Modern Neural Networks for Small Tabular Datasets: The New Default for Field-Scale Digital Soil Mapping?

Viacheslav Barkov[a,b], Jonas Schmidinger[a,b], Robin Gebbers[b] and Martin Atzmueller[a,c]

[a]Osnabrück University, Joint Lab Artificial Intelligence and Data Science, Osnabrück, Germany
[b]Leibniz Institute for Agricultural Engineering and Bioeconomy (ATB), Department of Agromechatronics, Potsdam, Germany
[c]German Research Center for Artificial Intelligence (DFKI), Research Department Cooperative and Autonomous Systems (CAS), Osnabrück, Germany

ARTICLE INFO

Keywords:
Machine Learning
TabPFN
Benchmarking
Artificial Neural Networks
Digital Soil Mapping
Pedometrics

ABSTRACT

In the field of pedometrics, tabular machine learning is the predominant method for predicting soil properties from remote and proximal soil sensing data, forming a central component of digital soil mapping. At the field-scale, this predictive soil modeling (PSM) task is typically constrained by small training sample sizes and high feature-to-sample ratios in soil spectroscopy. Traditionally, these conditions have proven challenging for conventional deep learning methods. Classical machine learning algorithms, particularly tree-based models like Random Forest and linear models such as Partial Least Squares Regression, have long been the default choice for field-scale PSM. Recent advances in artificial neural networks (ANN) for tabular data challenge this view, yet their suitability for field-scale PSM has not been proven. We introduce a comprehensive benchmark that evaluates state-of-the-art ANN architectures, including the latest multilayer perceptron (MLP)-based models (TabM, RealMLP), attention-based transformer variants (FT-Transformer, ExcelFormer, T2G-Former, AMFormer), retrieval-augmented approaches (TabR, ModernNCA), and an in-context learning foundation model (TabPFN). Our evaluation encompasses 31 field- and farm-scale datasets containing 30 to 460 samples and three critical soil properties: soil organic matter or soil organic carbon, pH, and clay content. Our results reveal that modern ANNs consistently outperform classical methods on the majority of tasks, demonstrating that deep learning has matured sufficiently to overcome the long-standing dominance of classical machine learning for PSM. Notably, TabPFN delivers the strongest overall performance, showing robustness across varying conditions. We therefore recommend the adoption of modern ANNs for field-scale PSM and propose TabPFN as the new default choice in the toolkit of every pedometrician.

1. Introduction

Soil maps are relevant for a range of environmental and agricultural applications (Keesstra et al., 2016). Especially at the field-scale, they play a crucial role in supporting high crop yields while considering negative environmental impacts (Gebbers and Adamchuk, 2010; Viscarra Rossel and Bouma, 2016). Sensors enable rapid and cost-effective generation of high-resolution soil data needed for creating these soil maps. This includes off-site sensing in the laboratory or direct in-situ measurements in field-conditions through proximal-soil sensing (PSS) or remote sensing (RS) (Viscarra Rossel and Bouma, 2016). However, these sensors do not directly measure actual soil properties relevant for agriculture but proxies related to them (Gebbers, 2019). The translation of sensor signals into useful soil property estimates requires a set of soil samples to train a site-specific prediction model. This predictive soil modeling (PSM) framework falls in the field of pedometrics and is used for digital soil mapping (McBratney et al., 2003, 2019).

Tabular machine learning (ML) has emerged as the predominant approach in contemporary PSM (Heuvelink and Webster, 2022). However, the application of ML for field-scale PSM faces specific challenges, distinguishing it from other ML domains. These challenges include high noise, complexity of soil as a medium, variable measurement conditions, differences in measurement footprints and spatial autocorrelation. Most importantly, training sample sizes are generally rather low as soil sampling and soil analyses in laboratory rapidly becomes cost- and labor-prohibitive (Söderström et al., 2016). This creates a fundamental tension: while sample sets cannot be too small as this would result in poorly fitted prediction models, economic considerations constrain the size of training sample sets (Schmidinger et al., 2024). Additionally, the feature dimensionality in PSM can vary drastically from very low dimensional data generated by common in-situ PSSs like electrical conductivity sensors to very high dimensional data obtained from mid- and near-infrared spectroscopy. In soil spectroscopy, the feature dimensionality usually exceeds the number of samples, making modeling prone to overfitting (dos Santos et al., 2023; Wang and Wang, 2022; Wang et al., 2022), a challenge known as the curse of dimensionality (Bellman et al., 1957).

These dataset characteristics favored classical ML approaches like tree-based algorithms and linear models, which became the default choices for PSM (Wadoux et al., 2020; Barra et al., 2021; Ding et al., 2025). In particular, Random Forest dominates due to its straightforward application, good performance with default hyperparameters (Probst et al., 2019), fast training times, and general effectiveness even with low training sample sizes (Ma et al., 2020; Schmidinger et al., 2024). Recent reviews consistently report Random Forest as the most frequently used algorithm for PSM (Ding et al., 2025; Wadoux et al., 2020). In soil spectroscopy applications specifically, linear models, usually combined with linear feature transformations such as principal component analysis (PCA) or partial least squares, have established themselves as fundamental methods (Barra et al., 2021). They demonstrate robust performances in contemporary spectroscopy studies, even when evaluated alongside popular ML algorithms such as Random Forest or Artificial Neural Networks (ANN) (e.g. Xue et al., 2023; Schmidinger et al., 2025).

Deep learning applications in the form of various ANN architectures, notably Multilayer Perceptrons (MLP) and Convolutional Neural Networks, have attracted increasing interest for PSM (Wadoux, 2025), but have mostly shown success for contexts where abundant training data is available. This includes large-area soil mapping at a regional up to continental level utilizing RS (e.g. Yang et al., 2021) or vast spectral libraries (e.g. Ng et al., 2019). However, this success did not extend to field- or smaller regional-scale applications where sample sizes are commonly lower, with ANNs showing relatively poor performances (e.g. Xue et al., 2023; Oukhattar et al., 2025). Moreover, even in large-area applications, deep learning did not always show improvements over classical methods (e.g. Sun et al., 2023). These limitations, particularly in smaller area high resolution soil mapping, are further emphasized by the methodological comparison of Khaledian and Miller (2020), who argue that classical ML methods provide more stable results than ANNs in cases of limited training data. In fact, Khaledian and Miller (2020) explicitly recommend using classical ML algorithms when working with fewer than 1,000 soil samples. Overall, the effectiveness of deep learning for PSM remains uncertain, and it has not replaced classical ML methods.

The historical dominance of classical ML for PSM aligns with broader trends in the tabular data domain, where classical ML approaches have maintained a competitive edge over ANNs until recent years. Following deep learning's widespread success in computer vision and natural language processing, early attempts to apply ANNs to tabular data generated considerable optimism, with several studies reporting promising results (Arik and Pfister, 2021; Katzir et al., 2020). However, subsequent benchmarking studies revealed that classical ML, particularly tree-based ensemble methods like gradient-boosted decision trees (GBDTs), consistently outperformed these early deep learning approaches (Shwartz-Ziv and Armon, 2022). In fact, Shwartz-Ziv and Armon (2022) observed that the early claims of ANN superiority were largely driven by limited benchmarking practices, with studies relying on favorable results from a narrow selection of datasets, failing to generalize across broader evaluations. This was subsequently confirmed by other comprehensive tabular benchmarks (Grinsztajn et al., 2022; McElfresh et al., 2023). Notably, these benchmarks confirmed classical ML superiority in conditions with limited sample sizes and high dimensionality. McElfresh et al. (2023) observed that the performance gap between tree-based models and ANNs widened considerably as the dataset size decreased, specifically noting that ANNs performed comparatively worse when the feature-to-sample ratio became larger.

Recent developments in deep learning for tabular data challenge this perspective. Over the past year, multiple ANN approaches claiming to surpass classical ML methods have been proposed for tabular data (Hollmann et al., 2025; Holzmüller et al., 2024; Gorishniy et al., 2025; Ye et al., 2025c; Cheng et al., 2024). In contrast to previous deep learning models, recent tabular benchmarks also support this shift (Erickson et al., 2025; Ye et al., 2025b). These advances extend far beyond incremental improvements to the classically used ANN architectures like MLPs. Entirely new tabular ANN approaches have emerged, including attention-based models, retrieval-based approaches, and in-context learning foundation models (Ye et al., 2025a). We refer to these recent architectural innovations, from the past year, collectively as modern ANN approaches throughout this manuscript. These modern ANNs hold particular promise for addressing PSM-specific challenges. In-context learning models such as TabPFN (Hollmann et al., 2025) claim to achieve particularly strong performance on small datasets, making them potentially strong choices for the small-area PSM domain. Retrieval architectures such as ModernNCA (Ye et al., 2025c) exploit sample neighborhood relationships, and may offer unique advantages for soil property prediction given the spatially correlated nature of soil data.

Preliminary ML benchmarking studies showed impressive performance on public domain-independent benchmarks with these modern ANN approaches (Ye et al., 2025a; Erickson et al., 2025; Hollmann et al., 2025; Ye et al., 2025b). However, these studies did not adequately represent the specific constraints of field-scale PSM, as their dataset properties differed substantially from those encountered in field-scale PSM (see Appendix C). For example, datasets in the benchmark of Erickson et al. (2025) and Ye et al. (2025a) only included datasets with at least 500 samples. While this sample size is considered small in the broader ML context, it still largely exceeds the maximum sample size commonly available in field-scale PSM (Schmidinger et al., 2025). Additionally, the aspect of high-dimensionality and unfavorable feature-to-sample ratios is often dismissed or marginally discussed. Consequently, there remains a crucial need to examine the applicability and efficacy of these advanced deep learning paradigms in the context of the inherent constraints of field-scale PSM.

Applying and comparing these promising architectures in field-scale PSM is non-trivial. As the broader tabular data literature already warns (Shwartz-Ziv and Armon, 2022), proper evaluation of the new deep learning methods is itself a significant methodological challenge, and inadequate benchmarking practices can substantially affect conclusions about their relative performance. Methodological problems of previous tabular benchmarking studies have been largely overlooked in pedometrics, leading to critical limitations that may produce incomplete or misleading results (Schmidinger et al., 2025). Recent work by Schmidinger et al. (2025) revealed that over 95% of PSM benchmarking studies relied on a single dataset, with the largest number of datasets used in any reviewed study being only three. As demonstrated by Shwartz-Ziv and Armon (2022), reliance on a single dataset may inadvertently result in overinterpretation of incidental findings, obscuring the actual strengths and limitations of deep learning methods for PSM. Other confounding factors, like the choice of hyperparameters, can also introduce significant bias (Nießl et al., 2022). In some PSM studies, hyperparameters of ANNs were carefully selected, whereas the competing tree-based models have not been tuned or no information has been given (e.g. Ng et al., 2019; Yang et al., 2021). Given that fewer than 10% of PSM benchmarking studies provided open datasets and less than 5% shared their code (Schmidinger et al., 2025), it is usually not possible to reproduce and verify the results. These shortcomings highlight the critical need for comprehensive benchmarking studies in PSM that systematically assess novel methods across multiple diverse datasets in a fair and rigorous way. Different training regimes, parameter optimization approaches, validation strategies, and other methodological choices must be implemented consistently to ensure fair comparisons (Nießl et al., 2022).

In this work, we benchmark modern deep learning methods that have shown promise in tabular data applications but remain unexplored in the domain of PSM. Our study makes several key contributions. We are the first to systematically introduce and assess state-of-the-art deep learning methods for PSM at the field-scale with its distinct dataset challenges regarding sample size and feature dimensionality. Second, we establish a fair evaluation framework with consistent hyperparameter optimization, training regimes, and validation strategies across all methods. Third, we conduct experiments across multiple diverse field-scale soil datasets, avoiding the single-dataset evaluation dominating in previous benchmarking studies in pedometrics. Finally, we ensure complete reproducibility by making all datasets, code, and experimental configurations publicly available (see Code and data availability). Through this analysis, we aim to provide empirical evidence on whether these advanced deep learning methods can overcome the longstanding dominance of classical ML in pedometric applications, particularly under the small-sample constraints typical for farm-scale PSM.

2. Materials and methods

2.1. Datasets

We utilized Precision Liming Soil Datasets (LimeSoDa), a collection of 31 field- to farm-scale datasets spanning multiple countries and diverse agricultural contexts, representing a broad spectrum of soil mapping scenarios encountered in precision agriculture (Schmidinger et al., 2025). Their global distribution is illustrated in Figure 1. Each dataset contains three target soil properties: soil organic matter (SOM) or soil organic carbon (SOC), pH, and clay content, yielding a total of 93 regression tasks for model evaluation. Individual dataset sizes range from 30 to 460 samples, providing varied data scenarios typical of real-world PSM projects for small-area soil mapping. All datasets are openly accessible and have been standardized in tabular format to facilitate reproducible research through the LimeSoDa repository (see Code and data availability).

Features are dataset-specific and were obtained by different sensing technologies, including laboratory-based spectroscopy, in-situ PSS, and RS. This includes a variety of common sensor modalities, such as apparent electrical resistivity, gamma-ray spectrometry, ion selective electrodes, digital elevation models, multispectral RS data, vegetation indices, and X-ray fluorescence derived elemental concentrations. Several datasets incorporate high-dimensional optical spectroscopy data from visible and near-infrared (vis-NIR), near-infrared (NIR), and mid-infrared (MIR) spectrometers. These measurements result in feature sets with up to 2,489 variables, as the reflectance is typically recorded across numerous continuous wavelength bands. There are no categorical features present across any of the datasets. See Appendix C for comparisons to feature sizes and feature-to-sample ratios reported in prior benchmarks.

Given the computational challenges posed by high-dimensional spectral data, we categorized the datasets into two groups based on feature characteristics and dimensionality for our analysis (Figure 2).

The first group, referred to as "High-Dimensional", includes datasets with feature-to-sample ratios >1, containing vis-NIR, NIR, or MIR spectroscopy features. These datasets require dimensionality reduction because the unfavorable feature-to-sample ratio would otherwise lead to severe overfitting (Schmidinger et al., 2025). Dimensionality reduction was performed during preprocessing as described in Section 2.4.3.

The second group, referred to as "Low-Dimensional", includes datasets with lower-dimensional PSS features (e.g. apparent electrical resistivity) or RS features (e.g. vegetation indices). All of these datasets have a feature-to-sample ratio <1.

2.2. Classical machine learning models

We included several algorithms that have demonstrated strong performance in PSM as classical ML baselines. Specifically, we evaluated Linear Regression (including regularized variants), Partial Least Squares Regression (PLSR), Random Forest, and XGBoost. Implementation details are provided in Appendix E.

Linear Regression is a statistical method that models the relationship between features and target variables using a linear equation. When multiple independent features are used, it is defined as Multiple Linear Regression (MLR). Linear Regression, specifically MLR, serves as a fundamental baseline in PSM despite its simplicity, often serving as a primary baseline for comparison against other ML algorithms (Wadoux, 2025; Ding et al., 2025; Wadoux et al., 2020). In addition to MLR, we included its regularized variants in our evaluation. Specifically, we included Lasso regression (MLR with L1 norm regularization) and Ridge regression (MLR with L2 norm regularization).

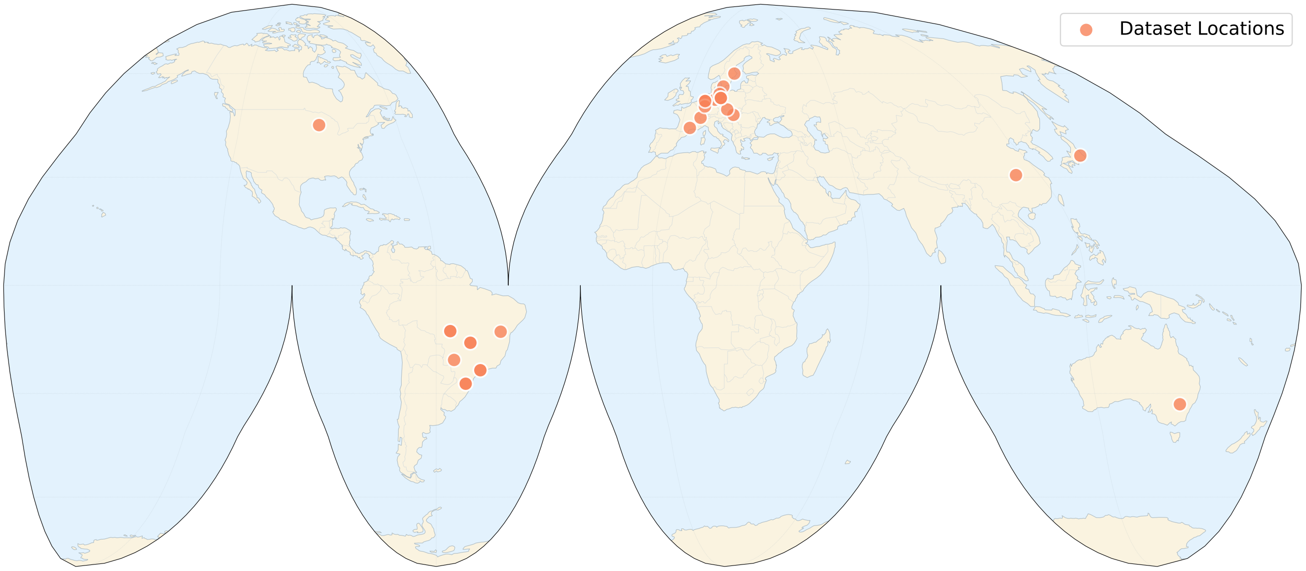


Figure 1: Global distribution of the 31 field- and farm-scale datasets used in this study. Background map is based on Homolosine projection. The datasets span multiple continents and diverse agricultural contexts, including locations across North and South America, Europe, Asia, and Australia.

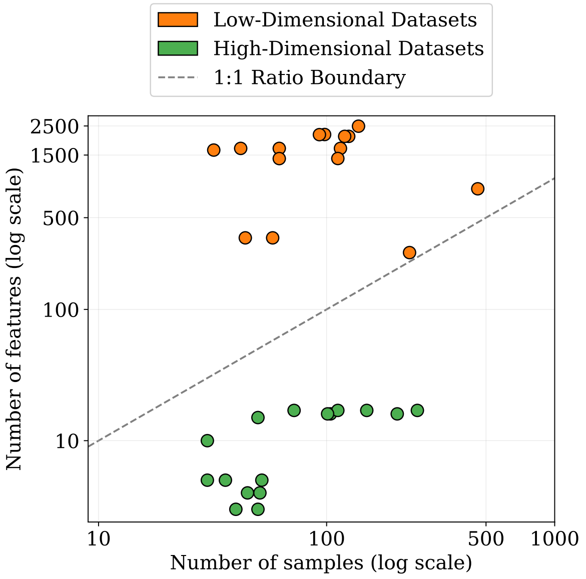


Figure 2: Distribution of datasets by number of features and samples (log scale), highlighting feature-to-sample ratios. Each point represents one dataset, with two groups shown: Low-Dimensional datasets (lower feature counts) and High-Dimensional datasets (higher feature counts due to spectral features).

PLSR is a statistical method that projects both features and target variables into a lower-dimensional latent space that maximizes their covariance and then fits a linear regression, aiming to enable modeling under severe multicollinearity and very large feature-to-sample ratios. We employed PLSR for High-Dimensional group of datasets, which requires dimensionality reduction prior to model fitting. For all other models we applied PCA as a preprocessing step for these datasets (see Section 2.4.3).

Random Forest (Breiman, 2001) is an ensemble of decision trees (DT) that uses bootstrap aggregating (bagging) to reduce the variance of a single DT by employing a randomized ensemble. It is the most widely adopted ML algorithm in PSM (Ding et al., 2025; Wadoux et al., 2020). Wadoux et al. (2020) found that 80% of reviewed studies employed at least one tree-based algorithm, with Random Forest being the most popular.

XGBoost (Chen and Guestrin, 2016) is a popular implementation of GBDTs that has exhibited strong performance in many tabular data domains, including PSM. Unlike Random Forest, which trains DTs in parallel, GBDTs work by sequentially training DTs, with each new tree correcting the residuals of the previous ensemble. The optimization is performed by minimizing a differentiable loss function using gradient descent. GBDTs, and XGBoost in particular, have historically outperformed classical ML and early neural network approaches in general tabular regression benchmarks (Grinsztajn et al., 2022), although seeing more limited application in PSM compared to the widespread adoption of Random Forest (Wadoux, 2025).

2.3. Artificial neural networks

The landscape of deep learning for tabular data has evolved substantially in recent years, with various architectural innovations attempting to bridge the performance gap with traditional tree-based methods (Hollmann et al., 2025; Holzmüller et al., 2024; Gorishniy et al., 2025, 2021; Ye et al., 2025c). We included four groups of modern ANN architectures for tabular data, selecting established baseline models from each group alongside recent improvements introduced within the past year. References and details on code, implementation and dependencies of ANN models employed in this study are described in Appendix E.

2.3.1. MLP-based models

MLP represents a classical ANN baseline for tabular data. MLPs are not new to the domain of PSM and have been used and evaluated in previous studies, although generally underperforming compared to classical ML methods in field-scale PSM. We included a standard MLP as a baseline along with two recent architectural improvements, RealMLP and TabM.

MLP architecture consists of fully connected layers with non-linear activation functions. As noted by Gorishniy et al. (2025) and further elaborated by Holzmüller et al. (2024), MLPs remain competitive baselines for tabular data, often matching or outperforming more complex attention- and retrieval-based architectures when evaluated under consistent experimental protocols. According to Gorishniy et al. (2025), "MLPs [. . .] form a line of stronger and more practical models compared to attention- and retrieval-based architectures".

TabM (Gorishniy et al., 2025) extended the MLP architecture with new scaling parameter-efficient layers. The idea is to produce multiple predictions per input sample by training specialized weight-sharing components within a single model, retaining most parameters in common while still maintaining diversity through distinct prediction pathways.

RealMLP (Holzmüller et al., 2024) took a different approach, with contributions not limited to model architecture. Holzmüller et al. (2024) introduced an improved MLP version employing neural tangent parametrization, parametric Mish activation functions, and specialized scaling layers, with an important contribution of their own training regime and pre- and post-processing pipeline. The whole range of these contributions, which are not limited to ML architectural decisions, may significantly improve MLPs in the context of tabular data.

2.3.2. Retrieval-based models

We included retrieval-based models that utilize learned embedding spaces to find relevant training examples for guiding predictions, drawing inspiration from classical nearest neighbor approaches. We evaluated TabR, which pioneered this concept, and ModernNCA, introduced last year as an advancement of the approach.

TabR (Gorishniy et al., 2024) implemented this concept by searching for K nearest neighbors in a learned embedding space, computing their contributions based on both feature and label representations. These contributions are then aggregated with attention weights determined by distances in embedding space, effectively creating a differentiable variant of k-nearest neighbors with adaptive representation learning.

ModernNCA (Ye et al., 2025c) improved on the retrieval-based approach of TabR by revisiting Neighborhood Components Analysis (NCA) in a modern context. ModernNCA learns a deep embedding with a soft nearest-neighbor objective and predicts by distance-weighted averaging of neighbor labels in the learned embedding space. Training is made scalable with stochastic neighborhood sampling that evaluates neighbors against a random subset per mini-batch and the full set at inference.

2.3.3. Attention-based models

Inspired by the success of the transformer architecture in natural language processing (Vaswani et al., 2017), a range of attention-based models has emerged for tabular data, seeking to explicitly model feature interactions via self-attention. We included AutoInt as an early baseline, FT-Transformer as a more principled adaptation of the transformer architecture, and recent incremental improvements including T2G-Former, AMFormer, and ExcelFormer.

AutoInt (Song et al., 2019) was among the first to successfully adopt multi-head self-attention in a tabular setting. Each feature is treated as a separate token, and multiple self-attention layers capture higher-order interactions. Residual connections preserve the original feature representations, while deep stacking of attention layers iteratively refines the learned relationships among features.

FT-Transformer (Gorishniy et al., 2021) provided a more principled adaptation of the transformer architecture to tabular data. Introducing a Feature Tokenizer module, it encodes numerical and categorical inputs into embeddings suitable for transformer processing. Unlike the original transformer, FT-Transformer introduced feature-specific biases and embedding schemes for numerical data. The rest of the architecture remains close to the standard transformer, featuring multi-head self-attention and feed-forward blocks.

Successive works on attention-based models have introduced multiple incremental upgrades to FT-Transformer. T2G-Former (Yan et al., 2023) incorporated feature-relation graphs to guide attention, while AMFormer (Cheng et al., 2024) integrated additive and multiplicative operations into the attention blocks, accounting for arithmetic feature interactions that commonly appear in tabular data. Meanwhile, ExcelFormer (Chen et al., 2024) presented a transformer-based architecture with semi-permeable attention mechanisms, gating modules, and also contributed an improved training protocol, introducing new tabular data augmentation techniques.

2.3.4. In-context learning

We evaluated TabPFN, a foundation model that leverages in-context learning and fundamentally differs from traditional ANN approaches by eliminating the need for dataset-specific training.

TabPFN (Hollmann et al., 2025) introduced the concept of Prior-data Fitted Network (PFN) for tabular data, pretraining on a vast collection of synthetic datasets sampled from a carefully designed prior. At inference time, TabPFN processes an entire dataset, including both training and test samples, in a single forward pass, effectively performing in-context learning. This approach is claimed to be particularly effective for small to medium-sized datasets (up to 10,000 samples), where TabPFN could achieve state-of-the-art performance without any dataset-specific training. Since TabPFN was originally released as a proof-of-concept classifier without regression support (Hollmann et al., 2023), its initial application for PSM became possible only after Barkov et al. (2024) introduced a target discretization approach to adapt it to regression tasks. The most recent TabPFN version (Hollmann et al., 2025) introduced several improvements, including a native regression capability, but has not been employed or evaluated for PSM prior to the present study.

2.4. Experimental design

2.4.1. Fair architectural comparison

Our experimental protocol aimed to provide a fair comparison between different models. Recent advances in tabular deep learning often combine architectural innovations with complementary methodological improvements. For example, ExcelFormer introduces custom tabular-specific data augmentations, while RealMLP (Holzmüller et al., 2024) presents a comprehensive "bag-of-tricks" that extends well beyond architectural decisions to include specialized preprocessing (e.g. robust scaling with smooth clipping), postprocessing (e.g. output clipping), specific training procedures and best-epoch selection strategies. While these contributions are valuable, they are mostly model-agnostic and can theoretically be applied to any approach. We focused on unifying the evaluation protocol by either applying these improvements universally to every model, or excluding them from evaluation. This follows established conventions of fair comparison in tabular deep learning benchmarks, e.g. Gorishniy et al. (2025) who excluded custom data augmentations when evaluating against ExcelFormer performance to maintain experimental consistency.

Specifically, we applied robust scaling and universal numerical embeddings (Gorishniy et al., 2021) to all models whenever possible. We used robust scaling for data preprocessing. We excluded interpolation-based augmentations employed by ExcelFormer. We did not include output clipping implemented in the postprocessing of RealMLP, as we do not want to prevent the models from extrapolating beyond the training data target variable range. We also employed systematic early stopping with a separate validation set. We provide further details of these decisions in the following sections. These choices allow us to benefit from modern training insights while preserving a balanced playing field for our model comparisons.

2.4.2. Numerical feature embeddings

PSM commonly employs continuous numerical features derived from proximal and remote sensors. Traditionally, deep learning models represent these features as raw scalars. However, Gorishniy et al. (2022) demonstrated that first transforming each scalar into a trainable vector, which they refer to as a numerical feature embedding, can significantly improve the performance of a plain MLP and allows even simple ANNs to become competitive with GBDTs on tabular benchmarks. As a consequence, modern tabular ANNs such as RealMLP, TabM, TabR and ModernNCA all implement numerical embeddings (usually piecewise-linear embeddings or their derivatives) as a core element of the network architecture. To eliminate confounding factors, we adopted a single embedding scheme for all models in this study that employ numerical feature embeddings. We used a piecewise-linear embedding module with a data-dependent number of quantile bins, implemented according to the reference code of Gorishniy et al. (2022).

2.4.3. Data preprocessing

We employed a robust scaling procedure that prevents the outliers from affecting the inlier scaling. Specifically, we preprocessed each numerical feature as follows. Let $x_1, \dots, x_n \in \mathbb{R}$ be the values of a given numerical column k and let q_p be the p-quantile of $(x_1, \dots, x_n)$ for $p \in [0, 1]$.

For every entry j in the column k we set

$$\tilde{x}_j = s_k\big(x_j - q_{1/2}\big),$$
$$s_k = \begin{cases} \dfrac{1}{q_{3/4} - q_{1/4}}, & \text{if } q_{3/4} \neq q_{1/4}, \\ 1, & \text{otherwise.} \end{cases}$$

Thus each feature is first centred by subtracting the median $q_{1/2}$ and then scaled by the reciprocal interquartile range (IQR) whenever the IQR is non-zero. If the IQR vanishes, the scale factor is set to one to do centering only, avoiding division by zero.

For datasets whose spectroscopic feature dimensionality exceeds the number of samples we additionally applied PCA. PCA combined with MLR has repeatedly shown competitive performance in soil spectroscopy (Barra et al., 2021; Schmidinger et al., 2025). Schmidinger et al. (2025) identified PCA as the most effective preprocessing strategy when compared to using raw spectra or correlation-based feature selection in soil spectrometry. We therefore transformed the High-Dimensional datasets with PCA before model fitting, while keeping the raw variables for Low-Dimensional datasets. The number of retained components was treated as a hyperparameter and was optimized jointly with the model-specific parameters during the search described in Section 2.4.5.

All scaling statistics and PCA transformations were computed from the training set within each inner and outer fold to prevent data leakage.

2.4.4. Validation

We applied nested cross-validation with 5 outer folds for evaluation and 5 inner folds for hyperparameter selection, consistent with common ML practices in PSM (Kasraei et al., 2021).

Following previous works, (e.g. Ye et al., 2025c; Gorishniy et al., 2024), we used patience-based early stopping on the inner validation set, while keeping the inner fold test set strictly for scoring. Prior studies typically used a patience threshold of 16 consecutive epochs without validation set improvement. We extended the threshold to 40 epochs to account for the very small dataset nature in field-scale PSM, since each epoch iteration would include fewer training samples. Some architectures, such as RealMLP, train for a fixed number of epochs, typically 256, and then select the best epoch according to the validation score, which effectively mirrors a higher-patience early-stopping scheme. Following this, for final model training, we used 256 training epochs with the best epoch determined by outer validation set, while keeping the outer fold test set strictly for final evaluation. This allows us to keep the advantages of patience-based stopping serving as a run pruning strategy during hyperparameter search, and have a final model trained to escape potential early stopping patience selection bias.

For early stopping, we used Root Mean Squared Error (RMSE):

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

where y_i are the true values, $\hat{y}_i$ are the predicted values, and n is the number of samples.

For performance evaluation and ranking, we used R^2 (coefficient of determination):

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

where $\bar{y}$ is the mean of the true values.

We ranked models based on their R^2 scores for each regression task, and then averaged these ranks across datasets to obtain the overall performance metrics.

2.4.5. *Training and hyperparameter optimization*

For hyperparameter optimization, we employed the tree-structured Parzen Estimator (TPE) algorithm (Bergstra et al., 2011) with 100 rounds of optimization and an initial 20-round random search warm-up. This approach follows established best practices in recent tabular deep learning research. For example, Gorishniy et al. (2025) used TPE with 50-100 iterations across different model configurations. Our choice of 100 TPE rounds aligns with their findings, as the authors showed that 50-100 iterations typically provide adequate convergence for tabular models. We specifically utilized the multivariate TPE implementation (Ozaki et al., 2022) which accounts for dependencies between hyperparameters and has been shown to outperform independent parameter optimization in complex neural architectures.

We optimized parameters of all models (including classical ML algorithms) except for TabPFN, which does not require hyperparameter tuning. For ANN architectures, we explored hyperparameters such as the number of layers, hidden dimensions, dropout rates, and additional parameters dedicated to attention or retrieval mechanisms. For tree-based methods (XGBoost, Random Forest), we tuned maximum depth, subsampling ratios, and other regularization parameters. The number of trees was set to 1,000 for Random Forest, and XGBoost employed early stopping. The detailed hyperparameter grids can be found in Appendix D.

For datasets containing high-dimensional spectral features (vis-NIR, NIR, or MIR), we applied PCA and searched over 2 to 32 principal components, incorporating the choice of the number of components directly into the hyperparameter search. This consistent dimensionality reduction step was applied to all models to mitigate overfitting when the feature-to-sample ratio was highly unfavorable. Specifically for PLSR, we searched for the ideal number of components using the same range as for PCA.

All ANNs were trained using AdamW as the optimizer with mean squared error (MSE) loss. All random seeds were fixed to ensure reproducibility. We trained with a constant learning rate, which was optimized during hyperparameter search.

2.4.6. *Deep ensembles*

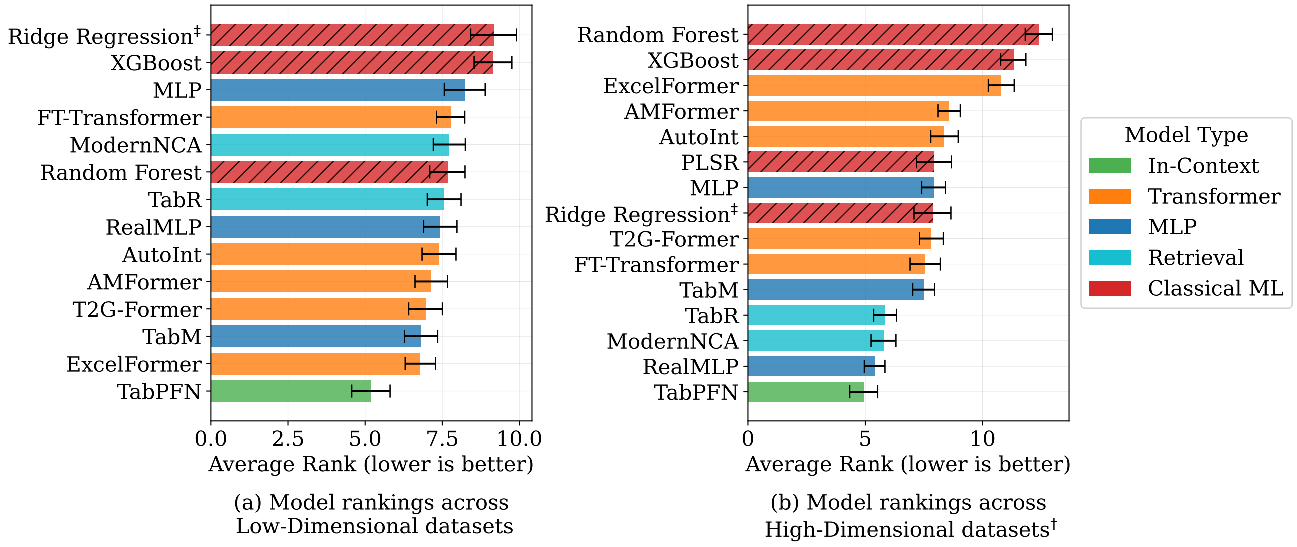
We adopted deep ensembles in our evaluation framework. State-of-the-art ANN methods routinely report deep ensemble results rather than single ANN performance. For instance, FT-Transformer and TabM report results as ensembles by averaging predictions from models trained with 15 different random seeds. Similarly, ExcelFormer uses 5 runs with different seeds per dataset, while TabPFN enforces ensembling by default using 8 estimators trained with feature permutations. Beyond ensuring fair evaluation, deep ensembles have strong theoretical foundations, with recent work demonstrating that independently trained ANNs improve test accuracy when data exhibits multi-view structure (Allen-Zhu and Li, 2023) and that diversity among ensemble members is crucial for generalization (Jeffares et al., 2023b).

Furthermore, in the PSM context with extremely small datasets, we propose deep ensembles as a methodological solution to address data scarcity. Early stopping with a separate validation set requires sacrificing another portion of already limited training data. In our setup, this meant withholding 20% of the data from training set within both inner and outer folds (see Section 2.4.5). Deep ensembles enable training multiple models not only with different weight initialization and sample batching, but also on different validation splits, effectively ensuring that no single subset of data is excluded from training the deep ensemble while maintaining proper validation procedures for each individual model.

For fair evaluation across all methods, we fixed the ensemble size to 16 members for all ANN models. We conducted ablation studies to further analyze the contribution of deep ensembles to overall performance of ANNs in a PSM context (see Appendix A).

3. Results

The following results summarize model performance across 93 regression tasks spanning 31 datasets. The Low-Dimensional group comprised 16 datasets and 48 regression tasks, while the High-Dimensional group comprised 45 regression tasks from datasets with vis-NIR, NIR, and MIR spectroscopy features.



(a) Model rankings across Low-Dimensional datasets

(b) Model rankings across High-Dimensional datasets†

†*Models in the High-Dimensional group other than PLSR use PCA for dimensionality reduction*

‡*Best ranked linear model out of Linear Regression, Lasso Regression, Ridge Regression*

Figure 3: Performance comparison of ANN architectures versus classical ML baselines on PSM tasks. Models are grouped by architectural approach: MLP-based (MLP, TabM, RealMLP), Attention-based (AutoInt, T2G-Former, ExcelFormer), Retrieval-based (TabR, ModernNCA), and In-Context Learning (TabPFN). Lower rank values indicate better performance. Error bars show ±1 standard error of the mean of each model's rank across the datasets. The in-context learning model TabPFN consistently achieved the best performance across both Low-Dimensional and High-Dimensional datasets.

The performance of classical ML approaches and ANN architectures assessed through average R^2 ranks is shown in Figure 3. Lower rank values indicate better performance, with the best possible rank being 1 and the worst being 14 for the Low-Dimensional group and 15 for the High-Dimensional group. Since we employed different regularization schemes for MLR (see Section 2.2), the figure demonstrates the best-ranking linear model for each group. A detailed comparison of MLR, its regularized variants, and PLSR is provided in Figure B1 of Appendix B.

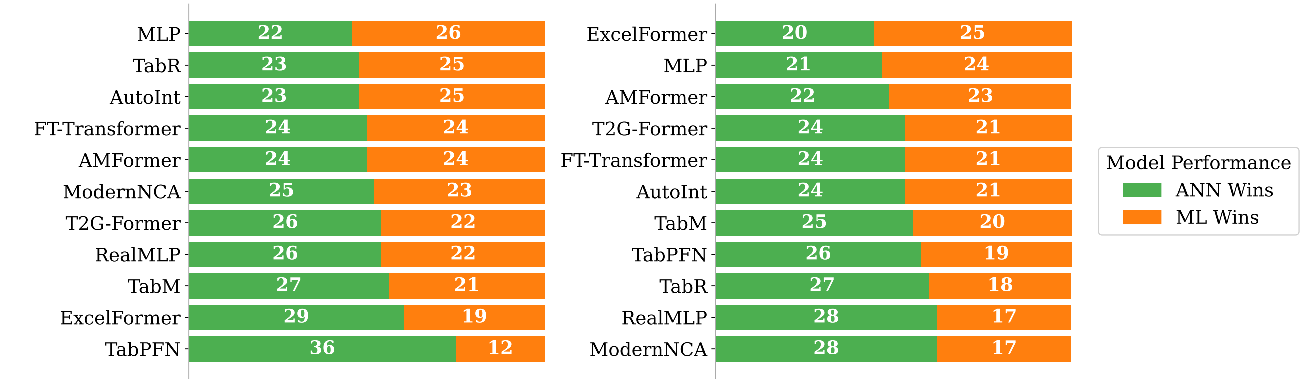
In the Low-Dimensional group, the in-context learning model TabPFN achieved the best overall performance. The strongest classical ML baseline was Random Forest. The attention-based ExcelFormer and MLP-based TabM emerged as the second-best performers. Other attention-based models, including T2G-Former, AMFormer, and AutoInt, also showed strong performance, outperforming Random Forest. However, three models failed to outperform Random Forest: MLP, FT-Transformer, and ModernNCA.

For High-Dimensional datasets, we additionally evaluated PLSR, while all other models in the High-Dimensional group used PCA as a preprocessing step (see Section 2.4.3). The best-performing classical ML baseline for this group was MLR with Ridge penalty. While ANNs generally showed weaker absolute performance on these PCA-preprocessed datasets, most modern models still outperformed classical methods. TabPFN, similar to the Low-Dimensional group, showed the strongest performance. RealMLP ranked second after TabPFN, and retrieval-based models TabR and ModernNCA demonstrated notably stronger performance than on Low-Dimensional datasets. MLP and most attention-based models underperformed when compared to MLR with Ridge regularization.

Figure 4 presents head-to-head comparisons between each ANN and the best classical baseline for its respective group. TabPFN achieved a 75% win rate against Random Forest on Low-Dimensional tasks (36 wins, 12 losses). Other strong performers included ExcelFormer, TabM, RealMLP, and T2G-Former. On High-Dimensional tasks, three models achieved 62% win rates against Ridge Regression: TabR, ModernNCA, and RealMLP. TabPFN followed closely with a 58% win rate. Standard MLP consistently underperformed across both groups.

Figure 5 reveals insights about the sample size–dependent performance. To examine this, we partitioned datasets into three categories based on sample size: ≤50 samples, 51–120 samples, and >120 samples. We employed two-stage ranking to avoid bias from varying numbers of models per model group. It shows that TabPFN maintained superior performance across nearly all scenarios, with classical ML surpassing it only for High-Dimensional datasets containing ≤50 samples. Retrieval-based models showed clear improvement with increasing dataset size in both groups, while MLP-based models exhibited stable performance regardless of size. Attention-based models showed degraded performance when PCA preprocessing was applied.

Overall, these results demonstrate that modern deep learning approaches, and TabPFN in particular, broadly surpass classical ML on the majority of field-scale PSM tasks.

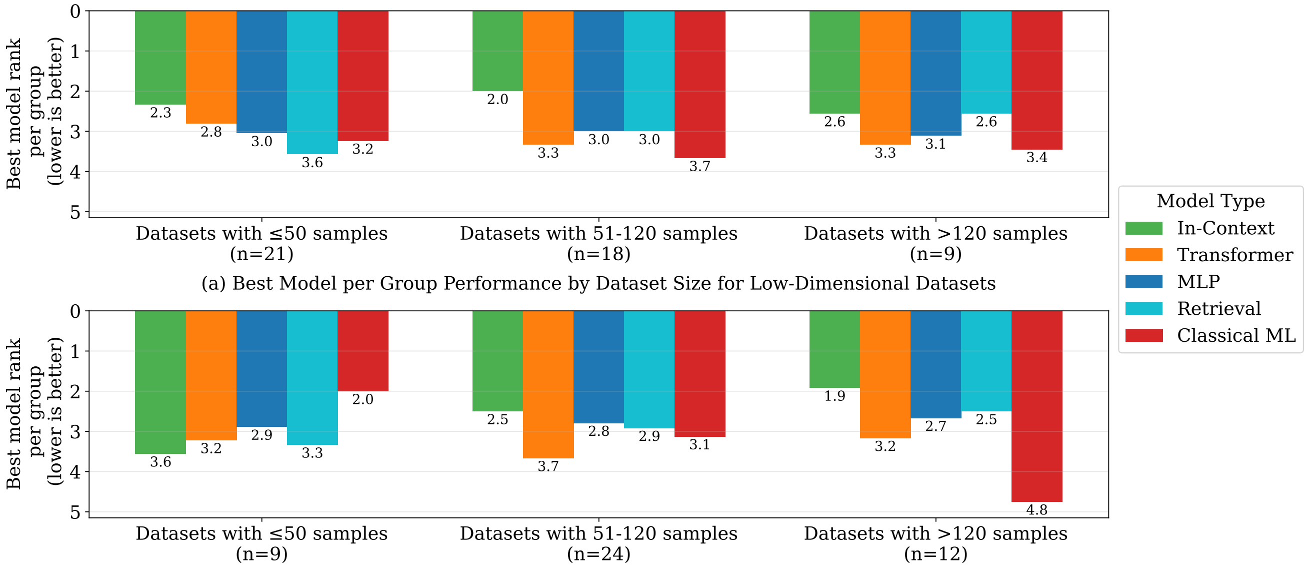


(a) ANNs perfomance against Random Forest baseline on Low-Dimensional datasets

(b) ANNs perfomance against Ridge Regression baseline on High-Dimensional datasets†

†*Models in the High-Dimensional group use PCA for dimensionality reduction*

Figure 4: Head-to-head comparison of ANN architectures against the best-performing classical ML baseline for each dataset group. For Low-Dimensional datasets, models are compared against Random Forest; for High-Dimensional datasets, against Ridge Regression. A higher number of wins indicates superior performance. Modern deep learning approaches, particularly TabPFN, substantially outperformed classical baselines on the majority of tasks.



(a) Best Model per Group Performance by Dataset Size for Low-Dimensional Datasets

(b) Best Model per Group Performance by Dataset Size for High-Dimensional Datasets

Figure 5: Model performance rankings grouped by dataset size, demonstrating the relationship between sample size and model effectiveness. Rankings are computed at the architectural group level using a two-stage process: first identifying the best model within each group, then ranking groups by their best model's performance. In-context learning (TabPFN) maintained superior performance across nearly all dataset sizes, with classical ML only prevailing on the smallest High-Dimensional datasets with PCA preprocessing.

4. Discussion

4.1. Deep learning versus classical machine learning

Our results demonstrated, that modern ANNs outperformed classical ML methods on the majority of prediction tasks in LimeSoDa. While our results align with recent tabular benchmarking studies (Ye et al., 2025b; Erickson et al., 2025), we extend their conclusions to the small-sample regime inherent to field-scale PSM, using datasets with as few as 30 samples(Table C1). These small-sample conditions were historically perceived as unfavorable for ANNs (Khaledian and Miller, 2020; McElfresh et al., 2023), making our findings particularly notable for field-scale PSM.

The consistent performance of modern ANNs across both Low-Dimensional and High-Dimensional datasets directly challenges the established dominance of Random Forest and Linear Regression in pedometrics. Nonetheless, our results do not contradict previous findings that favored classical ML over ANN (Shwartz-Ziv and Armon, 2022; Grinsztajn et al., 2022; McElfresh et al., 2023; Shmuel et al., 2024) as standard MLP still underperformed compared to classical ML algorithms in field-scale PSM. The critical distinction between our findings and earlier ANN evaluations lies in modern ANN architectures that either enhance the MLP framework through architectural innovations (TabM, RealMLP) or introduce fundamentally different mechanisms such as attention (ExcelFormer, T2G-Former), retrieval (TabR, ModernNCA), or in-context learning (TabPFN).

Although ANNs outperformed classical ML on both Low-Dimensional and High-Dimensional datasets, the performance gap was notably smaller for High-Dimensional data. One key distinction in our High-Dimensional experiments was the use of PCA for dimensionality reduction, which was a straightforward solution to handle the unfavorable feature-to-sample ratios characteristic of spectroscopic data. While PCA has proven robust and superior to using full raw spectral data in LimeSoDa (Schmidinger et al., 2025), alternative dimensionality reduction strategies based on feature transformation (e.g. autoencoders (Rumelhart et al., 1986)), feature reduction (e.g., stochastic gates (Yamada et al., 2020)) or ensemble-based feature subsampling (Ye et al., 2025b) could further enhance ANN performance. However, comprehensive evaluation of such techniques for soil spectroscopy warrants dedicated investigation beyond the scope of this study. Regardless of these avenues for improvement, ANNs maintained their performance advantage and demonstrated general compatibility with dimensionality reduction techniques like PCA. This versatility across different data characteristics represents another key strength of selecting modern ANN architectures for PSM. It is a key advantage over tree-based models, which struggle with linearly transformed features such as those produced by PCA (Menze et al., 2009; Grinsztajn et al., 2022; Schmidinger et al., 2025).

Our analysis extended the preliminary LimeSoDa benchmark of Schmidinger et al. (2025), which excluded ANNs due to their historically poor performance in tabular benchmarks. Their conclusion that different models excel in different scenarios remains partially valid. For example, we observed classical ML to be superior in the case of High-Dimensional datasets with extremely limited samples (≤50), as shown in Figure 5. However, beyond this exception, modern ANNs demonstrated remarkable consistency across various field-scale PSM scenarios. TabPFN in particular maintained superior performance across nearly all dataset sizes and dimensionalities.

4.2. On the success of TabPFN

TabPFN, a pioneer in tabular in-context learning, consistently achieved the highest performance across our experiments, emerging as the top-ranked model for both Low-Dimensional and High-Dimensional datasets. This aligns with Hollmann et al. (2025) claims that TabPFN specifically excels on small to medium-sized datasets. Nonetheless, previous tabular benchmarks demonstrated TabPFN's effectiveness mainly on datasets containing several hundred to thousands of samples (Table C1).

The practical advantages of TabPFN extend beyond predictive performance. Unlike conventional ML approaches, TabPFN requires neither hyperparameter optimization nor dataset-specific training. These characteristics eliminate two of the most time-consuming and technically challenging aspects of ML workflows (Hollmann et al., 2025). For pedometricians, this simplicity rivals that of Random Forest, which has long been valued for its reliable default performance and ease of application. The combination of superior accuracy and operational simplicity positions TabPFN as an ideal modeling approach for field-scale PSM. We therefore recommend TabPFN as a new default modeling choice for every pedometrician.

Additionally, TabPFN's success points toward opportunities for creating pedology-specific foundation models. Since the current model was trained on synthetic data, fine-tuning or retraining with real soil datasets could produce models that better understand the distinctive patterns in pedological data (Rubachev et al., 2025b; Thomas et al., 2024). Such soil-specific foundation models could combine in-context learning capabilities with learned representations of actual pedological relationships.

While TabPFN excels in field-scale PSM contexts, we acknowledge scenarios where its performance may be limited. These include datasets with very large training sets, high-dimensional feature spaces without dimensionality reduction, and multi-class classification tasks (Hollmann et al., 2025; Rubachev et al., 2025a; Ye et al., 2025b). These constraints, however, fall outside the typical scope of field-scale PSM.

Although TabPFN emerged as the best overall performer, we emphasize that the broader success of modern ANN architectures extends beyond TabPFN alone, and should not be overlooked. Retrieval-based models like TabR and ModernNCA, MLP-based TabM and RealMLP, and attention-based models such as T2G-Former and ExcelFormer all consistently outperformed both classical ML baselines and classical ANN architectures like MLP. Each architecture brings unique strengths and design principles that could prove valuable for specific PSM applications or inform the design of future soil-specific models. This diversity of successful approaches indicates that the integration of deep learning into PSM represents a broad methodological advance rather than dependence on any single model or architectural innovation.

4.3. Future considerations

The emergence of tabular in-context learning has catalyzed development of additional tabular foundation models, including MotherNet (Mueller et al., 2024) and TabICL (Jingang et al., 2025). We did not include these models in our study, as they currently only support classification. However, they may be adapted to regression tasks employing the discretization approach proposed by Barkov et al. (2024).

Although this is beyond the scope of our current study, it represents a promising avenue for future research.

Beyond predictive accuracy, uncertainty quantification remains critical for PSM (Breure et al., 2022; Barkov et al., 2024). While our benchmark focused on point predictions, models like TabPFN can naturally provide probabilistic predictions for uncertainty estimation, since they perform regression using target discretization (Hollmann et al., 2025; Barkov et al., 2024). Furthermore, other ANN-specific model-agnostic methods for uncertainty quantification can be evaluated, such as Laplace approximation (Kristiadi et al., 2021) or Monte Carlo dropout (Huang et al., 2025).

Additionally, since we propose modern ANNs as a safe default choice for future pedometrical studies, this opens the field to ANN-specific improvements that could further boost model performance, such as model souping (Wortsman et al., 2022) or specialized regularization frameworks like TANGOS (Jeffares et al., 2023a).

Furthermore, the spatial nature of soil data presents compelling research directions beyond our tabular framing (Heuvelink and Webster, 2022). Future work could explore ways to incorporate spatial information into these high-performing tabular models, potentially through hybrid architectures that combine the demonstrated effectiveness of models like TabPFN with explicit spatial correlation structures.

These opportunities, alongside the proven performance of modern architectures, demonstrate that ANNs have matured sufficiently to not only advance field-scale PSM but also unlock the rapidly evolving deep learning methodological landscape for future advancements.

5. Conclusion

This study provides the first systematic assessment of modern ANNs for small tabular datasets from digital soil mapping. We implemented and evaluated a comprehensive range of recent state-of-the-art ANN architectures, including MLP-based models (MLP, TabM, RealMLP), attention-based transformers (AutoInt, FT-Transformer, T2G-Former, ExcelFormer, AMFormer), retrieval-based approaches (TabR, ModernNCA), and in-context learning models (TabPFN). We compared these architectures against established classical ML baselines (Random Forest, XGBoost, Linear Regression) across 31 diverse datasets from LimeSoDa. Through this multi-dataset and fully reproducible benchmark, we establish a new comprehensive benchmarking standard for PSM.

Our results demonstrate that many modern ANNs consistently outperform classical ML methods in field-scale PSM tasks, even under challenging small-sample conditions. The in-context learning model TabPFN emerged as a particularly strong method, surpassing Random Forest, Linear Regression, and PLSR, the long-standing default methods in PSM. We therefore recommend TabPFN as the new default model and baseline for field-scale PSM.

While classical ML retains advantages in specific areas, such as extremely small and high-dimensional datasets with fewer than 50 samples where PCA preprocessing is employed, its superiority diminishes rapidly as dataset size increases. Even for datasets containing only 50 to 100 samples, modern neural architectures can demonstrate clear performance advantages, challenging former views about sample size requirements for deep learning in general.

Our findings indicate that the recent wave of tabular deep learning research translates into tangible benefits for PSM. These results mark a substantial step forward for PSM, and as both deep learning and pedometrics fields continue to evolve, future advancements hold promise for further enhancing map accuracy and reliability in precision agriculture applications.

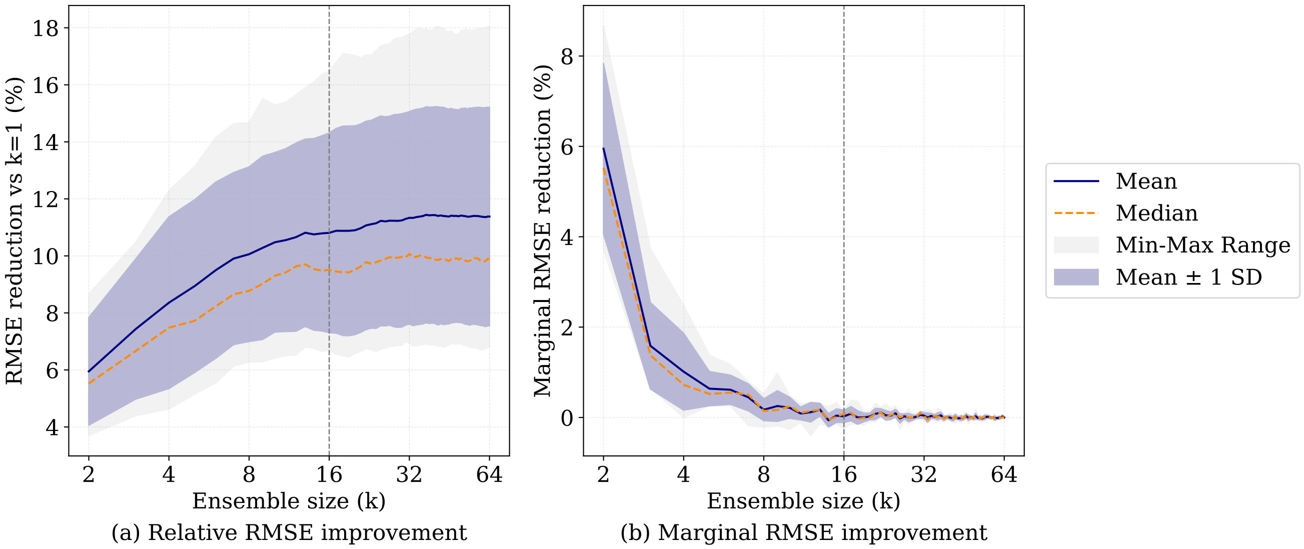
Code and data availability

All the experiments are reproducible, and the source code, including the training and evaluation scripts, is open and publicly available at `https://github.com/slavabarkov/smalltabnets`. All the data used in this study is publicly available in the LimeSoDa repository (Schmidinger et al., 2025) at `https://zenodo.org/records/14936177`.

Acknowledgements

This work was supported by the Lower Saxony Ministry of Science and Culture (MWK), via the zukunft.niedersachsen program of the Volkswagen Foundation (ZN4072). Compute resources were funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project number 456666331.

A. Deep ensembles ablation study



(a) Relative RMSE improvement

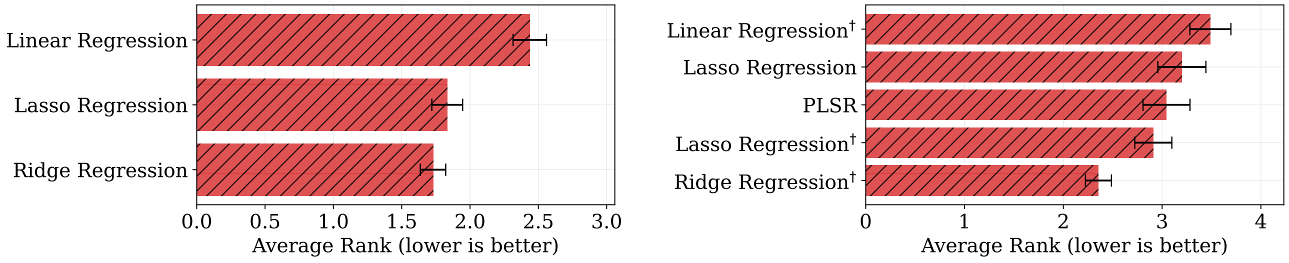
(b) Marginal RMSE improvement

Figure A1: Impact of deep ensemble size on predictive performance, showing both absolute RMSE reduction and marginal improvements. Performance gains plateau after approximately 16 ensemble members, validating our choice of ensemble size and demonstrating that even small ensembles provide substantial benefits over single models in PSM.

To validate our choice of ensemble size, we conducted ablation studies varying the number of ensemble members from 1 to 64. Figure A1 presents these results.

Deep ensembles consistently improved predictive accuracy, but with diminishing returns beyond 16 members. Marginal gains decay rapidly after $k \approx 16$ and plateau after $k \approx 32$. This empirically supports the ensemble size used throughout the main experiments. Even small ensembles provide significant performance improvements over single models, highlighting the inherent instability of individual ANN training runs and demonstrating the value of deep ensembles for tabular data for a fair model evaluation.

B. Detailed comparison of linear models



(a) Model rankings across Low-Dimensional datasets

(b) Model rankings across High-Dimensional datasets

[†]*Models using PCA for dimensionality reduction*

Figure B1: Average rank comparison of the four linear baselines across all 93 regression tasks. Lower rank values indicate better performance (best = 1). Panel (a) shows results for Low-Dimensional datasets, panel (b) for High-Dimensional datasets that were reduced with PCA. Error bars denote ± one standard error of the mean.

The Figure B1 summarizes how the linear baselines rank across all 93 regression tasks used in this study.

C. Benchmark studies review

Table C1 summarizes the dataset properties used in prior influential tabular benchmarking studies. As shown, most benchmarks focused on datasets with significantly larger sample sizes and typically very low feature-to-sample ratios. Among

Table C1
Summary statistics of dataset properties from other previously mentioned benchmarks, compared to those used in our study, denoted as "ours". Most other benchmarks consist of datasets with larger sample sizes and low dimensional features.

Benchmark	Sample			Feature			Feature-to-sample ratio		
	Min	Median	Max	Min	Median	Max	Min	Median	Max
Shwartz-Ziv and Armon (2022)	7k	500k	1M	10	32	2k	<0.001	0.001	0.009
Grinsztajn et al. (2022)	2.6k	17.4k	940k	4	12	613	<0.001	0.001	0.085
Shmuel et al. (2024)	43	4.8k	245k	4	13	267	<0.001	0.004	0.116
Ye et al. (2025a)	506	5.1k	153k	3	16	308	<0.001	0.003	0.161
Hollmann et al. (2025)	240	2.1k	10k	3	20	376	0.001	0.011	0.517
Ye et al. (2025b)*	50	151	7k	2k	5.4k	22.3k	0.714	41.807	262.153
Ye et al. (2025b)†	21k	626k	10M	8	54	784	<0.001	<0.001	0.025
Erickson et al. (2025)	748	6.5k	150k	5	18	1.8k	<0.001	0.005	0.474
High Dimensional (ours)	32	98	460	272	1.7k	2.5k	1.178	17.350	51.156
Low Dimensional (ours)	30	52	250	3	14	17	0.060	0.126	0.433

*, †: Ye et al. (2025b), expands on Ye et al. (2025a) with an additional high-dimensional setting (*) and a large-sample setting (†).

Table D1
Search spaces for modern ANN architectures. Continuous intervals were sampled uniformly, logarithmic intervals log-uniformly, and width/dimension parameters were sampled from the categorical sets shown.

Model	Learning rate (log)	Weight decay (log)	Batch size	# Blocks / layers	Hidden width (d)	Attention heads	Dropout† range
AMFormer	$[10^{-4}, 5\times10^{-3}]$	$[10^{-6}, 10^{-2}]$	{4,8,16,32}	1–4	16–64	2,4	0–0.3
AutoInt	$[10^{-4}, 5\times10^{-3}]$	$[10^{-6}, 10^{-2}]$	{4,8,16,32}	1–4	16–64	2,4	0–0.3
ExcelFormer	$[10^{-4}, 5\times10^{-3}]$	$[10^{-6}, 10^{-2}]$	{4,8,16,32}	1–4	16–64	2,4	0–0.3
FT-Transformer	$[10^{-4}, 5\times10^{-3}]$	$[10^{-6}, 10^{-2}]$	{4,8,16,32}	1–4	16–64	2,4	0–0.3
T2G-Former	$[10^{-4}, 5\times10^{-3}]$	$[10^{-6}, 10^{-2}]$	{4,8,16,32}	1–4	16–64	2,4	0–0.3
MLP	$[10^{-4}, 5\times10^{-3}]$	$[10^{-6}, 10^{-2}]$	{4,8,16,32}	2–4	64–512	—	0–0.3
RealMLP	$[10^{-4}, 5\times10^{-3}]$	$[10^{-6}, 10^{-2}]$	{4,8,16,32}	2–4	64–512	—	0–0.3
TabM	$[10^{-4}, 5\times10^{-3}]$	$[10^{-6}, 10^{-2}]$	{4,8,16,32}	2–4	64–512	—	0–0.3
ModernNCA	$[10^{-4}, 5\times10^{-3}]$	$[10^{-6}, 10^{-2}]$	{4,8,16,32}	2–4	64–512	—	0–0.3
TabR	$[10^{-4}, 5\times10^{-3}]$	$[10^{-6}, 10^{-2}]$	{4,8,16,32}	Enc. 0–4, Pred. 1–4	64–512	—	0–0.3

†: All dropout-type hyperparameters (e.g. attn_dropout, ffn_dropout, residual_dropout, context_dropout).

these, only Ye et al. (2025b) incorporated datasets that closely resemble our High-Dimensional setting as part of a broader benchmark. We did not include the dataset properties from McElfresh et al. (2023), because we only found public available information from a subset of datasets used in that benchmark.

D. Hyperparameter search space

D.1. Hyperparameters of artificial neural networks

Table D1 summarizes the hyperparameter search spaces of all neural architectures considered in this study, with the exception of TabPFN, which is a foundation model that requires no hyperparameter optimization. Parameters that are shared across similar models are listed under a common name. During hyperparameter optimization, all continuous ranges were sampled uniformly, all the ranges indicated with sub-script "log" were sampled log-uniformly, and all width and dimension parameters were drawn from the categorical sets shown in the table. For the High-Dimensional data group we applied PCA in a joint search, sampling the number of components uniformly from 2–32.

Some model-specific hyperparameters that are not listed in Table D1 were additionally optimized. For AMFormer these were number of prompt tokens $\in \{1, 2, 4\}$, additive and multiplicative attention tokens $\in \{1, 2, 4\}$. For ModernNCA these were retrieval parameters, including embedding dimensionality 64–256 and soft-NN temperature 0.1–5.0 with log-uniform sampling. For TabR these were retrieval parameters, including context size $\{2, 4, 8, 16\}$ and context dropout 0–0.3.

Table D2
Hyperparameter grids for classical baselines. Bracketed intervals were sampled uniformly; "log" indicates log-uniform sampling.

Algorithm	Parameter	Search grid
XGBoost	learning_rate	$[0.001, 0.3]_{\text{log}}$
	max_depth	$\{3, \dots, 10\}$
	min_child_weight	$\{1, \dots, 10\}$
	subsample	$[0.5, 1.0]$
	colsample_bytree	$[0.5, 1.0]$
	gamma	$[0, 5]$
	reg_alpha	$[10^{-6}, 10]_{\text{log}}$
	reg_lambda	$[0.1, 10]_{\text{log}}$
Random Forest	max_depth	$\{3, \dots, 30\}$
	min_samples_split	$\{2, \dots, 10\}$
	min_samples_leaf	$\{1, \dots, 10\}$
	max_features	$[0.6, 1.0]$
Lasso Regression	alpha	$[10^{-4}, 100]_{\text{log}}$
Ridge Regression	alpha	$[10^{-4}, 100]_{\text{log}}$

D.2. Hyperparameters of classical machine learning algorithms

Table D2 lists the search spaces of all classical ML algorithms considered in this study. Sampling rules and the PCA or partial least squares component search (2–32) are identical to those used for the ANNs.

E. Implementation details

All ANN models employed in this study utilize reference implementations of their core architectures. For the model architecture implementations, we use the authors' reference code provided with the respective publications for ModernNCA (Ye et al., 2025c), TabR (Gorishniy et al., 2024), TabM (Gorishniy et al., 2025), FT-Transformer (Gorishniy et al., 2021), T2G-Former (Yan et al., 2023), AMFormer (Cheng et al., 2024), and ExcelFormer (Chen et al., 2024). For TabPFN (Hollmann et al., 2025), we use the authors' implementation including the pre-trained model weights. MLP follows the implementation from Gorishniy et al. (2025) and AutoInt follows the PyTorch reimplementation from Gorishniy et al. (2021). RealMLP was reimplemented based on the tabular benchmarking framework from Holzmüller et al. (2024), excluding their custom training pipeline. Numerical feature embeddings (see Section 2.4.2) were based on `rtdl-num-embeddings 0.0.12`, using an implementation of Quantile-Based Piecewise Linear Encoding that corresponds to the "Q-L" variation from Table 2 in Gorishniy et al. (2022). The benchmarking framework, including our unified model training interface, training loop with early stopping, data preprocessing pipelines (scaling, PCA, numerical embeddings), optimization utilities, cross-validation structure, YAML-based configuration system, and experimental logging, was implemented as part of this work to ensure consistent evaluation across all models.

The evaluation framework was developed in `Python 3.10` using `PyTorch 2.8.0` (Paszke et al., 2019) for deep learning models, `scikit-learn 1.6.1` (Pedregosa et al., 2011) for preprocessing and classical methods, `XGBoost 3.0.3` (Chen and Guestrin, 2016) for GBDT implementation, and `Optuna 4.4.0` (Akiba et al., 2019) for hyperparameter optimization. The complete implementation, including hyperparameters and experimental results, is openly available (see Code and data availability).

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M., 2019. Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Association for Computing Machinery. pp. 2623–2631.

Allen-Zhu, Z., Li, Y., 2023. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning, in: The Eleventh International Conference on Learning Representations.

Arik, S.Ö., Pfister, T., 2021. Tabnet: Attentive interpretable tabular learning, in: Proceedings of the AAAI conference on artificial intelligence, pp. 6679–6687.

Barkov, V., Schmidinger, J., Gebbers, R., Atzmueller, M., 2024. An efficient model-agnostic approach for uncertainty estimation in data-restricted pedometric applications, in: 2024 International Conference on Machine Learning and Applications (ICMLA), IEEE. pp. 198–205.

Barra, I., Haefele, S.M., Sakrabani, R., Kebede, F., 2021. Soil spectroscopy with the use of chemometrics, machine learning and pre-processing techniques in soil diagnosis: Recent advances–a review. TrAC Trends in Analytical Chemistry 135, 116166.

Bellman, R., Bellman, R., Corporation, R., 1957. Dynamic Programming. Rand Corporation research study, Princeton University Press.

Bergstra, J., Bardenet, R., Bengio, Y., Kégl, B., 2011. Algorithms for hyper-parameter optimization, in: Advances in Neural Information Processing Systems, Curran Associates, Inc.

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Breure, T.S., Haefele, S., Hannam, J.A., Corstanje, R., Webster, R., Moreno-Rojas, S., Milne, A., 2022. A loss function to evaluate agricultural decision-making under uncertainty: a case study of soil spectroscopy. Precision Agriculture 23, 1333–1353.

Chen, J., Yan, J., Chen, Q., Chen, D.Z., Wu, J., Sun, J., 2024. Can a deep learning model be a sure bet for tabular prediction?, in: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery. pp. 288–296.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery. pp. 785–794.

Cheng, Y., Hu, R., Ying, H., Shi, X., Wu, J., Lin, W., 2024. Arithmetic feature interaction is necessary for deep tabular learning. Proceedings of the AAAI Conference on Artificial Intelligence 38, 11516–11524.

Ding, Z., Liu, K., Grunwald, S., Smith, P., Ciais, P., Wang, B., Wadoux, A.M.C., Ferreira, C., Karunaratne, S., Shurpali, N., et al., 2025. Advancing soil organic carbon prediction: A comprehensive review of technologies, AI, process-based and hybrid modelling approaches. Advanced Science , e04152.

Erickson, N., Purucker, L., Tschalzev, A., Holzmüller, D., Desai, P.M., Salinas, D., Hutter, F., 2025. TabArena: A living benchmark for machine learning on tabular data. arXiv:2506.16791.

Gebbers, R., 2019. Proximal soil surveying and monitoring techniques.. Burleigh Dodds Science Publishing Limited, Cambridge. pp. 29–78.

Gebbers, R., Adamchuk, V.I., 2010. Precision agriculture and food security. Science 327, 828–831.

Gorishniy, Y., Kotelnikov, A., Babenko, A., 2025. TabM: Advancing tabular deep learning with parameter-efficient ensembling, in: The Thirteenth International Conference on Learning Representations.

Gorishniy, Y., Rubachev, I., Babenko, A., 2022. On embeddings for numerical features in tabular deep learning, in: Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 24991–25004.

Gorishniy, Y., Rubachev, I., Kartashev, N., Shlenskii, D., Kotelnikov, A., Babenko, A., 2024. TabR: Tabular deep learning meets nearest neighbors, in: The Twelfth International Conference on Learning Representations.

Gorishniy, Y., Rubachev, I., Khrulkov, V., Babenko, A., 2021. Revisiting deep learning models for tabular data, in: Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 18932–18943.

Grinsztajn, L., Oyallon, E., Varoquaux, G., 2022. Why do tree-based models still outperform deep learning on typical tabular data? Advances in neural information processing systems 35, 507–520.

Heuvelink, G.B., Webster, R., 2022. Spatial statistics and soil mapping: A blossoming partnership under pressure. Spatial statistics 50, 100639.

Hollmann, N., Müller, S., Eggensperger, K., Hutter, F., 2023. TabPFN: A transformer that solves small tabular classification problems in a second, in: The Eleventh International Conference on Learning Representations.

Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S.B., Schirrmeister, R.T., Hutter, F., 2025. Accurate predictions on small data with a tabular foundation model. Nature 637, 319–326.

Holzmüller, D., Grinsztajn, L., Steinwart, I., 2024. Better by default: Strong pre-tuned MLPs and boosted trees on tabular data, in: Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 26577–26658.

Huang, Y.C., Padarian, J., Minasny, B., McBratney, A.B., 2025. Using Monte Carlo conformal prediction to evaluate the uncertainty of deep-learning soil spectral models. SOIL 11, 553–563.

Jeffares, A., Liu, T., Crabbé, J., Imrie, F., van der Schaar, M., 2023a. TANGOS: Regularizing tabular neural networks through gradient orthogonalization and specialization, in: The Eleventh International Conference on Learning Representations.

Jeffares, A., Liu, T., Crabbé, J., van der Schaar, M., 2023b. Joint training of deep ensembles fails due to learner collusion, in: Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 13559–13589.

Jingang, Q., Holzmüller, D., Varoquaux, G., Le Morvan, M., 2025. TabICL: A tabular foundation model for in-context learning on large data, in: Forty-second International Conference on Machine Learning.

Kasraei, B., Heung, B., Saurette, D.D., Schmidt, M.G., Bulmer, C.E., Bethel, W., 2021. Quantile regression as a generic approach for estimating uncertainty of digital soil maps produced from machine-learning. Environmental Modelling & Software 144, 105139.

Katzir, L., Elidan, G., El-Yaniv, R., 2020. Net-dnf: Effective deep modeling of tabular data, in: International conference on learning representations.

Keesstra, S.D., Bouma, J., Wallinga, J., Tittonell, P., Smith, P., Cerdà, A., Montanarella, L., Quinton, J., Pachepsky, Y., Van Der Putten, W.H., et al., 2016. Forum paper: The significance of soils and soil science towards realization of the un sustainable development goals (SDGS). Soil Discussions 2016, 1–28.

Khaledian, Y., Miller, B.A., 2020. Selecting appropriate machine learning methods for digital soil mapping. Applied Mathematical Modelling 81, 401–418.

Kristiadi, A., Hein, M., Hennig, P., 2021. Learnable uncertainty under Laplace approximations, in: Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, PMLR. pp. 344–353.

Ma, T., Brus, D.J., Zhu, A.X., Zhang, L., Scholten, T., 2020. Comparison of conditioned Latin hypercube and feature space coverage sampling for predicting soil classes using simulation from soil maps. Geoderma 370, 10–1016.

McBratney, A., de Gruijter, J., Bryce, A., 2019. Pedometrics timeline. Geoderma 338, 568–575.

McBratney, A.B., Santos, M.M., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52.

McElfresh, D., Khandagale, S., Valverde, J., Prasad C, V., Ramakrishnan, G., Goldblum, M., White, C., 2023. When do neural nets outperform boosted trees on tabular data? Advances in Neural Information Processing Systems 36, 76336–76369.

Menze, B.H., Kelm, B.M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., Hamprecht, F.A., 2009. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC bioinformatics 10, 213.

Mueller, A.C., Curino, C.A., Ramakrishnan, R., 2024. Mothernet: Fast training and inference via hyper-network transformers, in: The Thirteenth International Conference on Learning Representations.

Ng, W., Minasny, B., Montazerolghaem, M., Padarian, J., Ferguson, R., Bailey, S., McBratney, A.B., 2019. Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra. Geoderma 352, 251–267.

Nießl, C., Herrmann, M., Wiedemann, C., Casalicchio, G., Boulesteix, A.L., 2022. Over-optimism in benchmark studies and the multiplicity of design and analysis options when interpreting their results. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 12, e1441.

Oukhattar, M., Gadal, S., Robert, Y., Saby, N., Houmma, I.H., Keller, C., 2025. Variability analysis of soil organic carbon content across land use types and its digital mapping using machine learning and deep learning algorithms. Environmental Monitoring and Assessment 197, 535.

Ozaki, Y., Tanigaki, Y., Watanabe, S., Nomura, M., Onishi, M., 2022. Multiobjective tree-structured parzen estimator. Journal of Artificial Intelligence Research 73, 1209–1250.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems, Curran Associates, Inc.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Édouard Duchesnay, 2011. Scikit-learn: Machine learning in python. Journal of Machine Learning Research 12, 2825–2830.

Probst, P., Wright, M.N., Boulesteix, A.L., 2019. Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: data mining and knowledge discovery 9, e1301.

Rubachev, I., Kartashev, N., Gorishniy, Y., Babenko, A., 2025a. TabReD: Analyzing pitfalls and filling the gaps in tabular deep learning benchmarks, in: The Thirteenth International Conference on Learning Representations.

Rubachev, I., Kotelnikov, A., Kartashev, N., Babenko, A., 2025b. On finetuning tabular foundation models. arXiv:2506.08982.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning internal representations by error propagation. MIT Press, Cambridge, MA, USA. pp. 318–362.

dos Santos, E.P., Moreira, M.C., Fernandes-Filho, E.I., Demattê, J.A.M., dos Santos, U.J., da Silva, D.D., Cruz, R.R.P., Moura-Bueno, J.M., Santos, I.C., Sampaio, E.V.d.S.B., 2023. Improving the generalization error and transparency of regression models to estimate soil organic carbon using soil reflectance data. Ecological Informatics 77, 102240.

Schmidinger, J., Schröter, I., Bönecke, E., Gebbers, R., Ruehlmann, J., Kramer, E., Mulder, V.L., Heuvelink, G.B., Vogel, S., 2024. Effect of training sample size, sampling design and prediction model on soil mapping with proximal sensing data for precision liming. Precision Agriculture 25, 1529–1555.

Schmidinger, J., Vogel, S., Barkov, V., Pham, A.D., Gebbers, R., Tavakoli, H., Correa, J., Tavares, T.R., Filippi, P., Jones, E.J., et al., 2025. LimeSoDa: A dataset collection for benchmarking of machine learning regressors in digital soil mapping. Geoderma 459, 117337.

Shmuel, A., Glickman, O., Lazebnik, T., 2024. A comprehensive benchmark of machine and deep learning across diverse tabular datasets. arXiv:2408.14817.

Shwartz-Ziv, R., Armon, A., 2022. Tabular data: Deep learning is not all you need. Information Fusion 81, 84–90.

Söderström, M., Sohlenius, G., Rodhe, L., Piikki, K., 2016. Adaptation of regional digital soil mapping for precision agriculture. Precision Agriculture 17, 588–607.

Song, W., Shi, C., Xiao, Z., Duan, Z., Xu, Y., Zhang, M., Tang, J., 2019. AutoInt: Automatic feature interaction learning via self-attentive neural networks, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Association for Computing Machinery. pp. 1161–1170.

Sun, Y., Ma, J., Zhao, W., Qu, Y., Gou, Z., Chen, H., Tian, Y., Wu, F., 2023. Digital mapping of soil organic carbon density in China using an ensemble model. Environmental research 231, 116131.

Thomas, V., Ma, J., Hosseinzadeh, R., Golestan, K., Yu, G., Volkovs, M., Caterini, A., 2024. Retrieval & fine-tuning for in-context tabular models, in: Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 108439–108467.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.

Viscarra Rossel, R.A., Bouma, J., 2016. Soil sensing: A new paradigm for agriculture. Agricultural Systems 148, 71–74.

Wadoux, A.M.C., 2025. Artificial intelligence in soil science. European Journal of Soil Science 76, e70080.

Wadoux, A.M.C., Minasny, B., McBratney, A.B., 2020. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth-Science Reviews 210, 103359.

Wang, J., Liu, T., Zhang, J., Yuan, H., Acquah, G.E., 2022. Spectral variable selection for estimation of soil organic carbon content using mid-infrared spectroscopy. European Journal of Soil Science 73, e13267.

Wang, L., Wang, R., 2022. Determination of soil pH from Vis-NIR spectroscopy by extreme learning machine and variable selection: A case study in lime concretion black soil. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 283, 121707.

Wortsman, M., Ilharco, G., Gadre, S.Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A.S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., et al., 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time, in: International conference on machine learning, PMLR. pp. 23965–23998.

Xue, J., Zhang, X., Chen, S., Lu, R., Wang, Z., Wang, N., Hong, Y., Chen, X., Xiao, Y., Ma, Y., et al., 2023. The validity domain of sensor fusion in sensing soil quality indicators. Geoderma 438, 116657.

Yamada, Y., Lindenbaum, O., Negahban, S., Kluger, Y., 2020. Feature selection using Stochastic Gates, in: Proceedings of the 37th International Conference on Machine Learning, PMLR. pp. 10648–10659.

Yan, J., Chen, J., Wu, Y., Chen, D.Z., Wu, J., 2023. T2G-FORMER: Organizing tabular features into relation graphs promotes heterogeneous feature interaction. Proceedings of the AAAI Conference on Artificial Intelligence 37, 10720–10728.

Yang, L., Cai, Y., Zhang, L., Guo, M., Li, A., Zhou, C., 2021. A deep learning method to predict soil organic carbon content at a regional scale using satellite-based phenology variables. International Journal of Applied Earth Observation and Geoinformation 102, 102428.

Ye, H.J., Liu, S.Y., Cai, H.R., Zhou, Q.L., Zhan, D.C., 2025a. A closer look at deep learning methods on tabular datasets. arXiv:2407.00956.

Ye, H.J., Liu, S.Y., Chao, W.L., 2025b. A closer look at TabPFN v2: Understanding its strengths and extending its capabilities. arXiv:2502.17361.

Ye, H.J., Yin, H.H., Zhan, D.C., Chao, W.L., 2025c. Revisiting nearest neighbor for tabular data: A deep tabular baseline two decades later, in: The Thirteenth International Conference on Learning Representations.