Eduvest – Journal of Universal Studies
Volume 6 Number 4, April, 2026
p- ISSN 2775-3735 - e-ISSN 2775-3727

Bank Customer Churn Prediction Using a Hybrid Ensemble Soft Voting Approach Based on Tabnet and XGBOOST

Mohamad Syazimmi Hersyaputra[1], Shintami Chusnul Hidayati[2]
Institut Teknologi Sepuluh Nopember, Indonesia[1,2]
Email: 6025241017@student.its.ac.id[1], shintami@its.ac.id[2]

Keywords:
Prediksi Churn Nasabah Bank, Hybrid Ensemble, Soft Voting, TabNet, XGBoost

ABSTRACT

In an increasingly competitive banking industry, the ability to predict potential customer churn is a strategic factor in maintaining business profitability and sustainability. Churn has a direct impact on a bank's revenue and operational efficiency; therefore, a prediction model is needed that is not only accurate but also stable and adaptive to variations in customer data. This study proposes a hybrid ensemble soft voting approach based on TabNet and XGBoost to improve the performance and robustness of churn prediction. TabNet, with its sequential attention mechanism, can selectively identify important features, while XGBoost excels at handling nonlinear relationships and controlling overfitting through gradient boosting regularization. The two models are combined using a probability-based soft voting mechanism to produce more balanced and consistent predictions. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied so that the data distribution is more proportional and churn patterns can be better represented. The experimental results show that the proposed approach achieves optimal performance, with an accuracy of 96.74%, precision of 90.09%, recall of 89.53%, and an F1-score of 89.81%. These values indicate that the model is able to maintain a balance between accurate churn detection and the minimization of misclassification. This hybrid ensemble soft voting approach has proven to be superior to single models in terms of predictive stability and generalization capability, making it an effective framework to support data-driven customer retention strategies in the banking sector.

INTRODUCTION

In an increasingly competitive banking industry, the ability to retain customers is a strategic factor that determines the sustainability and profitability of financial institutions (de Lima Lemos et al., 2022). The phenomenon of customer churn has a direct impact on bank revenue and operational efficiency because the cost of acquiring new customers is generally much higher than the cost of retaining existing customers (Alizadeh et al., 2023; Joy et al., 2024). In addition, it can also indicate problems with service quality, customer satisfaction, and the company's competitiveness in facing external pressures (Galal et al., 2022; Rana et al., 2025; Singh et al., 2024). Therefore, early identification of potential churn is an important aspect of customer retention strategies, allowing banks to carry out targeted preventive interventions.

Various Machine Learning (ML) approaches have been widely used to predict customer churn, such as Logistic Regression (Mansoor et al., 2023), Decision Trees (Pulkundwar et al., 2023), Random Forests (Muneer et al., 2022), XGBoost (Gala et al., 2024), and Artificial Neural Networks (ANNs) (Rudd et al., 2022). Although these methods have yielded good

results and are relatively easy to interpret, single models still have limitations in recognizing nonlinear patterns and interactions between features in diverse datasets (Ljubičić et al., 2023; Manzoor et al., 2024; Edwine et al., 2022). In the context of the banking industry, customer behavior is influenced by various factors, such as transaction frequency, account tenure, and digital activity patterns, which do not always have a linear relationship with the probability of churn. In addition, churn datasets generally have an imbalanced class distribution, where the number of non-churn customers is much larger than that of churn customers. This imbalance causes models to be biased toward the majority class, resulting in overall accurate predictions but weak performance in detecting customers who are at risk of churn. Therefore, an approach is needed that can address data imbalance, effectively learn nonlinear relationships between features, and maintain stability and predictive accuracy on varied banking data.

Several previous studies have shown that churn prediction in the banking sector has evolved, but has not yet fully achieved the optimal combination of accuracy, stability, and generalizability. De Lima Lemos et al. (2022) applied a machine learning approach to financial institutions and showed that churn behavior can be effectively predicted from a combination of transaction variables and customer profiles. Muneer et al. (2022) also prove that machine learning approaches in the banking industry can provide strong predictive capabilities. However, challenges remain in balancing model complexity with interpretability and generalizability across different banking contexts. On the other hand, Singh et al. developed a banking churn prediction approach combined with analytical visualization to support management in understanding risky customer behavior. Zheng and Zhang, (2024) further demonstrated that XGBoost has strong potential for predicting bank credit card user churn because it can model complex relationships in tabular data with high performance. Another study by Imani et al. (2025), also confirms that improved accuracy in churn prediction can be achieved through probabilistic approaches and probability calibration, highlighting the importance of high-quality probability estimation in business decision-making contexts.

This study proposes a hybrid ensemble soft voting approach based on TabNet and XGBoost as a solution to improve churn prediction performance in the banking sector, as discussed in Bank Customer Churn Prediction Using a Hybrid Ensemble Soft Voting Approach Based on TabNet and XGBoost. XGBoost is a gradient boosting decision tree algorithm known for its computational efficiency (Khan et al., 2025), strong generalization capability, and robustness when handling large-scale tabular data, including datasets with missing values (Zheng & Zhang, 2024; Silalahi et al., 2023). Meanwhile, TabNet is a deep learning architecture designed specifically for tabular data, implementing a sequential attention mechanism that allows the model to identify and focus on the most relevant features from the dataset without manual feature engineering (Kanász et al., 2024; Chowdhury et al., 2025).

The two models are combined through a soft voting mechanism, which is an ensemble learning approach that aggregates the predicted probabilities of each model to produce a final decision (Nair et al., 2025). In contrast to hard voting, which only considers the majority class predictions, soft voting considers the probability distribution of each model (Zheng et al., 2025), allowing the final decision to reflect the averaged predictive confidence contributed by each model (Sen & Verma, 2023). This approach enables synergy between the representation

learning capability of TabNet and the tree-based boosting power of XGBoost, producing predictions that are more accurate, stable, and generalizable (Salur & Aydın, 2022).

With the integration of SMOTE and hybrid ensemble learning, this research contributes to the development of an accurate and stable churn prediction model suitable for competitive banking environments. Most previous studies relied on single models that were less effective at capturing nonlinear patterns. Therefore, this study proposes a hybrid ensemble soft voting approach to produce a more adaptive and consistent model. The main contribution of this research lies in integrating two models: TabNet, which excels in adaptive feature selection through a sequential attention mechanism, and XGBoost, which efficiently processes tabular data using gradient boosting to capture nonlinear patterns reliably. Furthermore, this study applies a hybrid soft voting mechanism to combine the output probabilities of both models, improving stability and generalization across diverse datasets. This research also provides a systematic and reproducible methodological framework that can be applied in real-world banking environments.

Based on this description, the urgency of this research lies in the need for a churn prediction model that not only achieves high accuracy but is also stable, adaptive, and capable of effectively detecting minority classes in complex banking datasets. This research offers novelty through the integration of two models with complementary strengths—TabNet, which excels in adaptive feature selection using sequential attention, and XGBoost, which is effective in modeling nonlinear relationships in tabular data—combined through a hybrid ensemble soft voting mechanism. Unlike previous studies that primarily focused on single models or simple algorithm comparisons, this study develops a hybrid framework specifically designed to improve predictive stability and generalization. In addition, it incorporates a class imbalance handling strategy using SMOTE, making the model more sensitive in detecting potential churn customers.

Thus, the purpose of this study is to develop and evaluate a bank customer churn prediction model using a hybrid ensemble soft voting approach based on TabNet and XGBoost and to assess its effectiveness using accuracy, precision, recall, and F1-score metrics. Academically, this research is expected to enrich the literature on ensemble learning and deep learning for tabular data in the domain of churn prediction within the banking sector. Practically, the findings are expected to assist banking institutions in building more accurate decision support systems for customer retention strategies, customer relationship management, and targeted loyalty program allocation. Furthermore, this study provides a systematic, reproducible, and adaptive methodological framework applicable to various tabular data-based churn prediction cases, both in banking and other service sectors.

RESEARCH METHOD

Figure 1 shows the flow of the research methodology starting from data collection and data preprocessing. Furthermore, the data were used to build two main models, namely TabNet and XGBoost, which were then combined through an ensemble soft voting mechanism. The final stage was the evaluation of the model to assess the overall performance of the proposed approach.

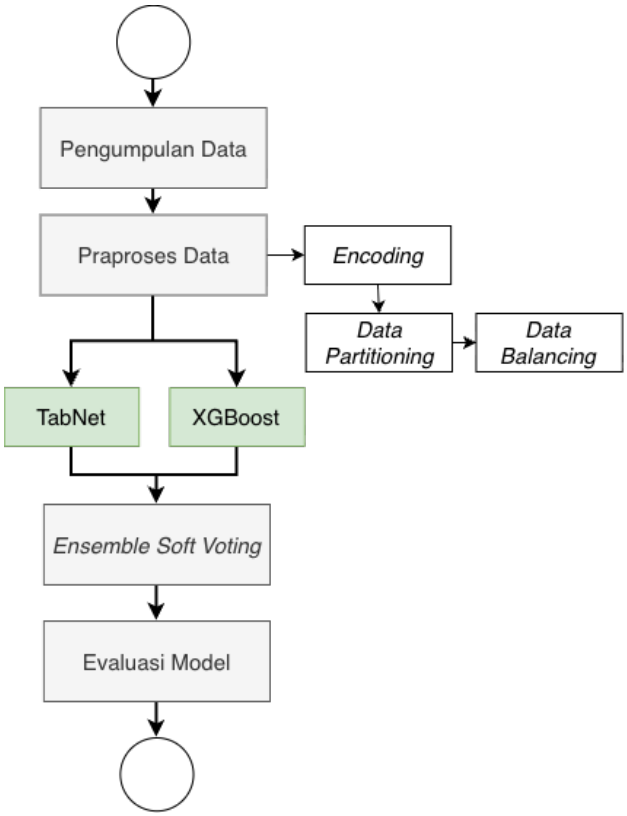


Image. 1 Research methodology
Source: Processed by the author (2026)

The data collection stage in this study was carried out by utilizing the public dataset of Credit Card Customers available on the Kaggle platform. This dataset was chosen because it representative describes the behavioral characteristics and profiles of bank customers that are relevant for churn analysis. The dataset contains 10,127 data entries with 21 variables that include demographic information, credit card usage behavior, and churn status which are the target variables.

In general, this dataset is supervised classification, where the target label is an Attrition_Flag attribute that indicates the customer's churn status, namely whether the customer is still active (existing customer) consisting of 8500 data or churn (attrited customer) consisting of 1627 data. This dataset provides a strong empirical basis for the development of ML-based churn prediction models, in particular to evaluate the effectiveness of the hybrid ensemble approach proposed in this study.

Table 1 shows the demographic features that play an important role in describing customer profiles and influencing the tendency of churn behavior.

Table 1. Demographic Category Features

No	Feature Name	Data Type	Description
1	*Customer_Age*	Numerical	Age
2	*Gender*	Categorical	Gender
3	*Dependent_count*	Numerical	Number of dependents
4	*Education_Level*	Categorical	Last level of education
5	*Marital_Status*	Categorical	Marital status

6	*Income_Category*	Categorical	Annual income category
7	*Card_Category*	Categorical	Types of credit cards owned

Source: Processed by the author based on the Credit Card Customers dataset from Kaggle (2026)

Table 2 describes the features that represent the customer's relationship with the bank as well as the account activity that identifies the level of customer involvement.

Table 2. Relational Category and Account Activity Features

No	Feature Name	Data Type	Description
1	*Months_on_book*	Numerical	Account ownership length (months)
2	*Total_Relationship_Count*	Numerical	Total products used
3	*Months_Inactive_12_mon*	Numerical	Number of inactive months
4	*Contacts_Count_12_mon*	Numerical	Number of interactions

Source: Processed by the author based on the Credit Card Customers dataset from Kaggle (2026)

Table 3 describes the features that represent the transaction activities and financial profile of the customer.

Table 3. Transaction and Financial Category Features

No	Feature Name	Data Type	Description
1	*Credit_Limit*	Numerical	Maximum credit limit
2	*Total_Revolving_Bal*	Numerical	Saldo *revolving*
3	*Avg_Open_ To_Buy*	Numerical	Average balance that can still be used
4	*Avg_Utilization _Ratio*	Numerical	Average credit utilization ratio
5	*Total_Trans _Amt*	Numerical	Total transaction amount
6	*Total_Trans_Ct*	Numerical	Number of transactions
7	*Total_Amt _Chng_Q4_Q1*	Numerical	Total transaction change ratio
8	*Total_Ct_Chng_Q4_Q1*	Numerical	Ratio of change in the number of transactions

Source: Processed by the author based on the Credit Card Customers dataset from Kaggle (2026)

The data preprocessing stage aimed to prepare the data for use in the model training process by ensuring consistency, cleanliness, and appropriate representation for the ML models used. This process was carried out on the dataset with a focus on handling categorical data types, data splitting, and data balancing. The initial stage was conducted through a data quality checking process to identify missing values, duplication, or data type mismatches. The results of the exploration showed that the dataset had no missing values or duplicates; therefore, all data could be used. Distribution checks were also performed to ensure that each feature had reasonable variation and that there were no outliers that could affect the model training results.

The next step was to encode the categorical features so that they could be processed by the ML models. At this stage, two types of methods were used according to the characteristics

of each feature. First, One-Hot Encoding was applied to the Gender and Marital_Status features because these variables were nominal and had no inherent hierarchical order. This technique allowed the model to recognize each category separately without assuming any order among its values. Second, Label Encoding was applied to the Education_Level, Income_Category, and Card_Category features, which had ordinal or naturally sortable category values.

After the encoding process, data partitioning was carried out to divide the dataset into two subsets, namely the training set and the testing set, with a ratio of 80:20. This division was performed in a stratified manner so that the proportion of churn and non-churn classes remained balanced in each subset, thereby avoiding potential bias in the model learning process. However, the comparison between the two classes still showed a significant imbalance, where the number of non-churn customers was much higher than that of churn customers. To address this issue, this study applied SMOTE as an oversampling method to the training data. SMOTE generated synthetic data for the minority class (churn) by creating new samples based on a linear combination of the nearest neighbors in the feature space, thereby increasing the representativeness of the churn pattern without duplicating the data.

Through this series of stages, the data preprocessing process produced a clean, structured, and balanced dataset that was ready to be used for training the TabNet- and XGBoost-based hybrid ensemble soft voting models. This approach ensured that the models utilized high-quality and appropriately processed data, resulting in more accurate and consistent churn prediction performance, particularly in the banking context.

Ensemble Soft Voting

The ensemble soft voting approach was used in this study to combine two models with different but complementary learning characteristics, namely TabNet and XGBoost, with the aim of improving the accuracy, stability, and generalizability of customer churn predictions. Unlike hard voting which only looks at the majority results, soft voting calculates the average probability of each model for each class, so that the final decision is more balanced and takes into account the confidence level of both models. Mathematically, the mechanism of this ensemble can be expressed as :

$$P(y = c) = \frac{1}{n}\sum_{i=1}^{n} P_i(y = c) \quad (1)$$

With P_i(y=c) is the probability of the prediction of the model i to class c, and n is the number of models combined. This approach ensures models with higher prediction confidence have a greater contribution to the final decision, resulting in more robust predictions of data variation (Salur & Aydın, 2022).

The first model used is TabNet. TabNet implements a single deep learning architecture that works through multiple sequential processing stages with sequential attention mechanisms and optimization processes using gradient descent (McDonnell et al., 2023). As shown in the architecture in Figure 2, each feature transformer extracts a feature representation through a split and mask mechanism controlled by the transformer's attentive module (Yu et al., 2024). This process allows the model to select the most relevant subset of features at each step, enabling effective interpretability without sacrificing predictive power (Liu et al., 2023).

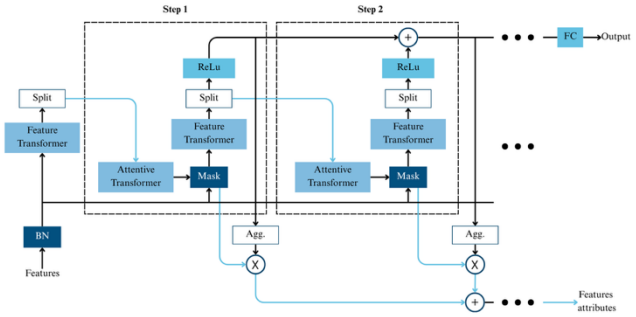


Figure 2. TabNet Architecture

Source: Adapted from Arik and Pfister (2021), reworked by the author (2026)

Mathematically, feature selection is controlled by a matrix of masks calculated as (Liu, 2023):

$$M[i] = Sparsemax\left(\prod_{j=1}^{i-1}(\gamma - M[j]) \cdot (h_i(a[i-1]))\right) \quad (2)$$

Description:

M[i]: Matrix mask in step i that indicates which features the attentive transformer chose to use in that step

Sparsemax(·): An activation function that makes only a few features active with high values, while others remain low or inactive

∏_(j=1)^(i-1)⟦(γ-M[j])⟧ : Cumulative multiplication operation of all masks in the previous step to ensure that previously selected features are not reused in the next step

γ: A parameter that governs how many features are excluded in each step, usually slightly greater than 1 to keep the feature selection process stable

h_i (a[i-1]): The feature transformer function in step i processes the output of the previous step (a[i-1]) to generate a representation of the new feature before the mask is applied

To maintain the sparsity of the features used, TabNet adds regularization with the function (Liu, 2023):

$$L_{sparse} = \frac{1}{N_{steps} \cdot B} \sum_{i=1}^{N_{steps}} \sum_{b=1}^{B} \sum_{j=1}^{D} -M_{b,j}[i] \log\left(M_{b,j}[i] + \in\right) \quad (3)$$

Description:

L_sparse: Regularization function to control feature sparsity so that TabNet selects only important features

N_steps: Total of decision steps

B: Batch size or number of samples processed on each iteration of the training

D: The total number of features (feature dimensions) in the input after the encoding process

M_(b,j) [i]: The mask matrix value of the attentive transformer in step i, indicates the level of attention to the feature of the j in the sample b

∈: Small constants to maintain numerical stability so that errors do not occur in logarithmic operations (log(0))

TabNet balances complexity and regularization by studying complex patterns while still focusing on informative features, thus preventing overfitting and preserving the generalization capabilities of the model (Fares & Abd Elaziz, 2025). In this study, the TabNet hyperparameter tuning process was carried out systematically by testing several learning rate values (1e^(-3) to 5e^(-4)) as well as mask function types (entmax and sparsemax). The training process uses the Adam optimizer and StepLR scheduler to keep learning stable.

The second model, namely XGBoost, is a tree-based ensemble algorithm that uses the principle of additive boosting to minimize the loss function gradually (Azim Mim et al., 2024). Each new tree is built to correct prediction errors from previous models by optimizing objective functions (Zheng & Zhang, 2024):

$$Obj = \sum_i l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i) + \Omega(f_t)\right) \quad (4)$$

where l is the loss function and Ω(f_t) is the regularization to control the complexity of the model.

This regularization makes XGBoost more resistant to overfitting than conventional boosting algorithms. In this study, hyperparameter tuning was carried out using a randomized search cross-validation approach (RandomizedSearchCV) with variations of parameters such as n_estimators (50, 100, 200, and 300), learning_rate (0.01, 0.05, 0.1, and 0.2), max_depth (3, 5, 7, and 10), subsample (0.6, 0.8, and 1.0), and colsample_bytree (0.6, 0.8, and 1.0). The best parameters are selected based on the highest F1-score, which reflects the balance between precision and recall.

The two optimized models were then combined using a soft voting ensemble mechanism, where the probability results from TabNet and XGBoost were averaged to determine the final class. This approach helps improve predictive outcomes over single models and makes model performance more stable on data with complex patterns and unbalanced class distributions.

Model Evaluation

The model evaluation was carried out to objectively measure the performance of bank customers' churn prediction systems using four main metrics, namely accuracy, precision, recall, and f1-score. These four metrics were chosen because they were able to provide a comprehensive picture of the model's ability to classify churn and non-churn customers, especially in unbalanced dataset conditions such as in this study.

Mathematically, the evaluation metric is calculated based on the results of a confusion matrix consisting of four components, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

Accuracy measures the proportion of correct predictions to all test data calculated through equations (5).

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

Precision measures the accuracy of the model in predicting the actual churn customers among all churn predictions. The high precision value indicates that the model gives a slight false positive prediction (minimal FP). Precision measured through equations (6).

$$precision = \frac{TP}{TP+FP} \quad (6)$$

Recall or sensitivity shows the model's ability to detect all customers who are truly churn. A high recall value means an effective model in identifying customers who are at risk of churn. The value of the recall is measured through the equation (7).

$$recall = \frac{TP}{TP+FN} \quad (7)$$

F1-score merupakan rata-rata harmonik antara precision dan recall yang memberikan keseimbangan antara keduanya. Metrik ini penting untuk kasus class imbalance karena membantu menilai kinerja model secara lebih adil terhadap kelas minoritas [32].

$$f1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (8)$$

RESULT AND DISCUSSION

This section discusses the results of this research. The analysis focused on the effectiveness of the TabNet and XGBoost-based hybrid ensemble soft voting methods. The discussion included the influence of the encoding process, SMOTE, and parameter tuning results on model performance measured using accuracy, precision, recall, and F1-score metrics.

Data Preporgestion

The data preporgestion stage in this study is a fundamental process in producing a clean, representative, and ready-to-use dataset for churn prediction model training. Table IV shows the results of the One-Hot Encoding process on the Gender feature which originally had two categories, namely M (male) and F (female). After the transformation, each category is represented in the form of two new binary variables, namely Gender_M and Gender_F. The value "1" indicates the existence of the category in a particular observation, while "0" indicates its absence. For example, a client with a male gender has a value of 1 in column Gender_M and 0 in column Gender_F. This transformation ensures that the model can process categorical variables numerically without assuming ordinal relationships between categories, thereby improving the accuracy and stability of the model training process.

Table 4. One-Hot Encoding Results of Gender Feature

No	Before *Encoding*	After *Encoding*	
	Gender	*Gender_M*	*Gender_F*
1	*M*	1	0
2	*F*	0	1

Source: Processed by the author (2026)

Table 5 shows the results of the Label Encoding process in the Card_Category feature, which originally had four nominal categories, namely Blue, Silver, Gold, and Platinum. Each category is converted into a numerical representation in the order of 0 to 3. These transformations have a sequential nature that reflects the card tier based on value, so the resulting numbers still reflect the relationship between the card categories.

Table 5. Label Encoding Card Category Feature Results

No	Before *Encoding*	After *Encoding*
1	*Blue*	0
2	*Silver*	1
3	*Gold*	2
4	*Platinum*	3

Source: Processed by the author (2026)

Figure 3 shows the changes in class distribution before and after the implementation of SMOTE. Prior to balancing, the churn class had a much smaller amount of data than the non-churn class, which could cause the model to tend to be biased towards the majority class. Once SMOTE is implemented, the amount of data in both classes becomes balanced. This suggests that synthetic oversampling processes are successful in improving minority class representation, so that the model has a better chance of recognizing churn patterns and reducing misclassification of customers at high risk for churn.

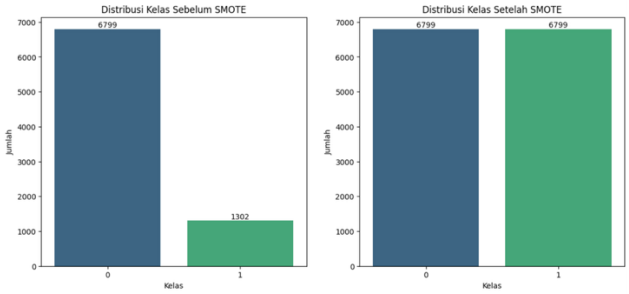


Figure 3. Balancing data results using SMOTE

Source: Processed by the author using the results of research data processing (2026)

Modeling

The modeling stage aims to build a churn prediction model using a TabNet approach that is optimized through a hyperparameter tuning process to achieve the best performance. This model is implemented using the TabNetClassifier of the TabNet PyTorch library, which combines the advantages of deep learning and attention mechanisms for tabular data

processing. TabNet works by extracting features gradually using sequential attention, where each step of the decision learns the most relevant features through the masking process of the transformer's attentive module. This process is carried out end-to-end so that the model can self-adjust the most influential features in the classification process.

The model was trained using the Adam optimizer because it is stable and efficient in accelerating convergence, with learning rate parameters determined dynamically through the StepLR scheduler scheduling scheme ($\gamma = 0.9$, step_size = 10). To control the training process, an early stop with a patience of 20 epoches is used to automatically stop training when the model's performance is no longer significantly improved, and a batch size of 28 to maintain processing efficiency on the CPU. The training process is carried out for a maximum of 100 epoches which are evaluated using test data. The hyperparameter tuning process is carried out by testing several combinations of learning rate values and mask function types to determine the best configuration with the results shown in Table 6.

Table 6. Results of Hyperparameter Tuning Tabnet

No	Parameter Name	Value
1	*learning_rate*	0.0005
2	*mask_type*	*entmax*

Source: Processed by the author based on experimental results (2026)

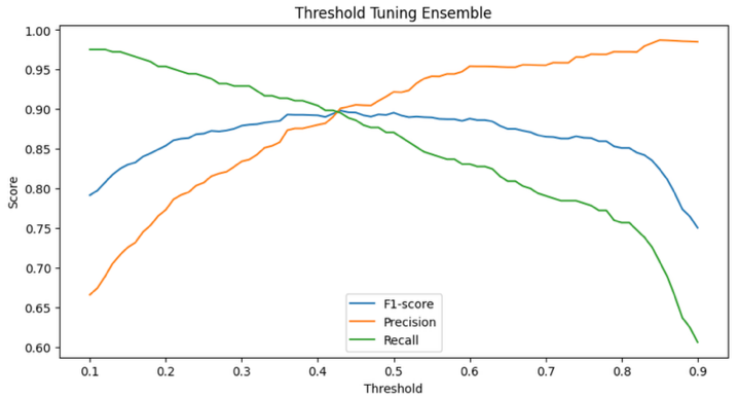
The XGBoost model in this study was built using the Randomized Search Cross-Validation approach to perform hyperparameter tuning efficiently. This process aims to find the best combination of parameters that are able to maximize model performance in predicting customer churn. Some of the key parameters tested included the number of trees (n_estimators), learning rate (learning_rate), tree depth (max_depth), proportion of training data used in each iteration (subsample), and proportion of features used in each tree (colsample_bytree). Based on the tuning results, the best configuration obtained is shown in Table 7.

Table 7. Xgboost Tuning Hyperparameter Results

No	Parameter Name	Value
1	*subsample*	0.8
2	*n_estimators*	300
3	*max_depth*	7
4	*learning_rate*	0.05
5	*colsample_bytree*	0.6

Source: Processed by the author based on experimental results (2026)

Figure 4 shows the threshold tuning process in the ensemble soft voting mechanism. This graph shows the relationship between threshold values and evaluation metrics such as precision, recall, and F1-score. As threshold values increase, precision tends to go up while recalls decrease, as models become more rigorous in classifying churn. The F1-score value is used to balance the two, and from the tuning results an optimal value is obtained at the threshold of 0.43. This point reflects the best balance between the model's ability to detect churn customers without generating too many misclassifications.



Gambar. 4 Threshold tuning ensemble soft voting

Source: Prepared by the author based on the results of model testing (2026)

Performa Model

The results of the model evaluation as shown in Table 8 use a hybrid ensemble soft voting approach. These results illustrate that the combination of TabNet and XGBoost is able to provide a balance between precision and recall, demonstrating consistent performance in recognizing churn customers without too many errors in the non-churn class. The soft voting approach makes the two models complement each other, and TabNet is able to select the most relevant features through sequential attention mechanisms and sparse feature masking, so that the model focuses more on significant attributes in distinguishing churn and non-churn customers. On the other hand, XGBoost effectively captures non-linear relationships and complex interactions between features with gradient boosting and regularization processes that prevent overfitting.

Table 8. Performance Metrics

No	***Performance Metric(s)***	Value
1	*Accuracy*	96.74%
2	*Precision*	90.09%
3	*Recall*	89.53%
4	F1-*Score*	89.81%

Source: Processed by the author based on the results of model evaluation (2026)

As visualized in Figure 5, the confusion matrix shows that the model can correctly classify the majority of customers, with TN (1669) and TP (291) dominating over FP (32) and FN (34). This distribution signifies that the model has strong churn detection capabilities without sacrificing accuracy in the majority class. This reflects the robustness and generalization power of the proposed ensemble learning approach, making it a reliable method in proactively supporting customer retention strategies in the banking sector. Overall, these results suggest that the combination of different (hybrid) architectures with the right data balancing process can result in a balanced model in detecting churn.

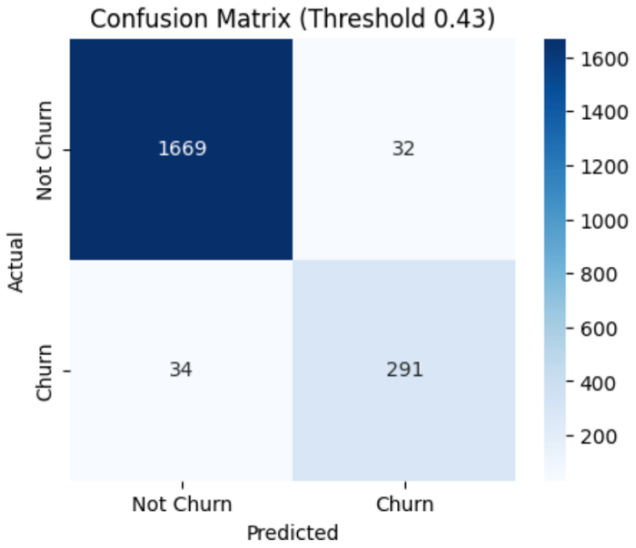


Figure 5. Confusion matrix

Source: Processed by the author based on the results of model evaluation (2026)

CONCLUSION

This study demonstrated that a hybrid ensemble soft voting approach based on TabNet and XGBoost can produce a more accurate and stable bank customer churn prediction model than single-model approaches by combining TabNet's sequential attention–based feature selection with XGBoost's strength in modeling nonlinear relationships and controlling complexity through regularization. The soft voting mechanism effectively integrated the probability outputs of both models, resulting in balanced predictions with improved adaptability to new data, reduced overfitting risk, and a strong balance between precision and recall, even in imbalanced datasets typical of churn scenarios. Consequently, this approach offers a robust solution for tabular data–based churn prediction and has practical potential for application in banking Customer Relationship Management (CRM) systems to support data-driven strategies for customer retention and loyalty. Future research could extend this work by incorporating temporal features and real-time transaction data for streaming prediction, as well as exploring customer relationship structures based on behavioral similarity to enhance model responsiveness to evolving churn dynamics.

REFERENCES

Alizadeh, M., Zadeh, D. S., Moshiri, B., & Montazeri, A. (2023). Development of a customer churn model for banking industry based on hard and soft data fusion. *IEEE Access, 11*, 29759–29768. https://doi.org/10.1109/ACCESS.2023.3257352

Azim Mim, M., Majadi, N., & Mazumder, P. (2024). A soft voting ensemble learning approach for credit card fraud detection. *Heliyon, 10*(3). https://doi.org/10.1016/j.heliyon.2024.e25466

Chowdhury, M. N. H., et al. (2025). Deep learning for early detection of chronic kidney disease stages in diabetes patients: A TabNet approach. *Artificial Intelligence in Medicine, 166*. https://doi.org/10.1016/j.artmed.2025.103153

Das, D. K. (2024). Exploring the symbiotic relationship between digital transformation,

infrastructure, service delivery, and governance for smart sustainable cities. *Smart Cities, 7*(2), 806–835.

Edwine, N., Wang, W., Song, W., & Ssebuggwawo, D. (2022). Detecting the risk of customer churn in telecom sector: A comparative study. *Mathematical Problems in Engineering, 2022*. https://doi.org/10.1155/2022/8534739

Fares, I. A., & Abd Elaziz, M. (2025). Explainable TabNet transformer-based on Google Vizier optimizer for anomaly intrusion detection system. *Knowledge-Based Systems, 316*. https://doi.org/10.1016/j.knosys.2025.113351

Galal, M., Rady, S., & Aref, M. (2022). Enhancing customer churn prediction in digital banking using ensemble modeling. In *Proceedings of the 4th Novel Intelligent and Leading Emerging Sciences Conference (NILES 2022)* (pp. 21–25). IEEE. https://doi.org/10.1109/NILES56402.2022.9942408

Gala, D. M., Pawar, B., Band, G., Dua, P., Mohanty, B. R., & Patil, B. V. (2024). Enhancing predictive accuracy for customer churn in digital banking: A multi-model analysis. In *Proceedings of the IEEE Delhi Section Flagship Conference (DELCON 2024)*. IEEE. https://doi.org/10.1109/DELCON64804.2024.10866346

Imani, M., Joudaki, M., Beikmohammadi, A., & Arabnia, H. R. (2025). Customer churn prediction: A systematic review of recent advances, trends, and challenges in machine learning and deep learning. *Machine Learning and Knowledge Extraction, 7*(3), 105.

Kanász, R., Drotár, P., Gnip, P., & Zoričák, M. (2024). Clash of titans on imbalanced data: TabNet vs XGBoost. In *Proceedings of the IEEE Conference on Artificial Intelligence (CAI 2024)* (pp. 320–325). IEEE. https://doi.org/10.1109/CAI59869.2024.00068

Khan, M. I., et al. (2025). XGBoost-TabNet ensemble model for prediction of methylene blue adsorption on activated carbon geopolymer composite. In *Proceedings of the International Conference on Innovation in Artificial Intelligence and Internet of Things (AIIT 2025)*. IEEE. https://doi.org/10.1109/AIIT63112.2025.11082811

de Lima Lemos, R. A., Silva, T. C., & Tabak, B. M. (2022). Propensity to customer churn in a financial institution: A machine learning approach. *Neural Computing and Applications, 34*(14), 11751–11768. https://doi.org/10.1007/s00521-022-07067-x

Liu, Z. (2023). A new porosity prediction method based on deep learning of TabNet algorithm. In *Proceedings of the IEEE International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA 2023)* (pp. 1681–1685). IEEE. https://doi.org/10.1109/EEBDA56825.2023.10090680

Liu, Z., Fan, Q., Wang, Z., & Cai, Y. (2023). A novel algorithm for credit default prediction using TabNet. In *Proceedings of the International Conference on Electronic Information Engineering and Computer Science (EIECS 2023)* (pp. 24–27). IEEE. https://doi.org/10.1109/EIECS59936.2023.10435411

Ljubičić, K., Merćep, A., & Kostanjčar, Z. (2023). Churn prediction methods based on mutual customer interdependence. *Journal of Computational Science, 67*. https://doi.org/10.1016/j.jocs.2022.101940

Manzoor, A., Qureshi, M. A., Kidney, E., & Longo, L. (2024). A review on machine learning methods for customer churn prediction and recommendations for business practitioners. *IEEE Access, 12*, 70434–70463. https://doi.org/10.1109/ACCESS.2024.3402092

McDonnell, K., Murphy, F., Sheehan, B., Masello, L., & Castignani, G. (2023). Deep learning in insurance: Accuracy and model interpretability using TabNet. *Expert Systems with Applications, 217*. https://doi.org/10.1016/j.eswa.2023.119543

Muneer, A., Ali, R. F., Alghamdi, A., Taib, S. M., Almaghthawi, A., & Ghaleb, E. A. A. (2022). Predicting customers churning in banking industry: A machine learning approach. *Indonesian Journal of Electrical Engineering and Computer Science, 26*(1), 539–549. https://doi.org/10.11591/ijeecs.v26.i1.pp539-549

Nair, A. S., Krishna, A., Gupta, S. T., & Susan, S. (2025). Credit card fraud detection using soft voting ensemble with imbalance treatment. In *Proceedings of the International Conference on Intelligent Technologies* (pp. 1–5). IEEE. https://doi.org/10.1109/CONIT65521.2025.11167470

Pulkundwar, P., Rudani, K., Rane, O., Shah, C., & Virnodkar, S. (2023). A comparison of machine learning algorithms for customer churn prediction. In *Proceedings of the IEEE International Conference on Advances in Science and Technology (ICAST 2023)* (pp. 437–442). IEEE. https://doi.org/10.1109/ICAST59062.2023.10455051

Rana, M. S., et al. (2025). AI-driven predictive modeling for banking customer churn: Insights for the US financial sector. *Journal of Ecohumanism, 4*(1), 3478–3497. https://doi.org/10.62754/joe.v4i1.6188

Salur, M. U., & Aydın, İ. (2022). A soft voting ensemble learning-based approach for multimodal sentiment analysis. *Neural Computing and Applications, 34*(21), 18391–18406. https://doi.org/10.1007/s00521-022-07451-7

Singh, P. P., Anik, F. I., Senapati, R., Sinha, A., Sakib, N., & Hossain, E. (2024). Investigating customer churn in banking: A machine learning approach and visualization app for data science and management. *Data Science and Management, 7*(1), 7–16. https://doi.org/10.1016/j.dsm.2023.09.002

Zheng, C., Chen, Y., & Du, X. (2025). A robust soft voting ensemble of the isolation forest model, extended isolation forest model and generalized isolation forest model for multivariate geochemical anomaly recognition. *Ore Geology Reviews, 185*. https://doi.org/10.1016/j.oregeorev.2025.106787

Zheng, Y., & Zhang, N. (2024). Research on prediction bank credit card user churn based on XGBoost. In *Proceedings of the International Conference on Electronic Technology and Information Science (ICETIS 2024)* (pp. 690–693). IEEE. https://doi.org/10.1109/ICETIS61828.2024.10593724