

Article

Performance Evaluation of Publicly Funded Agricultural Research Projects with Light-TabNet[†]

Zelin Liu^{1,‡}, Lu Fan^{2,‡}, Qiulian Chen^{1,*}, Haipeng Li¹ and Ailan Wei²¹ College of Computer and Electronic Information, Guangxi University, Nanning 530004, China; lz1996@st.gxu.edu.cn (Z.L.); fu@alu.gxu.edu.cn (H.L.)² Guangxi Academy of Agricultural Sciences, Nanning 530007, China; ffffxl@foxmail.com (L.F.); weiailan70@foxmail.com (A.W.)

* Correspondence: chenql@gxu.edu.cn

[†] This paper is an extended version of our paper published in the 2025 6th International Conference on Big Data Economy and Information Management (BDEIM 2025), Chengdu, China, 19–21 December 2025.[‡] These authors contributed equally to this work.

Abstract

This study focuses on the performance evaluation of publicly funded agricultural research projects in a structured tabular-data setting characterized by small sample size and heterogeneous features. We construct a project-level performance evaluation dataset covering 24 provincial agricultural research institutions in China, with $n = 280$ samples. The target variable is the project self-evaluation score, reflecting overall annual target completion rather than a fixed explicit transformation of the input indicators. To address the limitations of manual evaluation—including subjectivity, poor inter-rater consistency, and potential bias—we propose Light-TabNet, which enhances the model's fitting capability in small-sample scenarios while preserving interpretability. Interpretability is achieved through sparse decision masks and aggregated feature-attribution analysis, with partial cross-model support from comparison with XGBoost-SHAP rankings. Compared with 13 deep learning and traditional machine learning baselines, Light-TabNet achieves improved accuracy in terms of mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of determination (R^2) (MAE 4.9765, RMSE 8.8140, R^2 0.8891). In a preliminary real-world validation on eight projects from a provincial agricultural research institution, the model's predicted scores were overall close to ratings provided by a third-party organization, suggesting preliminary practical usefulness in a similar management setting. The results suggest that Light-TabNet can serve as a decision-support tool for the performance evaluation of publicly funded agricultural research projects by providing an objective, traceable, and interpretable quantitative reference.

Keywords: performance evaluation; structured tabular data; lightweight deep learning; TabNet; small sample size; agricultural research; public funding



Academic Editor: Paolo Renna

Received: 10 February 2026

Revised: 13 March 2026

Accepted: 13 March 2026

Published: 27 March 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Agriculture plays a critical role in ensuring food security, maintaining social stability, and promoting economic development. Agricultural scientific research, in turn, drives sustainable agricultural development by fostering innovation and productivity, enabling societies to better respond to natural and market fluctuations. Public funding constitutes a substantial proportion of the budget for agricultural R&D projects. With tightened government debt constraints and expanding public expenditures, the use and effectiveness

of fiscal appropriations have attracted increasing attention. In this context, governments aim to improve funding efficiency, strengthen fiscal governance, correct problems in a timely manner, and enhance accountability by conducting performance evaluations of publicly funded agricultural R&D projects. However, in many institutions, such evaluation practices are still at an early stage and leave considerable room for improvement. Reliance on purely manual evaluation is not only costly but also susceptible to subjective judgment. Consequently, researchers have begun to leverage machine learning to extract information from project performance data, enabling effective score prediction and more rational recommendation of performance indicators.

At present, project performance evaluation is still dominated by manual (expert-based) review. For example, the Standard Evaluation Process of the European Union's Horizon Europe involves independent expert assessment [1]; the U.S. National Science Foundation (NSF) typically relies on 3–10 external experts through ad hoc reviews, panel reviews, or a combination of both [2]; in the United Kingdom, funding decisions of the Medical Research Council (MRC) are made by integrating external reviewer comments with the deliberations of research councils and expert panels [3]; the Australian Research Council (ARC) assigns assessors to provide scores and incorporates bias-testing procedures in the peer-review process [4]; and Japan's JST SATREPS collects evidence through research reporting meetings and site surveys, and reaches conclusions via external expert review meetings [5]. Nevertheless, manual-dominant evaluation often suffers from strong subjectivity, insufficient inter-rater consistency, vulnerability to bias and conflicts of interest, and high labor costs, making it difficult to meet institutional goals of transparency, standardization, and scientific rigor. These governance-oriented requirements also imply a corresponding technical demand: a practical evaluation model should not only provide sufficiently accurate score prediction, but also remain traceable, robust under small-sample heterogeneous tabular data, and usable in routine management settings with incomplete indicators.

To mitigate these issues, prior studies have attempted to introduce structured evaluation methodologies. Garefalakis et al. proposed a contextualized ESG–BSC integrated model for small municipalities [6]; Shim and Kim used the analytic hierarchy process (AHP) to weight and prioritize multi-domain indicators for local fiscal investment projects [7]; and Shin et al. applied data envelopment analysis (DEA), complemented by Tobit regression, to assess government expenditure efficiency [8]. However, these approaches still face practical limitations. In BSC/ESG settings, indicator specification and weight assignment remain highly dependent on manual design. AHP fundamentally relies on expert pairwise comparisons, so the resulting weights are sensitive to subjective judgment and the consistency of comparison matrices. DEA, although useful for relative efficiency assessment, may lose discriminative power when the number of indicators is large relative to the number of evaluated projects, because many decision-making units can appear similarly efficient near the frontier; its results may also be sensitive to variable specification and measurement noise. These issues are particularly relevant in real-world project performance datasets that are small in sample size, heterogeneous in indicator composition, and partially incomplete.

In recent years, deep learning models for structured tabular data have developed rapidly, leading to a series of methods that can effectively model feature interactions and achieve strong performance on public benchmarks (e.g., NODE, TANGOS, TabM, and Mam-bular) [9–12]. However, applying these models to performance evaluation of agricultural R&D projects still presents several challenges. First, there is a lack of publicly available real-world datasets with relatively consistent definitions and measurement criteria, which hinders rigorous evaluation. Second, under small-sample and high-dimensional conditions, both machine learning and deep learning models may exhibit degraded predictive performance. Third, if a model outputs only predicted scores without interpretable evidence

on feature contributions, it is difficult to support downstream decisions such as indicator design and weight adjustment. Among deep tabular models, TabNet is attractive because it combines competitive predictive modeling with built-in feature-selection masks that can provide step-wise interpretability. However, the original TabNet architecture was not specifically designed for small-sample project-evaluation scenarios with heterogeneous indicators and partial missingness. This motivates a lightweight and more stable adaptation.

Accordingly, this study is guided by three research questions: (RQ1) Can a lightweight TabNet-based model achieve more accurate score prediction than representative machine-learning and deep tabular baselines on a small-sample agricultural R&D project dataset? (RQ2) Can the model provide interpretable feature-attribution evidence that is useful for understanding score formation and informing indicator-system refinement? (RQ3) Does the proposed approach show preliminary usability in a real-world management scenario beyond the development dataset?

To address these challenges, this study proposes a performance evaluation approach for publicly funded agricultural research projects. Specifically, we: (i) construct a project-level evaluation dataset covering multiple provincial agricultural research institutions in China and develop a dedicated preprocessing and standardization pipeline; (ii) propose Light-TabNet, which improves predictive performance for project evaluation scores; and (iii) leverage the built-in interpretability mechanism of Light-TabNet to explain why projects receive specific scores and to provide actionable recommendations for subsequent indicator-system design.

This paper is organized as follows. Section 1 presents the background and contributions. Section 2 introduces the dataset and preprocessing pipeline, and then describes the proposed Light-TabNet together with its interpretability formulation for indicator recommendation. Section 3 provides comprehensive evaluations, including benchmark comparisons, ablation, interpretability results, and an external validation in a real management scenario. Section 4 concludes the study and outlines limitations and future directions.

2. Experimental Setup and Data

2.1. Overall Experimental Workflow

Figure 1 summarizes the overall workflow of our study for data-driven score prediction and indicator-design recommendation in support of performance evaluation tasks for publicly funded agricultural research projects. The pipeline consists of three stages: (i) public data collection and preprocessing, (ii) predictive model implementation and optimization, and (iii) interpretability analysis and indicator recommendation based on Light-TabNet.

2.1.1. Stage 1: Data Preparation

We first collect publicly available data and construct the dataset through cleaning and preprocessing.

2.1.2. Stage 2: Model Development

We then propose and integrate two improved modules within the TabNet framework—Tabular Dynamic Tanh Layer (TabDyT) and Light Feature Transformer (LightFT)—and demonstrate the reliability of our Light-TabNet model through comparative experiments and ablation studies.

2.1.3. Stage 3: Interpretation and Recommendation

Finally, by using Light-TabNet's native capability to aggregate feature contributions across different decision steps, we obtain the model's overall feature importance and then provide recommendations for future indicator design.

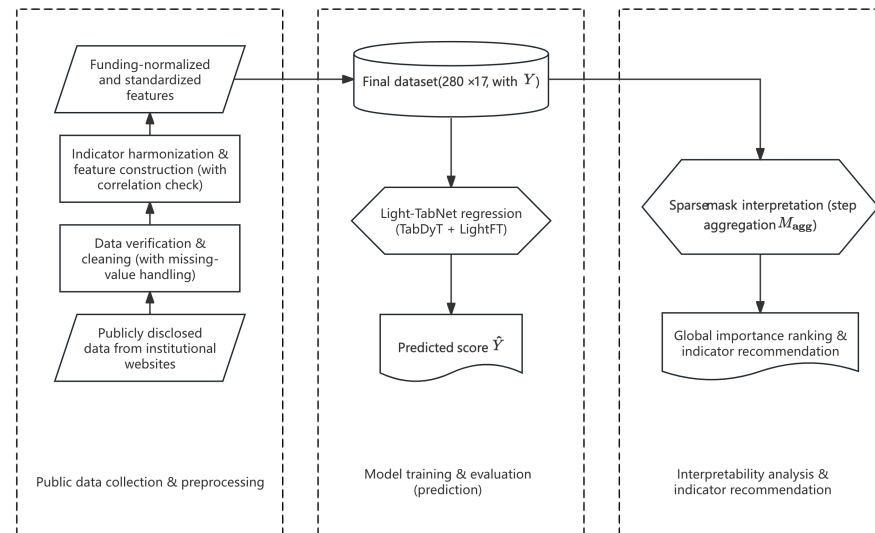


Figure 1. Overall workflow of this study.

2.2. Public Data Collection and Preprocessing

The data used in this study are obtained from publicly disclosed performance-evaluation information of agricultural research projects released on the official websites of provincial agricultural research institutions in China. After collecting the raw samples, we manually verified each record and removed entries that were evidently low-quality or violated common sense. To address inconsistencies in indicator definitions across institutions, we standardized semantically similar indicators and merged their quantities, thereby forming a unified indicator system.

For feature construction, the dataset contains 280 project samples. The input consists of 17 performance-evaluation indicators, and the output is the project self-evaluation score (Y). The 17 input indicators cover research outputs (e.g., publications, patents and software copyrights, standards and protocols, and new varieties/technologies/devices), platform and base construction, talent cultivation and academic exchange, awards, fund execution ratio, technology demonstration and dissemination, as well as achievement transformation, and beneficiary satisfaction. To mitigate scale differences induced by varying project budgets, all indicators are normalized by project funding and converted into the number of outputs per 100,000 CNY of funding.

Different projects may emphasize different indicators, and some indicators are missing for certain samples. In this study, we interpret missing indicators as the project not specifying or not emphasizing the corresponding assessment dimension, and we apply zero imputation to keep the modeling pipeline simple and consistent. Considering that some models are sensitive to feature scales, after zero imputation and funding normalization, we further standardize all input features to improve training stability and enhance comparability across models.

Regarding variable distributions, several indicators exhibit pronounced right-skewness and long-tail behavior, i.e., a small number of projects take substantially large values on some indicators after normalization; the fund execution ratio is generally high with rela-

tively low dispersion; and indicators related to technology demonstration, dissemination, and achievement transformation show higher dispersion, reflecting more substantial heterogeneity across project types in the “extension/transfer” dimension. We also conducted correlation analysis and found that overall linear correlations between feature pairs are weak, with no clustered high-correlation structure observed. This suggests a limited risk of severe pairwise multicollinearity; therefore, we retain the complete feature set for subsequent modeling, prediction, and interpretability analysis.

2.3. Model Training and Evaluation

Prior studies have shown that TabNet, an interpretable deep model for tabular data, has been applied to tasks such as insurance risk/pricing prediction, software cost estimation, and analysis of accident risk factors for agricultural machinery vehicles, and has demonstrated competitive performance in both predictive accuracy and interpretability [13–15].

Motivated by these findings, we adopt TabNet as the backbone and propose a lightweight variant, Light-TabNet, tailored to the project-performance regression setting characterized by “small samples + missing features”, and use it for subsequent experimental evaluation.

Light-TabNet consists of three components: (i) the TabNet decision-mask backbone (retaining the sequential decision process and sparse instance-wise feature selection to preserve interpretability); (ii) TabDyT (replacing the input BN to improve training stability under small batch sizes); (iii) LightFT (lightweighting the feature transformation module to reduce parameter count and alleviate overfitting risk in small-sample regimes).

2.3.1. TabNet

In the project performance evaluation task, it is necessary to capture nonlinear relationships between input features and project scores while maintaining a certain degree of interpretability. TabNet [16] is a deep model designed for tabular data, achieving interpretability via sequential decision steps and sparse feature-selection masks, and has shown competitive results on multiple tabular benchmarks. However, under the small-sample and partially missing feature conditions considered in this study, directly using the default TabNet configuration yields suboptimal performance (see Section 3).

2.3.2. Tabular Dynamic Tanh Layer (TabDyT)

In small-sample tabular settings with small batch sizes, input Batch Normalization (BN) relies on batch statistics and may introduce batch-to-batch fluctuations, thereby impairing training stability. To address this issue, we introduce a per-feature dynamic hyperbolic activation layer, TabDyT, at the input of TabNet. TabDyT is a variant of Dynamic Tanh (DyT) [17] adapted to tabular inputs, and is used to replace BN.

Given an input $\mathbf{x} \in \mathbb{R}^{B \times d}$, the forward computation of TabDyT is

$$\text{TabDyT}(\mathbf{x}) = \mathbf{x} + \gamma \odot \tanh(\alpha \odot \mathbf{x}) + \beta, \quad (1)$$

where $\alpha, \gamma, \beta \in \mathbb{R}^d$ are learnable per-feature parameters, and $\odot$ denotes element-wise multiplication.

Functionality and properties: (i) no batch statistics: the behavior is consistent between training and inference, avoiding variance fluctuations caused by small batches; (ii) bounded nonlinear squeezing: $\tanh(\cdot)$ suppresses extreme values and is approximately linear near the origin; (iii) per-feature adaptivity: α, γ, β enable feature-wise scaling and shifting; (iv) identity approximation and stable gradients: the residual path $\mathbf{x}$ preserves an information highway, facilitating an initial near-identity mapping and stable optimization.

In implementation, TabDyT is introduced specifically as a replacement for the input Batch Normalization used in the original TabNet encoder, while the remaining backbone structure and hyperparameter settings are kept unchanged. Therefore, the comparison between TabNet and TabNet + TabDyT in the ablation study is intended to isolate the effect of replacing the original input BN with TabDyT under the same architectural setting.

2.3.3. Light Feature Transformer (LightFT)

In TabNet, the feature transformer is typically composed of several sub-blocks shared across decision steps and several decision-step-dependent sub-blocks, stacked together to enhance the model's representation capability for complex tasks. However, in small-scale tabular-data scenarios, excessively deep transformation stacks may introduce too many parameters and thus increase the risk of overfitting.

To reduce model complexity while retaining the essential gating and feature-transformation capabilities, we propose LightFT, which keeps only one shared sub-block and one decision-step-dependent sub-block. Both sub-blocks adopt a sequential FC–GBN–GLU structure (FC: fully connected layer; GBN: Ghost Batch Normalization; GLU: Gated Linear Unit). A scaled residual connection is introduced within the decision-step-dependent sub-block, and an additional FC layer is appended at the end to linearly recombine the learned feature representations. Without altering the original decision-mask mechanism of TabNet, this design reduces the transformation depth and parameter scale, thereby alleviating the risk of overfitting under small-sample settings. To make this lightweight claim more explicit, the trainable parameter counts of different ablation variants are quantitatively reported in Section 3.

The structure of the proposed Light Feature Transformer is illustrated in Figure 2.

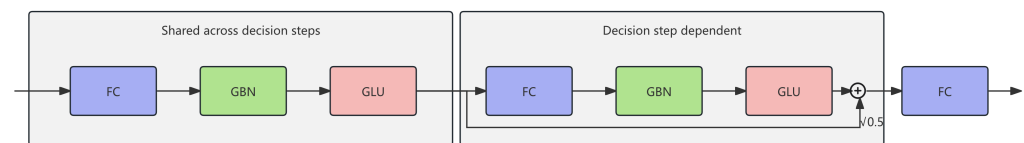


Figure 2. Light Feature Transformer (LightFT).

2.3.4. Light-TabNet Architecture

The overall architecture of Light-TabNet is illustrated in Figure 3.

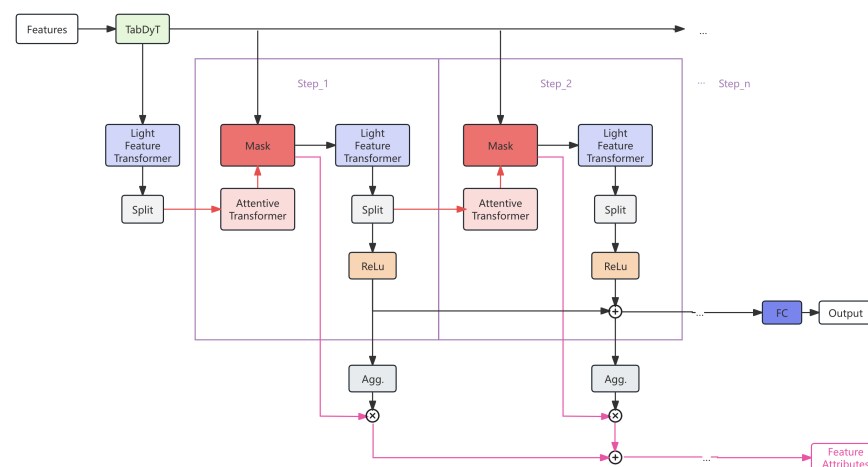


Figure 3. Overall architecture of Light-TabNet. TabDyT replaces input BN; per decision step uses a lightweight feature transformer (LightFT) and an attentive transformer with sparse instance-wise masks, whose contributions are aggregated to form the final regression output.

The model consists of multiple decision steps executed sequentially.

Given an input sample, Light-TabNet first applies TabDyT at the input stage to perform per-feature dynamic scaling and nonlinear compression. This operation does not rely on batch statistics, which helps mitigate fluctuations under small-batch training and enables feature-wise adaptation. The resulting representation serves as the unified input to the subsequent decision process.

Each decision step involves the following components. First, the information passed from the previous step is fed into the Attentive Transformer, which constructs a sparse set of feature weights via an attention mapping with a prior factor. A sparse normalization strategy is employed to ensure that the model focuses only on the most relevant features. Next, the input features are reweighted by this mask. One part of the mask is directly accumulated into the feature contribution scores to quantify feature importance at the current step, while the other part selects candidate features that are forwarded to LightFT. LightFT adopts a lightweight feature transformation module to extract the higher-order representations required for decision making. The output of LightFT is then split into two branches: one branch is activated by ReLU, accumulated into the overall decision vector, and mapped to the final prediction through a fully connected layer; the other branch is propagated as the input to the next decision step. Meanwhile, the ReLU-activated outputs at all steps are also aggregated into the feature contribution scores to produce the global explanation.

In addition, before the first decision step, the input features undergo an initial LightFT preprocessing and splitting operation to initialize the decision chain and the accumulation of feature contributions.

Through this step-wise decision process, Light-TabNet not only outputs the final performance score prediction but also provides the relative contributions of each feature to the evaluation for individual samples.

2.4. Interpretability Analysis and Indicator Recommendation

Light-TabNet employs a sparse feature selection mechanism and outputs feature attribution values. As shown in Figure 3, for any sample b , at the i -th decision step, the attentive module produces a feature mask $M_b[i]$, where $M_{b,j}[i]$ denotes the relative contribution weight of the j -th feature to sample b at this step. When $M_{b,j}[i] = 0$, the j -th feature can be interpreted as making no contribution to the decision at this step. The mask is normalized along the feature dimension, so $M_b[i]$ can be viewed as the “relative dependency strength on each feature” at step i .

Because different decision steps may contribute unequally to the final output, the step-wise masks should be aggregated with appropriate weights. Light-TabNet introduces a decision-step contribution coefficient $\eta_b[i]$:

$$\eta_b[i] = \sum_{c=1}^{N_d} \text{ReLU}(d_{b,c}[i]), \quad (2)$$

where $d_{b,c}[i]$ is the c -th component of the decision representation at step i , and N_d is the dimensionality of the decision representation. Intuitively, if the decision representation at a step is overall negative (and thus truncated by ReLU), that step should contribute zero to the final linear combination; a larger $\eta_b[i]$ indicates a higher weight of step i in the final prediction.

Accordingly, we define the aggregated feature-importance mask across decision steps as

$$M_{\text{agg-}b,j} = \sum_{i=1}^{N_{\text{steps}}} \eta_b[i] M_{b,j}[i], \quad (3)$$

and further normalize it to obtain the global relative contribution:

$$\tilde{M}_{\text{agg-}b,j} = \frac{\sum_{i=1}^{N_{\text{steps}}} \eta_b[i] M_{b,j}[i]}{\sum_{j=1}^D \sum_{i=1}^{N_{\text{steps}}} \eta_b[i] M_{b,j}[i]}. \quad (4)$$

Here, $M[i]$ reflects the feature-selection process of “step-wise decision making”, whereas M_{agg} (or its normalized form) characterizes the overall feature dependency structure after cross-step aggregation, thereby providing a traceable basis for explaining the model outputs.

After obtaining feature contributions, we further use them for recommending performance-indicator selection. On the one hand, we analyze $\tilde{M}_{\text{agg}}$ over historical projects to derive an overall importance ranking of indicators. On the other hand, we identify specific indicators that exhibit high local contributions only for certain project samples, thereby providing evidence for indicator specification in future projects of similar types.

3. Experimental Results and Analysis

This section evaluates the proposed Light-TabNet on the public agricultural R&D project dataset. We first describe the experimental setup, including the hardware/software environment, data split protocol, and evaluation metrics. We then compare Light-TabNet with representative deep tabular models and conventional machine learning baselines under the same preprocessing and evaluation pipeline. Next, we conduct an ablation study to quantify the contributions of TabDyT and LightFT. Finally, we present interpretability results based on the sparse decision masks and report a real-world external validation to demonstrate practical applicability. Unless otherwise specified, baseline methods are run using the recommended/default settings of their public implementations to ensure reproducibility.

3.1. Experimental Environment and Settings

All experiments were conducted on Ubuntu 24.04 with an Intel i5-8500 CPU, an RTX 5080 GPU (16 GB VRAM), and 24 GB RAM. The software environment includes Python 3.10.18, PyTorch 2.9.0, and CUDA 12.8.

The dataset was randomly split into training and test sets with an 8:2 ratio, and the same split was used for all compared models to ensure fairness. To prevent data leakage, preprocessing operations such as standardization were fitted only on the training set, and the learned transformations were then applied to the test set. In addition, to provide a more robust evaluation under the small-sample setting, we further conducted a five-fold cross-validation study with unified randomized search for representative conventional machine-learning baselines, as reported in Section 3.5.

3.2. Evaluation Metrics

The performance of our model was evaluated using the following four metrics: mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and the coefficient of determination (R^2). Lower MAE/MSE/RMSE and higher R^2 indicate better performance.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (5)$$

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (6)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (7)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (8)$$

3.3. Comparison of Deep Learning Models

We compared Light-TabNet with recent deep models designed specifically for tabular data, including Mambular [12], TabM [11], ResNet [18], NODE [9], MLP [19], Tangos [10], and ModernNCA [20]. All baselines were run using the default configurations of their public implementations. The results are reported in Table 1. On this small-sample dataset, Light-TabNet achieves the best performance across all four metrics (MAE, MSE, RMSE, and R^2). Compared with the strongest deep-learning baselines (NODE yields the best MAE, while Tangos yields the best MSE/RMSE/ R^2 among baselines), Light-TabNet reduces MAE by 23.12%, reduces MSE by 41.93%, reduces RMSE by 23.79%, and increases R^2 by 0.0800. These results indicate that, relative to heavier deep tabular models evaluated in this study, the streamlined design of Light-TabNet is more suitable and effective under the small-sample setting.

Table 1. Performance comparison of deep tabular models (MAE, MSE, RMSE, R^2 ; ↓ indicates lower is better and ↑ indicates higher is better).

Model	MAE ↓	MSE ↓	RMSE ↓	R^2 ↑
Mambular	9.0835	171.1310	13.0817	0.7557
TabM	9.5316	169.3347	13.0129	0.7583
ResNet	10.1438	203.4583	14.2639	0.7096
NODE	6.4731	138.5717	11.7716	0.8022
MLP	8.8383	153.3830	12.3848	0.7811
Tangos	8.2383	133.7740	11.5661	0.8091
ModernNCA	6.8336	157.4229	12.5468	0.7753
Light-TabNet (Ours)	4.9765	77.6874	8.8140	0.8891

3.4. Comparison of ML Models

We further compared Light-TabNet with commonly used conventional machine learning methods, including Linear Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), XGBoost [21], and LightGBM [22]. All models were implemented using Scikit-learn [23] or their official implementations, and the data split and evaluation protocol were identical to those in the previous subsection. As shown in Table 2, Light-TabNet consistently outperforms all classical baselines across the four metrics. Compared with the strongest conventional baseline, XGBoost, Light-TabNet reduces MAE by 15.08%, reduces MSE by 21.60%, reduces RMSE by 11.46%, and increases R^2 by 0.0305. Although traditional methods are often robust in small-sample regimes, the proposed lightweight optimizations enable Light-TabNet to surpass these conventional models under our evaluation setting. This performance provides preliminary empirical support for applying Light-TabNet as a decision-support tool in project performance evaluation tasks.

Table 2. Performance comparison between conventional machine learning models and Light-TabNet (MAE, MSE, RMSE, R^2 ; ↓ indicates lower is better and ↑ indicates higher is better).

Model	MAE ↓	MSE ↓	RMSE ↓	R^2 ↑
Linear Regression	10.9696	352.1820	18.7665	0.4973
Decision Tree	7.7863	203.4067	14.2621	0.7097
Random Forest	6.5418	126.2367	11.2355	0.8198
KNN	7.9352	224.7696	14.9923	0.6792
XGBoost	5.8602	99.0921	9.9545	0.8586
LightGBM	7.4194	114.8720	10.7178	0.8360
Light-TabNet (Ours)	4.9765	77.6874	8.8140	0.8891

3.5. Robustness and Fairness Analysis

To further examine the sensitivity of the results to missing-value handling, as well as the robustness of the proposed method and the fairness of baseline comparisons under the small-sample setting, we conducted two additional analyses: (1) a sensitivity analysis of different imputation strategies, and (2) a five-fold cross-validation study with a unified randomized-search protocol for representative conventional machine-learning baselines.

3.5.1. Sensitivity to Imputation Strategies

Because the dataset contains a non-negligible proportion of missing indicators, we further compared four imputation strategies, namely zero, mean, median, and KNN imputation. Table 3 reports the corresponding R^2 results for several conventional machine-learning models. The results show that the best imputation strategy is model-dependent, indicating that the influence of missing-value handling is non-negligible. However, for the two strongest tree-based baselines in our experiments, namely XGBoost and LightGBM, zero imputation achieves the best R^2 among the compared strategies.

Table 3. R^2 comparison of conventional machine-learning models under different imputation strategies.

Model	Zero	Mean	Median	KNN
Linear Regression	0.4973	0.4374	0.0889	0.5091
Decision Tree	0.7097	0.7038	0.5514	0.7503
Random Forest	0.8198	0.8267	0.8130	0.7599
KNN	0.6792	0.5059	0.5427	0.5474
XGBoost	0.8586	0.8364	0.8109	0.8184
LightGBM	0.8360	0.8102	0.8135	0.7644

From a task-semantic perspective, this observation is also consistent with the policy background of the dataset. In practical project evaluation documents, only the major indicators emphasized by a given project are typically listed explicitly. Therefore, for some indicators, non-reporting may partly reflect that the corresponding assessment dimension was not highlighted in that project, rather than purely random missingness. Under this interpretation, zero imputation avoids introducing artificial indicator values that may not belong to the project's actual assessment focus, whereas mean or median imputation may inject population-level values into project-specific indicator profiles. KNN imputation may better exploit local similarity across projects, but on this small and heterogeneous dataset, its empirical advantage is not consistently observed. Overall, the sensitivity analysis suggests that zero imputation is not only aligned with the task semantics of project-performance reporting, but also remains empirically competitive, especially for strong tabular baselines such as XGBoost and LightGBM. Therefore, we retain zero imputation in

the main experiments while explicitly acknowledging that alternative imputation strategies may still be worth exploring in future work.

3.5.2. Five-Fold Cross-Validation with Unified Randomized Search

To obtain a more robust and reliable comparison under the small-sample setting, we further evaluated several representative machine-learning baselines using five-fold cross-validation and a unified randomized-search protocol. The results are summarized in Table 4 in the form of mean $\pm$ standard deviation. Compared with the results obtained from a single 8:2 split, the cross-validation results are overall more conservative, which suggests that the performance estimates from a single partition may be affected by partition-specific variation. Nevertheless, Light-TabNet still achieves the best average performance among the compared models across all four metrics, with an MAE of 5.7300 ± 0.6338 , an MSE of 88.4510 ± 43.1257 , an RMSE of 9.1900 ± 2.2344 , and an R^2 of 0.7771 ± 0.0943 .

Table 4. Five-fold cross-validation results with unified randomized search (mean $\pm$ standard deviation; $\downarrow$ indicates lower is better and $\uparrow$ indicates higher is better).

Model	MAE $\downarrow$	MSE $\downarrow$	RMSE $\downarrow$	R^2 $\uparrow$
Decision Tree	7.3813 ± 1.0221	171.2960 ± 56.1543	12.9347 ± 2.2335	0.5050 ± 0.2274
Random Forest	6.5627 ± 0.8614	106.8092 ± 32.7939	10.2446 ± 1.5238	0.7086 ± 0.0678
KNN	7.0719 ± 0.7887	139.3577 ± 15.4909	11.7901 ± 0.6637	0.6020 ± 0.1112
XGBoost	6.0729 ± 1.0966	98.0508 ± 30.7189	9.8007 ± 1.5798	0.7022 ± 0.1551
LightGBM	6.4174 ± 1.3011	105.8683 ± 40.2812	10.1297 ± 2.0176	0.6819 ± 0.1879
Light-TabNet (Ours)	5.7300 ± 0.6338	88.4510 ± 43.1257	9.1900 ± 2.2344	0.7771 ± 0.0943

These results indicate that the superiority of Light-TabNet is not solely dependent on one favorable random split, but can still be observed under a more robust evaluation protocol with repeated training and testing. This additional analysis is intended to provide a more robust comparison under a unified evaluation protocol, rather than an exhaustive hyperparameter-optimization benchmark for all competing methods. A broader hyperparameter-benchmarking study with larger search spaces and more repeated trials remains an important direction for future work.

3.6. Ablation Study

In order to better understand the impact of the proposed improvements, we conducted an ablation study using four configurations on the same dataset. Specifically, we evaluated: (i) the baseline TabNet; (ii) TabNet with LightFT; (iii) TabNet with TabDyT; and (iv) the full Light-TabNet combining both LightFT and TabDyT. The results are reported in Table 5. Notably, since the original TabNet uses Batch Normalization at the encoder input, the comparison between TabNet and TabNet + TabDyT can be interpreted as a controlled substitution experiment between the default input BN and the proposed TabDyT, with the rest of the architecture kept unchanged.

Table 5. Ablation study results of Light-TabNet; $\downarrow$ indicates lower is better and $\uparrow$ indicates higher is better.

Model	Params	MAE $\downarrow$	MSE $\downarrow$	RMSE $\downarrow$	R^2 $\uparrow$
TabNet	9.834 K	6.317	112.275	10.596	0.8398
TabNet + LightFT	4.829 K	6.524	110.614	10.517	0.8421
TabNet + TabDyT	9.885 K	5.810	89.888	9.481	0.8717
Light-TabNet (Ours)	4.829 K	4.976	77.687	8.814	0.8891

As shown in Table 5, the full Light-TabNet achieves the best results across all metrics while maintaining a relatively small parameter count. Compared with the baseline TabNet, introducing LightFT reduces the number of trainable parameters from 9.834 K to 4.829 K, which supports the lightweight design claim. Although TabNet + LightFT alone yields only limited performance improvement, it substantially reduces model complexity. By contrast, TabNet + TabDyT increases the parameter count only marginally (from 9.834 K to 9.885 K) but improves all evaluation metrics, suggesting that TabDyT mainly enhances optimization stability rather than increasing model capacity. When both components are combined, Light-TabNet attains the best predictive performance with only 4.829 K trainable parameters, indicating that the observed gain is not merely due to a larger model size, but to the effectiveness of the proposed architectural modifications under the current setting.

3.7. Feature Interpretability

As illustrated in Figure 4, the model exhibits clear sparse feature-selection behavior at each decision step: only a limited subset of features becomes salient for certain samples, while the remaining features are largely suppressed at that step. This indicates that Light-TabNet does not rely uniformly on all indicators; instead, it forms a traceable decision path through sequential feature selection.

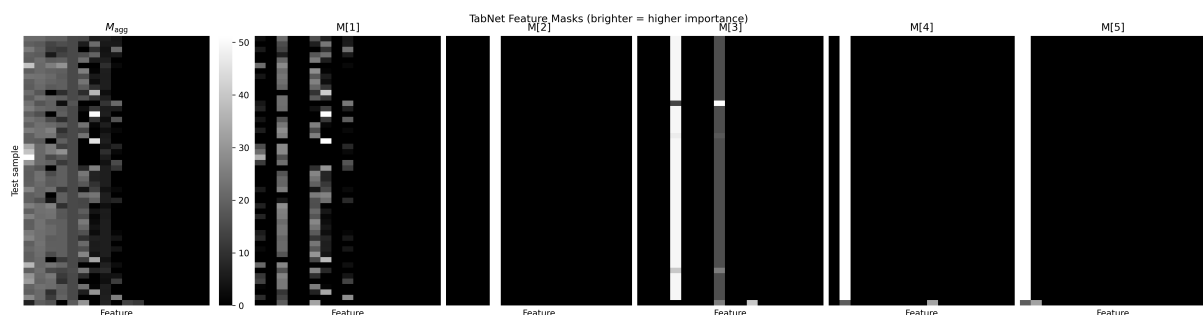


Figure 4. Visualization of feature-mask interpretability in Light-TabNet. Each row corresponds to a test sample, and each column corresponds to an input feature; brighter colors indicate a higher relative contribution (mask weight) for that feature in the given sample. The left panel shows the aggregated global importance M_{agg} across decision steps; the five panels on the right show the instance-wise masks at decision steps 1–5, respectively.

From the aggregated mask M_{agg} , two broad patterns can be observed. First, features that remain relatively bright across many samples may reflect a more general dependence of the model on these indicators. Second, features that become salient only for a limited subset of samples may capture conditional effects associated with specific project contexts or subtypes. Together with the step-wise masks $M[i]$, these results provide an interpretable view of how feature usage evolves from earlier coarse screening to later fine-grained refinement.

Based on the global ranking in Figure 5, the model assigns relatively high importance to both research-output indicators and process-management indicators. For example, “new varieties/technologies/devices” and “publications” are among the most influential output-related features, while “fund execution ratio” also receives a high importance weight.

To provide a cross-model interpretability reference, Table 6 compares the top-ranked features identified by Light-TabNet with their corresponding ranks in the mean absolute SHAP ranking of an XGBoost model. The comparison shows partial consistency across model families. In particular, three indicators, namely “fund execution ratio”, “publications”, and “new varieties/technologies/devices”, were consistently ranked among the top contributors by both methods, although their exact ordering differed. Moreover, several additional indicators, such as “patents & software copyrights”, “reports/research/material

collection/tests”, “talent cultivated/introduced”, and “standards/protocols/systems”, also appeared at relatively high positions in both rankings. These overlaps provide convergent evidence for a small set of core indicators in the current dataset.

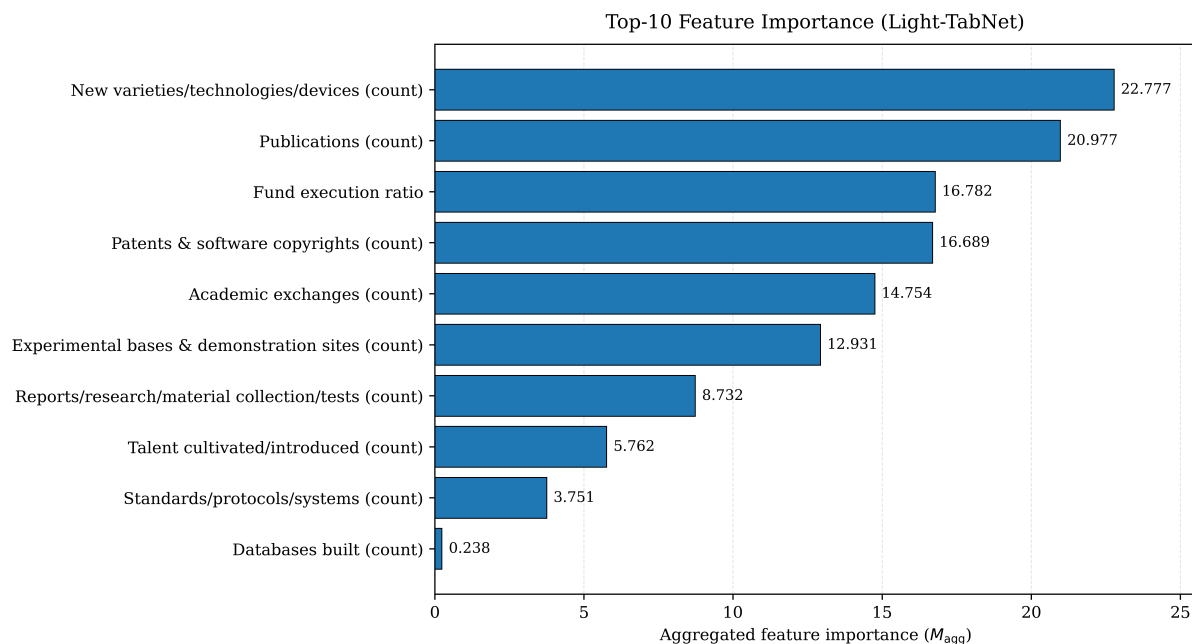


Figure 5. Top-10 feature importance ranked by the aggregated mask M_{agg} of Light-TabNet.

Table 6. Cross-model comparison of top-ranked features between Light-TabNet and XGBoost-SHAP.

ID	Light-TabNet Rank	SHAP Rank	Shared Top-10	Feature Name
8	1	3	Yes	New varieties/technologies/devices (count)
5	2	2	Yes	Publications (count)
18	3	1	Yes	Fund execution ratio
6	4	7	Yes	Patents & software copyrights (count)
15	5	13	No	Academic exchanges (count)
10	6	11	No	Experimental bases & demonstration sites (count)
11	7	6	Yes	Reports/research/material collection/tests (count)
12	8	10	Yes	Talent cultivated/introduced (count)
7	9	9	Yes	Standards/protocols/systems (count)
9	10	15	No	Databases built (count)

At the same time, noticeable discrepancies remain for several mid-ranked and lower-ranked features. For example, Light-TabNet assigns relatively higher importance to “academic exchanges” and “experimental bases and demonstration sites”, whereas XGBoost-SHAP gives more weight to “service recipient satisfaction”, “achievement transformation contracts”, and “technology demonstration/promotion/guidance”. A plausible explanation is that, although all samples belong to agricultural R&D projects, the dataset still contains heterogeneous project subtypes, such as basic research, breeding, pest-control studies, technology extension, and talent-oriented projects. Under such heterogeneous conditions, some fine-grained performance indicators may be more sensitive to specific project categories, which can lead to differences in feature ranking across models. As the sample size increases and project-type annotation becomes more refined in future work, it may become possible to conduct stratified modeling or subgroup interpretability analysis at a finer granularity, thereby enabling more precise examination of indicator weights and functional roles across different types of agricultural R&D projects.

Overall, the mask-based ranking of Light-TabNet is better interpreted as a traceable, model-based interpretability result that provides exploratory evidence on potential indicator relevance, rather than as a definitive policy-priority ranking. The partial agreement with XGBoost-SHAP suggests that several core indicators may have relatively stable importance in the current dataset, while the remaining discrepancies indicate that feature-importance rankings should be interpreted with caution in the absence of further expert validation or stability analysis.

3.8. Real-World External Validation

To further examine the practical applicability of Light-TabNet, we conducted a real-world external validation using eight agricultural R&D projects (anonymized as A–H) from a provincial agricultural research institute. Importantly, these validation cases were drawn from the institute’s 2024 performance-evaluation cycle, whereas the model in the previous sections was trained and tested on publicly collected 2023 project data. Therefore, the external-validation set does not overlap with the training or test data used in the main experiments.

Although the institutional context is similar to that of the main dataset, this setting is still meaningful for practical validation. The purpose of this experiment is not to demonstrate broad cross-domain generalization, but to assess whether the learned mapping between project performance indicators and final evaluation scores can remain usable in a temporally independent yet realistically relevant management scenario. In addition, these eight projects were not arbitrarily selected cases; rather, they correspond to the complete set of agricultural R&D projects of that institute in 2024. Under the local budgeting and project-management framework, these eight major projects were formed by aggregating 128 sub-projects. Thus, the validation reflects the annual performance output of an entire institute rather than a few isolated cases.

For each project, we compared the score predicted by Light-TabNet with the score assigned by a third-party evaluation organization. The project-wise comparison is shown in Table 7. To provide a more rigorous quantitative assessment, we further computed several error-based statistics. Across the eight projects, MAE, MSE, and RMSE were 2.9838, 15.0627, and 3.8811, respectively, while the mean bias was 0.9762, indicating that the predictions were overall close to the third-party ratings.

Table 7. Project-wise comparison and absolute error (0–100 scale).

Project	Third-Party Score	Light-TabNet	$ \Delta $
A	99.00	98.82	0.18
B	95.00	97.87	2.87
C	99.00	96.57	2.43
D	90.00	98.76	8.76
E	95.00	95.39	0.39
F	99.00	96.01	2.99
G	91.89	95.71	3.82
H	100.00	97.57	2.43

From the perspective of practical agreement, 75.0%, 87.5%, and 100.0% of the projects fell within tolerance intervals of ± 3 , ± 5 , and ± 10 points, respectively. Moreover, for 7 out of the 8 projects, the absolute error was below 4 points, suggesting that the predicted scores were generally close to the third-party ratings. The only relatively large deviation occurred in Project D, with an absolute error of 8.76 points. After case inspection, this discrepancy appears to be associated mainly with score deductions related to budget-preparation compliance rather than with poor completion of performance targets themselves. This

result suggests that the proposed model captures the relationship between performance indicators and evaluation scores reasonably well, while also indicating that some institution-specific scoring factors beyond indicator completion may still affect the final rating.

Overall, this external validation provides preliminary evidence that Light-TabNet has practical potential in real agricultural research-management scenarios. At the same time, the sample size remains limited, and the validation cases come from a similar institutional setting. Therefore, these findings should be interpreted cautiously as an initial real-world feasibility check rather than definitive proof of broad external generalizability. Future work will further extend the validation to larger-scale, multi-institutional, and cross-regional datasets.

4. Conclusions

In this study, we proposed a project performance evaluation method named Light-TabNet, which integrates TabDyT and a lightweight feature transformer within TabNet. This method employs a per-feature Dynamic Tanh activation in place of input batch normalization, alleviating small-batch instability and facilitating stable optimization. By incorporating a lightweight feature transformer composed of one shared and one decision-step-dependent FC–GBN–GLU block into the TabNet backbone, it improves the modeling of nonlinear interactions under limited data compared to selected baselines. Furthermore, the model preserves instance-wise sparse decision masks, which maintain the interpretability of feature selection during prediction. Experimental results demonstrate that the Light-TabNet model proposed in this paper is more capable of capturing complex relationships in agricultural R&D project data compared to other models, producing higher predictive accuracy.

Despite these promising results, several limitations remain. Our dataset currently contains 280 valid samples and is skewed toward traditional research-oriented projects, with relatively limited coverage of technology transfer and talent development, which may constrain the learning and interpretation of such patterns. In addition, the proposed model should be viewed as a decision-support tool rather than a replacement for expert-based evaluation, especially in practical settings where institution-specific rules, contextual judgments, and governance considerations remain important. Moreover, residual missingness and bias may still affect model performance even after manual screening, and the use of default hyperparameters for some baselines may underestimate their fully tuned upper bounds.

In future work, we will further expand the dataset in both scale and project-type diversity, and perform validation across multiple institutions and regions to better assess the generalizability of the proposed approach. We will also conduct more systematic robustness analyses under alternative preprocessing strategies. In addition, we plan to carry out broader comparative benchmarking under more harmonized hyperparameter tuning budgets to provide a fairer assessment of model performance across competing methods.

Author Contributions: Conceptualization, Z.L., A.W. and L.F.; methodology and software, Z.L. and H.L.; validation, Z.L., L.F. and H.L.; formal analysis and investigation, Z.L. and L.F.; data curation and resources, Z.L. and L.F.; writing—original draft preparation, Z.L.; writing—review and editing, Q.C. and Z.L.; visualization and project administration, Z.L.; supervision, Q.C.; funding acquisition, Z.L. and L.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study were collected from publicly available performance-evaluation information released on the official websites of provincial agricultural research institutions in China. The processed dataset supporting the findings of this study is available from the first author upon reasonable request.

Acknowledgments: The authors would like to thank the authors of all references used in the paper, the editors, and the anonymous reviewers for their detailed comments and suggestions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. European Commission. Standard Briefing Slides for Experts (Horizon Europe). Official Expert Briefing for Horizon Europe Evaluations. 2023. Available online: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/experts/standard-briefing-slides-for-experts_he_en.pdf (accessed on 9 September 2025).
2. National Science Foundation. Merit Review at NSF: Intellectual Merit and Broader Impacts. NSF Official Page Describing Review Criteria and Panel Process. 2025. Available online: <https://new.nsf.gov/funding/merit-review> (accessed on 9 September 2025).
3. UK Research and Innovation (MRC). Guidance for Peer Reviewers. MRC Guidance Stating Peer Review as the Cornerstone of Funding. 2021. Available online: <https://www.ukri.org/wp-content/uploads/2021/08/MRC-0208212-Guidance-for-peer-reviewers-March-2021.pdf> (accessed on 9 September 2025).
4. Australian Research Council. The Peer Review Process—Assessment Process. ARC Assessment Criteria, Scoring, and Peer-Review Procedures. 2025. Available online: <https://www.arc.gov.au/funding-research/peer-review/assessment-process> (accessed on 9 September 2025).
5. Japan Science and Technology Agency. SATREPS: Guidelines for JST Terminal Evaluation. Terminal Evaluation Procedures Including Site Surveys and Expert Meetings. 2023. Available online: <https://www.jst.go.jp/global/english/kadai/hyouka/pdf/end-evaluation-procedure.pdf> (accessed on 9 September 2025).
6. Garefalakis, S.; Angelaki, E.; Spinthiropoulos, K.; Tsamis, G.; Garefalakis, A. The Implementation of ESG Indicators in the Balanced Scorecard—Case Study of LGOs. *Risks* **2025**, *13*, 154. [CrossRef]
7. Shim, H.; Kim, J. A Study on Project Prioritisation and Operations Performance Measurements by the Analysis of Local Financial Investment Projects in Korea. *Sustainability* **2023**, *15*, 5972. [CrossRef]
8. Shin, D.J.; Cha, B.S.; Kim, B.H. Efficient Expenditure Allocation for Sustainable Public Services?—Comparative Cases of Korea and OECD Countries. *Sustainability* **2020**, *12*, 9501. [CrossRef]
9. Popov, S.; Morozov, S.; Babenko, A. Neural Oblivious Decision Ensembles for Deep Learning on Tabular Data. *arXiv* **2019**, arXiv:1909.06312. [CrossRef]
10. Jeffares, A.; Liu, T.; Crabbé, J.; Imrie, F.; van der Schaar, M. TANGOS: Regularizing Tabular Neural Networks through Gradient Orthogonalization and Specialization. In Proceedings of the The Eleventh International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
11. Gorishniy, Y.; Kotelnikov, A.; Babenko, A. TabM: Advancing Tabular Deep Learning with Parameter-Efficient Ensembling. *arXiv* **2025**, arXiv:2410.24210.
12. Thielmann, A.F.; Kumar, M.; Weisser, C.; Reuter, A.; Säfken, B.; Samiee, S. Mambular: A Sequential Model for Tabular Deep Learning. *arXiv* **2025**, arXiv:2408.06291.
13. McDonnell, K.; Murphy, F.; Sheehan, B.; Masello, L.; Castignani, G. Deep learning in insurance: Accuracy and model interpretability using TabNet. *Expert Syst. Appl.* **2023**, *217*, 119543. [CrossRef]
14. Alhumam, A. Software cost estimation using TabNet and Harris Hawks Optimization. *Sci. Rep.* **2025**, *15*, 45434. [CrossRef] [PubMed]
15. Islam, M.M.; Liu, J.; Chakraborty, R.; Das, S. Evaluating crash risk factors of farm equipment vehicles on county and non-county roads using interpretable tabular deep learning (TabNet). *Accid. Anal. Prev.* **2025**, *217*, 108048. [CrossRef] [PubMed]
16. Arik, S.Ö.; Pfister, T. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021*; Association for the Advancement of Artificial Intelligence: Washington, DC, USA, 2021; Volume 35, pp. 6679–6687.
17. Zhu, J.; Chen, X.; He, K.; LeCun, Y.; Liu, Z. Transformers without Normalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 14901–14911.
18. Gorishniy, Y.; Rubachev, I.; Khrulkov, V.; Babenko, A. Revisiting deep learning models for tabular data. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 18932–18943.
19. Kadra, A.; Lindauer, M.; Hutter, F.; Grabocka, J. Well-tuned simple nets excel on tabular datasets. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 23928–23941.

20. Ye, H.J.; Yin, H.H.; Zhan, D.C.; Chao, W.L. Revisiting Nearest Neighbor for Tabular Data: A Deep Tabular Baseline Two Decades Later. *arXiv* **2025**, arXiv:2407.03257.
21. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
22. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3149–3157.
23. Fabian, P. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.