Assessing the Robustness of Tabular Prior-Data Fitted Network Classifier

Ali Nawaz [1] **Amir Ahmad** [1] **Shehroz S. Khan** [2]

Abstract

Label noise is a common and critical challenge in real-world machine learning, especially in tabular data settings where mislabeled instances can severely degrade model performance and generalization. The proposed study investigates the robustness of the Tabular Prior-Data Fitted Network (TabPFN), a transformer-based model under varying levels of label noise in binary classification tasks. Using 15 publicly available tabular datasets from OpenML, we systematically inject label noise at multiple levels (0%, 1%, 5%, 10%, 20%, 25%, and 30%) and evaluate TabPFN against seven traditional classifiers, including Random Forest (RF), Extreme Gradient Boosting (XGBoost), Light GBM (LGBM), Support Vector Machine (SVM), K-Nearest Neighbor (kNN), CatBoost, and Decision Tree (DT). All models are assessed using 2×5-fold stratified cross-validation, and their performance is reported in terms of average accuracy and AUC-ROC. Our experimental results reveal clear performance trends across classifier types. Boosting-based models are most sensitive to label noise. RF demonstrates moderate robustness and maintains relatively stable performance across noise levels. In contrast, TabPFN consistently exhibits superior resilience to noise. These findings confirm the potential of TabPFN as a robust and noise-tolerant solution for real-world tabular classification tasks.

1. Introduction

Tabular data is one of the most widely encountered data modalities in practical machine learning applications (Shwartz-Ziv & Armon, 2022). It forms the backbone of structured datasets in healthcare, banking, fraud detection, manufacturing, cybersecurity, and software analytics, etc. Unlike unstructured data types such as images or text, tabular data often lacks spatial or temporal continuity and presents unique challenges due to its mixed data types, missing values, and irregular feature distributions (Adnan & Akbar, 2019). One of the most pervasive issues in tabular datasets is label noise, or class noise, where incorrect target labels are introduced into the training data, which may result from human annotation errors, instrumentation faults, ambiguous class definitions, or procedural inconsistencies (Johnson & Khoshgoftaar, 2022). For example, in a credit approval dataset, an applicant labeled rejected may have met all approval criteria but was mislabeled due to a processing error. Such label noise distorts the learning patterns and impairs the ability of models to capture underlying patterns (Song et al., 2022). In supervised learning, label noise severely impacts model generalization, particularly in high-capacity classifiers such as boosting models or deep networks, which tend to memorize noise rather than generalize from clean patterns (Song et al., 2022). Traditional methods to counter label noise include robust loss functions, reweighting schemes, and noise filtering, but these often require additional assumptions or introduce computational complexity (Frénay & Verleysen, 2013).

Recently, the Tabular Prior-Data Fitted Network (TabPFN) has emerged as a promising architecture tailored for tabular classification. TabPFN, (Hollmann et al., 2025) is a transformer-based model trained on millions of synthetically generated tabular tasks. At inference time, it performs a single forward pass to compute the Bayesian posterior predictive distribution for classification, making it highly efficient and data-agnostic. Its prior-data fitting approach allows it to adapt quickly to new datasets with few training samples and no explicit hyperparameter tuning. Despite its excellent performance on datasets, the robustness of TabPFN under noisy label conditions has not been systematically evaluated. Our contribution is the first systematic empirical assessment of TabPFN under various levels of label noise on diverse, real-world tabular datasets, providing practical insight rather than introducing a novel methodology.

[1]College of Information Technology and Big Data Analytics Center, United Arab Emirates University, P.O. Box 15551, Al Ain, United Arab Emirates (UAE) [2]College of Engineering and Technology, American University of the Middle East, Egaila, 54200, Kuwait. Correspondence to: Amir Ahmad <amirahmad@uaeu.ac.ae>.

Proceedings of the 1st ICML Workshop on Foundation Models for Structured Data, Vancouver, Canada. 2025.

Objectives

To assess the robustness of TabPFN under label noise, we evaluate the performance of TabPFN alongside other classifiers under increasing levels of symmetric label noise (0%, 1%, 5%, 10%, 20%, 25%, 30%).

2. Related Work

Several studies have examined the sensitivity of traditional classifiers to synthetic and real-world label noise.

(Frénay & Verleysen, 2013) presented a comprehensive survey that categorizes label noise into symmetric and asymmetric, and reviewed techniques ranging from noise-tolerant algorithms to instance filtering strategies. They highlight that DT and K-Nearest Neighbor (kNN) are highly susceptible to noise, while ensemble methods like Random Forests (RF) provide moderate resilience due to bootstrapped aggregation.

(Patrini et al., 2017) proposed a label noise correction mechanism based on estimating the noise transition matrix, offering robustness in deep neural networks. However, such techniques require accurate noise estimation, which is nontrivial in real-world settings. (Han et al., 2018) introduced co-teaching, a training paradigm in which two neural networks collaboratively learn from clean samples, excluding suspected noisy labels during training . While effective, it is tailored for deep networks and assumes availability of large-scale data. Recent benchmarking by (Wei et al., 2021) revealed that gradient boosting methods such as Extreme Gradient Boosting (XGBoost) and Light Gradient Boosting Method (LGBM) are among the most noise-sensitive models, especially when trained on high-capacity or small datasets. In contrast, RF exhibit greater robustness due to bagging and feature randomness.

The TabPFN, introduced by (Hollmann et al., 2025), represents a breakthrough by performing Bayesian posterior inference directly through a single forward pass (Hollmann et al., 2025). Trained on millions of synthetic tabular tasks drawn from plausible generative priors, TabPFN generalizes to unseen tasks by learning a universal prior over data distributions. This makes it highly sample-efficient and robust to noise, as it learns to ignore misleading features or examples if such patterns were seen during training.

While TabPFN has shown strong performance on several datasets, its application under noisy settings remains underexplored in the literature. Our study addresses this gap by systematically injecting class noise and evaluating its performance against conventional classifiers to examine the robustness of TabPFN.

3. Methodology

In this section, the detailed methodology is discussed along with datasets.

3.1. Datasets

We used 15 different binary classification datasets from the OpenML repository (ope, 2025) shown in Appendix A.1, ensuring diversity across different application domains. Only datasets with purely numerical features were selected to ensure compatibility with all classifiers, especially TabPFN.

3.2. Noise Induction

We applied symmetric label flipping to simulate class noise (Dietterich, 2000). Specifically, for a given noise rate $\eta \in \{1\%, 5\%, 10\%, 20\%, 25\%, 30\%\}$, a random subset of the training samples was selected and their labels were flipped ($0 \to 1$ or $1 \to 0$). The test set remained noise-free to allow fair evaluation.

This process was implemented using the equation 1:

$$y_{\text{noisy}}[i] = 1 - y[i], \quad \text{for randomly selected } i \tag{1}$$

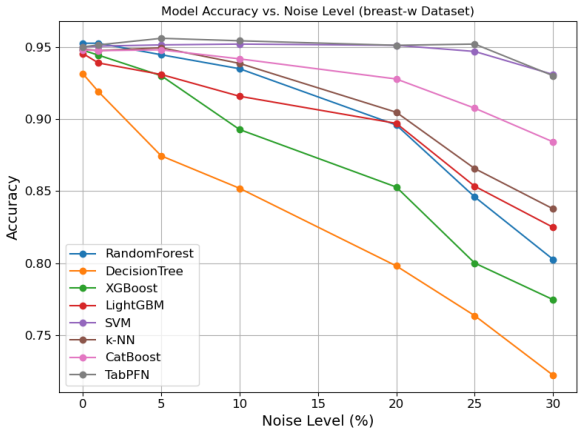
The mechanism mimics real-world label corruption scenarios and allows controlled experimentation on model robustness.

3.3. Models Applied

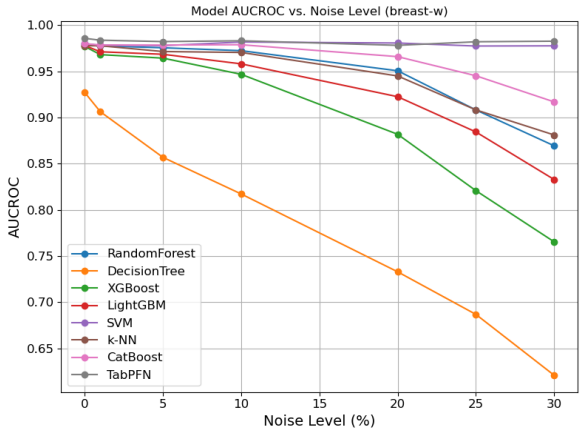
We evaluated the models, which are summarized in Table 1. Baseline models were run using common default hyperparameters. While this facilitates broad comparison, optimal performance may require dataset-specific tuning.

4. Experimental Results and Discussion

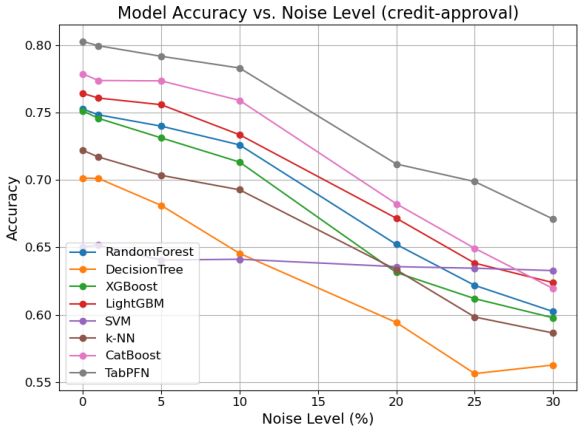
To assess the robustness of TabPFN under label noise, we conducted a comprehensive evaluation comparing TabPFN with other classifiers across increasing levels of symmetric label noise i.e., 0%, 1%, 5%, 10%, 20%, 25%, and 30%. Figures 1 demonstrate the impact of increasing symmetric label noise on model performance across different datasets using accuracy and AUC-ROC as evaluation metrics. In Figure 1 (a), the breast-w accuracy plot reveals that while all models perform well at 0% noise, accuracy drops noticeably for DT, XGBoost, and LGBM as noise increases, with TabPFN exhibiting minimal degradation and maintaining top performance throughout. Figure 1 (b) shows a similar trend for AUC-ROC on the same dataset, where boosting models and DT decline sharply, but TabPFN retains good class separation capabilities. Moving to credit-approval in Figure 1 (c), accuracy falls with increasing noise, especially for DT and boosting models, whereas TabPFN demonstrates strong



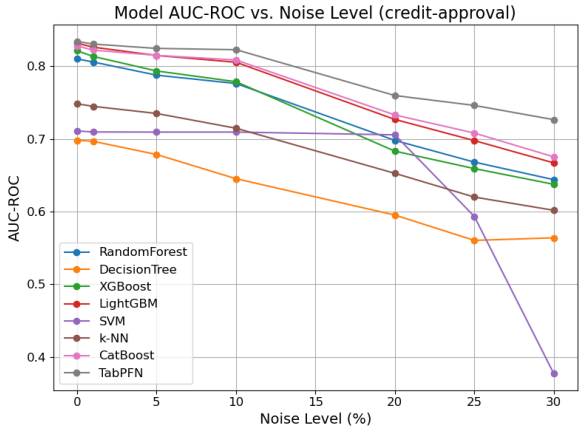
((a)) breast-w accuracy



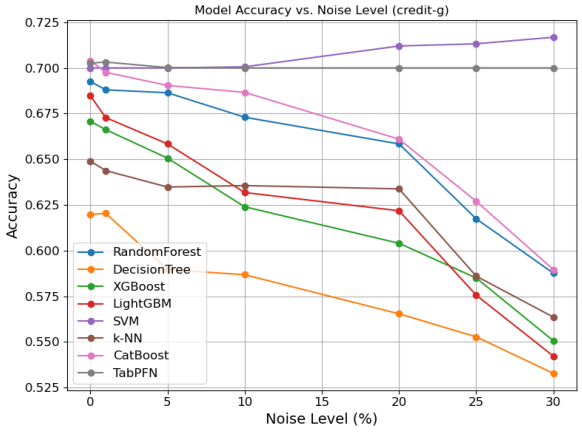
((b)) breast-w AUC-ROC



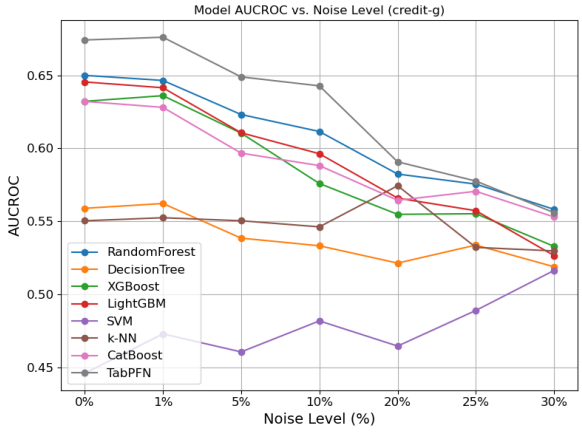
((c)) credit-approval accuracy



((d)) credit-approval AUC-ROC



((e)) credit-g accuracy



((f)) credit-g AUC-ROC

Figure 1. Accuracy and AUC-ROC results on a few datasets to illustrate the impact of increasing label noise on different classifiers.

Table 1. Summary of Classifiers: Definitions and Key Parameters

Model	Definition	Key Parameters / Characteristics
TabPFN	Transformer-based model that performs Bayesian posterior inference over tabular datasets.	Single forward-pass inference; robust to small data and noise.
Decision Tree (DT)	Classical CART model that splits features to maximize information gain.	criterion="gini", max_depth=None
RF	Ensemble ofDT built using bagging and random feature selection.	n_estimators=100, max_depth=None, bootstrap=True
XGBoost (XGB)	Gradient boosting framework that sequentially builds trees to minimize residual errors.	n_estimators=100, learning_rate=0.1, max_depth=6
LGBM	Fast, histogram-based leaf-wise boosting tree algorithm optimized for speed and accuracy.	boosting_type="gbdt", learning_rate=0.1, num_leaves=31
Support Vector Machine (SVM)	Margin-based classifier that finds the optimal separating hyperplane using kernels.	kernel="rbf", C=1.0, probability=True
k-Nearest Neighbors (k-NN)	Non-parametric algorithm that assigns class based on majority vote among k-nearest neighbors.	n_neighbors=5, metric="minkowski"
Category Boosting (CatBoost)	Gradient boosting library with native categorical support and ordered boosting.	iterations=100, depth=6, learning_rate=0.1, verbose=0

resilience. Figure 1 (d) further highlights the sensitivity of AUC-ROC in credit-approval, where models like k-NN collapse under noise and boosting methods degrade significantly, while TabPFN maintains a clear advantage. Figure 1 (e) displays credit-g accuracy trends, where TabPFN outperforms other models as most, including DT, XGBoost, and LBGM, show a sharp decline. Finally, Figure 1 (f) presents AUC-ROC for credit-g, confirming that TabPFN sustains its discriminative strength even as k-NN, DT, and other models deteriorate under noise. Across all figures, TabPFN consistently proves to be the most robust and reliable model under varying levels of label noise.

The results on other datasets are described in Appendix A.2.1. Our experimental results indicate distinct performance patterns among different classifier types. Boosting-based models such as XGBoost, LGBM, and CatBoost were notably the most sensitive to label noise, exhibiting significant performance degradation across accuracy and AUC-ROC metrics, especially visible in datasets like credit-approval, qsar-biodeg, and diabetes. RF demonstrated moderate robustness, maintaining relatively stable performance despite increasing noise, as evidenced by its performance on datasets such as pc3, banknote-authentication, and steel-plates-fault. In contrast, TabPFN consistently exhibited superior resilience, maintaining minimal performance reductions across all datasets tested. This was particularly apparent in datasets wdbc, breast-w, and climate-model-simulation-crashes, where TabPFN sustained high accuracy and AUC-ROC values even under high noise conditions.

TabPFN's robustness arises from its universal prior-data fitting during pretraining and Bayesian inference approach, which regularizes predictions and makes it less sensitive to noisy labels. Its architecture, exposed to synthetic noisy data, enables it to generalize patterns that avoid overfitting to spurious signals.

5. Conclusion and Future Work

In our experiments, we demonstrate that TabPFN demonstrates superior robustness to label noise when compared to traditional machine learning classifiers. It performs Bayesian posterior prediction over a wide range of synthetic datasets, allowing it to generalize well even under increasing levels of label corruption. We found out that TabPFN is more robust to noise than other classification models. In future work, we aim to extend this research by examining TabPFN's behavior under more complex noise settings, such as asymmetric or instance-dependent label noise. Additionally, we intend to apply the proposed ensemble approach to real-world domains such as medical diagnostics and anomaly detection, where label noise is inherently present and robustness is critical for deployment.

References

Openml repository,. https://www.openml.org/search?type=data&sort=runs&status=active, 2025. Accessed: 2025-05-08.

Adnan, K. and Akbar, R. Limitations of information extraction methods and techniques for heterogeneous unstructured big data. *International Journal of Engineering Business Management*, 11:1847979019890771, 2019.

Dietterich, T. G. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, 40:139–157, 2000.

Frénay, B. and Verleysen, M. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869, 2013.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., and Sugiyama, M. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018.

Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., Schirrmeister, R. T., and Hutter, F. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

Johnson, J. M. and Khoshgoftaar, T. M. A survey on classifying big data with label noise. *ACM Journal of Data and Information Quality*, 14(4):1–43, 2022.

Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., and Qu, L. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1944–1952, 2017.

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

Song, H., Kim, M., Park, D., Shin, Y., and Lee, J.-G. Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems*, 34(11):8135–8153, 2022.

Wei, J., Zhu, Z., Cheng, H., Liu, T., Niu, G., and Liu, Y. Learning with noisy labels revisited: A study using real-world human annotations. *arXiv preprint arXiv:2110.12088*, 2021.

A. Appendix

A.1. Used Datsets

The utilized datasets is collected from (ope, 2025) and are highlighted in Table 2.

Table 2. Dataset Mapping

Code	Dataset Name
D1	breast-w
D2	credit-approval
D3	credit-g
D4	diabetes
D5	pc4
D6	pc3
D7	kc2
D8	pc1
D9	banknote-authentication
D10	blood-transfusion-service-center
D11	ilpd
D12	qsar-biodeg
D13	wdbc
D14	steel-plates-fault
D15	climate-model-simulation-crashes

A.2. Experimental results

All classifiers were evaluated using 2×5-fold stratified cross-validation, i.e., 5-fold cross-validation repeated twice. The final performance metrics (accuracy and AUC-ROC) were reported as the average across all 10 folds, which reduces variance and improves the stability of the reported results.

A.2.1. Robustness analysis under different noise levels

Our experimental results across a diverse set of datasets highlight the effects of increasing label noise on model performance, measured using both accuracy and AUC-ROC. The figures 2 to 25 collectively demonstrate that boosting-based models such as XGBoost, LGBM, and CatBoost are among the most vulnerable to label noise. This degradation is especially evident in datasets such as qsar-biodeg (Figures 10 and 22), and diabetes (Figures 2 and 15), where these models show significant drops in both metrics as noise levels increase.

In contrast, RF exhibits a moderate level of robustness, maintaining stable accuracy and AUC-ROC on datasets such as pc3 (Figures 3 and 15), banknote-authentication (Figures 9 and 24), and steel-plates-fault (Figures 12 and 24). Notably, TabPFN consistently demonstrates superior resilience across all datasets and metrics. It maintains high accuracy and AUC-ROC even at higher noise levels, as seen in wdbc (Figures 11 and 23), and climate-model-simulation-crashes (Figures 13 and 25).

In the steel-plates-fault (Figure 24), TabPFN sustains near-perfect AUC-ROC despite increasing noise, while other models, particularly DT and SVM, deteriorate sharply. Similarly, in climate-model-simulation-crashes in Figure 25, most models show steep declines in AUC-ROC, with boosting models and SVM being heavily affected. Yet, TabPFN remains notably stable compared to all others, reinforcing its robustness under noisy conditions.

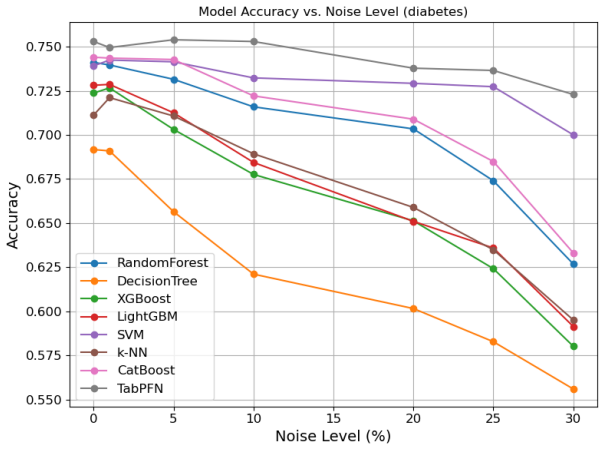


Figure 2. diabetes accuracy

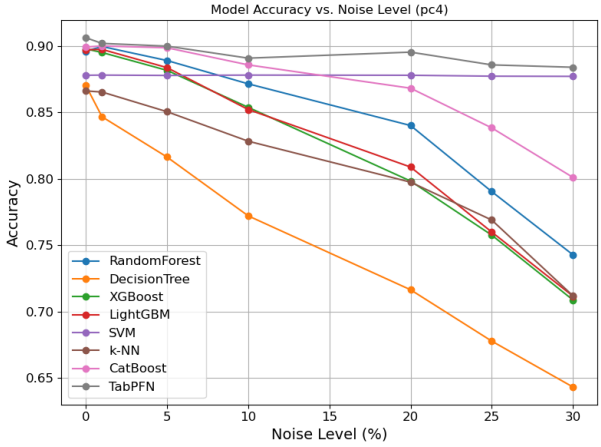


Figure 3. pc4 accuracy

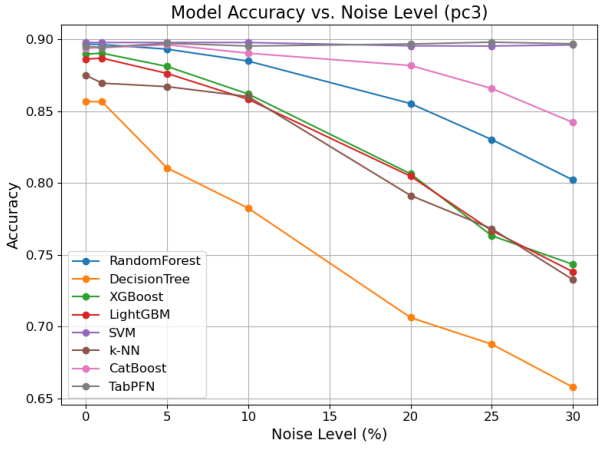


Figure 4. pc3 accuracy

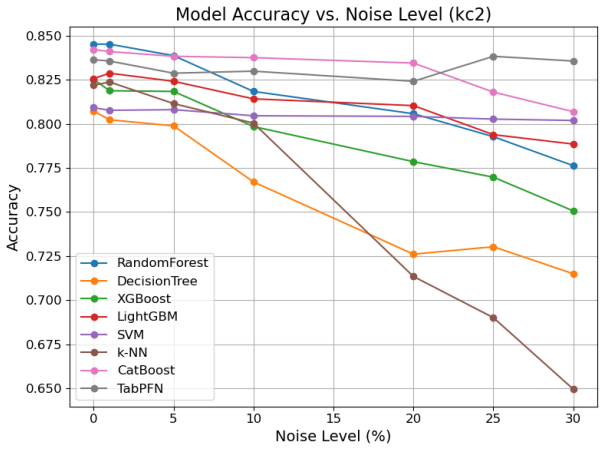


Figure 5. kc2 accuracy

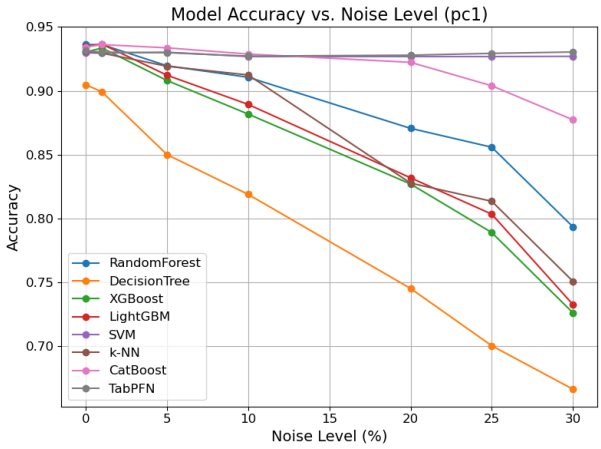


Figure 6. pc1 accuracy

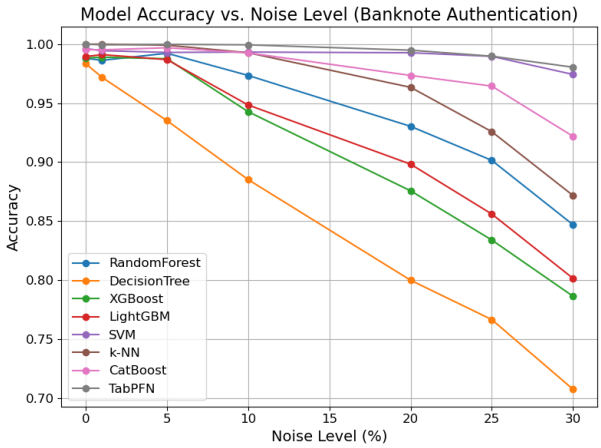


Figure 7. banknote-authentication accuracy

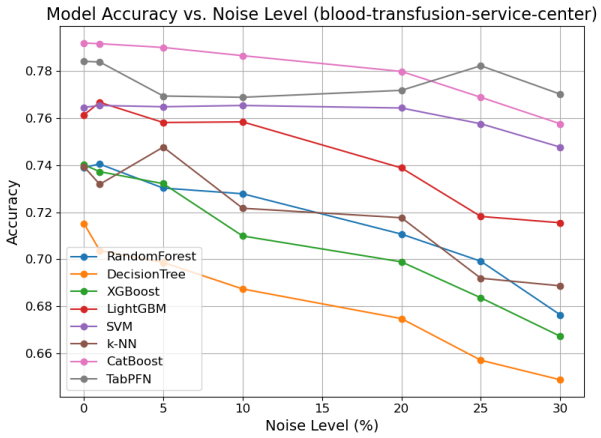


Figure 8. blood-transfusion-service-center accuracy

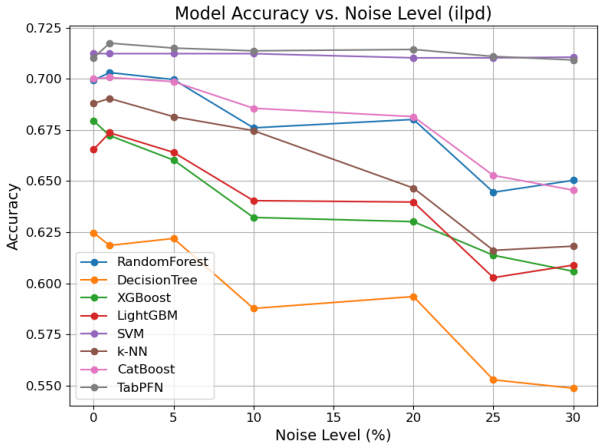


Figure 9. ilpd accuracy

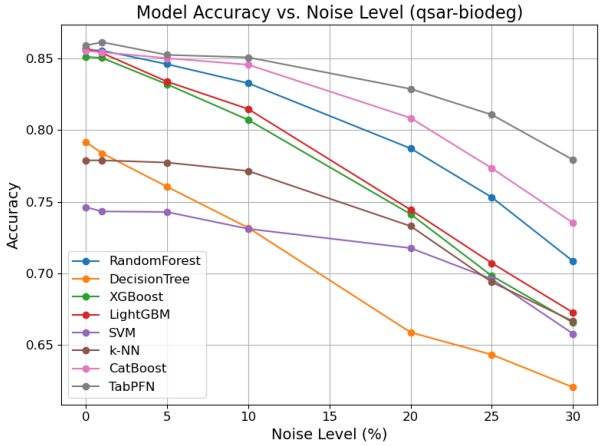


Figure 10. qsar-biodeg accuracy

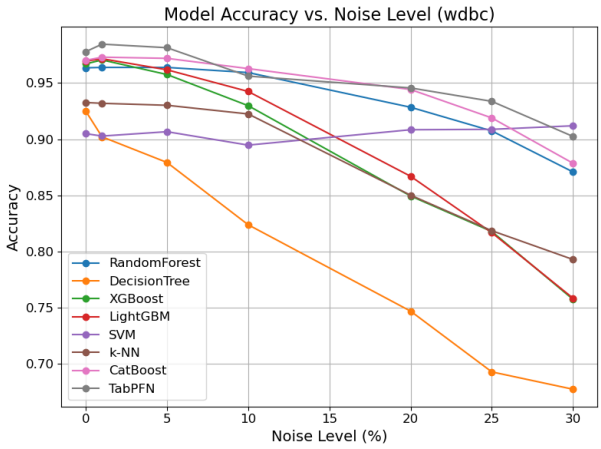


Figure 11. wdbc accuracy

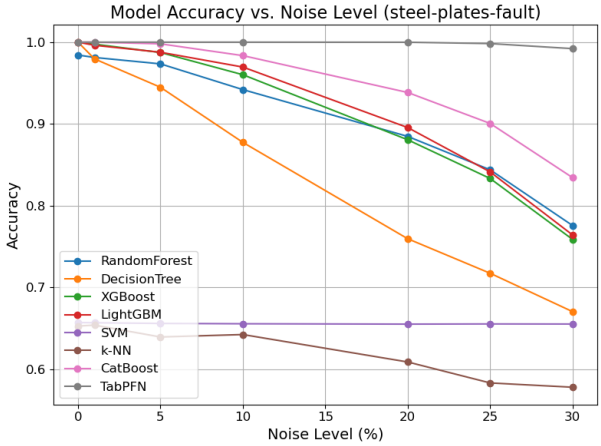


Figure 12. steel-plates-fault accuracy

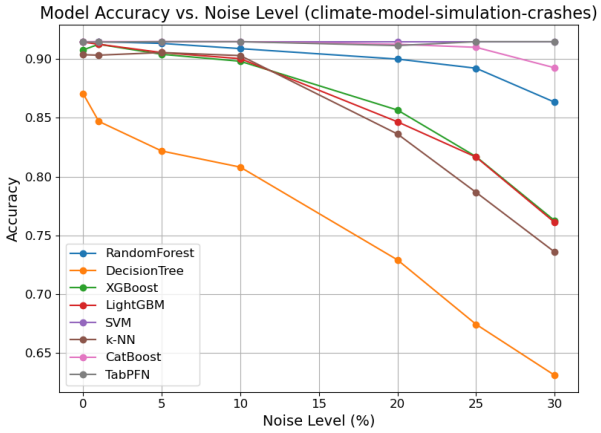


Figure 13. climate-model-simulation-crashes accuracy

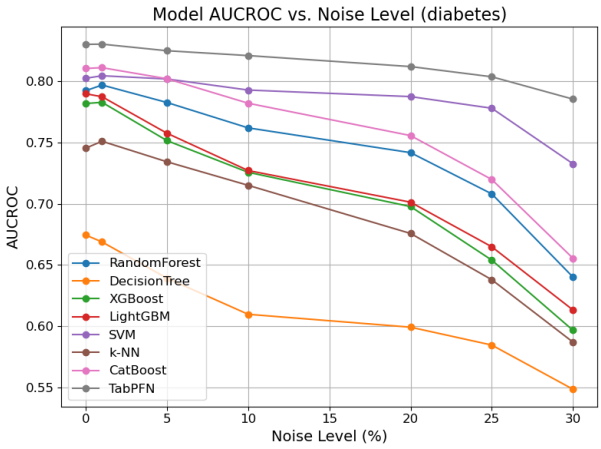


Figure 14. diabetes AUC-ROC

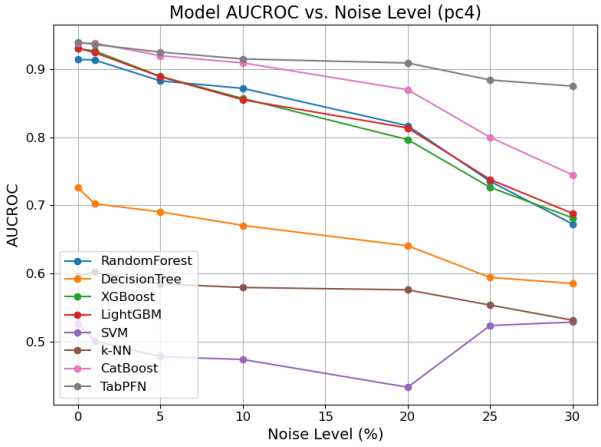


Figure 15. pc4 AUC-ROC

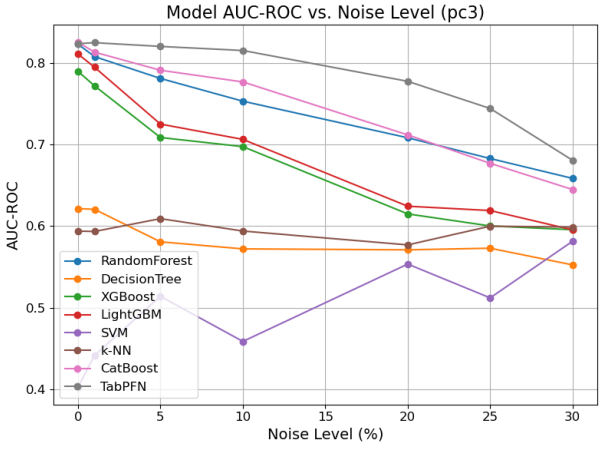


Figure 16. pc3 AUC-ROC

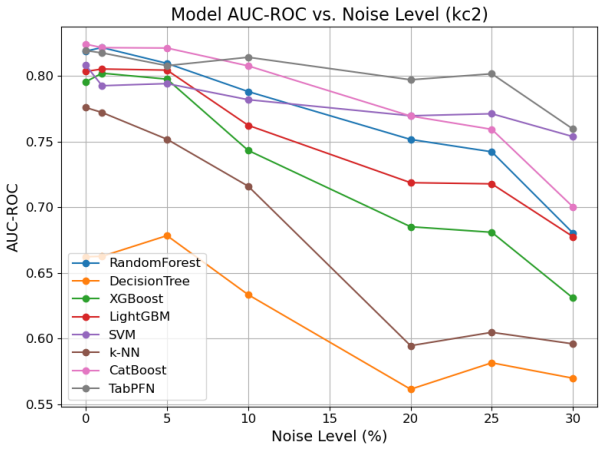


Figure 17. kc2 AUC-ROC

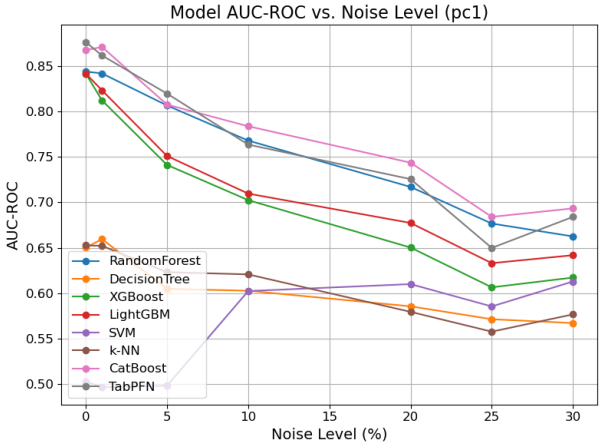


Figure 18. pc1 AUC-ROC

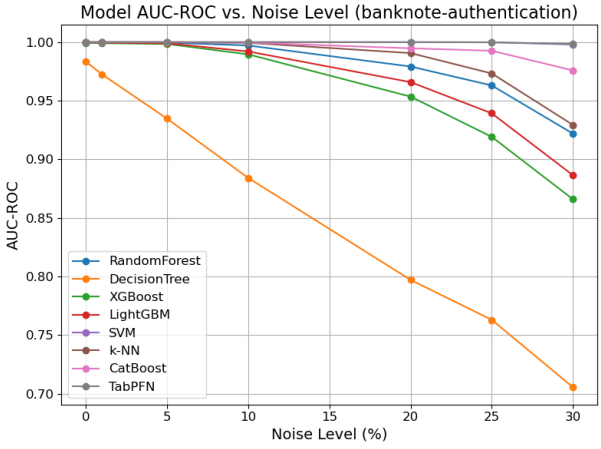


Figure 19. banknote-authentication AUC-ROC

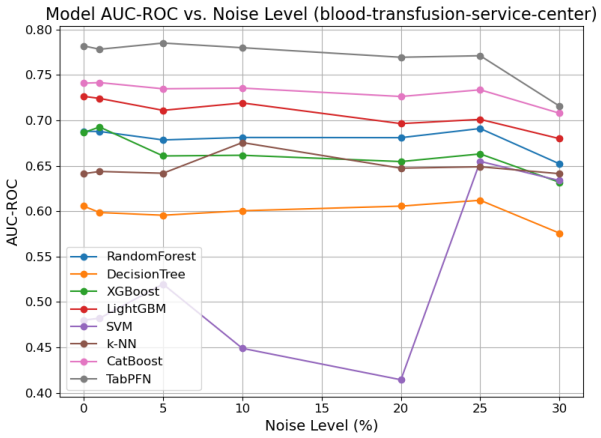


Figure 20. blood-transfusion-service-center AUC-ROC

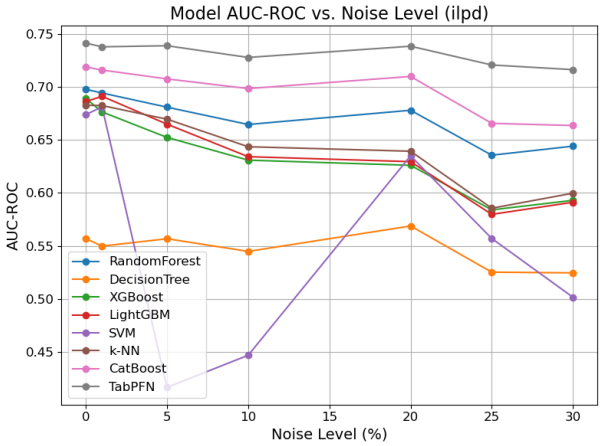


Figure 21. ilpd AUC-ROC

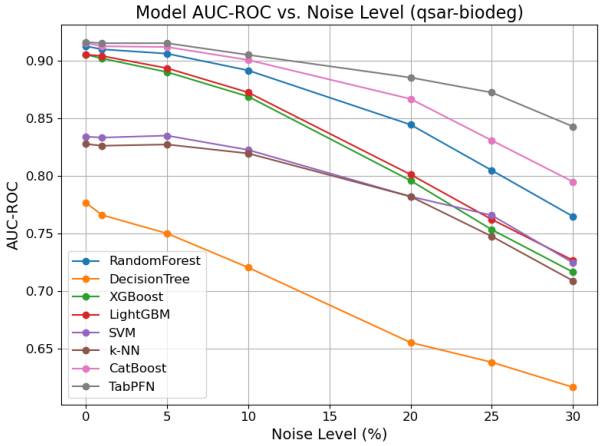


Figure 22. qsar-biodeg AUC-ROC

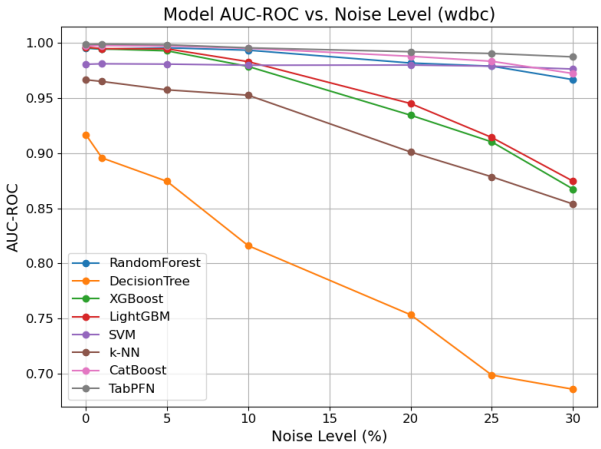


Figure 23. wdbc AUC-ROC

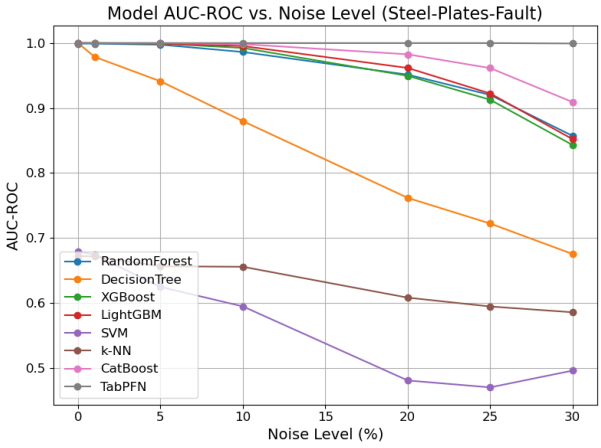


Figure 24. steel-plates-fault AUC-ROC

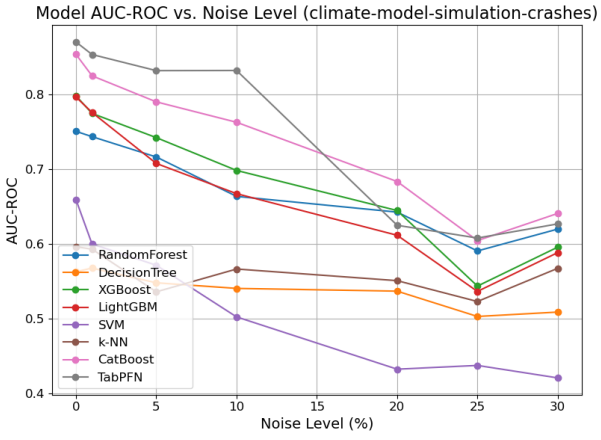


Figure 25. climate-model-simulation-crashes AUC-ROC