Open Access ISSN: 3007-1437 (Online), ISSN: 3007-1429 (Print)

Alkadhim Journal for Computer Science (KJCS)

Journal Homepage: https://alkadhum-col.edu.iq/JKCEAS

Zero-training: Pretrained Foundation Model for Limited Healthcare Tabular Datasets

[1]Mohammed F. Zamil*

1 University of Information Technology and Communications (UOITC), Baghdad, Iraq

Article information

Article history:
Received: June, 04, 2026
Accepted: July, 07, 2026
Available online: September, 25, 2026

Keywords:
Tabular Data Classification,
Zero-Training,
TabPFN,
Data Scarcity,
Tabular Foundation Models

Corresponding Author:
Mohammed F. Zamil
mfadhil@uoitc.edu.iq

DOI:
https://doi.org/10.61710/9bwz3424

Abstract

Tabular data is central to many machine learning applications, as most enterprises use databases to store their data as structured or semi-structured data. While traditional machine learning models such as gradient boosting trees, Random Forest models dominate this domain, pre-trained tabular models like Tabular Prior-data Fitted Network (TabPFN) have shown ability to build models using limited training data or no training data. The Study shows transfer learning can be deployed in tabular data. Implemented models are evaluated using F1-score to ensure the balance between recall and precision. Results show that TabPFN outperforms models from 100 to 10k sample sizes datasets, TabPFN achieved (76.0% F1-score) in 5k sample size. However, as the training size increases, the performance gap decreased. At 10k sample size XGBoost eventually matches TabPFN and then starts outperforming TabPFN in above 10k sample sizes. These findings highlight pretrained foundation models in building machine learning models in domains that lack enough labeled tabular datasets.

1. Introduction

Tabular data remains one of the most widely used data formats across diverse domains, including healthcare, finance, and education [1], [2]. Both structured and semi-structured data represent the primary format for enterprise and business applications. Structured datasets are difficult to collect and limited in size due to strict schema requirements. In healthcare specifically, EHR-derived datasets are mainly small sized structured tabular formats due to the data privacy and protection protocols. Building smart systems perform better with high quality and large size data [3]. Therefore, many studies have worked on finding best ML algorithms that are capable to extract insights and patterns form small size datasets. Although deep learning advances rapidly, traditional machine learning algorithms such as tree-based models (bagging and boosting) continue to dominate tabular data prediction tasks due to their less complex structure making them top models for tasks that have less complex training datasets [4].

Despite deep learning superiority in fields such as computer vision and natural language processing, but in structured and less complex features interactions do not thrive [5]. Such deep and complex models perform better with complex feature interactions and large sample sizes when compared to classical ML algorithms [6].

Nevertheless, recently neural networks and deep learning methods through pretrained tabular foundation models, such as TabPFN [7] and Tab-Net [8], have proven to be a promising alternative for building models from limited size tabular datasets. TabPFN is pretrained using huge synthetic tabular on broad simulated tasks, which enable it to not need a real dataset for training, instead it uses the training set just to predict the task context. However, while these pretrained models have demonstrated impressive results on small tabular data, there is still a limited understanding of how their performance changes with varying the amount of training data size. Specifically, the question remains unresolved whether their superiority remains on all range of sizes of tabular datasets or decreases compared to traditional ML approaches.

This motivates the need for a study and analyze zero-training and pretrained models scaling behavior and to validate results we compared it against leading models on small to medium tabular datasets reported in literature such as XGBoost, Random Forest, and Logistic Regression [9], [10], [11]. Evaluating model performance using different training dataset sizes can help in proving models' validity in domains where data are limited. In most real-world applications, especially healthcare and security, face a significant shortage of available labelled tabular datasets, which make need for learning approaches capable of effectively training and building models with small tabular datasets [12].

The remainder of this paper is organized as follows. Section II reviews related studies that build machine learning models using small scale datasets. Section III describes the methodology, starting by describing BRFSS dataset, the experimental setup, and the models evaluated. Section IV presents and discusses the experimental results. Finally, Section V concludes the study and outlines its future work.

2. Related Work

Tabular data scarcity remains a great challenge that many studies have tried to propose methods to overcome. In most recent studies, tree-based ML algorithms have been performed as one of the top models for tasks using tabular data training. Specifically, ensemble tree-based models like XGBoost and random forest are widely deployed in previous proposed methods and frameworks due to their reliable and high performance in wide range of domains that have no sufficient data size for training such as healthcare [7]. Besides small size, most healthcare datasets are tabular and their features are heterogeneous.

A diabetes prediction study reported that Artificial Neural Networks (ANN) achieved the best performance on the PIMA dataset with approximately 99.4% in accuracy, while RF was ranked first when applied to early diabetes detection dataset (99.4%). Besides, selecting RF study also highlights the impact model tuning by selecting a sub-group of features in order to enhance predictive performance. Both studies' datasets used small-scale datasets under 1k rows [13]. In addition, a clinically relevant study used MIMIC-IV dataset for sepsis mortality prediction. Moreover, the study reported that RF outperformed other models, achieving an AUROC of 94% while also providing interpretable prediction by providing a rank list of features that influence model decision [14]. In general, healthcare datasets have several kinds of data types and their features are heterogenetic, thus tree-based models often show capability to manage diverse features distributions because of their intrinsic through hierarchical partitioning [11].

Conversely, deep learning models have achieved notable success in unprocessed, complex data such as images, text, and time series. It is highly effective in large-scale data where performance increases as dataset size grows. However, their performance on tabular data has been less consistent, particularly when data size is small or even moderate in some complex tasks [15].

In breast cancer detection, a study that applied multiple convolutional neural network architectures found that fine-tuned ResNet50 achieved superior performance (97.54% accuracy), highlighting the effectiveness of transfer learning in improving diagnostic accuracy in medical imaging tasks [16]. In histopathological breast cancer

classification, a study introduced Cell-Sage, a lightweight attention-enhanced CNN, which outperformed larger models such as ResNet-50 and Vision Transformers while achieving highest performance, demonstrating that efficient architecture can balance predictive performance and computational cost in medical imaging tasks [17]. In Alzheimer's disease monitoring and prediction, a study proposing a framework (DeTrAs) utilizing RNN-based models alongside CNN and NLP components demonstrated up to 20% improvement in accuracy compared to conventional ML models [18].

Also, a recent study integrated a deep learning model (CTGAN) to generate synthetic data via to be used in training traditional machine learning models. Although datasets contain 309 instances and sixteen attributes, RF achieved superior performance (98.93% accuracy), outperforming models such as SVM and k-NN, especially in addressing class imbalance and improving prediction robustness [19].

Despite their long-standing stability and widespread use, both classical machine learning (e.g., Random Forest and Gradient Boosting) and deep learning models (e.g., RNNs and CNNs) are still both less effective when training datasets is small. In general, all ML algorithms perform better as more training data provided, however, not all models have the ability to learn from limited data. Thus, many studies have studied building small tabular in domains such as healthcare, as models suffer from issues such as overfitting [20]. Notably, tree-based models in last year's literature demonstrate superior performance on tabular heterogeneous non-temporal data; nevertheless, they still suffer from limitations, including high sensitivity to noise, outliers, and overfitting. Thus, it requires careful hyper-parameter tuning.

Such challenges have motivated researchers to deploy pretrained foundation models, for example a recent study shows that TabPFN is able to perform reasonable performance on small to medium size datasets [7], [21]. Such models have the ability to perform classification/regression tasks with no training set, as they used prior learning on synthetic datasets to predict the output [22]. A comprehensive comparative study clearly showed that TabPFN outperforms traditional ML models such as tree-based, LR, and SVM [7]. Pretrained foundation models show a promising solution to bridge the gap between performance and generalization in small data training. Therefore, this study aims to systematically investigate the effectiveness of TabPFN in particular, in comparison to traditional ML models across different tabular data scales.

3. Methodology

This section outlines the experiment on using pretrained techniques on limited datasets. We use pretrained tabular foundation models against three traditional ML models that were built using various training sample sizes. Experiments aim to determine the dataset size at which pretrained tabular models perform better and to understand how their performance evolves relative to established ML models.

Consequently, we construct 11 training subsets of varied sizes starting from 100 to 50k, while keeping a fixed test set for evaluation. For each subset, models are trained and evaluated on the unified test set to analyze how their predictive performance evolves as the number of training samples increases. Unifying training and testing set across models are to provide a systematic and fair assessment of traditional models (XGBoost, Random Forest and Logistic Regression) and pretrained foundation model TabPFN.

The experiments were conducted using real-world and publicly available healthcare tabular datasets. The dataset is a sample derived from the Behavioral Risk Factor Surveillance System (BRFSS), an annual large-scale telephone survey conducted by the Centers for Disease Control and Prevention (CDC), which collects health-related information on risk behaviors, chronic diseases states, and what preventive practices conducted.

The dataset sample includes more than 70k rows and 21 features covering a wide range of health-related attributes, and diabetes as a target variable. Dataset independent variables include patient clinical features (e.g., BMI, blood pressure, and cholesterol level), lifestyle habits such as smoking, physical activity, and diet quality. Besides, general mental and physical conditions and socioeconomic and demographic factors such as age, sex, education level, and salary.

3.1. Experimental Design

Figure 1 illustrates the overall experimental workflow. As the dataset is preprocessed and divided using a stratified method producing train/test sets, in order to maintain the class distribution as same as original full dataset. Then the training set subsampled into varying sets sizes, but all training sets used with the same fixed test set across all models.

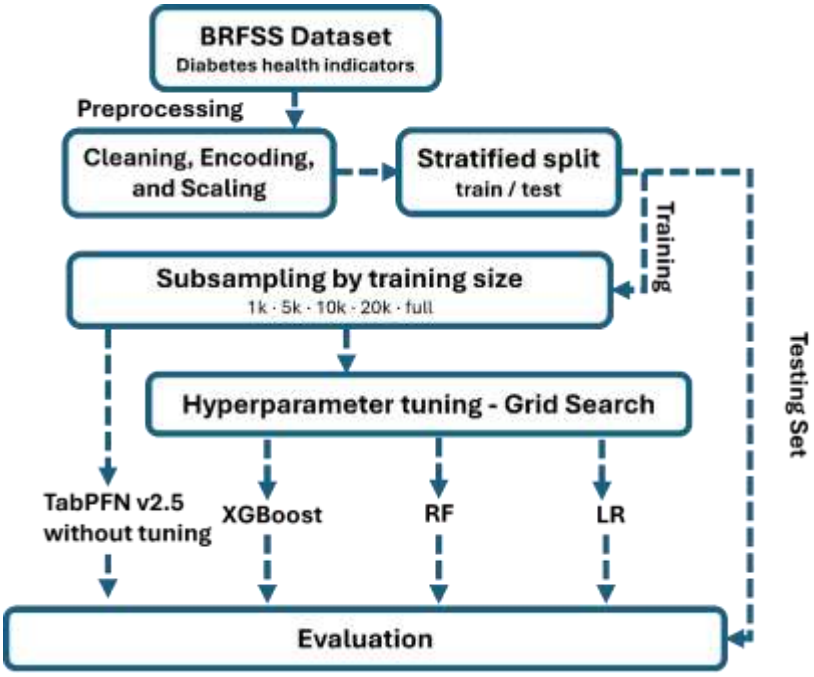

Figure (1): Methodology workflow.

To conduct a systematic assessment of model performance across different data ranges, we implement a controlled experimental design that varies the size of the training dataset while maintaining a fixed test data sample for evaluation. The original dataset is sampled into training and testing using a stratified sampling method to preserve class balance. The test set is kept fixed throughout all experiments to have fair comparisons across all models and training sizes. To study how models perform with increasing datasets, multiple subsets size ranges generated using stratified sampling five times each sample size. The training ranges from small to moderately large sets (100, 250, 500, 1k, 2k, 5k, 10k-50k samples). Also, to mitigate variability arising from sampling, five subsets were generated for each data size using different random seeds.

Model performance is primarily evaluated using the F1-score, it is suitable metric for classification performance evaluation due to its ability to show the real model performance without favoring class on another and balances precision and recall effectively. Besides, F1 has the ability to handle class imbalance better than accuracy. All reported results are the average of all repeated experiments conducted using the five random samples for each sample size, to show a reasonable estimate of model behavior.

3.2. Models Hyper-parameter Settings

To set hyper-parameters in a way that suits each algorithm, we used a grid search method for tuning XGBoost, RF, and LR using a unified subset. Table 1, shows selected hyper-parameters settings used for all training dataset sizes. Also, the table shows hyper-parameters vary depending on the model type and what resulted from the search results. Furthermore, to avoid producing biased and over-fitted models the tuning models conducted separately for each data sample size used in the experiments. The key hyper-parameter settings for all models used in the experiments are summarized in Table 1.

Table (1): Hyper-parameter configurations used for each model.

Hyper-parameter	TabPFN v2.5	XGBoost	Random Forest	Logistic Regression
Model Type	Pretrained	Boosting	Bagging	Linear
n_estimators	—	300	200	Solver: LBFGS
max_depth	—	6	None	Penalty: L2
subsample	—	0.9	Min samples leaf: 10	C (Regularization): 1.0
Scaling	—	no	no	Standard

It is worth noting, and can be considered an advantage, that TabPFN v2.5 was used designed to work without hyper-parameter tuning because it is pretrained designed to operate without task-specific tuning. Besides, pretrained foundation models require small dataset just to sense the context of the task and dataset. However, traditional machine learning algorithms used in this study' experiments tuned using grid search method implemented by scikit-learn.

4. Results and Discussion

This section presents all experiments results evaluating the performance of TabPFN and classical machine learning models as training dataset size varies. As the results and analysis focus on understanding how model performance evolves as more training data becomes available, by comparing tree-based models with pretrained model TabPFN. Results are reported using mean and standard deviation of five times experiments repeated using random seeds to ensure that the results are not biased toward a specific data sample.

Table (2): Comparison of model performance across selected training sample sizes using F1-score (mean and standard deviation). The best result for each training size is shown in bold.

Train Size	Logistic Regression	Random Forest	TabPFN	XGBoost
100	0.689 ± 0.024	0.721 ± 0.015	0.718 ± 0.010	0.709 ± 0.016
250	0.727 ± 0.011	0.745 ± 0.008	0.747 ± 0.009	0.729 ± 0.013
500	0.741 ± 0.006	0.748 ± 0.004	0.752 ± 0.003	0.733 ± 0.003
1k	0.742 ± 0.005	0.749 ± 0.005	0.755 ± 0.005	0.741 ± 0.005
2k	0.749 ± 0.002	0.752 ± 0.002	0.758 ± 0.003	0.751 ± 0.002
5k	0.749 ± 0.001	0.756 ± 0.004	0.760 ± 0.002	0.758 ± 0.002
10k	0.751 ± 0.001	0.757 ± 0.001	**0.760 ± 0.002**	**0.760 ± 0.003**
20k	0.750 ± 0.001	0.758 ± 0.002	**0.759 ± 0.001**	**0.762 ± 0.000**
30k	0.751 ± 0.001	0.759 ± 0.002	**0.761 ± 0.001**	**0.763 ± 0.001**
40k	0.750 ± 0.001	0.759 ± 0.002	**0.760 ± 0.001**	**0.764 ± 0.001**
50k	0.750 ± 0.000	0.757 ± 0.001	**0.761 ± 0.000**	**0.764 ± 0.000**

Table 2, summarizes the F1-score performance of all models using all sample scales. At the smallest training size (Training Size = 100 to 2000), TabPFN demonstrates competitive and highest performance in 500 and 2000, outperforming XGBoost and Logistic Regression, while achieving results comparable to Random Forest in 100 sample size. As the training size increases to 500 and 2,000 samples, TabPFN consistently maintains superior or near-superior performance.

However, with further increase in training datasets, the performance gap between TabPFN and other models get less. At training size (10,000), TabPFN and XGBoost show very comparable results, implying a convergence point where pretraining models' benefit becomes marginal. As training size reaches 50,000 sample size, XGBoost starts to outperform TabPFN, while RF and LR remain below both models.

Despite the increase in dataset size and overall performance improvement, LR remained almost stable with no significant improvement in its performance above the 10k sample size. This could be due to model capability as there is an under-fitting situation clearly shown when models cannot capture more complex and deeper patterns in datasets even when more data are given no noticeable enhancement in model performance. In contrast, tree-based models' behavior shows steady enhancement till the 30k sample size. After that, only XGBoost gets benefits from increased data, showing its capability to capture more uncaptured complex feature interactions.

Figure 2 illustrates how models perform using F1 metric as datasets size increases. The learning curve shows that TabPFN achieves the highest performance in small data samples when compared to traditional models. At under 5k sample sizes, the difference between TabPFN and XGBoost is observable, indicating the advantage of the pretrained model when training data are limited, as pertaining show as an advantage. As the training size increases, from 10k to 30k all models improve, but XGBoost starts to lead from the 20k sample size to 50k. XGBoost benefits from increased data by capturing more complex patterns. In contrast, Logistic Regression consistently remains below the tree-based methods, showing no enhancement even when more data is presented.

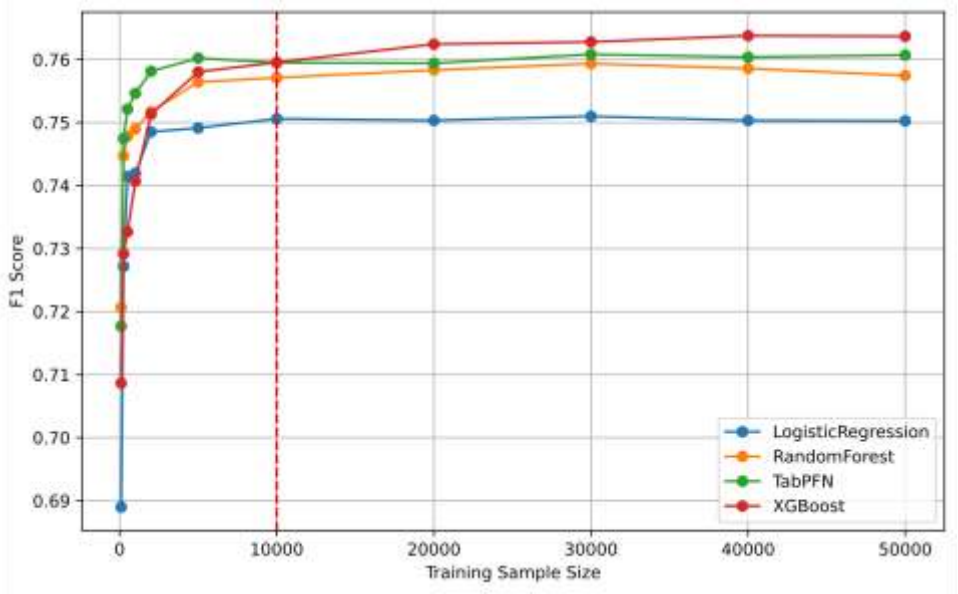

Figure (2): Models performance using F1 metric using 11 training sample sizes. The vertical dashed line indicates the approximate transition point where the performance advantage of XGBoost starts to outperform TabPFN.

Figure 3 provides a clearer view of the difference in performance (ΔF1) results between the top two models (TabPFN and XGBoost) by plotting the difference in F1 metric. The positive data points below 10k indicate that TabPFN has better performance compared to XGBoost. However, the difference decreases steadily towards zero, which means their performance is almost identical and to the negative as the number of training samples grows. The 10k sample size becomes a transition point, and beyond this point, the curve becomes slightly negative, demonstrating that XGBoost starts to outperform TabPFN as training size increases.

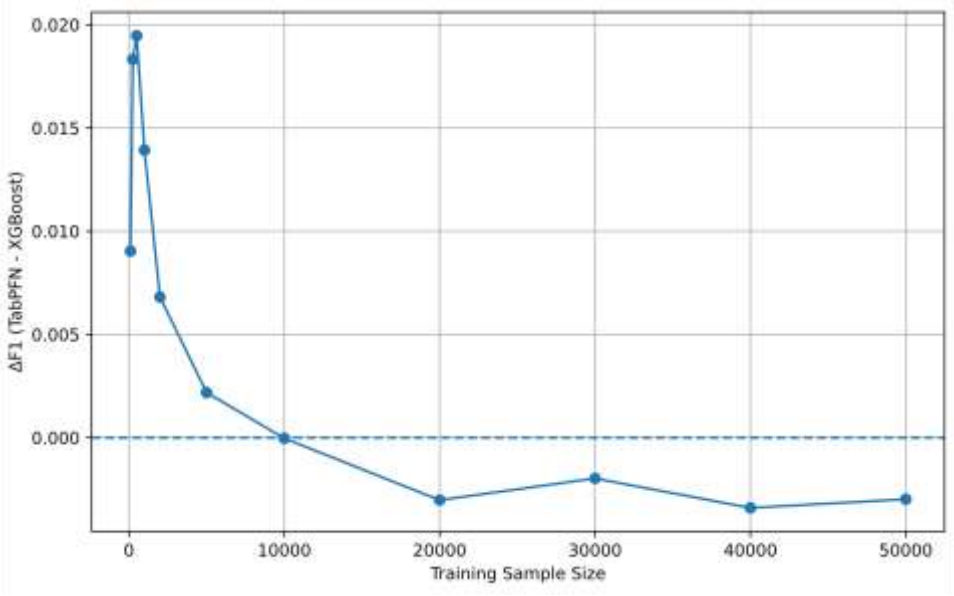

Figure (3): ΔF1 (TabPFN - XGBoost) vs sample size.

Results clearly show that pretrained models (TabPFN) have a great advantage with smaller dataset samples, whereas as more training data presented XGBoost becomes increasingly competitive (10k) and better in 20k and above.

The study result provides recommendations in model selection where dataset size is limited and has structured type, especially in healthcare domain or any domain with small tabular scall dataset size. Empirically, pretrained foundation models demonstrate strong suitability for small to medium range dataset sample size, thus making it well-suited for in domains where data expansion could be operationally challenging or expensive. However, as data becomes larger to a point where tree-based models and XGBoost in particular become more suitable for building using real sufficient training dataset. In summary, considering dataset size as a key factor in model selection for tabular machine learning tasks.

5. Conclusions and Future work

The results of this study highlights models have the ability to learn from limited tabular datasets, and this is proved by a systematic evaluation of a pretrained tabular foundation model (TabPFN) and classical machine learning models that are well known in literature such as XGBoost, Random Forest, and Logistic Regression. The results demonstrate that TabPFN achieves superior performance on small training datasets, outperforming tree-based models and LR up to a crossover point between 10k and 20k samples. However, as the number of training samples increases, the performance gap between TabPFN and classical models decreases, beyond 20k scales XGBoost outperforming TabPFN.

Pretrained foundation models excel on small datasets, making them highly suitable for healthcare, where data are limited. Pretrained models show advantage over classical machine learning models as they require no hyper-parameters for tuning and show better performance in small-scale datasets.

A key limitation of this study is that all experiments were conducted on (BRFSS) diabetes health indicators dataset. Although, it help to control experiments and focus on set size and model performance, generalizing findings required more investigations and experiments on other domains. Thus, as future work, further analysis required multiple datasets from different domains.

Also, to explore hybrid architecture pretrained models such TabPFN and regular models, hybrid model could complement the strengths of both approaches. TabPFN, with it pretrained in-context learning captures the

underlying data distribution and produces calibrated estimates on small data. TabPFN output calibrated with other models using stacking, voting, or as a feature augmentation.

Conflict of Interest: The authors declare that there are no conflicts of interest associated with this research project. We have no financial or personal relationships that could potentially bias our work or influence the interpretation of the results.

References

[1] H. O'Brien Quinn, M. Sedky, J. Francis, and M. Streeton, "Literature Review of Explainable Tabular Data Analysis," Oct. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/electronics13193806.

[2] R. Marcinkevičs and J. E. Vogt, "Interpretable and explainable machine learning: A methods-centric overview with concrete examples," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 13, no. 3, May 2023, doi: 10.1002/widm.1493.

[3] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Brief. Bioinform.*, vol. 19, no. 6, pp. 1236–1246, Nov. 2018, doi: 10.1093/BIB/BBX044.

[4] N. Hollmann *et al.*, "Accurate predictions on small data with a tabular foundation model," *Nature*, vol. 637, no. 8045, pp. 319–326, Jan. 2025, doi: 10.1038/s41586-024-08328-6.

[5] S. Somvanshi, S. Aaqib Javed, G. Antariksa, A. Hossain, and S. Das, "A Survey on Deep Tabular Learning," *2024. A Survey on Deep Tabular Learning*, vol. 1, no. 1, p. 43, 2024, doi: 10.48550/arXiv.2410.12034.

[6] A. A. Adegun, S. Viriri, and J. R. Tapamo, "Review of deep learning methods for remote sensing satellite images classification: experimental survey and comparative analysis," *J. Big Data*, vol. 10, no. 1, Dec. 2023, doi: 10.1186/s40537-023-00772-x.

[7] L. Grinsztajn, K. Flöge, O. Key, F. Birkel, P. Jund, B. Roof, B. Jäger, D. Safaric, S. Alessi, A. Hayler, M. Manium, R. Yu, F. Jablonski, S. B. Hoo, A. Garg, J. Robertson, M. Bühler, V. Moroshan, L. Purucker, C. Cornu, L. C. Wehrhahn, A. Bonetto, B. Schölkopf, S. Gambhir, N. Hollmann, and F. Hutter, "TabPFN-2.5: Advancing the state of the art in tabular foundation models," arXiv preprint arXiv:2511.08667, 2025.

[8] S. Sercan¨, S. Arık, and T. Pfister, "TabNet: Attentive Interpretable Tabular Learning," 2021. [Online]. Available: www.aaai.org

[9] A. Al-Nafjan, A. Aljuhani, A. Alshebel, A. Alharbi, and A. Alshehri, "Artificial Intelligence in Predictive Healthcare: A Systematic Review," Oct. 01, 2025, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/jcm14196752.

[10] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, 2022, doi: 10.1016/j.inffus.2021.11.011..

[11] H.-J. Ye, S.-Y. Liu, H.-R. Cai, Q.-L. Zhou, and D.-C. Zhan, "A closer look at deep learning methods on tabular datasets," arXiv preprint arXiv:2407.00956, 2024.

[12] M. Y. Ng *et al.*, "Perceptions of Data Set Experts on Important Characteristics of Health Data Sets Ready for Machine Learning: A Qualitative Study," *JAMA Netw. Open*, vol. 6, no. 12, p. E2345892, Dec. 2023, doi: 10.1001/jamanetworkopen.2023.45892.

[13] Q. W. Khan, K. Iqbal, R. Ahmad, A. Rizwan, A. N. Khan, and D. H. Kim, "An intelligent diabetes classification and perception framework based on ensemble and deep learning method," *PeerJ Comput. Sci.*, vol. 10, 2024, doi: 10.7717/peerj-cs.1914.

[14] J. Gao, Y. Lu, N. Ashrafi, I. Domingo, K. Alaei, and M. Pishgar, "Prediction of sepsis mortality in ICU patients using machine learning methods," *BMC Med. Inform. Decis. Mak.*, vol. 24, no. 1, Dec. 2024, doi: 10.1186/s12911-024-02630-z.

[15] A. Nazir, A. Hussain, M. Singh, and A. Assad, "Deep learning in medicine: advancing healthcare with intelligent solutions and the future of holography imaging in early diagnosis," *Multimed. Tools Appl.*, vol. 84, no. 17, pp. 17677–17740, May 2025, doi: 10.1007/s11042-024-19694-8.

[16] M. Chaieb, M. Azzouz, M. Ben Refifa, and M. Fraj, "Deep learning-driven prediction in healthcare systems: Applying advanced CNNs for enhanced breast cancer detection," *Comput. Biol. Med.*, vol. 189, p. 109858, May 2025, doi: 10.1016/j.compbiomed.2025.109858.

[17] K. Avazov *et al.*, "Bridging the Gap Between Accuracy and Efficiency in AI-Based Breast Cancer Diagnosis from Histopathological Data," *Cancers (Basel).*, vol. 17, no. 13, Jul. 2025, doi: 10.3390/cancers17132159.

[18] S. Sharma, R. K. Dudeja, G. S. Aujla, R. S. Bali, and N. Kumar, "DeTrAs: deep learning-based healthcare framework for IoT-based assistance of Alzheimer patients," *Neural Comput. Appl.*, vol. 37, no. 28, pp. 23017–23029, Oct. 2025, doi: 10.1007/s00521-020-05327-2.

[19] A. Alzahrani, "Early Detection of Lung Cancer Using Predictive Modeling Incorporating CTGAN Features and Tree-Based Learning," *IEEE Access*, vol. 13, pp. 34321–34333, 2025, doi: 10.1109/ACCESS.2025.3543215.

[20] S. Bang, M. O. Aarvold, W. J. Hartvig, N. O. E. Olsson, and A. Rauzy, "APPLICATION OF MACHINE LEARNING TO LIMITED DATASETS: PREDICTION OF PROJECT SUCCESS.," *Journal of Information Technology in Construction*, vol. 27, p. 732, Jan. 2022, doi: 10.36680/J.ITCON.2022.036.

[21] C. Varghese, E. Habermann, K. Hanson, A. Choudhary, H. Salehinejad, and C. Thiels, "Tabular foundation models as a new portable standard in local surgical risk prediction," *Surgery*, vol. 192, p. 110078, Apr. 2026, doi: 10.1016/j.surg.2025.110078.

[22] P. García, J. de Curtò, I. de Zarzà, J. C. Cano, and C. T. Calafate, "Foundation Models for Cybersecurity: A Comprehensive Multi-Modal Evaluation of TabPFN and TabICL for Tabular Intrusion Detection," *Electronics (Switzerland)*, vol. 14, no. 19, Oct. 2025, doi: 10.3390/electronics14193792.