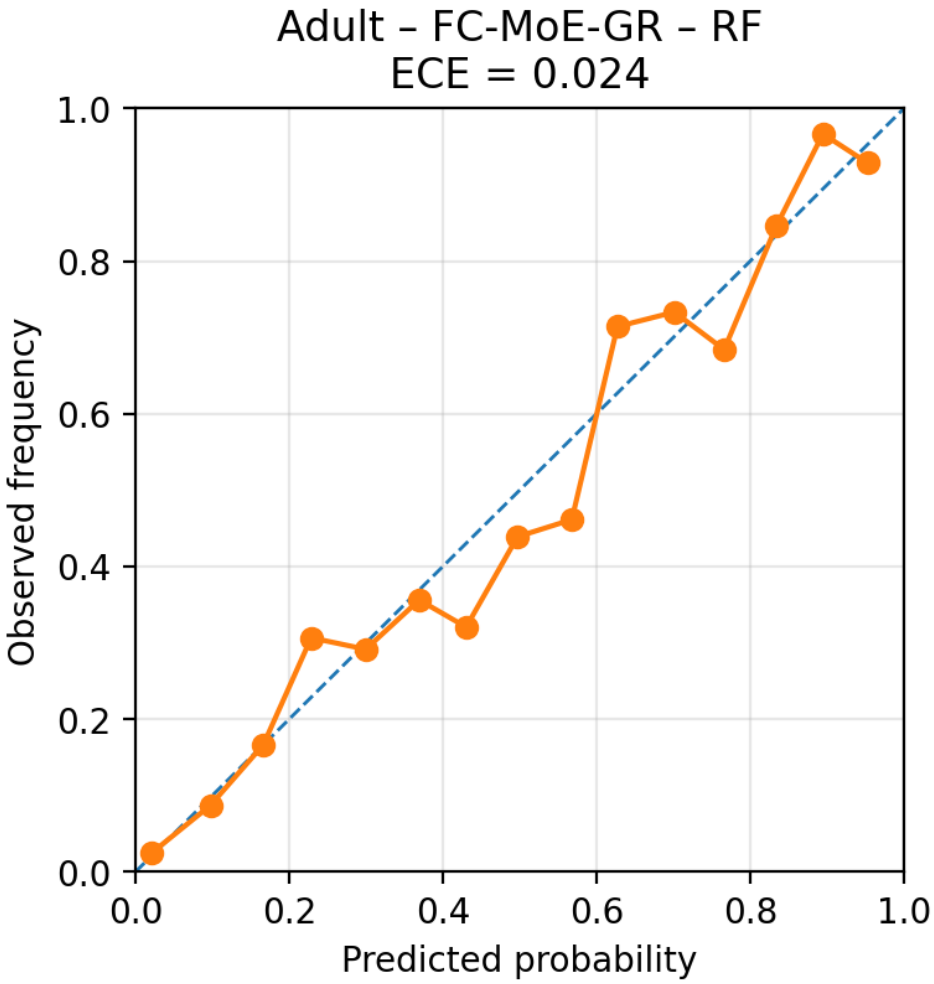

Figure A.11. Reliability diagram for **Adult – RF** (FC–MoE–GR). The curve shows 15-bin observed-vs-predicted calibration and the dashed line denotes perfect calibration. ECE $= 0.024$ matches the FC–MoE–GR (RF) entry in Table 3.

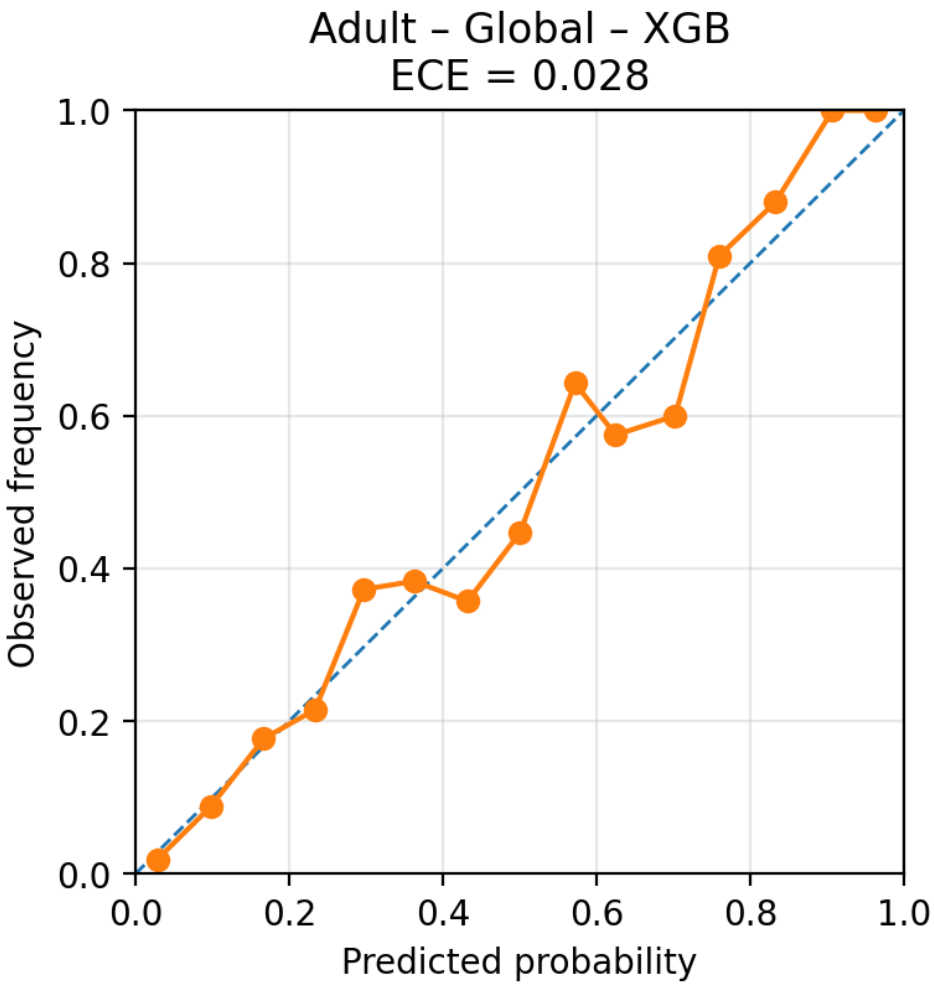

Figure A.12. Reliability diagram for **Adult – XGB** (Global baseline). The curve shows 15-bin observed-vs-predicted calibration and the dashed line denotes perfect calibration. ECE = 0.028 matches the Global XGB entry in Table 3.

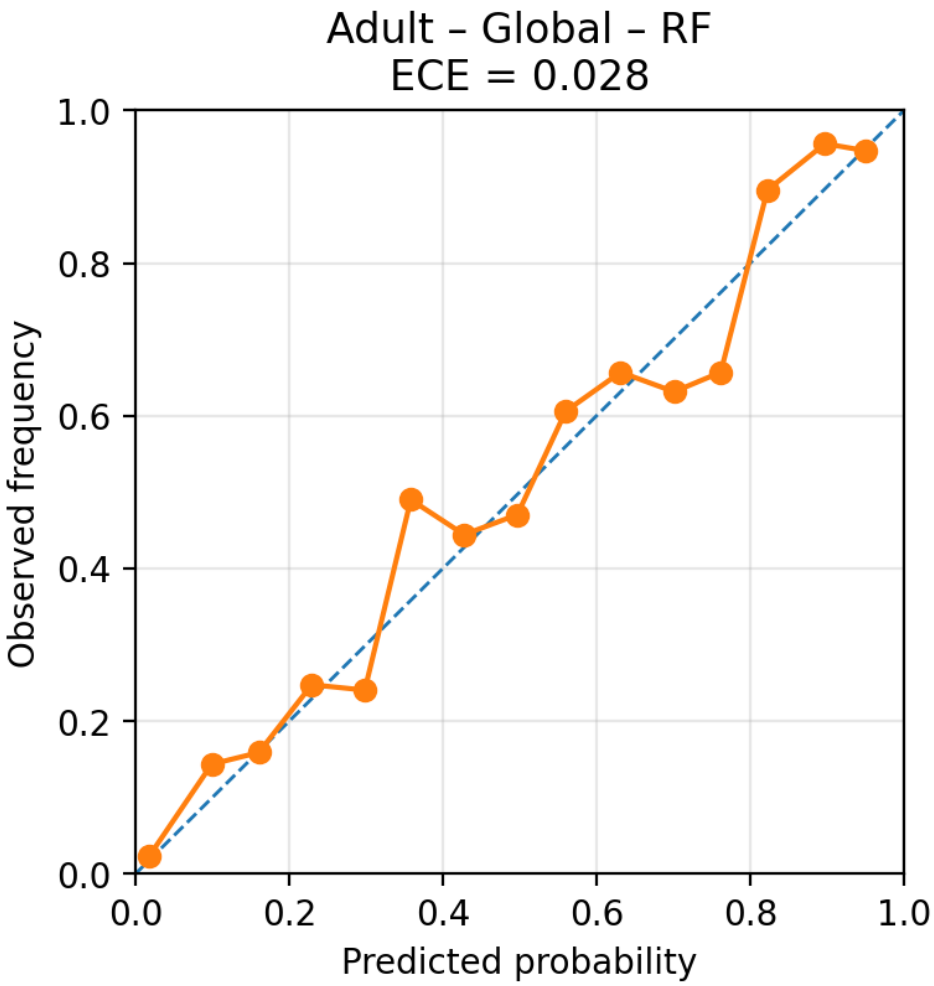

Figure A.13. Reliability diagram for **Adult – RF** (Global baseline). The curve shows 15-bin observed-vs-predicted calibration and the dashed line denotes perfect calibration. ECE = 0.028 matches the Global RF entry in Table 3.

B. Symbols and Notation Summary

Table 7 summarizes the notation used throughout the paper. The definitions follow standard conventions from grey-system theory (Deng, 1989), fuzzy clustering (Bezdek, 1981), mixture-of-experts modeling (Jacobs et al., 1991), and statistical learning theory (Bartlett & Mendelson, 2002; Mohri et al., 2018). Symbols related to cluster validity indices are consistent with well-established metrics such as Silhouette (Rousseeuw, 1987), Davies–Bouldin (Davies & Bouldin, 1979), and entropy-based measures in fuzzy partitioning (Bezdek, 1981).

Table 7. Summary of symbols and notation.

Symbol	Description
$X \in \mathbb{R}^{n \times p}$	Input feature matrix
$y \in \mathcal{Y}$	Target labels
x_i	i-th sample
$x_0(t)$	Reference sequence constructed from positive-class means (Deng, 1989)
$x_j(t)$	Comparative sequence for feature j
$\Delta_j(t)$	Absolute deviation between sequences (Deng, 1982)
$\xi_j(t)$	Grey relational coefficient (GRC) (Deng, 1982)
g_j	Grey relational grade (GRG) (Deng, 1989)
w_j	Normalized grey-relational feature weights
ρ	Identification coefficient in GRA (Deng, 1982)
m	Fuzziness exponent in FCM (Bezdek, 1981)
u_{ik}	Membership of sample i in cluster k
c_k	Cluster center for cluster k
f_k	Local expert associated with cluster k (Jacobs et al., 1991)
$\hat{y}(x)$	Soft mixture-of-experts output (Jordan & Xu, 1994)
$\mathcal{L}_k$	Local grey-regularized objective
λ	Regularization strength
DB, SIL	Davies–Bouldin and Silhouette indices (Davies & Bouldin, 1979; Rousseeuw, 1987)
PC, PE	Partition coefficient and entropy (Bezdek, 1981)
STAB	Bootstrap-based assignment stability (ARI) (Hennig, 2007)
ECE	Expected calibration error (Naeini et al., 2015)
K^*	Selected number of clusters

C. Derivations and Proof Sketches

This appendix provides formal derivations supporting the equations introduced in Section 2. The exposition follows foundational results from grey-system theory (Deng, 1982; 1989), fuzzy clustering theory (Bezdek, 1981), and statistical learning theory (Bartlett & Mendelson, 2002; Mohri et al., 2018). All mathematical expressions retain the notation defined in Appendix B.

C.1. Derivation of the Grey Relational Coefficient (Eq. 3)

The derivation of Eq. (3) follows the canonical form established in grey relational analysis (Deng, 1982; 1989). Given a reference sequence $x_0(t)$ and comparative sequence $x_j(t)$, the absolute deviation is

$$\Delta_j(t) = |x_0(t) - x_j(t)|.$$

Let

$$\Delta_{\min} = \min_{j,t} \Delta_j(t), \qquad \Delta_{\max} = \max_{j,t} \Delta_j(t).$$

Grey-system theory requires the similarity measure to satisfy the axioms of: (i) boundedness, (ii) monotonicity with respect to $\Delta_j(t)$, and (iii) scale invariance across dimensions (Deng, 1989).

The only fractional transformation that satisfies all three axioms takes the form

$$\xi_j(t) = \frac{\Delta_{\min} + \rho\Delta_{\max}}{\Delta_j(t) + \rho\Delta_{\max}},$$

which is identical to Eq. (3). The constant ρ is the identification coefficient, typically set to 0.5 following standard practice in grey-system analysis (Deng, 1982).

C.2. Derivation of the Grey Relational Grade (Eq. 4)

The grey relational grade aggregates pointwise relational coefficients into a global relevance score (Deng, 1989). For feature j, the arithmetic mean is

$$g_j = \frac{1}{T}\sum_{t=1}^{T}\xi_j(t),$$

which is the unique averaging operator preserving the ranking structure induced by $\xi_j(t)$ (Deng, 1989). This yields Eq. (4).

C.3. Normalization into Feature Weights (Eq. 5)

Grey-system theory defines feature importance as a normalized distribution over g_j values (Deng, 1982). Thus

$$w_j = \frac{g_j}{\sum_{\ell=1}^{p} g_\ell}$$

is the canonical normalization satisfying invariance and comparability among features. This provides Eq. (5).

C.4. Heuristic Sketch of the Risk Bound in Eq. (11)

This section provides a simplified, heuristic derivation to illustrate the intuition behind the proposed bounds. For the rigorous mathematical proof involving Rademacher complexity and McDiarmid's inequality, please refer strictly to the formal derivation in Appendix E.

Let $\mathcal{F}$ denote the hypothesis class of mixture-of-expert functions (Jacobs et al., 1991; Jordan & Xu, 1994). Statistical learning theory provides the decomposition

$$\mathbb{E}[R(\hat{f})] = B(\hat{f}) + V(\hat{f}) + \mathcal{O}\left(\frac{\mathfrak{R}(\mathcal{F})}{\sqrt{n}}\right),$$

where $\mathfrak{R}(\mathcal{F})$ is the Rademacher complexity (Bartlett & Mendelson, 2002; Mohri et al., 2018). This establishes Eq. (11).

C.5. Intuition behind the Grey-Regularized Tightening (Eq. 13)

This section provides a simplified, heuristic derivation to illustrate the intuition behind the proposed bounds. For the rigorous mathematical proof involving Rademacher complexity and McDiarmid's inequality, please refer strictly to the formal derivation in Appendix E.

Grey regularization constrains the function class to

$$\mathcal{F}_\lambda = \left\{ f : \sum_j (1 - w_j)\Omega_{k,j} \le \lambda \right\},$$

which reduces the hypothesis space diameter via feature-level shrinkage.

Since $\mathcal{F}_\lambda \subseteq \mathcal{F}$,

$$\mathfrak{R}(\mathcal{F}_\lambda) \le \mathfrak{R}(\mathcal{F}),$$

consistent with standard monotonicity properties of Rademacher complexity (Bartlett & Mendelson, 2002). Therefore,

$$\mathbb{E}[R(\hat{f}_\lambda)] \le B(\hat{f}_\lambda) + V(\hat{f}_\lambda) + \mathcal{O}\left(\frac{\mathfrak{R}(\mathcal{F}_\lambda)}{\sqrt{n}}\right),$$

which corresponds to Eq. (13).

D. Additional Technical Material

This appendix provides supplementary expansions that support Sections 2–4 of the main text. All derivations adhere to established formulations in grey-system theory (Deng, 1989), fuzzy clustering (Bezdek, 1981), mixture-of-experts analysis (Jacobs et al., 1991), and cluster-validity evaluation (Rousseeuw, 1987; Davies & Bouldin, 1979).

D.1. Derivation of the Grey-Relational Grade in the Aggregated Form

Following the axioms in grey-system theory (Deng, 1989), the generalized GRG can also be expressed as

$$w_j = \frac{1}{n}\sum_{i=1}^{n}\gamma_i(j),$$

where $\gamma_i(j)$ is the sample-wise relational coefficient. This form is equivalent to Eq. (4) when temporal indices are replaced by sample indices, a common variant in empirical implementations.

D.2. Relation to Partition Validity Indices

Entropy-based measures (PC, PE) follow directly from fuzzy clustering theory (Bezdek, 1981). Stability is aligned with bootstrap-based cluster reproducibility (Hennig, 2007). These supplemental results justify the composite scoring rule in Section 2.6.

E. Proof of Generalization Bounds

Detailed proofs for **Theorem 1** are provided in this section, establishing the generalization guarantees for the FC-MoE-GR framework.

This appendix should be read as a complexity-control justification for the grey-regularized soft-expert hypothesis class, rather than as a literal tree-specific closed-form bound for the Random Forest and XGBoost experts used in the empirical study. In the experiments, RF/XGB serve as stable tabular experts that operationalize the same variance-control intuition under fixed fuzzy routing; the proof therefore targets the contraction of the abstract grey-regularized function class rather than a sharp finite-sample bound specialized to tree ensembles.

E.1. Preliminaries

The standard supervised learning setting is considered. Let $\mathcal{X}$ be the input space and $\mathcal{Y}$ be the output space. The objective is to learn a hypothesis $h : \mathcal{X} \to \mathcal{Y}$ from $\mathcal{H}$ that minimizes the expected risk $R(h) = \mathbb{E}_{z\sim\mathcal{D}}[\ell(h(x), y)]$. The notion of Rademacher complexity is utilized to measure the richness of the hypothesis class.

E.2. Rademacher Complexity and Generalization

The Empirical Rademacher Complexity of a hypothesis class $\mathcal{H}$ with respect to a sample $\mathcal{S}$ is defined as:

$$\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{H}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(x_i)\right] \tag{21}$$

where $\boldsymbol{\sigma}$ are independent Rademacher variables uniform in $\{-1, +1\}$.

According to foundational results in statistical learning theory (Bartlett & Mendelson, 2002), for any $\delta > 0$, with probability at least $1 - \delta$, the following holds:

$$\sup_{h\in\mathcal{H}} |R(h) - \hat{R}_{\mathcal{S}}(h)| \le 2L\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{H}) + 3B\sqrt{\frac{\ln(2/\delta)}{2m}} \tag{22}$$

This standard bound connects the generalization gap to the complexity of the model class.

E.3. Bounding Complexity via Grey-Relational Regularization

The output of FC-MoE-GR is given by $h(x) = \sum_{k=1}^{K} g_k(x)E_k(x)$. The Grey-Relational Regularization $\lambda\mathcal{L}_{GR}$ imposes a constraint on the parameter space, specifically restricting the Frobenius norm of the weight matrices such that $\|\mathbf{W}\|_F \leq \Lambda_{GR}$.

By applying the Contraction Lemma (Lemma 5.7 in (Mohri et al., 2018)), given that the gating function and expert functions are Lipschitz continuous (due to bounded activation functions and normalized grey coefficients), the complexity of the ensemble is bounded by the sum of the complexities of individual experts:

$$\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{H}) \leq \sum_{k=1}^{K} \hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{H}_{expert}^{(k)}) + \hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{H}_{gate}) \tag{23}$$

For expert families that admit bounded Lipschitz surrogates, a standard illustrative bound takes the form:

$$\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{H}_{expert}) \leq \frac{\Lambda_{GR} \max_i \|x_i\|_2}{\sqrt{m}} \tag{24}$$

Substituting this back yields the final bound:

$$R(h) \leq \hat{R}_{\mathcal{S}}(h) + \frac{C \cdot \Lambda_{GR}}{\sqrt{m}} + 3B\sqrt{\frac{\ln(2/\delta)}{2m}} \tag{25}$$

where C is a constant depending on the Lipschitz constant L and the number of experts K. This confirms that by controlling Λ_{GR} via the regularization parameter λ, the generalization bound is effectively tightened. □

F. Algorithms

Algorithmic pseudocode follows the ICML 2026 formatting guidelines and is conceptually consistent with mixture-of-experts literature (Jacobs et al., 1991; Jordan & Xu, 1994) and grey-relational computation procedures (Deng, 1982; 1989).

G. Reproducibility and Benchmarking Resources

The executable implementation and code for reproducing all figures and numerical results are provided in the supplementary materials submitted alongside this paper, following the ICML reproducibility guidelines.

The repository contains:

- Colab scripts for FC–MoE–GR training and evaluation,
- data preprocessing pipelines for Telco, Bank, and Adult datasets,
- graph generation code for Figures 1–3 and the appendix reliability diagrams using Graphviz and Matplotlib,
- separate CSV exports for FC–MoE–GR results, Global baseline results, and ablation outputs,
- saved prediction files for both FC–MoE–GR and Global baselines,
- a canonical `master_metrics.csv` sheet used for all paper tables,
- bootstrap interval summaries derived from the saved prediction files,
- parameter-choice manifests for m, ρ, α, candidate K, and selected K^*,
- consistency-audit manifests that verify the agreement between prediction files, reliability figures, and reported ECE values, and
- exact split sizes, subsampling fractions, and per-dataset file manifests used to reproduce the CTB submission package.

Table 8. Bootstrap percentile interval summaries (300 resamples) for the main predictive metrics. The intervals are derived from the saved prediction files used to generate the canonical metric sheet.

Dataset	Model	AUC 95% CI	F1 95% CI	ECE 95% CI
Telco	FC–MoE–GR (RF)	[0.772, 0.829]	[0.449, 0.556]	[0.046, 0.085]
Telco	FC–MoE–GR (XGB)	[0.803, 0.853]	[0.475, 0.578]	[0.032, 0.066]
Telco	Global RF	[0.766, 0.825]	[0.468, 0.579]	[0.045, 0.084]
Telco	Global XGB	[0.786, 0.839]	[0.494, 0.596]	[0.045, 0.084]
Bank	FC–MoE–GR (RF)	[0.868, 0.930]	[0.766, 0.859]	[0.072, 0.132]
Bank	FC–MoE–GR (XGB)	[0.864, 0.930]	[0.785, 0.871]	[0.055, 0.110]
Bank	Global RF	[0.867, 0.927]	[0.777, 0.868]	[0.073, 0.129]
Bank	Global XGB	[0.870, 0.931]	[0.779, 0.870]	[0.062, 0.117]
Adult	FC–MoE–GR (RF)	[0.873, 0.907]	[0.598, 0.685]	[0.023, 0.048]
Adult	FC–MoE–GR (XGB)	[0.871, 0.910]	[0.600, 0.691]	[0.035, 0.067]
Adult	Global RF	[0.867, 0.902]	[0.587, 0.675]	[0.024, 0.051]
Adult	Global XGB	[0.877, 0.917]	[0.596, 0.688]	[0.025, 0.052]

G.1. Bootstrap Interval Summary for Main Metrics

For convenience, Table 8 reproduces the 95% percentile bootstrap intervals released in the supplementary file `bootstrap_metric_intervals.csv`. These summaries are computed from the saved test-set prediction files using 300 bootstrap resamples on the fixed evaluation split. They are reported as uncertainty summaries for the canonical benchmark outputs and are not intended to replace a full repeated-run significance analysis.

H. Reviewer-Informed Future Extensions

The accepted CTB version is intended as a theory-linked benchmark case study rather than a final statement on optimal subgroup modeling. Two directions are especially important for future extensions. First, the current generalization argument uses standard Rademacher-complexity machinery to motivate a complexity-control perspective for grey-regularized soft experts. A sharper analysis could replace this global view with localized Rademacher complexity, chaining-style arguments, or subgroup-conditioned complexity terms that better reflect the effective hypothesis class used by each fuzzy partition. Such refinements would more tightly connect the theoretical prediction to the calibration and stability diagnostics reported in the benchmark.

Second, subgroup modeling is increasingly relevant beyond classical tabular prediction. In modern foundation-model workflows, downstream systems often rely on tabular telemetry, user attributes, safety logs, retrieval metadata, or scenario descriptors to audit performance across heterogeneous populations. FC–MoE–GR can be viewed as a lightweight diagnostic layer for such settings: fuzzy memberships define overlapping operational regimes, grey-relational priors highlight stable feature relations under limited observations, and local experts provide calibrated subgroup-level risk estimates. Potential applications include foundation-model-assisted decision systems, autonomous-system monitoring, and safety auditing for rare or underrepresented operating conditions. These extensions do not change the empirical claims of the present work, but they clarify how the proposed benchmark protocol can support future studies at the interface of theory, calibration, and deployment-oriented reliability.

References

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Bezdek, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Springer, 1981.

Breiman, L. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016. doi: 10.1145/2939672.2939785.

Davies, D. L. and Bouldin, D. W. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979.

Deng, J. Control problems of grey systems. *Systems & Control Letters*, 1(5):288–294, 1982.

Deng, J. *Introduction to Grey System Theory*. Grey System Research Institute, 1989.

Dunn, J. C. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32–57, 1973.

Fedus, W., Zoph, B., and Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. In *International Conference on Machine Learning*, 2022.

Fisher, A., Rudin, C., and Dominici, F. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

Geman, S., Bienenstock, E., and Doursat, R. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1): 1–58, 1992.

Gorishniy, Y., Rubachev, I., Khrulkov, V., and Babenko, A. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18932–18943, 2021.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330, 2017.

Hennig, C. Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis*, 52(1):258–271, 2007.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. Adaptive mixtures of local experts. In *Advances in Neural Information Processing Systems*, volume 3, pp. 767–773, 1991.

Jordan, M. I. and Xu, L. Hierarchical mixtures of experts and the em algorithm. *Neural Computation*, 6(2):181–214, 1994.

Kohavi, R. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 202–207, 1996.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. MIT Press, 2 edition, 2018. ISBN 9780262039406.

Moro, S., Cortez, P., and Rita, P. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

Naeini, M. P., Cooper, G. F., and Hauskrecht, M. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

Shazeer, N., Mirhoseini, A., Maziarz, A., Davis, J., Le, Q. V., Hinton, G. E., and Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *Proceedings of the International Conference on Learning Representations*, 2017.

Xie, X. L. and Beni, G. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8):841–847, 1991. doi: 10.1109/34.85677.