Information Dynamics and Applications
https://www.acadlore.com/journals/IDA

From Data to Knowledge: A Denoising Autoencoder and Stacking-Based Framework for Customer Retention

Zhaohe Liu*

School of Management and Engineering, Capital University of Economics and Business, 100070 Beijing, China

* Correspondence: Zhaohe Liu (19337857550@163.com)

Received: 11-25-2025 **Revised:** 12-07-2025 **Accepted:** 12-22-2025

Citation: Z. H. Liu, "From data to knowledge: A denoising autoencoder and stacking-based framework for customer retention," *Inf. Dyn. Appl.*, vol. 4, no. 4, pp. 224–237, 2025. https://doi.org/10.56578/ida040404.

Abstract: In a highly competitive telecommunications environment, customer behavior data has become an important source of organizational knowledge for service innovation and strategic decision-making. The ability to transform large-scale user data into actionable knowledge is essential for effective customer retention and sustainable business development. This study develops a knowledge discovery framework that integrates a denoising autoencoder with an enhanced stacking learning strategy to support customer retention innovation. The denoising autoencoder is employed to extract latent behavioral representations from complex and noisy user data, enabling the identification of underlying patterns that are difficult to capture through conventional statistical features. These latent representations are further combined with structured indicators and integrated through a stacking ensemble composed of decision trees, random forests, and XGBoost to achieve robust knowledge fusion. Empirical results show that the proposed framework provides more reliable identification of high-risk customers and improves decision support quality in terms of accuracy and area under curve (AUC). The study demonstrates how artificial intelligence can serve as a mechanism for organizational knowledge creation and offers practical implications for data-driven service innovation and resource allocation in the telecommunications sector.

Keywords: Knowledge discovery; Customer retention innovation; Denoising autoencoder; Stacking learning; Organizational decision support

1 Introduction

In recent years, with the rapid advancement of global communication technologies and the widespread adoption of mobile internet, the telecommunications industry has transitioned from a phase of incremental market expansion to one of intense competition within an established market. On one hand, the homogenization of basic communication services has intensified, with price wars among operators continuously squeezing profit margins. On the other hand, the implementation of number portability policies has further lowered barriers to switching providers, subjecting operators to unprecedented pressure from customer churn. Within this competitive landscape, research indicates that acquiring new customers typically costs 5 to 8 times more than retaining existing ones. Customer churn not only directly undermines a company's profitability but also leads to shrinking market share and diminished brand value. Industry reports reveal that reducing customer churn by just 5% can potentially increase a company's profits by over 25%. Consequently, accurately identifying high-risk churn users and developing proactive intervention strategies have become critical for telecom operators to achieve sustainable growth.

The massive scale and digitization of telecom user behavior data present new opportunities for churn prediction. Beyond their operational function, these data resources gradually constitute a form of organizational knowledge that reflects user preferences, consumption habits, and interaction patterns. The transformation of such raw data into meaningful knowledge has become a central task for service innovation and strategic decision making. Traditional prediction methods primarily rely on structured statistical indicators such as call duration, data consumption, and bill amounts, often employing models like Logistic Regression [1, 2], Support Vector Machines (SVM) [3, 4], or single decision trees. However, these traditional approaches face two major limitations when applied to modern telecom data. First, telecom data is often characterized by high-dimensional sparsity and noise interference. Traditional methods heavily depend on manual feature engineering, making it difficult to capture deep nonlinear behavioral patterns from damaged or noisy data, resulting in insufficient model robustness. Second, single models often struggle with the

trade-off between bias and variance, making it challenging to simultaneously achieve strong generalization capabilities across different user segments. From a knowledge perspective, these limitations restrict the conversion of behavioral traces into reliable organizational insights.

To overcome the aforementioned challenges, this paper proposes a customer churn prediction framework based on Denoising Autoencoders (DAE) and an improved stacking ensemble learning approach. The framework is not only a technical tool for prediction but also a mechanism for knowledge discovery and integration. First, to address the inadequacy of manual feature extraction, this study incorporates the DAE from deep learning. By injecting random noise into the input layer and forcing the network to reconstruct the original data, the DAE automatically learns robust, latent low-dimensional feature representations within the data. These representations can be understood as distilled knowledge about customer behavior, which significantly enhances the model's tolerance to noisy data and its feature expression capabilities. Second, to overcome the performance limitations of single models, this paper constructs a heterogeneous Stacking ensemble model incorporating decision trees, random forests, and XGBoost. This model employs a meta-learner to adaptively fuse predictions from base models, achieving complementary strengths and forming an integrated decision knowledge structure.

The proposed study aims to connect algorithmic modeling with organizational innovation. By translating large-scale behavioral records into structured and interpretable knowledge, telecom operators are able to redesign customer management strategies, allocate service resources with greater precision, and develop proactive engagement measures. Such a process reflects a broader movement from data processing to knowledge-driven innovation in the digital economy, where artificial intelligence functions as an enabler of learning and organizational renewal rather than a mere computational instrument.

Our contributions include:

(1) An end-to-end solution combining unsupervised deep feature extraction with supervised strong ensemble learning has been proposed, which supports the transformation from raw behavioral data to operational knowledge.

(2) DAE effectively resolves noise interference issues in telecommunications data, uncovering non-linear leakage signals imperceptible to traditional statistical methods and enriching the knowledge base for customer understanding.

(3) Experimental results demonstrate that this model significantly outperforms both traditional single models and conventional ensemble methods in terms of accuracy and area under curve (AUC) metrics, providing operators with a scientifically sound and actionable basis for precisely identifying churn-prone users and guiding service innovation.

2 Literature Review

With the rapid iteration of artificial intelligence technologies, the research focus in telecommunications customer churn prediction has shifted from traditional single machine learning models towards deep representation learning and complex ensemble strategies. This transformation reflects not only a technical evolution but also a gradual change in how customer behavior data is understood as a source of organizational knowledge. Existing research primarily concentrates on three areas: enhancements to machine learning-based churn prediction models, the application of deep learning in feature engineering, and strategies for handling imbalanced data. These strands of research collectively shape the foundation for converting fragmented behavioral records into structured insights that can support service innovation.

Traditional statistical learning methods dominated early research. In recent years, variants of gradient-boosted decision trees (GBDT) have garnered significant attention due to their outstanding performance. Asadi Ejgerdi and Kazerooni [5] employed a stacked ensemble learning approach to forecast customer lifetime value (CLV), utilising random forest, XGBoost, and LightGBM as base learners alongside linear regression as the meta-learner. They observed this methodology outperformed deep neural networks and other standalone machine learning models in predictive accuracy. Sikri et al. [6] further demonstrated that combining refined feature engineering with machine learning models can effectively enhance telecom customer retention rates, indicating the potential of analytical techniques to inform managerial decisions. Duru et al. [7] enhanced prediction accuracy by integrating traditional machine learning models (random forests, XGBoost, CatBoost) with one-dimensional convolutional neural networks, illustrating how hybrid approaches can expand the knowledge space extracted from data. A comparative study by Rivaldo et al. [8] also demonstrated that when processing large-scale telecom data, LightGBM significantly outperforms traditional GBDT in training speed due to its histogram optimization algorithm. Meanwhile, CatBoost effectively avoids the curse of dimensionality by handling categorical features without requiring one-hot encoding, which facilitates the utilization of complex user information.Shaikhsurab and Magadum [9] proposed an adaptive ensemble learning framework integrating XGBoost, LightGBM, and SVM, demonstrating the effectiveness of multi-model fusion in enhancing prediction robustness and broadening the interpretative horizon of customer analysis.

Although tree models excel at handling structured data, their capabilities are limited when processing high-dimensional sparse data and noisy interference. From the perspective of knowledge construction, these limitations restrict the extraction of deeper behavioral meanings. To address this, researchers have begun incorporating deep learning for higher-order feature extraction. Saha et al. [10] proposed the ChurnNet architecture, which integrates a

one-dimensional convolutional layer with residual blocks, squeeze blocks, activation blocks, and a spatial attention module, significantly improving prediction accuracy and enabling a more nuanced representation of user activities. Mirza et al. [11] designed an Optimal Depth-Based Convolutional Autoencoder Prediction (ODCCAEP) model tailored for highly competitive customer-dependent application domains to determine the nature of customer churn, highlighting the role of representation learning in revealing hidden consumption logic. Hasumoto and Goto [12] proposed a method of extracting latent features from purchase histories as explanatory variables for churn prediction using a variational autoencoder with the actual customer distribution as a prior, which further illustrates how generative models can transform historical traces into conceptualized knowledge. Li et al. [13] proposed a multimodal autoencoder-decoder framework for customer churn prediction model to better deal with the heterogeneity and consistency problems in the acquired multimodal data. These studies provide a solid theoretical foundation for introducing DAE into this paper for robust feature extraction and for treating model outputs as forms of organizational insight rather than mere numerical results.

Single models often face the dilemma of balancing bias and variance. Stacking ensemble learning, through the combination of multiple layers of models, has become a key technique for resolving this issue and for integrating diverse analytical perspectives. Oladimeji et al. [14] used K-Nearest Neighbor (KNN), Classification and Regression Trees (CART), and Naive Bayes as base classifiers, with logistic regression as the meta-classifier. After multiple evaluations of model accuracy, the final output results demonstrated a model accuracy rate as high as 83%, suggesting that layered integration can enhance decision reliability. Adnan and Awang [15] proposes an enhanced ensemble stacking method designed to improve the predictive performance of ensemble approaches. Compared to other ensemble methods and single classifiers, the stacking ensemble method demonstrates superior generalization capabilities and accuracy, which is essential for translating algorithmic outcomes into dependable managerial knowledge. De and Prabu [16] proposes a novel sampling stacking framework named "Unbalanced Learning Sampler Stacking." This framework integrates the predictive capabilities of sampling schemes to stimulate information gain from meta-features within ensemble models, emphasizing the role of model architecture in knowledge enrichment. Adnan and Awang [17] proposes a multi-level stacking ensemble model enhanced by Soft Set Theory to improve the accuracy and efficiency of customer churn prediction. The proposed model leverages Soft Set Theory to eliminate redundant classifiers via the analysis of the indiscernibility matrix, increasing classifier diversity and ensemble generalization, which contributes to more comprehensive understanding of customer states.

Telecom attrition data typically exhibits severe class imbalance. From an innovation viewpoint, imbalance not only affects accuracy but also distorts the representation of minority user experiences. Imani et al. [18] emphasized the decisive impact of data balance on model performance and validated the effectiveness of synthetic minority over-sampling technique (SMOTE) in addressing the complexity of imbalanced datasets and intricate interactions among features. Suguna et al. [19] noted in their case study on customer churn that incorporating undersampling techniques can effectively mitigate the impact of imbalanced datasets on model prediction accuracy, ensuring that analytical knowledge reflects real operational risks. Gore et al. [20] further found that models employing the SMOTE sampling method demonstrated higher prediction accuracy compared to models without the SMOTE approach, reinforcing the necessity of balanced learning for credible decision support.

In summary, although existing research has made significant strides in algorithmic optimisation, limitations remain: firstly, traditional feature engineering struggles to handle noisy data, which constrains the depth of knowledge that can be obtained; secondly, current stacking models seldom account for data imbalance, reducing their value for practical innovation. This paper aims to address these issues by proposing an improved stacking model for predicting telecom customer churn based on deep feature extraction via DAE, and by positioning the model as a mechanism for transforming data resources into actionable organizational knowledge.

3 Overview of Relevant Theories

The core objective of this study is to construct a telecom customer churn prediction framework that combines high efficiency with high accuracy, empowering operators to identify high-risk churn users early and retain them with precision. Beyond the technical goal of prediction, the framework is intended to function as a channel through which dispersed behavioral data can be transformed into structured knowledge for organizational action. Given that customer churn directly threatens operators' revenue base and market share, traditional prediction methods often face bottlenecks of insufficient feature extraction and limited generalization performance when dealing with massive, high-dimensional, and sparse telecom data. These bottlenecks not only reduce computational performance but also restrict the depth of understanding that organizations can obtain from customer information.

To address these challenges, this paper establishes the following key research directions. First, leveraging the powerful representation capabilities of deep learning to extract robust deep features from complex data, so that implicit behavioral meanings hidden in noisy records can be expressed in a stable form. Second, employing heterogeneous ensemble learning strategies to effectively integrate the decision advantages of different strong classifiers, thereby enhancing both prediction accuracy and model stability and forming a comprehensive view of customer states.

Through this dual pathway, the study attempts to bridge the gap between algorithmic modeling and practical knowledge use, enabling analytical outcomes to support managerial judgment rather than remaining isolated technical results. This approach provides operators with scientifically grounded data-driven decision support for optimizing resource allocation and minimizing churn-related losses, and at the same time contributes to the cultivation of a knowledge-oriented service culture.

To achieve the aforementioned research objectives, this study employs three core methodological modules: data balancing and preprocessing, deep feature extraction based on denoising autoencoders, and the construction of a stacking ensemble model. These modules correspond respectively to the stages of knowledge preparation, knowledge generation, and knowledge integration. Data balancing and preprocessing ensure that the information environment reflects real user diversity; the denoising autoencoder serves as a mechanism for discovering latent representations of customer behavior; and the stacking ensemble organizes multiple analytical perspectives into a coherent decision structure. Together, they form a systematic process through which raw telecom data evolves into operational knowledge that can guide retention innovation.

3.1 Data Imbalance Handling and the SMOTE Algorithm

In telecom customer churn prediction tasks, the number of churned users is typically far smaller than that of retained users. This severe class imbalance causes models to overemphasize the majority class during training, thereby neglecting the identification of minority churn users. To address this issue, this study employs the SMOTE. Unlike simple random oversampling, SMOTE generates new synthetic samples by interpolating minority class samples in the feature space. Its fundamental principle is as follows: for each minority class sample x_i, compute its x_i nearest neighbors; randomly select one sample $\hat{x}_i$ from these neighbors; and then generate a new sample x_{new} by randomly interpolating along the line connecting x_i and $\hat{x}_i$. The formula is calculated as Eq. (1):

$$x_{\text{new}} = x_i + \text{rand}(0, 1) \times (\hat{x}_i - x_i) \tag{1}$$

Through SMOTE processing, the class distribution of the dataset is balanced, enabling the model to learn the characteristics of both user categories more equitably. This enhances the recall rate for identifying potential churn users.

3.2 Deep Feature Extraction Based on Denoising Autoencoder

After data preprocessing and balancing, this study introduced the DAE [21–23] to delve deeper into the latent value within the data. Traditional feature engineering primarily relies on manual expertise to construct statistical features, making it difficult to capture higher-order nonlinear relationships between variables and rendering it sensitive to noise. DAE is an unsupervised learning algorithm based on neural networks, whose core idea is to learn robust feature representations by distorting the input data.

In the feature extraction mechanism based on denoising autoencoders, random noise is first injected into the original input vector x to generate corrupted data $\tilde{x}$. Subsequently, the encoder maps this corrupted data to a low-dimensional latent space h, while the decoder attempts to reconstruct the original undamaged data x from h. During this process, by minimizing reconstruction error, the model compels the hidden layer h to learn intrinsic, interference-resistant nonlinear features inherent in the data distribution. This mechanism not only achieves data denoising but also effectively generates deep latent features rich in semantic information. These features are subsequently fused with original statistical features to serve as enhanced inputs for subsequent prediction models, significantly boosting their discriminative capabilities.

3.3 Improved Stacking Ensemble Learning Framework

Model construction constitutes the core component of machine learning. Individual models often face the trade-off dilemma between bias and variance, struggling to simultaneously accommodate generalisation capabilities across diverse sample types. Stacking represents a hierarchical model fusion strategy that constructs more robust predictive systems by combining multiple heterogeneous base learners. The improved stacking framework developed in this study is illustrated in Figure 1.

At the first layer of the stacking framework, to ensure diversity in integration while balancing bias and variance, this study selected three tree model algorithms with distinct mechanisms:

Decision Tree A decision tree is a supervised learning algorithm based on a tree structure, serving as the building block for numerous ensemble models. This algorithm employs a top-down recursive approach, selecting optimal splitting features by calculating metrics such as information gain or Gini impurity. It maps the dataset into a series of mutually exclusive rule sets. While a single decision tree offers high interpretability and can visually reveal key customer churn pathways, it is prone to overfitting when handling high-dimensional complex data. Within the stacking framework of this study, the decision tree primarily serves as a high-variance baseline model. Its purpose is to

capture strong local features within the data and provide foundational decision boundary information for subsequent meta-learners.

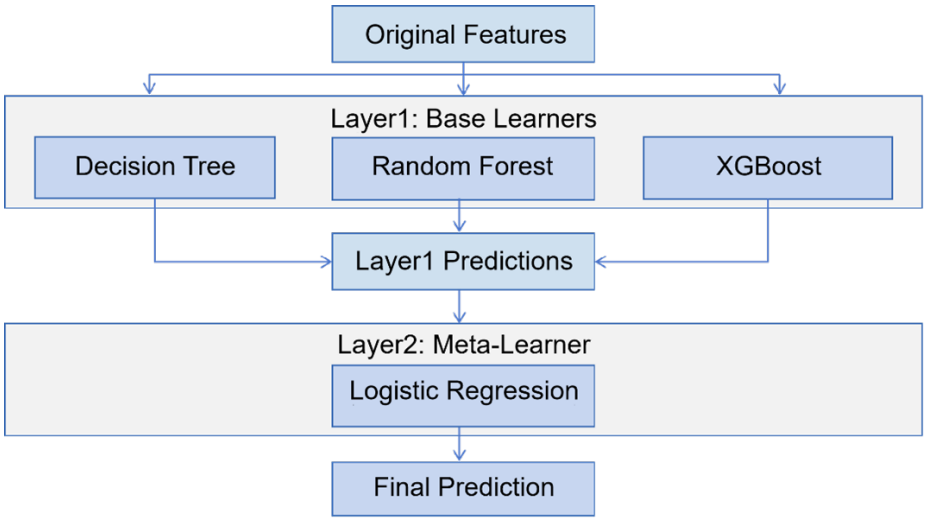

Figure 1. Architecture of the improved stacking ensemble model

Random Forest To overcome the overfitting tendency of individual decision trees, this study introduces the Random Forest algorithm based on the Bagging strategy. This algorithm concurrently constructs multiple decision trees, generating independent training subsets for each tree via Bootstrap Sampling. Additionally, a random feature selection mechanism is incorporated during node splitting, further reducing correlation among trees. Final predictions are obtained through voting or averaging across all base decision trees. By reducing overall model variance, Random Forest significantly enhances robustness to noisy data. It is particularly well-suited for handling high-dimensional sparse features common in telecom data, effectively improving model generalization on unseen samples.

XGBoost is an efficient gradient boosting decision tree algorithm based on the Boosting strategy [24–26]. Unlike the parallel mechanism of Random Forests, XGBoost employs sequential training, where each new tree is constructed to fit the prediction residuals of the previous model, thereby progressively reducing model bias. Compared to traditional Gradient Boosted Decision Trees (GBDT), XGBoost performs a second-order Taylor expansion of the loss function, accelerating convergence by utilizing first- and second-order derivative information. Simultaneously, it explicitly incorporates a regularization term into the objective function, effectively controlling model complexity to prevent overfitting. In telecom customer churn prediction, XGBoost leverages its exceptional feature capture capabilities and computational efficiency to accurately identify nonlinear churn signals hidden within long-tail data.

4 Customer Churn Prediction Model Based on Denoising Autoencoder and Improved Stacking Integration

This paper proposes a telecommunications customer churn prediction model that integrates the DAE with an improved stacking ensemble learning approach. The overall framework of the model is illustrated in Figure 2. The model primarily consists of four core components: data preprocessing, deep feature engineering based on DAE, stacking model construction and training, and model performance evaluation.

Step 1: Data Preprocessing. Input the raw telecom dataset D_0, perform data cleaning and missing value handling, remove irrelevant features (e.g., user ID), address class imbalance using SMOTE technology, and encode class features.

Step 2: Deep Feature Engineering. This constitutes the core innovation of this study. Unlike traditional statistical feature construction, we build a deep neural network based on the DAE. By injecting noise into the original data and forcing reconstruction, robust low-dimensional deep latent features are extracted. These new features are then fused with the original structured features to form an enhanced multimodal dataset.

Step 3: Model Training. An improved stacking ensemble framework is constructed. The first layer employs decision trees, random forests, and XGBoost as heterogeneous base learners; The second layer employs logistic regression as a meta-learner to adaptively weight and fuse predictions from the base models.

Step 4: Model Evaluation. Accuracy, Recall, and AUC values serve as core evaluation metrics to validate the model's effectiveness in identifying potential churn users. Experimental results demonstrate that this model architecture exhibits outstanding feature capture capabilities and predictive accuracy.

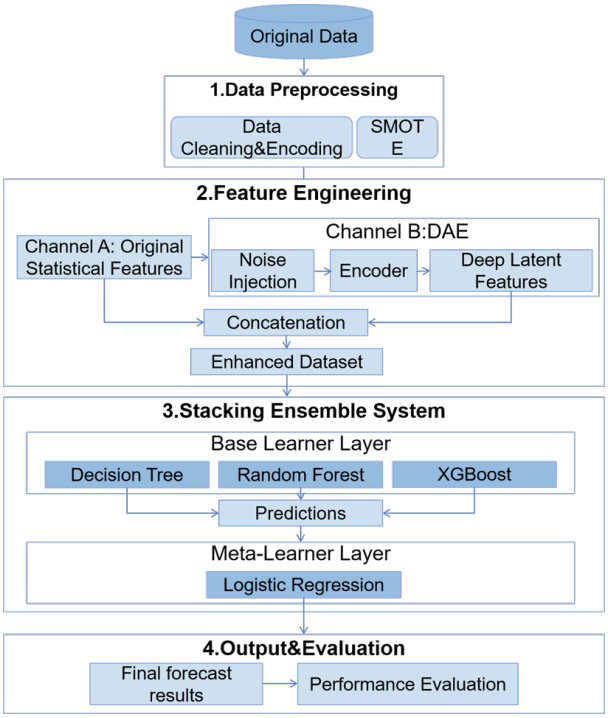

Figure 2. Framework of customer churn prediction modelling integrating DAE and stacking

4.1 Data Pre-Processing

Data preprocessing includes data cleaning, feature selection, categorical encoding, and SMOTE-based data balancing. First, check for missing values and outliers in the dataset. For missing values, numerical features are imputed using mean imputation, while categorical features use mode imputation. In the dataset used for this study, no significant missing values were detected. To enhance computational efficiency, identifying features irrelevant to churn prediction (e.g., CustomerID, PhoneService_ID) are removed. Second, addressing the common category imbalance in telecom data (where churned users typically constitute a low proportion), direct training would cause the model to skew toward the majority class. This study employs SMOTE. By generating synthetic samples through linear interpolation of minority class samples in the feature space, the ratio of positive to negative samples is balanced. Finally, categorical features like "Plan Type" and "Payment Method" are converted into numerical formats using Label Encoding or One-Hot Encoding for subsequent processing by neural networks and tree models. The complete data preprocessing workflow is illustrated in Figure 3.

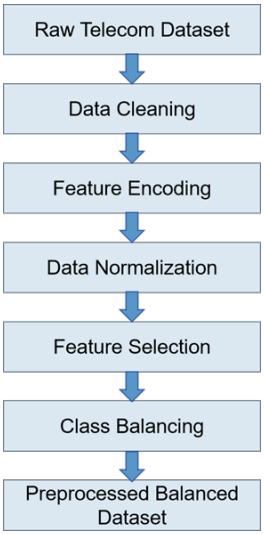

Figure 3. Data preprocessing flowchart

4.2 Feature Engineering Based on DAE

In studies predicting telecom customer churn, traditional feature engineering is often constrained to raw statistical attributes, making it difficult to capture deep nonlinear interactions between variables. To address this, this paper designs a deep feature extraction module based on the DAE, with the workflow illustrated in Figure 4.

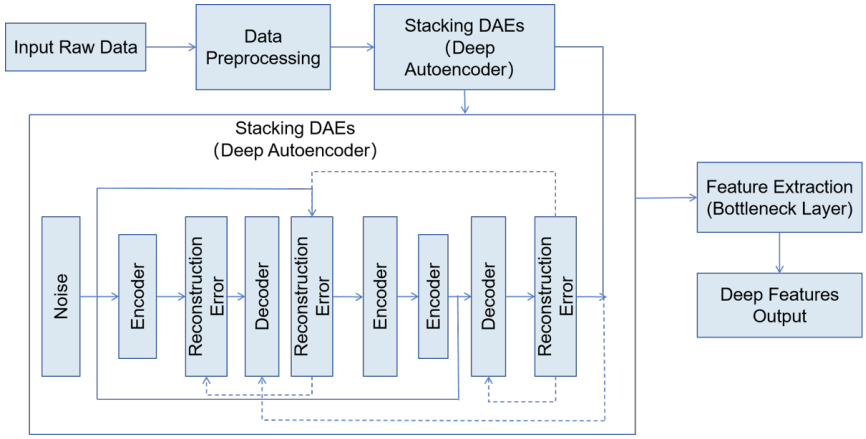

Figure 4. Flowchart of deep feature extraction based on DAE

The core concept of this module is to move beyond reliance on manual rules and instead automatically extract robust representations within the data through unsupervised deep learning. The specific steps are as follows: First, the preprocessed and standardized data x is input. Next, a noise injection mechanism is introduced to simulate real-world data interference. Subsequently, an encoder-decoder network is trained, with the output from the bottleneck layer extracted as new features. Finally, these "deep latent features" are concatenated with the original features to construct an enhanced feature set containing both deep and shallow information.

This process not only enhances the richness of the data but also endows the model with resilience against noise and outliers, laying the groundwork for subsequent high-precision predictions.

4.2.1 Noise injection and data reconstruction

To force the model to learn robust features inherent to the data rather than merely replicating the input, the DAE first applies Gaussian noise with a distribution $N\left(0, \sigma^2\right)$ to the original input vector x during training, generating corrupted data $\tilde{x}$. The formula is calculated as Eq. (2):

$$\tilde{x} = x + \epsilon, \epsilon \sim N(0, 0.1) \tag{2}$$

Subsequently, the corrupted data $\tilde{x}$ is fed into a multi-layer neural network. The network aims to reconstruct the original, uncorrupted data x from $\tilde{x}$ by minimizing the reconstruction error. This process forces the network to ignore noise interference while capturing the underlying patterns and nonlinear correlations within the data.

4.2.2 Deep hidden feature generation and fusion

The DAE network consists of an Encoder and a Decoder. As shown in Figure 4, the Encoder maps high-dimensional inputs to a low-dimensional latent space, generating latent vectors; the Decoder is responsible for mapping these latent vectors back to the original dimensions. In this study, after the DAE network converges during training, we extract the latent layer vectors output by the Encoder and define them as deep latent features. These deep features represent a highly abstracted representation of user behavior. Finally, these deep features are horizontally concatenated with the original structured features to form the final training dataset. This original features and deep features fusion strategy enables the model to retain statistically meaningful business insights while gaining the deep neural network's ability to perceive complex patterns, significantly enhancing the model's upper limit of feature representation.

4.3 Improving Stacking Model Construction and Training

During the model training phase, to overcome the limitations of weak generalization and susceptibility to overfitting in single models, this paper constructs an improved stacking ensemble learning framework. First, the base learners are defined: decision trees, random forests, and XGBoost.

Second, the meta-learner is defined. The prediction probabilities output by the base models are treated as meta-features and fed into a Logistic Regression model. The meta-learner learns the weights of different base models, performs linear correction and integration on the prediction results, and ultimately outputs the customer churn

probability. During training, 5-fold cross-validation is employed to generate meta-features, preventing data leakage and evaluating model robustness. The construction process of the improved stacking model is illustrated in Figure 1.

Through this hierarchical ensemble strategy, the model fully leverages the strengths of different algorithms. It achieves high accuracy while significantly enhancing adaptability to various churn scenarios.

5 Experiments and Analysis of Results

This chapter will detail the sources, feature distributions, and preprocessing procedures of experimental data based on real operational data provided by China Mobile Fujian Company. Building upon this foundation, core evaluation metrics for the model will be defined. Through multiple comparative experiments, the analysis will focus on the feature enhancement effects achieved using the DAE and the performance advantages of the improved stacking ensemble model over single models. This will validate the effectiveness of the proposed framework in the task of predicting telecom customer churn.

5.1 Data Sources and Description

5.1.1 Dataset overview

The data for this study was sourced from sample data provided by China Mobile Fujian Company for a specific month in 2018. The dataset encompasses monthly consumption and behavioral records of 50,000 users, comprising a total of 29 original variables. The target variable is "blacklisted customer status" (in telecommunications, blacklisted customers typically refer to high-risk churn groups with long-term unpaid bills or abnormal service suspensions). A label of '1' indicates churn/high risk, while "0" indicates retention/normal status. Variables are presented as shown in Table 1.

To comprehensively characterize user profiles, data features span five major dimensions:

(1) Identity Characteristics: Includes real-name verification status, college student status, age group, and network tenure.

(2) Spending Capacity: Includes average spending over the past 6 months, current month's bill amount, account balance, overdue status, and sensitivity to call charges.

(3) Interpersonal Connections: Core metric is "number of contacts in the current month's calling network," reflecting social activity.

(4) Location Trajectory: Records offline activity scenarios such as "frequent mall visits," "monthly visits to Wanda Plaza/Sam's Club in Cangshan District, Fuzhou," and "visits to sports venues," providing valuable insights into users' lifestyle tiers.

(5) App Usage: Detailed records of app usage frequency across categories including e-commerce, logistics/delivery, finance/investment, video streaming, and transportation (air/rail).

5.1.2 Data cleaning and preprocessing

Following the EDA, we performed the following cleaning and preprocessing on the raw data:

Outlier handling: Removed outlier samples with ages of 0 and 118; excluded extreme samples where monthly call interaction circles exceeded 600 individuals, as these outliers likely stemmed from machine behavior or data entry errors.

Feature selection and discretization: Removed the "User ID" column, which had no predictive value. Discretized continuous variables 'Age' and "Network tenure" using equal-frequency binning. For example, age was divided into five intervals: (0,18], (18,30], (30,35], (35,45], (45,100). Analysis indicates that the 35–45 age group has the highest number of blacklisted users (815 individuals), exhibiting significant risk characteristics.

Data Standardization: The Min-Max normalization method was applied to map all numerical features to the [0, 1] range, eliminating the impact of differing feature scales (e.g., spending amount vs. app usage frequency) on model convergence speed.

Sample Imbalance Handling: The original dataset exhibited severe imbalance between blacklisted (churned) users and whitelisted users. To prevent model overfitting to the majority class, this study employed SMOTE (Synthetic Minority Over-sampling Technique). By calculating k-nearest neighbors for minority samples and generating interpolated samples, the positive-to-negative sample ratio was adjusted to 1:1. The processed training dataset expanded to 66,136 samples, with 33,068 blacklist and 33,068 whitelist users.

5.1.3 Feature mining

To enhance the model's ability to capture complex behaviors, we performed feature derivation based on preprocessing. We introduced new metrics reflecting user stability (e.g., "telecom bill stability," "payment stability") and consumption potential indicators. Correlation analysis, as shown in Figure 5, revealed that all feature correlation coefficients were below 0.8, indicating no significant multicollinearity issues. A total of 21 feature variables were ultimately incorporated into the model, encompassing multidimensional data such as identity, consumption, trajectory, and application behavior, as detailed in Table 2.

Table 1. Summary of dataset information

	Variable	Type
	User ID	Character type
Identity Characteristics	Whether user passed real name verification	Boolean type: 1 for true, 0 for false
	Whether university student customer	Boolean type: 1 for true, 0 for false
	Whether blacklist customer	Boolean type: 1 for true, 0 for false
	Whether 4G unhealthy customer	Boolean type: 1 for true, 0 for false
	Age Group	Numeric type
	User's length of service (months)	Numeric type
Purchasing Power	Time elapsed since user's last payment (months)	Numeric type
	The amount of the most recent payment made by the paying customer (yuan)	Numeric type
	Average expenditure per user over the past six months (yuan)	Numeric type
	Total charges for the current month on the customer's bill (yuan)	Numeric type
	Current month's account balance (yuan)	Numeric type
	Is the paying subscriber currently in arrears with their payments?	Numeric type
Interpersonal Relations	Number of contacts in the calling network for the current month	Numeric type
	Whether frequent mall visitor	Boolean type: 1 for true, 0 for false
Position Trajectory	Average monthly mall visits in last 3 months	Numeric type
	Whether visited Fuzhou Cangshan Wanda this month	Boolean type: 1 for true, 0 for false
	Whether visited Fuzhou Sams Club this month	Boolean type: 1 for true, 0 for false
	Whether watched movie this month	Boolean type: 1 for true, 0 for false
	Whether visited tourist attractions this month	Boolean type: 1 for true, 0 for false
	Whether spent at sports venues this month	Boolean type: 1 for true, 0 for false
Applied Behaviour	Online shopping app usage count this month	Numeric type
	Logistics and express app usage count this month	Numeric type
	Financial management app total usage count this month	Numeric type
	Video streaming app usage count this month	Numeric type
	Airline app usage count this month	Numeric type
	Train app usage count this month	Numeric type
	Travel information app usage count this month	Numeric type

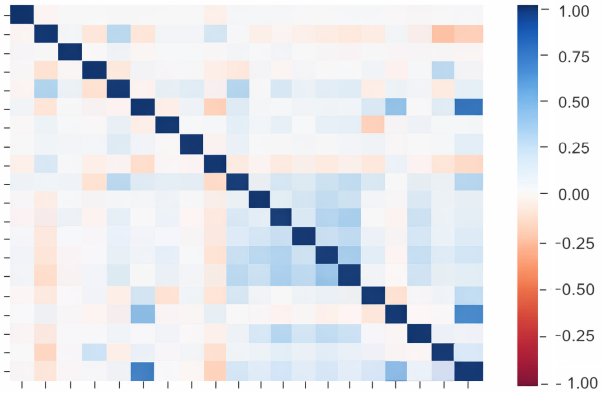

Figure 5. Feature correlation coefficient heatmap

Table 2. Variable status following feature extraction

Variable	Type
Whether user passed real name verification	Boolean type: 1 for true, 0 for false
Whether university student customer	Boolean type: 1 for true, 0 for false
Whether blacklist customer	Boolean type: 1 for true, 0 for false
Whether 4G unhealthy customer	Boolean type: 1 for true, 0 for false
Age Group	Numeric type
User's length of service (months)	Numeric type
Time elapsed since user's last payment (months)	Numeric type
Total charges for the current month on the customer's bill (yuan)	Numeric type
Is the paying subscriber currently in arrears with their payments?	Numeric type
User sensitivity to mobile phone charges	Numeric type
Number of contacts in the calling network for the current month	Numeric type
Whether frequent mall visitor	Boolean type: 1 for true, 0 for false
Average monthly mall visits in last 3 months	Numeric type
Whether watched movie this month	Boolean type: 1 for true, 0 for false
Whether visited tourist attractions this month	Boolean type: 1 for true, 0 for false
Whether spent at sports venues this month	Boolean type: 1 for true, 0 for false
Telephone charges remain stable	Numeric type
Stable payment	Numeric type
Consumption potential	Numeric type

5.2 Evaluation Indicators

To objectively evaluate model performance, this paper selected the confusion matrix, accuracy, recall, and ROC-AUC values as evaluation metrics.

Confusion Matrix: Used to display the detailed distribution of classification results, as shown in Table 3.

Table 3. Confusion matrix

	Predicted to be in the Positive Category	**Forecasts are in the Negative Category**
Actual Positive Class	TP (True Positive)	FN (False Negative)
Actual Negative Category	FP (False Positive)	TN (True negative)

Among these, due to discrepancies between predicted results and actual values, the following four scenarios may occur:

TP (True Positive): The true category of the sample is 1, and the model also predicts it as 1.

FN (False Negative): The true category of the sample is 1, but the model predicts it as 0.

FP (False Positive): The true category of the sample is 0, but the model predicts it as 1.

TN (True Negative): The true category of the sample is 0, and the model predicts it as 0.

Core Metric Formula:

Accuracy: The proportion of correctly predicted samples to the total number of samples. The accuracy is calculated as Eq. (3):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3}$$

Recall: The proportion of all actual churn (blacklisted) users that the model successfully predicts. This metric is critical for risk control. The recall is calculated as Eq. (4):

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

AUC: The area under the ROC curve, with a value range of [0.5, 1.0]. The closer the AUC value is to 1, the stronger the model's ability to distinguish between whitelisted and blacklisted users.

5.3 Experimental Results

The experiment divided the dataset into training and test sets at a 3:1 ratio. To validate the superiority of the proposed "DAE and Improved Stacking Fusion Model," we compared it against Decision Tree, Random Forest, and XGBoost as standalone models. All models were evaluated on the same test set comprising 14,907 samples.

5.3.1 Analysis of DAE feature enhancement effect

First, we validated the effectiveness of the DAE. By injecting Gaussian noise into the input layer and reconstructing the original 21-dimensional statistical features, the DAE extracted latent deep features. The comparison results of feature enhancement experiments are shown in Table 4. Experiments revealed that when relying solely on statistical features, the benchmark model's representational power was constrained by data sparsity (e.g., some users had zero app usage instances). However, integrating DAE-derived deep features enabled the model to better capture nonlinear patterns in user behavior, demonstrating the robustness of deep feature extraction in noisy telecom data.

Table 4. Comparison results of feature enhancement experiments

Feature Type	Accuracy	AUC
Original Features	0.86	0.70
DAE Feature Enhancement	**0.98**	**0.75**

5.3.2 Comparative analysis of model performance

The specific performance metrics of each model on the test set are shown in Table 5.

Table 5. Comparison of results of different modeling tests

	Accuracy	White List Prediction Accuracy Rate	Blacklist Prediction Accuracy Rate	AUC
Decision tree	65.55%	65.56%	43.34%	0.53
Random Forest	87.71%	89.92%	50.42%	0.67
XGBoost	96.63%	98.93%	63.17%	0.69
Stacking Model (Ours)	98.45%	99.12%	70.12%	0.75

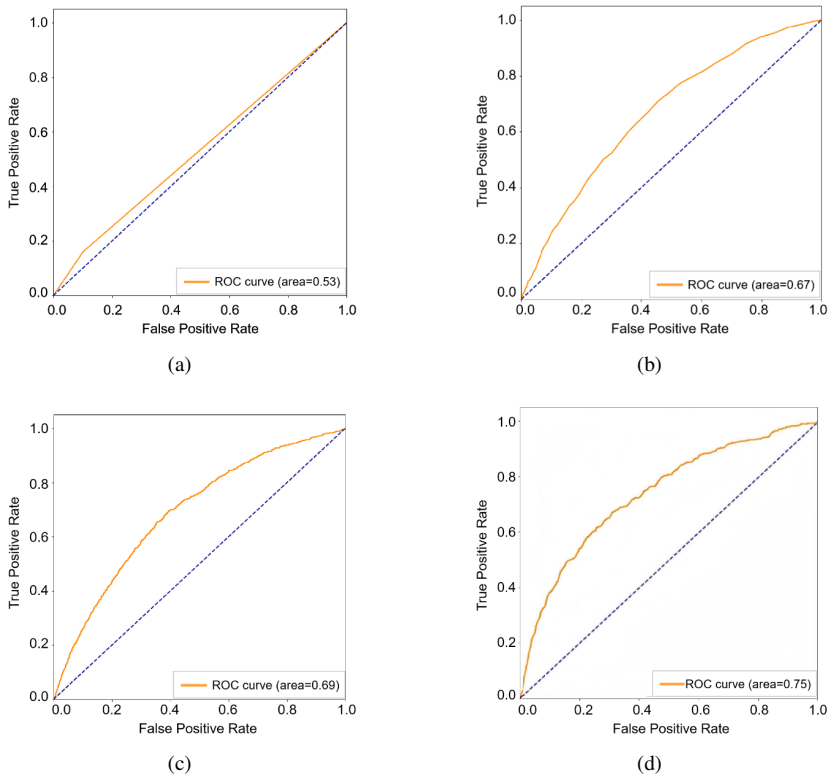

Figure 6. Model ROC curve

Table 5 clearly demonstrates that the proposed improved stacking model achieves the highest overall accuracy of 98.45%. Compared to decision trees (65.55%) and random forests (87.71%), the performance improvement is

significant. Even when compared to the high-performing standalone XGBoost model (96.63%), the stacking model still achieves an additional 1.82 percentage points. This demonstrates that by integrating the heterogeneous strengths of CatBoost, LightGBM, and XGBoost, the model effectively reduces the variance inherent in single algorithms and enhances the overall stability of predictions.

For blacklisted users (high-risk churn customers)—a critical concern for telecom operators—traditional models generally deliver suboptimal predictions. Decision trees and random forests achieved recall rates of only 43.34% and 50.42%, respectively, meaning nearly half of high-risk users were missed. While the standalone XGBoost model improved recall to 63.17%, room for further enhancement remained. The stacking model proposed in this paper, bolstered by SMOTE sample balancing and DAE deep feature extraction, further elevates the prediction accuracy (Recall) for blacklisted users to 70.12%. This demonstrates the model's heightened sensitivity to minority-class risk samples, enabling more precise identification of potential churn customers.

As shown in Figure 6, the stacking model achieves an AUC value of 0.75, outperforming both XGBoost (0.69) and Random Forest (0.67). A ROC curve that curves more sharply toward the upper-left corner indicates superior classification performance. The stacking model's curve encompasses the largest area, demonstrating consistent classification efficacy across different threshold settings.

In summary, the prediction model based on DAE deep feature extraction and enhanced stacking ensemble not only demonstrates outstanding overall accuracy but also achieves a substantial breakthrough in identifying critical blacklisted users. This validates the effectiveness of the "Deep Representation Learning and Heterogeneous Ensemble" strategy for processing large-scale, multidimensional telecom data.

6 Conclusion

To address data imbalance and noise interference in telecom customer churn prediction while enhancing forecasting accuracy, this paper proposes a customer churn prediction model integrating DAE with an improved stacking ensemble technique. The study first performs data preprocessing, SMOTE sample balancing, and deep feature engineering based on DAE, followed by model training and evaluation. Experimental results demonstrate that compared to untreated traditional single algorithms, this model achieves significant improvements in both accuracy (98.45%) and AUC (0.75) metrics. Consequently, the model effectively enhances the accuracy of telecom customer churn prediction, enabling operators to precisely identify potential churn customers. This facilitates optimized resource allocation and fosters long-term stable customer relationships.

Beyond the improvement of technical indicators, the study shows how large volumes of customer behavior data can be transformed into structured knowledge that supports organizational learning. The latent representations generated by the denoising autoencoder reveal patterns that are not directly observable in conventional statistical features, while the stacking mechanism integrates different analytical viewpoints into a coherent decision basis. Through this process, the predictive model functions not only as a computational tool but also as a means of producing actionable understanding for service management.

From a practical perspective, the framework provides telecom operators with a clearer foundation for designing retention strategies, prioritizing high-risk customers, and allocating marketing resources with greater precision. The approach contributes to the development of data-informed service innovation, in which managerial actions are guided by systematically extracted knowledge rather than isolated experience. Such a transition is essential for enterprises operating in an environment where competition depends increasingly on the ability to learn from digital traces of user behavior.

The study also indicates several directions for future work. Further research may explore the integration of additional behavioral sources to enrich the knowledge base, examine the interpretability of latent features to support transparent decision processes, and consider privacy-preserving mechanisms to ensure responsible use of customer information. Strengthening these aspects will help connect predictive analytics more closely with ethical governance and sustainable innovation in the telecommunications sector.

Funding

This work is supported by the China University Industry-University-Research Innovation Fund (Grant No.:2024AX023).

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares that they have no conflicts of interest.

References

[1] P. A. Sunarya, U. Rahardja, S. C. Chen, Y. M. Lic, and M. Hardini, "Deciphering digital social dynamics: A comparative study of logistic regression and random forest in predicting e-commerce customer behavior," *J. Appl. Data Sci.*, vol. 5, no. 1, pp. 100–113, 2024. https://doi.org/10.47738/jads.v5i1.155

[2] V. Vajrobol, B. B. Gupta, and A. Gaurav, "Mutual information based logistic regression for phishing URL detection," *Cyber Secur. Appl.*, vol. 2, p. 100044, 2024. https://doi.org/10.1016/j.csa.2024.100044

[3] A. Kurani, P. Doshi, A. Vakharia, and M. Shah, "A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting," *Ann. Data Sci.*, vol. 10, no. 1, pp. 183–208, 2023. https://doi.org/10.1007/s40745-021-00344-x

[4] Q. Wang, D. Chen, M. Li, S. Li, F. Wang, Z. Yang, W. Zhang, S. Chen, and D. Yao, "A novel method for petroleum and natural gas resource potential evaluation and prediction by support vector machines (SVM)," *Appl. Energy*, vol. 351, p. 121836, 2023. https://doi.org/10.1016/j.apenergy.2023.121836

[5] N. Asadi Ejgerdi and M. Kazerooni, "A stacked ensemble learning method for customer lifetime value prediction," *Kybernetes*, vol. 53, no. 7, pp. 2342–2360, 2024. https://doi.org/10.1108/K-12-2022-1676

[6] A. Sikri, R. Jameel, S. M. Idrees, and H. Kaur, "Enhancing customer retention in telecom industry with machine learning driven churn prediction," *Sci. Rep.*, vol. 14, no. 1, p. 13097, 2024. https://doi.org/10.1038/s41598-024-63750-0

[7] I. Duru, N. Ak, and R. Dede, "A comparative study of feature selection and resampling techniques for customer churn prediction with explainable AI in the telecommunications sector," in *2025 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA)*, Antalya, Turkiye, 2025, pp. 1–12. https://doi.org/10.1109/ACDSA65407.2025.11166474

[8] R. Rivaldo, R. Taufik, I. S. Ilman, and O. D. E. Wulansari, "A comparative study of XGBoost, LightGBM, and CatBoost models for customer churn prediction in the banking industry," *J. Pepadun*, vol. 6, no. 2, pp. 178–187, 2025. https://doi.org/10.23960/pepadun.v6i2.277

[9] M. A. Shaikhsurab and P. Magadum, "Enhancing customer churn prediction in telecommunications: An adaptive ensemble learning approach," *arXiv preprint arXiv:2408.16284*, 2024. https://doi.org/10.48550/arXiv.2408.16284

[10] S. Saha, C. Saha, M. M. Haque, M. G. R. Alam, and A. Talukder, "Churnnet: Deep learning enhanced customer churn prediction in telecommunication industry," *IEEE Access*, vol. 12, pp. 4471–4484, 2024. https://doi.org/10.1109/ACCESS.2024.3349950

[11] O. M. Mirza, G. J. Moses, R. Rajender, E. L. Lydia, S. Kadry, C. Me-Ead, and O. Thinnukool, "Optimal deep canonically correlated autoencoder-enabled prediction model for customer churn prediction," *Comput. Mater. Continua*, vol. 73, no. 2, 2022. https://doi.org/10.32604/cmc.2022.030428

[12] K. Hasumoto and M. Goto, "Predicting customer churn for platform businesses: Using latent variables of variational autoencoder as consumers' purchasing behavior," *Neural Comput. Appl.*, vol. 34, no. 21, pp. 18 525–18 541, 2022. https://doi.org/10.1007/s00521-022-07418-8

[13] Y. Li, G. Xia, S. Wang, and Y. Li, "A deep multimodal autoencoder-decoder framework for customer churn prediction incorporating ChatGPT," *Multimed. Tools Appl.*, vol. 83, no. 41, pp. 89 563–89 589, 2024. https://doi.org/10.1007/s11042-023-17715-6

[14] O. M. Oladimeji, A. R. Ajiboye, and F. E. Usman-Hamza, "An optimized stacking ensemble technique for creating prediction model of customer retention pattern in the banking sector," *Gadau J. Pure Allied Sci.*, vol. 2, no. 1, pp. 22–29, 2023. https://doi.org/10.54117/gjpas.v2i1.29

[15] N. N. bt Adnan and M. K. Awang, "Enhancing customer churn prediction across industries: A comparative study of ensemble stacking and traditional classifiers," *Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 1, 2025. https://doi.org/10.14569/ijacsa.2025.0160120

[16] S. De and P. Prabu, "A sampling-based stack framework for imbalanced learning in churn prediction," *IEEE Access*, vol. 10, pp. 68 017–68 028, 2022. https://doi.org/10.1109/ACCESS.2022.3185227

[17] N. N. Adnan and M. K. Awang, "A multi-level stacking ensemble model optimized by soft set theory for customer churn prediction," *Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 7, 2025. https://doi.org/10.14569/ijacsa.2025.0160775

[18] M. Imani, Z. Ghaderpour, M. Joudaki, and A. Beikmohammadi, "The impact of SMOTE and ADASYN on random forest and advanced gradient boosting techniques in telecom customer churn prediction," in *2024 10th International Conference on Web Research (ICWR)*, Tehran, Islamic Republic of Iran, 2024, pp. 202–209. https://doi.org/10.1109/ICWR61162.2024.10533320

[19] R. Suguna, J. Suriya Prakash, H. Aditya Pai, T. R. Mahesh, V. Vinoth Kumar, and T. E. Yimer, "Mitigating class imbalance in churn prediction with ensemble methods and SMOTE," *Sci. Rep.*, vol. 15, no. 1, p. 16256, 2025.

https://doi.org/10.1038/s41598-025-01031-0

[20] S. Gore, Y. Chibber, M. Bhasin, S. Mehta, and S. Suchitra, "Customer churn prediction using neural networks and SMOTE-ENN for data sampling," in *2023 3rd International Conference on Artificial Intelligence and Signal Processing (AISP)*, Vijayawada, India, 2023, pp. 1–5. https://doi.org/10.1109/AISP57993.2023.10134827

[21] W. H. Lee, M. Ozger, U. Challita, and K. W. Sung, "Noise learning-based denoising autoencoder," *IEEE Commun. Lett.*, vol. 25, no. 9, pp. 2983–2987, 2021. https://doi.org/10.1109/LCOMM.2021.3091800

[22] P. Singh and A. Sharma, "Attention-based convolutional denoising autoencoder for two-lead ECG denoising and arrhythmia classification," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–10, 2022. https://doi.org/10.1109/TIM.2022.3197757

[23] F. S. Alrayes, M. Zakariah, S. U. Amin, Z. I. Khan, and M. Helal, "Intrusion detection in IoT systems using denoising autoencoder," *IEEE Access*, 2024. https://doi.org/10.1109/ACCESS.2024.3451726

[24] J. Dong, Y. Chen, B. Yao, X. Zhang, and N. Zeng, "A neural network boosting regression model based on XGBoost," *Appl. Soft Comput.*, vol. 125, p. 109067, 2022. https://doi.org/10.1016/j.asoc.2022.109067

[25] M. Amjad, I. Ahmad, M. Ahmad, P. Wróblewski, P. Kamiński, and U. Amjad, "Prediction of pile bearing capacity using XGBoost algorithm: Modeling and performance evaluation," *Appl. Sci.*, vol. 12, no. 4, p. 2126, 2022. https://doi.org/10.3390/app12042126

[26] W. Liu, Z. Chen, and Y. Hu, "XGBoost algorithm-based prediction of safety assessment for pipelines," *Int. J. Press. Vessel. Pip.*, vol. 197, p. 104655, 2022. https://doi.org/10.1016/j.ijpvp.2022.104655