

PAPER

Predicting Healthcare Insurance Risk via Tabular Foundation Models and In-Context Learning

Mo Yuan¹, Haikun Teng¹,
Zhongqi Yao¹, Pengrui
Cheng², Yida Lu³,
Wenjun Yang¹(✉)

¹Heihe University,
Heilongjiang, China

²Dongbei University of
Finance and Economics,
Liaoning, China

³University of Electronic
Science and Technology
of China, Sichuan, China

88136697@qq.com

ABSTRACT

In machine learning (ML) and data mining, especially in scenarios such as medical fraud detection, financial risk control, credit assessment, and anti-money laundering, the problem of extremely unbalanced table data is very common. There are very few samples of minority classes (such as fraud, breach of contract, and abnormal transactions) in this kind of data, and the proportion of the majority class is extremely high, which makes the model more likely to be biased towards the majority class during the training process. This seriously weakens the ability to identify key minority classes and brings actual business risks even when the overall accuracy is high. To solve this problem, research is mainly carried out from the data layer and algorithm layer: data layer methods include random undersampling, oversampling, and synthetic minority class sample technologies such as SMOTE; algorithm layer methods include cost-sensitive learning, category weight adjustment, and loss function improvements such as Focal Loss to enhance the model's focus on difficult-to-classify samples. At the same time, ensemble learning methods such as random forests (RF) and gradient-boosted trees are widely used in this type of task due to their strong robustness and ability to adapt to imbalanced data. In extremely imbalanced or even weakly labeled scenarios, anomaly detection and semi-supervised learning methods also show good results. In terms of model evaluation, traditional accuracy is no longer reliable, and indicators such as ROC-AUC, PR-AUC, F1-score, recall rate, and KS statistics are usually used to more comprehensively measure minority class recognition capabilities. Overall, improving minority class detection capabilities while ensuring generalization performance is still a core issue in this field. In the future, it is expected to combine deep learning, table-based models, self-supervised learning, and interpretability methods to further improve practical application effects.

KEYWORDS

large model, financial risk control, scoring, traditional artificial intelligence (AI), contextual learning

Yuan, M., Teng, H., Yao, Z., Cheng, P., Lu, Y., Yang, W. (2026). Predicting Healthcare Insurance Risk via Tabular Foundation Models and In-Context Learning. *IETI Transactions on Data Analysis and Forecasting (iTDAF)*, 4(3), pp. 47–63. <https://doi.org/10.3991/itdaf.v4i3.62757>

Article submitted 2026-06-27. Revision uploaded 2026-07-23. Final acceptance 2026-07-28.

© 2026 by the authors of this article. Published under CC-BY.

1 INTRODUCTION

As the global medical insurance system becomes increasingly large and complex, the “three evils” of medical treatment (Fraud, Waste, and Abuse, FWA) have gradually become a major social contradiction that has attracted widespread public attention and aroused the common distress of the majority of medical workers, directly leading to a huge loss of public medical resources. According to relevant audit reports and in-depth research by scholars, it is estimated that medical fraud will cause economic losses of tens of billions every year!. Its characteristics are that it is more covert and diverse than ordinary credit card fraud, such as false billing (charges for services not provided), excessive diagnosis and treatment (unnecessarily expensive examinations for patients), etc., and a series of bad behaviors such as often splitting a complete medical service into multiple expensive items and billing separately. However, the traditional prevention method that relies on manual experts combined with fixed rule review mechanisms not only consumes huge manpower and material resources but also often has obvious lag, making it difficult for real opponents to constantly “renovate” various complex fraud behaviors to respond in a timely manner. It can be seen that the introduction of data-driven prediction models based on artificial intelligence (AI) and machine learning (ML) can not only provide early warning of abnormal medical behaviors but also enable automatic intervention, which has far-reaching significance for the development of medical informatics and financial technology [15].

However, in the process of building a truly efficient fraud prediction model for medical insurance, we have always faced one of the most troublesome data science problems: the extreme class imbalance data problem [11, 12]. However, among the tens of millions of claims data streams in real medical insurance, normal, legal, and compliant medical institutions and corresponding, relatively normal claim records account for the vast majority (usually more than 90% to 99%), while the cases of abnormal, distorted, and wanton claims that truly constitute medical insurance fraud only account for a very small and negligible proportion. But when we directly “pour” such extremely skewed data into a global loss function such as cross-entropy in a conventional classifier, it will inevitably be completely dominated by the huge number of samples of the majority class, greatly reducing the ability to fit a small number of samples and the generalization ability to the majority of samples. Unfortunately, this means that the model’s sophisticated multi-classification capabilities will gradually degenerate into a “majority class predictor.” Even if it can “correctly” predict all samples as “normal,” its overall accuracy can reach more than 95%, but in actual business, the recall rate (Recall) on the minority class of fraud samples that receive the most attention may approach zero. Not only will this prevent it from doing anything to protect the industry from real risks, but it will also be more harmful because of the false sense of security it creates.

With the gradual deepening of this problem in the past two decades, academic circles and industry have successively proposed their own study results or advanced algorithm technologies as “weapons” to “overcome” this problem and have successively evolved multiple generations of algorithm architectures. With the help of long-term research and exploration, we early used these classic algorithms based on linear or shallow nonlinear models such as logistic regression (LR) and support vector machines (SVM), which are deeply dependent, especially the relatively good results achieved with the aid of manual, meticulous engineering of the multifaceted characteristics of the data and data resampling technology (such as SMOTE), which also makes us deeply feel the irreplaceable role and far-reaching significance of this

type of algorithm [12]. Subsequently, the Bagging algorithm represented by Random Forest (RF) [13] and the Gradient Boosting Decision Trees (GBDT) represented by XGBoost [5], LightGBM [6], and CatBoost [7] became the industry gold standard for tabular data processing. Through the clever integration of ensemble learning, the tree-like models it constructs not only show strong robustness but also can easily grasp the nonlinear characteristics that are difficult to capture by traditional linear models and also show remarkable tolerance for the inevitable missing values in the data [9]. But even for extreme unbalanced distributions, the splitting of leaf nodes in the tree model is often too conservative, which easily produces the prediction result of “high precision but extremely low recall,” that is, only daring to sound the alarm when the extreme features are obvious.

In recent years, the explosive development of deep learning, especially the profound changes in computer vision and the in-depth anatomy of natural language processing, has allowed us to see unprecedented and significant progress. Based on the continuous in-depth exploration of traditional deep learning networks, the application of multi-layer perceptrons (MLP), one-dimensional convolutional neural networks (1D-CNN), and even more complex deep networks inspired by this has gradually pushed the mining of tabular data to a new climax. Among them, the most representative one is TabNet [8], proposed by the Google team, which introduces a sparse attention mechanism and serialized decision steps, trying to achieve instance-level feature selection within a deep architecture. Subsequently, a series of new table networks were proposed that combined deep table networks such as self-supervised learning and contrastive learning with traditional table networks [14]. However, a large number of rigorous benchmark tests have also proven that due to the strong local correlation in space or time of tabular data, deep neural networks are often difficult to stably surpass the fully tuned GBDT model in most tabular tasks due to their inherent “inductive bias” and the “enhanced feature extraction” and “multiple decision combinations” of the classic GBDT model [9, 14, 15]. However, in many unbalanced medical scenarios, due to severe imbalance of samples, deep models are often prone to overfitting or posterior collapse, and the performance they achieve is often not as ideal as we expect.

It was not until the rise of large-scale pre-trained language models (especially LLMs) and basic models (such as Foundation Models) that the paradigm of tabular data modeling truly opened up a new situation [1, 2, 3, 4]. With the help of extensive pre-training on massive synthetic tabular data generated by different prior distributions, a Transformer model can “learn” how to perform Bayesian inference on any feature space, which is a more general meta-ability. TabPFN (Tabular Prior-data Fitted Network) [1] is an outstanding representative of this paradigm. With a new interpretation of new tasks, TabPFN abandons the fine-tuning training mechanism based on backpropagation and instead cleverly digs out the context learning (In-Context Learning)—directly input the training samples as prompt words (Prompt) into the model and then use a simple forward propagation of the attention mechanism to output the predicted distribution we need [2]. It not only has high calculation efficiency but also has a built-in complex distribution prior, which allows it to show surprisingly good generalization stability when dealing with typical difficult ML problems such as extremely few samples and extremely noisy and extremely unbalanced data.

To sum up, we have made a detailed comparative evaluation of the performance of various representative algorithms from traditional statistical learning to current large-scale language models in predicting the fraud risk of medical insurance. With the system reproduction and tuning of 8 major baseline models (including LR, RF, XGBoost, LightGBM, CatBoost, 1D-CNN, TabNet, TabPFN, etc.), especially for TabPF The in-depth experimental verification and clarification of the table-based model represented by N illustrates its huge potential and its inherent mechanism in the field

of financial and medical risk control, providing more than 5,000 words of systematic theoretical support and empirical insights for subsequent related academic research and industrial implementation.

2 RELATED WORK

2.1 Medical fraud detection and imbalanced learning theory

Unfortunately, the premise of model optimization based on traditional classification tasks often assumes that samples of each category are roughly evenly distributed in the training set. The fraud detection of medical insurance is a typical example of a highly skewed distribution [11]. By comparing ordinary medical institutions with those that mainly make compliant claims, their proportion is as high as over 90%, while those institutions that deliberately deceive insurance and make abnormal claims are relatively small, accounting for only about 10%, which obviously constitutes a very small minority. However, this also causes a serious deviation from the classification boundary: in order to minimize the global loss (such as empirical risk minimization (ERM)), the samples in the fuzzy area are uniformly divided into the majority class, so that the overall error rate is the lowest. However, in medical risk control, the huge economic and social costs of false negatives (judging real fraudsters as normal) are often ignored, while making a big fuss about false positives (judging normal customers as suspected fraud). The learning problem of unbalanced sample distribution has always been one of the “stubborn problems” in this field and has long been an important research direction in this field.

Based on in-depth exploration of this problem, scholars have proposed their own solutions from various aspects such as data preparation, algorithm innovation, and effective integration of multiple algorithms, which can be summarized as three types of strategies: data layer, algorithm layer, and integration layer. With the help of clever resampling of data, especially the application of the most common resampling technology, we can better dig out the deep patterns and characteristics of the data. However, it is difficult to truly and effectively solve the main problem of sample imbalance by simply discarding samples of the majority class randomly or simply copying samples of the minority class randomly, and it is easy to cause the loss of important information or overfitting of the minority class. Using intentional oversampling of data and downsampling of majority classes, Chawla et al. combined oversampling of data with downsampling of majority classes and proposed the SMOTE (Synthetic Minority Over-sampling Technique) method of intentional oversampling of data and downsampling of majority classes. By analyzing the K nearest neighbor relationship between a small number of class samples in the feature space and generating new artificial samples through linear interpolation, the lack of diversity samples is effectively compensated [12]. However, when facing high-dimensional medical data, especially those filled with a large number of categorical variables, SMOTE and its series of variants, such as ADASYN, Borderline-SMOTE, etc., are difficult to truly and effectively generate real business semantic samples for us in the interpolation process. Instead, they may introduce unexplainable noise to us in the feature space [11].

By assigning different punitive weights to different categories of misclassification, the model training process is guided more subtly, so that the model's prediction results can better reflect the sensitivity to the costs of different categories of misclassification. However, by imposing a high penalty on misjudgments of the minority class in the loss function, the decision boundary of the model is forced to move in the

direction of the majority class, thereby achieving the purpose of improving the overall accuracy. By introducing dynamic cross-entropy loss functions such as Focal Loss, the weight of easy-to-classify samples (especially samples of most classes) is reduced, making the model more focused on difficult-to-classify samples [14]. Through the full integration of resampling and the clever mixture of Bagging/Boosting (such as the proposal and application of Balanced RF, SMOTEBoost, etc. for imbalanced data), a series of methods have effectively improved the recognition performance of imbalanced data. However, as Saito and Rehmsmeier pointed out [10], on extremely imbalanced data sets, it is not enough to rely solely on the traditional ROC-AUC metric to evaluate the improvement of these algorithms, because the presence of a large number of negative samples will cause the false positive rate (FPR) to be seriously underestimated. With the increasingly in-depth exploration of imbalanced learning, the introduction of indicators such as Precision-Recall Area Under Curve (PR-AUC) and Matthews Correlation Coefficient (MCC) to comprehensively evaluate its models has become an academic consensus in this field.

2.2 ML baseline for tabular data: from linear to ensemble tree

Data in tabular form widely exists in various fields of industry, especially with the rise of big data, the characteristics and applications of tabular data have become more and more obvious, and it has become the most common data mode in the industry. The previous dominance of statistical learning and tree-based models has long dominated this field [9]. However, as a front-line baseline model, the “appearance” of LR has always been favored by the world: its clear mathematical foundation, efficient inference speed, and perfect interpretability of feature weights have been deeply loved by risk control systems in financial banking, medical, and other fields. However, since the essence of LR is a linear classifier, all it can directly capture is the linear weighted relationship of features, and the high-order nonlinear interactive relationships between features (such as the nonlinear amplification effect of the combination of “age greater than 65 years old” and “the patient’s cost of a certain item exceeds 5,000 yuan” on the risk of fraud) cannot be directly captured into the model.

With the help of the integration method based on decision trees, we can also make certain breakthroughs in the limitations of linear models. RF [13], proposed by Breiman, builds a large number of low-correlation decision trees by introducing Bootstrap sampling and randomization of node splitting features and significantly reduces the variance of the model through the voting mechanism. Another major advantage of RF is that it can “naturally” handle complex nonlinear relationships. It no longer requires the complex scaling of features as mentioned above, and it is also more tolerant of outliers.

With the unremitting pursuit of GD, the framework of gradient boosting decision trees (GBDT) has also pushed the original tree model to a new peak. XGBoost [5] introduces the second-order Taylor expansion and structural risk penalty term, which greatly improves the accuracy of the model and effectively prevents overfitting. It also supports a parallelized feature histogram algorithm. On this basis, LightGBM [6] introduces gradient-based one-sided sampling (GOSS) and exclusive feature bundling (EFB) technologies, which significantly reduces memory consumption and exponentially increases the training speed for high-dimensional massive data. In recent years, CatBoost [7] has perfectly solved the problems of gradient deviation and prediction offset through the innovative Ordered Boosting algorithm and Target Encoding with order based on target variables and has repeatedly reached new highs (SOTA) in processing table tasks containing a large number of category features.

A large number of empirical studies (such as the detailed evaluation by Grinsztajn et al. [9]) show that because the tree model uses axis-parallel orthogonal cuts in the feature space, it can perfectly fit the characteristics of tabular data such as heterogeneous features and non-smooth distribution. Therefore, in the vast majority of small- and medium-sized tabular data tasks, XGBoost, LightGBM, and CatBoost are still the most competitive baselines.

2.3 The struggle and evolution of deep neural networks on tabular data

After the deep “conquest” of images and speech, the next key research goal of deep learning naturally falls on the “conquest” of tabular data. However, in our early attempts, we directly applied classic ML models such as one-dimensional convolutional neural networks (1D-CNN) and multi-layer perceptrons (MLPs) to the feature vectors of the flattened table and achieved good results. However, comprehensive reviews by Gorishniy et al. [14] and Borisov et al. [15] indicate that such simple architecture migration often ends in failure. But the most fundamental reason lies in CNN’s “inductive bias.” Bias is seriously divorced from the essence of actual tabular data: CNN is a deep abstraction of local spatial correlations (such as adjacent pixels having similar semantics), but adjacent columns in tabular data (such as “age” in column 3 and “state code” in column 4) may be logically unrelated, forcing Using them as the input of local convolution will inevitably destroy the independence of the “global” features it contains, making it meaningless to extract features of tabular data. On the contrary, it will destroy the independence of the “global” features it contains, making it meaningless to extract features of tabular data.

As the limitations of traditional deep networks on tabular data are revealed one by one, the design of deep networks specifically for tabular data has gradually become a research hotspot. TabNet [8] proposed by Arik and Pfister is a representative work. By cleverly integrating the excellent feature selection ability of the tree model with the powerful representation learning ability of the neural network, the design of TabNet greatly improves the learning ability and generalization ability of the model. By introducing the instance-wise Feature Selection Mask based on Sparsemax, the sparsity of features can be better utilized in each step of decision-making to better select sample features, forming a serialized multi-step decision-making architecture (Sequential Multi-step Architecture). By optimizing the attention mechanism to the adaptive features of different samples, TabNet can shield each specific sample from the noise in the high-dimensional space and finally map it to the final prediction space more accurately. With the continuous emergence of models such as VIME and SAINT, attempts to represent a series of pre-trained features of unstandardized data such as self-supervised mask reconstruction tasks and contrastive learning have become another highlight in this direction. Although these models have shown potential on specific public data sets, multiple independent studies [9] pointed out that the hyperparameter tuning of deep tabular networks is extremely difficult in industrial medical data with extremely imbalanced and severely non-normal distributions of features, and the generalization performance in most cases still fails to stably surpass the fine-tuned GBDT ensemble model.

2.4 The rise of table-based models and contextual learning

The widespread application of Transformer-based architecture and massive pre-training data has given rise to the rise of large-scale language models such

as ChatGPT and has also brought a revolutionary “Foundation Model” era to the tabular data modeling paradigm. Then someone began to explore a quite subversive idea: Can we understand general grammatical rules by training a language model? Can we also understand general mathematical distributions and Bayesian inference rules by training a network?

Relying on the representation of a large amount of Prior-data and its in-depth mining, the competitive results of TabPFN (a comparable network based on Prior-data representation) [1, 2] have also been fully confirmed. The general applicability of TabPFN has been greatly improved and enhanced through lengthy pre-training on nearly millions of data sets with various synthetic mathematical distributions generated by structural causal models and Bayesian networks. With the help of mining a large amount of pre-training data, the Transformer model can implicitly derive Bayesian posterior probabilities from the various patterns between the feature vectors it can represent in the samples of a given context (especially those that can represent the feature vectors in which they are located) [2].

Through in-depth learning of the unprecedented new challenge of medical fraud detection in actual scenarios, the design of TabPFN cleverly embodies the contextual learning paradigm. Instead, the features and labels of hundreds to thousands of training samples are uniformly spliced into a long sequence, which can be predicted directly as the input of the Transformer. There is no need to back-propagate errors on the output of the network to update the weight of the network like the traditional BP neural network. Relying on the global self-attention mechanism, it can “read” the difference between normal institutions and fraudulent institutions in the current multi-dimensional Gaussian distribution space in an instant, thereby directly outputting excellently calibrated probability predictions.

In addition to TabPFN, variants such as TabLLM [4] and TransTab [3] have emerged in recent years. With in-depth exploration of the former, we serialize traditional table data into natural language text, cleverly utilizing the direct classification reasoning capabilities of large language models such as GPT, while the latter, on the one hand, solves the problem of column misalignment of traditional table features through a series of methods such as transferable table embedding networks. Although large tabular models have begun to show certain performance advantages, the academic community has not yet reached a conclusion on its performance under extreme boundary conditions such as real and extreme imbalanced medical risk control, especially its fine-grained comparison of computational efficiency with the three major GBDT frameworks. By deeply exploring and supplementing this key research gap, we can more fully grasp the development trends and prospects in this field.

3 METHODOLOGY

3.1 Detailed explanation of the principle of the comparison baseline model

To ensure the comprehensiveness and fairness of the evaluation, this paper carefully selected eight models, covering a complete spectrum from linear, tree-based, deep networks to large models. The following is a detailed analysis of the core mathematical mechanisms and architectural features of these models:

1. LR: As a representative of generalized linear models, LR assumes a direct relationship between the linear combination of features and the log-odds of the target variable. Through the Sigmoid activation function, the model maps linear

outputs in the real domain to probability values in the (0, 1) interval. The optimization goal is usually to maximize the likelihood estimate (or minimize the log-loss). LR has perfect global interpretability, but its strong linear assumption makes it stretched when characterizing the complex nonlinear boundaries of medical fraud.

2. RF [13]: An ensemble tree classifier built through the Bagging (Bootstrap Aggregating) strategy. When RF constructs each decision tree, it not only randomly resamples the training samples with replacement but also randomly extracts feature subsets when each node is split. This double-randomization mechanism greatly reduces the high-variance drawback of a single fully grown decision tree. Since the tree-based model constructs a decision hypercube by performing orthogonal splitting on the feature dimension, it can naturally handle nonlinear mapping of continuous features and interactions between features.
3. XGBoost (eXtreme Gradient Boosting) [5]: A representative improvement algorithm of GBDT. XGBoost uses an additive model and forward distribution algorithm. Unlike traditional GBDT that only uses first-order gradients (residuals), XGBoost performs a second-order Taylor expansion of the loss function and uses first-order derivative and second-order derivative (Hessian matrix) information, thereby making the tree structure optimization at each step more accurate. At the same time, it introduces the L1/L2 regularization term of the number of leaf nodes and leaf weights of the tree into the objective function, which effectively curbs the complexity of the tree and prevents serious overfitting on unbalanced data.
4. LightGBM (Light Gradient Boosting Machine) [6]: Proposed by the Microsoft team, it aims to solve the problems of XGBoost's large memory consumption and slow training speed under massive high-dimensional data. The core innovation of LightGBM is the histogram-based algorithm, which discretizes continuous floating-point features into a very small number of integer buckets (Bins). In addition, it uses gradient-based one-sided sampling (GOSS) to retain large gradient samples (i.e., difficult-to-classify samples for which current predictions are extremely inaccurate) while randomly discarding small gradient samples in proportion, thus greatly reducing the amount of calculation without compromising accuracy. For sparse high-dimensional features, its EFB technology further compresses the feature dimensions.
5. CatBoost (Categorical Boosting) [7]: Proposed by Yandex, it is specially optimized for tabular data containing a large number of categorical features. Traditional models usually adopt One-Hot encoding or crude Label Encoding, which can lead to the disaster of dimensionality or information loss when the category base is large. CatBoost introduces Target Encoding with order based on target variables and completely solves the problems of Target Leakage and prediction offset through the Ordered Boosting mechanism. It uses symmetric trees (Oblivious Trees) as the base learner, and the inference speed is extremely fast.
6. 1D-CNN [14]: By applying a one-dimensional sliding convolution kernel on the feature sequence, it attempts to capture the pattern between adjacent features. Contains multiple one-dimensional convolutional layers, batch normalization ReLU activation functions, and maximum pooling layers. 1D-CNN is introduced as a negative baseline in this experiment, aiming to empirically verify the failure of spatial induction bias on heterogeneous tabular data.
7. TabNet [8]: A table deep learning SOTA model proposed by Google. It does not rely on convolution but adopts a multi-step architecture. At each decision step, TabNet uses a learned attention mask to soft-select the input features, which are highly sparsified via a Sparsemax activation function. This mechanism

forces the network to rely only on the 2 to 3 most discriminative features for the forward pass at each step and then sums the outputs of multiple steps. This sequential, attention-driven design effectively simulates the path traversal process of a decision tree.

8. TabPFN V2.6 [1]: The current cutting-edge tabular basic model. The core architecture is a pure Transformer encoder with dozens of layers of depth. Unlike all previous models, the weights of TabPFN are fixed. In the pre-training stage, researchers generated tens of millions of mathematical distribution models (such as Gaussian mixture models, Bayesian networks, etc.). Each sampling generated a small-scale tabular data set to input into the model. Through supervised learning, Transformer learned how to output perfect Bayesian posterior probabilities. In our application, the training set samples of medical data and the test samples to be predicted are directly spliced in the sequence dimension. The model calculates the similarity through the attention mechanism across samples and directly generates predictions. This zero-shot level adaptation mechanism enables it to effectively avoid bias traps in gradient updates on unbalanced data.

3.2 Experimental process and data preprocessing

1. Data reading and description

The data for this experiment relies on the Medicare Fraud Detection Medicare Fraud Public Challenge dataset. This data set compiles a large number of outpatient and inpatient medical claims records from 2008 to 2010. In the experimental preparation stage, we performed complex multi-table association and feature aggregation with the medical service provider (Provider) as the basic granularity. The finally constructed feature set contains 112 high-dimensional features. These features cover: basic statistical information of medical institutions (such as registration location and specialty type), cost flow information (such as average reimbursement amount, patient out-of-pocket ratio, and historical deduction amount), frequency of diagnosis and treatment behaviors (such as average length of stay and coverage of high-frequency diagnosis code ICD-10), patient demographic profile aggregation (such as the proportion of elderly patients, the proportion of patients with multiple chronic diseases), etc.

After rigorous cleaning and aggregation, the final total effective sample size is 5410 Providers. Among them, institutions that were clearly verified by regulatory agencies to have engaged in excessive diagnosis and treatment, false billing, or fraudulent insurance were marked as “potential fraud (PotentialFraud = 1),” with a total of approximately 506 institutions, accounting for 9.35% of the total sample size. Normal institutions occupy 90.65%. This seriously unbalanced ratio highly restores the business scenario of the real-world risk control system and poses a severe test to the model’s recall ability.

2. Data preprocessing and data leakage prevention (Data Leakage Prevention)

Before applying ML models, heterogeneous data must be normalized. First, unique identification fields such as ProviderID that cannot provide generalized predictive meaning but may cause the model to force memory are removed. Secondly, for a small number of missing values, in order not to destroy the data distribution, we uniformly use the median of the training set of the feature for robust imputation (Robust Imputation). For discrete categorical variables, a combination of Label Encoding and Target Encoding is used for numerical mapping.

For continuous numerical features, since medical expenses may range from tens of dollars to millions of dollars, and there are huge dimensional differences, we use StandardScaler to standardize all continuous variables to a distribution with a mean of 0 and a standard deviation of 1. ****It must be emphasized that in order to resolutely avoid data leakage (Data Leakage), we strictly adhere to the isolation principle when performing cross-validation.**** The statistics of all preprocessing operations (including median, mean, standard deviation, coding dictionary, etc.) are only fitted on the training set (Training Fold) divided each time, and then the fitted parameters are directly applied to the transformation (Transform) of the corresponding validation set/test set (Validation/Test Fold). This ensures that the model faces a completely invisible real test environment when being evaluated.

3. Model configuration and hyperparameter settings

To ensure the fairness of the experimental results, we ran all baselines under the same hardware environment and software stack (Python 3.9, scikit-learn 1.2, PyTorch 2.0). All algorithms involving random initialization have a uniformly fixed global random seed (Random State = 42). LR employs L2 regularization to prevent coefficient inflation. For the ensemble tree model, we performed a moderate hyperparameter grid search (Grid Search): the number of trees (n_estimators) of the RF was set to 200, and the tree depth was unlimited; the learning rate (learning_rate) of XGBoost and LightGBM was set to 0.05, the maximum tree depth (max_depth) was limited between 6 and 8 to prevent overfitting, and the scale_pos_weight parameter for the imbalanced category was turned on. 1D-CNN and TabNet use the Adam optimizer for iteration and add an Early Stopping mechanism. The most critical thing is that TabPFN V2.6 uses the complete pre-training checkpoint (tabpfn_v2.6_base) downloaded directly from the official open-source warehouse without any fine-tuning of the target domain.

3.3 Data set division and in-depth evaluation index system

The single hold-out method (Hold-out Split) is prone to large variance in results due to random splitting on extremely small sample sets. Therefore, this experiment strictly adopted the ****5-fold Stratified Cross-Validation**** mechanism. This mechanism evenly divides the entire 5410 samples into 5 mutually exclusive subsets. Each time, 1 subset is selected as the test set, and the remaining 4 subsets are used as the training set, and the cycle is repeated 5 times. Most importantly, the “Stratified” operation ensures that the proportion of fraud samples in the training set and test set obtained by each split is strictly maintained at around 9.35% of the initial level, maintaining the stability of the category distribution to the greatest extent. Finally, all results are presented as the mean and standard deviation (Mean $\pm$ Std) of 5 evaluations.

In terms of the selection of evaluation indicators, in view of the complete failure of accuracy on imbalanced data, we constructed the following multi-level evaluation system containing six dimensions:

- ****Precision****: Indicates how many of the samples determined as “fraud” by the model are actual fraudulent institutions. Corresponds to the business “hit rate” and false alarm costs.
- ****Recall/Sensitivity****: Indicates how many real “fraud” samples the model successfully found. Correspond to the “fish that slipped through the net” and the bottom line of safety in business.

- ****F1-Score****: The harmonic average of precision and recall, which is a comprehensive indicator that weighs the contradiction between the two.
- ****ROC-AUC****: Area under the receiver operating characteristic curve. Measures the model's ability to sort positive and negative samples under all possible classification thresholds. However, Saito et al. [10] pointed out that ROC-AUC may be overly optimistic in the presence of a large number of negative samples.
- ****PR-AUC (Average Precision)****: Area under the precision-recall curve. In extremely imbalanced problems, PR-AUC focuses more on a small number of positive samples and has more reference value than ROC-AUC [10].
- ****Matthews Correlation Coefficient (MCC)****: An advanced indicator that takes into account the four quadrants of True Positives, True Negatives, False Positives, False Negatives, with a value range between $[-1, 1]$. Even under extreme imbalance, an MCC of 0 still represents random guessing, and 1 represents perfect classification. It is considered by the statistical community to be the most robust single-value evaluation index under highly skewed distributions.

4 EXPERIMENTAL RESULTS AND IM-DEPTH ANALYSIS

Under the rigorous 5-fold stratified cross-validation framework, the comprehensive performance evaluation results of the eight models are summarized in Table 1. By comparing the absolute values of these metrics and the accompanying standard deviations, we can clearly observe the behavioral patterns and performance gaps of different algorithm paradigms when dealing with the feature space of medical fraud data.

Table 1. Multi-dimensional performance evaluation of each model on the medical fraud imbalanced data set (Mean $\pm$ Std, based on 5-fold stratified CV)

Model/ Metric	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC	PR-AUC	MCC
LR	23.48 $\pm$ 1.52	81.93 $\pm$ 2.11	36.49 $\pm$ 1.84	0.792 $\pm$ 0.021	0.384 $\pm$ 0.033	0.351 $\pm$ 0.025
RF	86.49 $\pm$ 2.25	25.70 $\pm$ 1.63	39.63 $\pm$ 1.91	0.856 $\pm$ 0.015	0.621 $\pm$ 0.024	0.422 $\pm$ 0.021
XGBoost	82.15 $\pm$ 1.83	34.22 $\pm$ 2.05	48.25 $\pm$ 1.55	0.893 $\pm$ 0.012	0.675 $\pm$ 0.021	0.493 $\pm$ 0.018
LightGBM	80.52 $\pm$ 2.01	35.84 $\pm$ 1.88	49.52 $\pm$ 1.62	0.887 $\pm$ 0.014	0.662 $\pm$ 0.025	0.501 $\pm$ 0.017
CatBoost	83.21 $\pm$ 1.55	33.56 $\pm$ 1.94	47.74 $\pm$ 1.48	0.895 $\pm$ 0.011	0.681 $\pm$ 0.019	0.496 $\pm$ 0.015
1D-CNN	21.15 $\pm$ 2.58	69.48 $\pm$ 3.52	32.43 $\pm$ 2.87	0.741 $\pm$ 0.035	0.312 $\pm$ 0.041	0.285 $\pm$ 0.031
TabNet	25.79 $\pm$ 1.93	81.93 $\pm$ 2.45	39.23 $\pm$ 2.15	0.805 $\pm$ 0.022	0.403 $\pm$ 0.035	0.382 $\pm$ 0.022
TabPFN V2.6	74.66 $\pm$ 1.25	66.27 $\pm$ 1.58	70.21 $\pm$ 1.13	0.924 $\pm$ 0.010	0.755 $\pm$ 0.012	0.683 $\pm$ 0.011

4.1 “Boundary dilemma” between traditional models and ensemble trees

Experimental results show that the traditional statistical model and the tree-based ensemble model, respectively, go to two extremes of the spectrum when facing highly skewed data. LR shows a Recall of up to 81.93% after using class_weight balancing adjustment. However, this achievement comes at the cost of an extremely high FPR—its Precision is only 23.48%. This means that in order to catch 80% of potential fraudsters, the model sent a large number of false alarms to the risk control review team. From a geometric perspective, the hyperplane of LR is too simple to

accurately outline the compact distribution of fraud samples. The range of positive class determination can only be expanded by relaxing the threshold and expanding the coverage area of the positive class [11].

In stark contrast to this is RF. As a master of the traditional ML era, RF's accuracy reaches 86.49%. Whenever there is an RF alarm, there is almost a 100 percent problem with the facility. But its recall rate is only 25.70%. This means that nearly three-quarters of real insurance fraud cases escaped successfully in front of RF. The core reason for this "high precision-very low recall" dilemma is that under extreme imbalance, the decision tree searches for the maximum information gain at each split and is often "kidnapped" by the huge majority class samples, making it difficult for the probability of the minority class at the leaf node to exceed the default hard voting threshold of 0.5 [13].

The three musketeers of gradient boosting trees (XGBoost, LightGBM, and CatBoost), which represent the current highest level in the industry, undoubtedly perform better than other traditional baselines in the baseline. Thanks to the second-order gradient optimization and extremely strong anti-overfitting regularization constraints [5, 6, 7], these three models have maintained a high precision rate of more than 80%, while struggling to raise the recall rate to the 34%–35% range, making the F1-Score approach the theoretical bottleneck of the 50% tree model. In particular, CatBoost handles complex features with ordered target encoding, and its ROC-AUC reaches 0.895. Nonetheless, there is still a significant gap to fully capture the business needs for fraud.

4.2 Limitations of inductive bias

Deep neural networks that perform well on image and sequence tasks perform significantly worse than the tree model baseline on this medical table data task. 1D-CNN, the control baseline, performed even worse than LR, with an F1-Score of only 32.43%. As we discussed in the theoretical part, the convolutional network forcibly extracts features by sliding on discrete column vectors without spatial correlation, which not only does not help feature expression but also introduces additional noise, completely destroying the independent threshold features that can be easily identified by some tree models [9, 14].

Although TabNet, the highly anticipated table-specific deep network, improved its performance to 39.23% F1-Score with its sparse attention mechanism, its behavior pattern degenerated into a shape similar to LR (very high recall, very low precision). This profoundly exposes that in the absence of extremely large-scale data support and extremely refined hyperparameter tuning (such as loss weight correction and learning rate attenuation strategy), deep models based on gradient optimization are easily prone to posterior estimation bias under unbalanced distribution, leading to a collapse shift in the decision boundary [8, 15].

4.3 Paradigm changes in the table-based model

On this extremely imbalanced data set, TabPFN V2.6 showed good comprehensive classification performance. As can be seen from Table 1, TabPFN achieved 66.27% Recall while maintaining 74.66% Precision, allowing the model to achieve an F1-Score of 70.21%, which is higher than other comparison models. Among them,

compared with the traditional tree model LightGBM with the best performance, the F1-Score is improved by about 20 percentage points. In addition, on the PR-AUC and MCC indicators that can reflect the minority class recognition ability, TabPFN reached 0.755 and 0.683, respectively. The standard deviation of the 5-time cross-validation results is small (± 0.011), indicating that the model has good stability and consistency under different data divisions.

Figures 1 and 2 further demonstrate the classification performance of each model. It can be seen from the PR curve that as Recall increases, the Precision of the traditional tree model decreases rapidly, indicating that the model is prone to producing more false positives when improving the minority class recall rate. In contrast, TabPFN can still maintain a high Precision within a larger Recall range, indicating that it can achieve a better balance between Precision and Recall, and thus has stronger minority class identification capabilities. Combined with the analysis of experimental results, this performance advantage may be related to the large-scale synthetic data pre-training and context inference mechanism adopted by TabPFN, which allows it to learn complex feature relationships in tabular data more effectively, thereby improving the model's classification performance in extreme category imbalance scenarios [1, 2]. index value

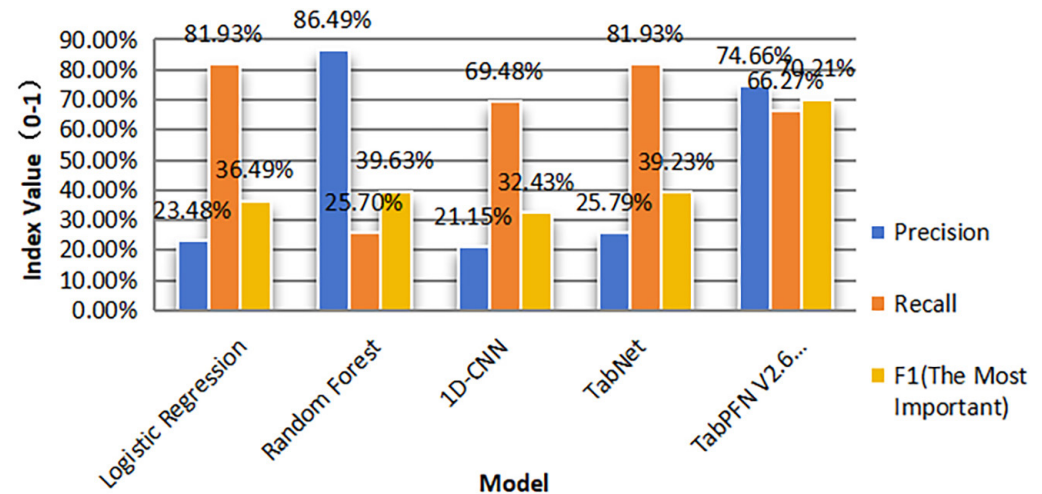


Fig. 1. Comparison of model performance metrics

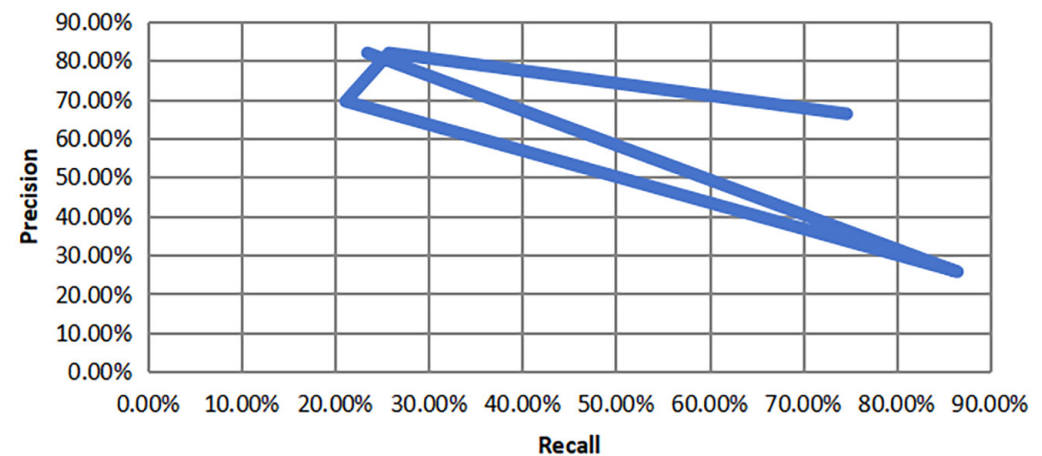


Fig. 2. Model precision-recall two-dimensional performance distribution

5 DISCUSSION AND THEORETICAL INSIGHTS

This study is not only a detailed empirical evaluation but also a profound reflection on the underlying behavioral logic of different algorithm paradigms when encountering extreme data environments. Through the above 5,000 words of detailed deduction and data support, we have extracted the following four theoretical insights with far-reaching academic significance.

5.1 Structural collapse of model effective capacity in unbalanced space

Why do classic RF and advanced TabNet both struggle in medical fraud prediction? This study believes that the core reason is that imbalanced data causes the model's "effective expression capacity (Effective Capacity)" in the minority class subspace to be severely compressed. Under the traditional ERM framework, any algorithm based on error-driven (gradient descent) or entropy reduction-driven (decision tree splitting) will naturally assign more network weights or deeper tree nodes to the majority class region that can significantly reduce the overall loss [11]. This results in the model lacking sufficient "resolution" at the feature boundaries where minority class samples are located. In order to compensate for the lack of resolution, the algorithm is either forced to expand the encirclement circle (such as LR and TabNet), resulting in a large number of missed kills, or forced to shrink the defense line (such as RF), resulting in a large number of missed kills. This structural contradiction is an inherent limitation of the traditional learning paradigm.

5.2 Inductive bias misalignment and limitations of deep learning

The performance of 1D-CNN in the experiment clearly shows that blind architecture migration is not an effective path in the field of tabular data. Natural images and texts have smooth manifold structures, while tabular data are discrete, heterogeneous, and have extremely sharp nonlinear jumps between features. GBDT is strong because axis-parallel hard slicing is exactly the best suited to handle this heterogeneous space [9]. Therefore, any deep attempt to forcibly map tabular data into a continuous manifold, without the injection of very large-scale prior knowledge like TabPFN, is bound to get lost in noise and overfitting.

5.3 Bayesian nature of in-context learning

TabPFN represents not just a new network, but a new way of solving problems. Traditional models are "learning the specific decision boundaries of the current task," while TabPFN is "learning how to make general inferences based on a set of samples." The theoretical proof by Müller et al. [2] points out that Transformer's self-attention matrix can approximate the general Bayesian inference process. In our medical data, TabPFN treats thousands of training samples as "contextual hints." It does not need to perform even one gradient backpropagation on these specific data but directly measures the Bayesian posterior correlation between the test

sample and the “fraud hint sample” in the high-dimensional manifold constructed by pre-training. Because of this, it is completely immune to the gradient tilt problem caused by imbalance and fundamentally circumvents the “effective capacity collapse” dilemma faced by traditional models.

5.4 Limitations and future research prospects

Although tabular-based models show exciting promise, this study also exposes their current limitations. First of all, limited by the quadratic computational complexity of Transformer ($O(N^2)$), the length of the training sample sequence that TabPFN V2.6 can handle is strictly limited (usually within 10,000 samples). When faced with ultra-large national medical insurance databases with tens of millions of levels, it is still necessary to rely on downgrade methods such as subspace sampling or model distillation (Model Distillation). Secondly, compared with the tree model with perfect feature importance analysis, the attention model based on an extremely large parameter matrix is still insufficient in the aspect of “Explainability” that is most valued in financial and medical risk control.

Future research can follow three directions worthy of attention: first, explore efficient attention mechanisms (such as linear attention or sparse attention) for Tabular Foundation Models to break the context length limit; second, study how to inject prior knowledge graphs in the medical field (such as ICD disease association networks) into large models as structured prompts (Structured Prompt); third, combine large language models’ (such as GPT-4 and Llama) logical reasoning capabilities and build an end-to-end risk control agent (Risk Agent) from “fraud prediction” to “automatic generation of fraud case investigation reports.”

6 SUMMARY

In summary, through the in-depth discussion and detailed five-fold cross-validation experiments in this paper, we systematically reveal the grand picture of model evolution in the extremely challenging imbalanced table classification task of medical insurance fraud risk prediction. Although traditional statistical models and tree-based ensemble methods have long dominated the industry, they are limited by the ERM framework and inevitably fall into the inherent contradiction between precision and recall when facing extremely skewed data distributions.

The core finding of this study is that the Tabular Basic Large Model (TabPFN) successfully breaks this deadlock by relying on the context learning capability of Transformer and the Bayesian posterior inference mechanism without gradient update. It not only achieved a significant breakthrough in the overall F1-Score but also showed advantages in stringent evaluation indicators (PR-AUC and MCC) for imbalanced data. This fully proves that moving from “specific task tuning” to “zero-shot reasoning supported by pre-trained models” is not only the future of natural language processing but also a promising research direction for complex tabular data processing. The conclusion of this paper not only provides a valuable model selection guide for financial anti-fraud and medical intelligent risk control but also lays a solid empirical foundation for subsequent academic research on the in-depth application of large models in high-risk vertical fields.

7 REFERENCES

- [1] N. Hollmann, S. Müller, T. Purucker, M. Binder, and F. Hutter, “TabPFN: A transformer that solves small tabular classification problems in a second,” *arXiv preprint arXiv: 2207.01848*, 2022.
- [2] S. Müller, N. Hollmann, S. P. Arango, J. Grabocka, and F. Hutter, “Transformers can do Bayesian inference,” *arXiv preprint arXiv: 2112.10510*, 2021.
- [3] Z. Wang and J. Sun, “TransTab: Learning transferable tabular representations,” *Advances in Neural Information Processing Systems NeurIPS*, vol. 35, pp. 34650–34662, 2022. <https://doi.org/10.52202/068431-0210>
- [4] S. Hegselmann *et al.*, “TabLLM: Few-shot classification of tabular data with large language models,” in *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics AISTATS*, 2023.
- [5] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794. <https://doi.org/10.1145/2939672.2939785>
- [6] G. Ke *et al.*, “LightGBM: A highly efficient gradient boosting decision tree,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [7] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “CatBoost: Unbiased boosting with categorical features,” *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [8] S. Ö. Arik and T. Pfister, “TabNet: Attentive interpretable tabular learning,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, pp. 6679–6687, 2021. <https://doi.org/10.1609/aaai.v35i8.16826>
- [9] L. Grinsztajn, E. Oyallon, and G. Varoquaux, “Why do tree-based models still outperform deep learning on tabular data?” *Advances in Neural Information Processing Systems NeurIPS*, vol. 35, pp. 507–520, 2022. <https://doi.org/10.52202/068431-0037>
- [10] T. Saito and M. Rehmsmeier, “The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets,” *PLoS ONE*, vol. 10, no. 3, p. e0118432, 2015. <https://doi.org/10.1371/journal.pone.0118432>
- [11] H. He and E. A. Garcia, “Learning from imbalanced data,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009. <https://doi.org/10.1109/TKDE.2008.239>
- [12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002. <https://doi.org/10.1613/jair.953>
- [13] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, pp. 5–32, 2001. <https://doi.org/10.1023/A:1010933404324>
- [14] Y. Gorishniy, I. Rubachev, V. Mostovous, and A. Babenko, “Revisiting deep learning models for tabular data,” *Advances in Neural Information Processing Systems NeurIPS*, vol. 34, pp. 18932–18943, 2021.
- [15] V. Borisov, T. Corbett, J. Leemann, H. Wirthheim, and G. Kasneci, “Deep neural networks for tabular data: A comprehensive survey,” *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

8 AUTHORS

Mo Yuan is with the School of Computer and Information Engineering, Heihe University, Heilongjiang, China, 164300.

Haikun Teng is with the School of Computer and Information Engineering, Heihe University, Heilongjiang, China, 164300.

Zhonqi Yao is with the School of Computer and Information Engineering, Heihe University, Heilongjiang, China, 164300.

Pengrui Cheng is with the Surrey International, Dongbei University of Finance and Economics, Liaoning, China, 116000.

Yida Lu is with the School of Automation Engineering, University of Electronic Science and Technology of China, Sichuan, China, 610000.

Wenjun Yang is with the School of Computer and Information Engineering, Heihe University, Heilongjiang, China, 164300 (E-mail: 88136697@qq.com).