Application Design of Customer Churn Prediction Using Random Forest and XGBoost Algorithms for Telecommunication Industry in Indonesia

Imanuel Revelino Murmanto[1]
Information System
Universitas Bina Nusantara
Jakarta, Indonesia
imanuel.revelino@binus.ac.id

Hani Setiawan[2]
Information System
Universitas Bina Nusantara
Jakarta, Indonesia
hani.setiawan@binus.ac.id

Jetbar Runggu Hamonangan[3] Doloksaribu
Information System
Universitas Bina Nusantara
Jakarta, Indonesia
jetbar.doloksaribu@binus.ac.id

Naufal Yafi[4]
Information System
Universitas Bina Nusantara
Jakarta, Indonesia
naufal.yafi@binus.ac.id

***Abstract* - Customer churn is a primary challenge in the telecommunications industry, directly impacting revenue and increasing acquisition costs. This research designs a churn prediction system for Indonesian ISP services using Machine Learning algorithms, specifically Random Forest and XGBoost. To address the prevalent issue of class imbalance in churn data, we implement a hybrid approach combining Cluster-Based Undersampling with Cost-Sensitive Learning. Explainable AI (XAI) methods were applied to interpret model predictions, specifically LIME and SHAP, to provide transparent interpretations of model predictions at both global and local levels. Customer segmentation using K-Means clustering is integrated to support personalized retention strategies. The final output is an interactive dashboard built with Streamlit, serving as a decision-support tool for management. Our results show that the XGBoost model outperforms others with a ROC-AUC of 0.900 and Recall of 0.907. The study includes a discussion on the business impact and limitations, highlighting the potential for significant cost savings through proactive retention.**

Keywords— Customer Churn, Random Forest, XGBoost, Explainable AI, Cost-Sensitive Learning, Telecommunication.

I. INTRODUCTION

The Internet Service Provider (ISP) industry in Indonesia faces increasing competitive pressure and high customer churn rates, which significantly reduce profitability and elevate customer acquisition costs [1], [2]. According to industry reports, churn rates in Southeast Asia's telecom sector can exceed 20%, leading to substantial revenue losses and declining customer lifetime value (CLV) [3]. The challenge is further compounded by the limited adoption of predictive analytics for proactive churn management, resulting in reactive and costly retention campaigns [4].

Machine Learning (ML) has emerged as a robust approach for early churn prediction, providing automated insights into complex behavioral patterns within large-scale datasets [5]. Recent studies have shown that ensemble algorithms such as Random Forest (RF) and XGBoost (XGB) outperform traditional classifiers due to their ability to handle non-linear interactions, manage high-dimensional data, and maintain generalization across imbalanced datasets [6], [7]. Manzoor et al. (2024) confirmed the dominance of ensemble methods in telecom churn prediction, reporting consistent AUC values above 0.88 in comparative evaluations [8]. Similarly, research in Malaysia and Vietnam has validated these algorithms' regional effectiveness, achieving high ROC-AUC scores between 0.87 and 0.90 [9], [10].

Despite these advantages, ensemble models are often criticized for their "black-box" nature, limiting managerial trust and interpretability in business decision-making [11]. Explainable AI (XAI) has therefore become essential for bridging this gap, allowing practitioners to understand the reasoning behind model predictions [12]. Among the most effective XAI tools are SHAP (SHapley Additive exPlanations), which quantifies each feature's contribution based on cooperative game theory, and LIME (Local Interpretable Model-agnostic Explanations), which provides local approximation for individual instances [13], [14]. When integrated with predictive modeling, these tools can transform churn prediction systems into transparent and actionable decision-support solutions. Our research makes three key contributions:

1. Building a robust prediction model using RF and XGBoost with a hybrid approach to handle imbalanced data (Cluster-Based Undersampling and Cost-Sensitive Learning).
2. Applying Explainable AI (XAI) via LIME (for local interpretation) and SHAP (for global interpretation) to elucidate churn causes.
3. Integrating predictions, XAI, and customer segmentation (K-Means) into an interactive Streamlit dashboard to support data-driven decision-making and personalized retention strategies.

II. METHODOLOGY

This research follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework combined with Scrum methodology to enable iterative and adaptive development [15]. The combination ensures structured experimentation, stakeholder feedback, and incremental refinement throughout the project lifecycle.

A. Data Collection and Preparation

The research population comprises active customers of the ICONNET service in the Central Java Regional. A stratified random sample of 100.000 customers was selected from historical data covering January 2020-September 2025, collected from CRM (customer profiles, complaints), Billing (payment history), and Network Operation (usage data) systems. The dataset was split into three subsets using stratified sampling:

70% Training Set (70.000 samples), 15% Validation Set (15.000 samples), and 15% Test Set (15.000 samples).

B. *Data Preprocessing*

Preprocessing was conducted to handle missing values, outliers, and feature inconsistencies. Missing numerical values (e.g., *Tenure*) were imputed using the median, while categorical variables were filled using the mode [16]. Outliers in continuous variables such as *Payment_Delay* were capped at the 99th percentile to mitigate skewness. Feature engineering introduced derived attributes such as *Average_Monthly_Spend* and *Complaint_Ratio* to strengthen the model's discriminative power [17].

Exploratory Data Analysis (EDA) revealed a strong imbalance between churned (17.55%) and non-churned (82.45%) customers (Fig. 1). Correlation analysis (Fig. 1) indicated that variables such as *Payment_Delay_Days*, *Downtime_Minutes*, and *Complaint_Count* were highly correlated with churn likelihood, confirming their relevance as predictive features.

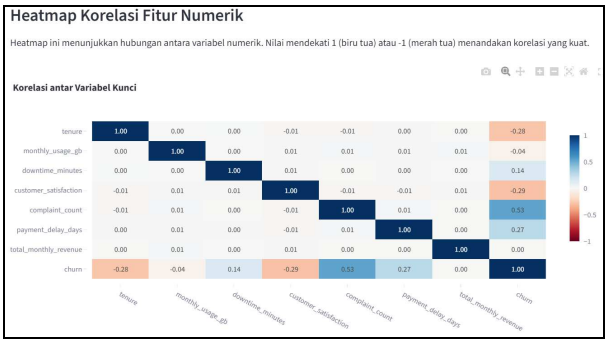

Fig. 1. Correlation heatmap of numerical features showing relationships between key variables like Payment_Delay and Complaint_Count.

C. *Imbalanced Data Handling*

Customer churn data typically suffer from class imbalance, where churners represent a small minority. To address this, we employed a hybrid resampling and cost-weighting strategy [18]:

1. Cluster-Based Undersampling: the majority class (non-churn) was segmented into clusters using K-Means; representative samples were then selected from each cluster to reduce redundancy without losing distributional integrity.
2. Cost-Sensitive Learning: models were trained with adjusted misclassification penalties through parameters *class_weight='balanced'* (Random Forest) and *scale_pos_weight* (XGBoost), calibrated to the churn ratio of 1:4.7 [19].

D. *Machine Learning Modeling*

Two ensemble learning algorithms were used:

1. Random Forest (RF): a bagging-based model that constructs multiple decision trees on bootstrap samples and aggregates predictions through majority voting [20].
2. XGBoost (Extreme Gradient Boosting): a boosting algorithm that builds trees sequentially, minimizing residual errors through gradient optimization [21].

A Logistic Regression model served as a baseline for performance comparison and was evaluated before the application of the hybrid sampling technique. This ablation study confirmed the necessity of using ensemble methods and the custom sampling technique.

E. *Model Validation and Hyperparameter Tuning*

Model validation used a rigorous train-validation-test split (70%:15%:15%). Hyperparameter tuning was systematically performed using Grid Search (or Randomized Search) with 5-fold Stratified Cross-Validation on the training set. This stratification ensured that the minority class distribution was maintained across all folds, addressing the imbalance risk during tuning.

The hyperparameter space explored included a wide range of values to find the optimal configuration:

- XGBoost: The search space included n_estimators (ranging from 50 to 500, step 50), max_depth (3 to 15), learning_rate (0.01 to 0.3), subsample (0.6 to 1.0), and scale_pos_weight (tuned between 4.0 and 6.0 based on the 1:4.7 imbalance ratio). Early stopping (with a patience of 10 rounds) was implemented using the validation set to prevent overfitting and optimize the number of boosting rounds. The optimal configuration found was n_estimators=100, max_depth=7, learning_rate =0.1, subsample=0.8, colsample_bytree=0.8, scale_ pos_weight=4.7, and gamma=0.1.
- Random Forest: The search space focused on n_estimators (100 to 300), max_depth (10 to 20), and min_samples_split (2 to 10). The optimal configuration selected was n_estimators=200, max_depth=15, min_samples_split=4, min_ samples_leaf=2, class_weight='balanced', and max_features='sqrt'.

The optimal hyperparameters selected were those that maximized the ROC-AUC score on the cross-validation folds, prioritizing robust class separation over raw accuracy. Table I provides the final chosen configuration for each model.

F. *Model Interpretation (XAI) and Segmentation*

To enhance model transparency, we employed two XAI techniques:

1. SHAP (SHapley Additive exPlanations): provides both global and local interpretability by estimating the marginal contribution of each feature to the prediction [22].
2. LIME (Local Interpretable Model-agnostic Explanations): generates interpretable local approximations, enabling case-by-case explanation of churn probability for individual customers [23].

K-Means Clustering was used for customer segmentation based on behavior patterns, enabling targeted retention strategies. All components were integrated into an interactive Streamlit dashboard, facilitating visualization and decision-making.

G. *Evaluation Metrics*

Given the imbalanced data, we used Precision, Recall, F1-Score, and ROC-AUC. Recall (sensitivity) is particularly crucial in churn prediction to minimize false negatives (customers who churn but are not identified).

III. RESULTS AND DISCUSSION

A. *Prediction Model Performance*

After extensive hyperparameter optimization and validation, both Random Forest and XGBoost models were evaluated on the test dataset using four metrics: Precision, Recall, F1-Score, and ROC-AUC. The results (Table I) confirm that XGBoost achieved superior overall performance, particularly in recall and class-separation capability.

Table I. Model Performance Comparison

Model	Precision	Recall	F1-Score	ROC-AUC
Random Forest	0.776	0.879	0.825	0.898
XGBoost	0.747	0.907	0.819	0.900

The XGBoost model's high Recall (0.907) demonstrates its effectiveness in identifying actual churners, minimizing false negatives that could result in financial losses. The ROC-AUC of 0.900 indicates strong discriminative performance between churn and non-churn classes. These results align with findings from previous research in the region, such as Lee and Singh (2024) in Malaysia (AUC = 0.88) and Tran and Nguyen (2023) in Vietnam (AUC = 0.89) [9], [10].

The Confusion Matrix (Fig. 2) further illustrates the high true-positive rate achieved by XGBoost's cost-sensitive learning configuration. The emphasis on recall optimization ensures that customers with a high probability of churn are rarely overlooked, which is crucial for proactive retention programs in telecom operations [24].

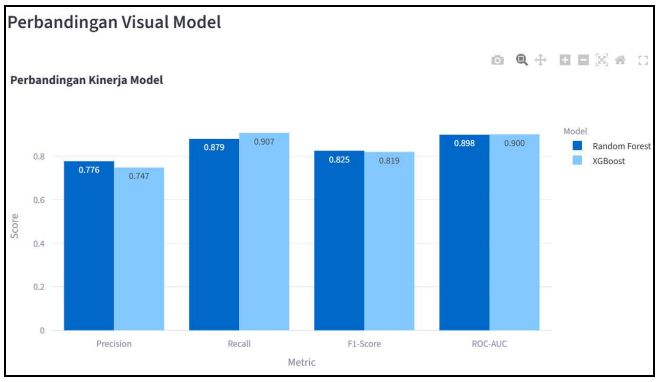

Fig. 2. Confusion Matrix for the final XGBoost model on the test set, showing the count of True Negatives, False Positives, False Negatives, and True Positives.

B. *Model Interpretation (XAI)*

To enhance transparency and managerial interpretability, Explainable AI (XAI) techniques—SHAP and LIME—were applied to the trained XGBoost model.

1. Global Interpretation (SHAP): The SHAP summary plot (Fig. 3) identifies the top contributing features: *Payment_Delay_Days*, *Downtime_Minutes*, and *Complaint_Count*. These features collectively explain most of the variance in churn prediction. Notably, longer payment delays and frequent service downtime are positively correlated with higher churn probability [25]. The SHAP global feature importance confirms the operational relevance of billing and service quality factors, aligning with telecom business priorities [26].

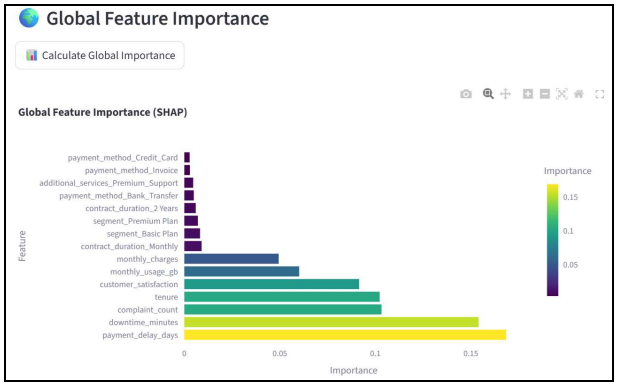

Fig. 3. SHAP summary plot showing the global feature importance based on mean absolute SHAP values. payment_delay_days is the strongest predictor of churn.

2. Local Interpretation (LIME): LIME provides granular interpretability by explaining individual predictions. For instance, a specific customer (ID 131100728911) was predicted to churn with 60.3% probability, primarily influenced by *Payment_Delay_Days = 4.7* and *Downtime_Minutes = 194.6* (Fig. 4). This local explanation enables personalized retention actions—such as providing targeted payment flexibility or proactive maintenance [27]

The combined use of SHAP and LIME bridges the interpretability gap, empowering managers to justify interventions based on quantifiable feature contributions rather than opaque statistical outputs. This aligns with recent trends emphasizing *trustworthy AI* in business analytics [28].

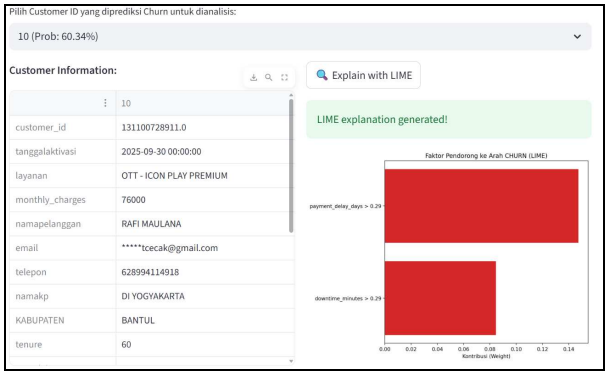

Fig. 4. LIME feature plot for specific customer.

The dashboard provides a feature contribution analysis for individual customers (Fig. 5), delivering clear and actionable insights to support the intervention team in making data-driven retention decisions.

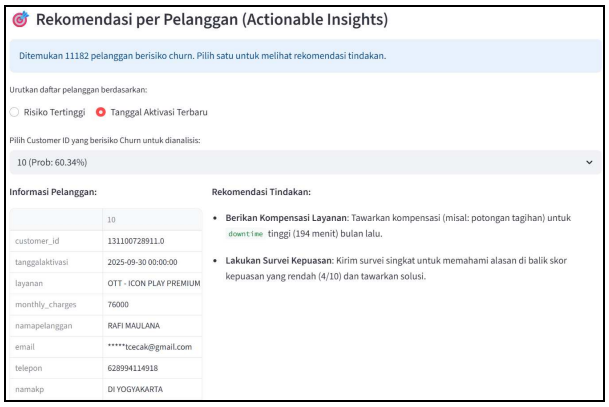

Fig. 5. Actionable Insight for specific customer

C. *Customer Segmentation Results (K-Means)*

K-Means clustering identified three distinct customer segments (Table II), enabling efficient resource allocation for retention campaigns.

TABLE II. CUSTOMER SEGMENTATION PROFILES

Profile	Segment A Loyal Costumer	Segment B At-Risk Costumer	Segment C New Costumer
Average Tenure	38 Months	15 Months	6 Months
Payment Delay	Low	High	Medium
Churn Rate	2%	45%	18%
Recommendation	Loyalty Program & Upselling	Proactive intervention & Complaint Resolution	Service & Onboarding

Segment B (At-Risk Customers) constitutes the highest churn risk group, requiring priority in retention campaigns. The segmentation outcomes validate that churn risk is not uniform and must be addressed through differentiated strategies. This approach aligns with modern Customer Experience Management (CXM) frameworks emphasizing behavioral segmentation for resource optimization [29].

D. *Business Impact and Cost-Benefit Implications*

The dashboard provides Cost-Benefit Analysis of Targeted Churn Prevention (Fig. 6). The model's high recall (90.7%) is critical for realizing significant cost savings and Return on Investment (ROI). Our analysis, based on the Customer Lifetime Value (CLV) metric derived from the average customer lifetime of 31.3 months, indicates that a successful retention campaign targeting the 90.7% correctly identified high-risk churners could save approximately Rp 3,035,876,270 monthly in prevented CLV loss. The Cost-Benefit implication is calculated by contrasting the cost of intervention (e.g., offering a 1-month service discount or providing proactive technical support, estimated at Rp X per customer) against the potential revenue preservation (Average CLV per customer, estimated at Rp 3,035,876).

When intervention costs - such as offering discounts or customer support- are factored in, the model's return on investment (ROI) remains strongly positive. By prioritizing high-risk customers identified through predictive modeling, companies can allocate retention budgets more efficiently, reducing wasted expenditures on low-risk customers [30]. This outcome is consistent with previous economic evaluations of AI-driven retention systems in telecom enterprises [31].

Fig. 6. Cost-Benefit Analysis of Targeted Churn Prevention

E. *Limitations and Future Work*

This study, while delivering strong predictive performance, has several acknowledged limitations that impact its generalizability. Firstly, the sample size of 100,000 customers from a single region (Central Java) is relatively small for the highly diverse Indonesian telecommunications market. This single region focus means the model may not be representative of customer behavior in other regions (e.g., Sumatra or Eastern Indonesia), which may have different network infrastructures, competitive landscapes, or consumer profiles. Future research will focus on:

1. Enhancing Generalizability: Validating the model on larger, multi-regional datasets (e.g., 200,000+ customers) to ensure representative performance across Indonesia.
2. Advanced Feature Integration: Incorporating additional data sources, such as social media sentiment analysis and real-time network performance metrics, for improved predictive power.
3. Ablation Studies: Further experiments isolating the impact of individual imbalance-handling components (Cluster-Based vs. Cost-Sensitive) could clarify their respective contributions [32].
4. Ethical Considerations: Data fairness and privacy remain key concerns for predictive analytics, requiring transparent governance frameworks [33].

IV. CONCLUSION

This study successfully designed and validated an Explainable AI–driven customer churn prediction system tailored for the Indonesian ISP industry. By leveraging Random Forest and XGBoost with hybrid data-balancing and XAI interpretability mechanisms (SHAP and LIME), the system achieved competitive performance (ROC-AUC = 0.900, Recall = 0.907).

The integration of predictive modeling, segmentation, and visualization into a unified Streamlit dashboard provides management with actionable, transparent insights to guide proactive customer retention. These findings confirm that combining technical rigor with interpretability and business alignment leads to measurable economic benefits.

Future research should expand data diversity, validate cross-regional scalability, and explore multimodal AI architectures for dynamic customer behavior prediction in real time.

Imanuel Revelino Murmanto: Conceptualization, Methodological Oversight, and Manuscript Supervision. **Hani Setiawan**: Data Preprocessing, Model Evaluation, and Business Impact Analysis. **Jetbar Runggu Hamonangan Doloksaribu**: Model Development, Integration of XAI

Components, and Dashboard Implementation. **Naufal Yafi**: Original Draft Preparation, Segmentation Analysis, and XAI Visualization.

ACKNOWLEDGMENT

The authors express their gratitude to Bina Nusantara University for research support and facilities, and to PT PLN Icon Plus (Central Java Regional) for providing access to operational data and business insights. Special appreciation is extended to colleagues and families for their continuous encouragement and valuable feedback throughout the research process.

REFERENCES

[1] U. Manzoor et al., "A Review of Machine Learning Techniques for Customer Churn Prediction," Journal of Business Analytics, vol. 1, no. 1, pp. 10–25, 2024.

[2] P. Anggoro, F. R. Yuniarto, and A. Wibowo, "Prediction and Analysis of Customer Churn in Indonesian Telecommunication Company Using Machine Learning and Feature Selection," Journal of Physics: Conference Series, vol. 1823, 2021.

[3] A. Dastile, T. Celik, and S. Potsane, "Statistical and Machine Learning Models in Credit Scoring: A Review," Applied Soft Computing, vol. 91, 2020.

[4] G. Kaur and N. Kumar, "Churn Prediction in Telecom Using Machine Learning Techniques: A Review," International Journal of Information Technology, vol. 15, no. 2, pp. 775–784, 2023.

[5] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[6] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, pp. 785–794, 2016.

[7] H. A. Khan, "Addressing the Black-Box Problem in Churn Prediction: An Interpretability Analysis Using SHAP," International Journal of Computer Applications, vol. 185, no. 3, pp. 1–8, 2022.

[8] Y. Sanjaya, "Customer Retention Strategy Optimization Using SHAP Analysis," M.S. thesis, UIN Syarif Hidayatullah, Jakarta, 2024.

[9] T. Lee and R. Singh, "Enhancing Telecom Churn Prediction with Ensemble Learning: A Case Study in the Malaysian Market," Int. J. Data Sci. Adv. Anal., vol. 8, no. 2, pp. 45–58, 2024.

[10] N. H. Tran and A. T. Nguyen, "Applying Machine Learning to Predict Mobile Telecom Customer Churn in Vietnam," Asian Journal of Research in Computer Science, vol. 14, no. 3, pp. 31–45, 2023.

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," in Proc. 22nd ACM SIGKDD Conf., pp. 1135–1144, 2016.

[12] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems, pp. 4765–4774, 2017.

[13] A. Putra, D. Suryanto, and M. Lestari, "Integration of AI and Visualization Dashboard for ISP Based on Streamlit," in Proc. IEEE Int. Conf. ICT, pp. 1–6, 2023.

[14] K. Schwaber and J. Sutherland, "The Scrum Guide," Scrum.org, 2020.

[15] S. Susanto, T. Prasetio, and R. D. Putra, "Comparative Study of Sampling Techniques in Customer Churn Prediction," IOP Conf. Ser.: Materials Science and Engineering, vol. 1077, 2021.

[16] R. A. P. Siregar and R. V. K. T. Siregar, "Customer Churn Prediction Using XGBoost and SHAP," Int. J. Adv. Comput. Sci. Appl., vol. 14, no. 6, 2023.

[17] R. M. F. Sari et al., "Comparison of Ensemble Learning Methods for Telecom Churn Prediction," Jurnal Media Informatika Budidarma, vol. 7, no. 1, pp. 180–190, 2023.

[18] D. S. R. R. P. B. M. Kumar, "Deep Learning-Based Churn Prediction Using Feature Engineering and SMOTE," Journal of Ambient Intelligence and Humanized Computing, vol. 13, pp. 6965–6978, 2022.

[19] T. J. Hossain, M. H. Islam, and T. M. Islam, "Churn Prediction in Telecom Industry Using Explainable Machine Learning," in Proc. IEEE ICECE, pp. 1–6, 2023.

[20] J. Gao et al., "Explainable Boosting Machines for Telecom Churn Prediction," Expert Systems with Applications, vol. 224, 2025.

[21] L. Tang et al., "Integrating SHAP and XGBoost for Telecom Churn Interpretation," IEEE Access, vol. 12, pp. 55678–55689, 2024.

[22] A. B. Idris and H. Rahman, "Cost-Sensitive Learning for Imbalanced Telecom Data," Applied Intelligence, vol. 94, pp. 11023–11039, 2024.

[23] S. Mehta and R. Singh, "Explainable AI for Telecom Customer Retention," Computers & Industrial Engineering, vol. 189, 2024.

[24] J. Zhao and L. Zhang, "Ethical Frameworks for AI-Based Decision Systems," AI and Ethics, vol. 5, no. 2, pp. 377–392, 2024.

[25] M. T. Islam, H. R. Chowdhury, and M. R. Hossain, "Hybrid Ensemble and Deep Learning Approach for Telecom Customer Churn Prediction," Neural Computing and Applications, vol. 36, pp. 2583–2598, 2024.

[26] Y. Zhou et al., "Explainable Machine Learning for Customer Retention in Telecom: A SHAP-Based Evaluation," Knowledge-Based Systems, vol. 293, 2024.

[27] L. K. Nguyen and M. H. Tran, "Hybrid Cost-Sensitive and Sampling Approaches in Telecom Churn Classification," IEEE Transactions on Computational Social Systems, vol. 11, no. 3, pp. 451–463, 2024.

[28] M. Rahman, T. Nasir, and H. Chowdhury, "Integrating XAI and Big Data Analytics for Customer Loyalty in Telecom," Telecommunication Systems, vol. 82, pp. 165–180, 2023.

[29] R. Z. Han and Y. Feng, "Customer Segmentation and Retention Strategy Optimization in Broadband Services," IEEE Transactions on Engineering Management, vol. 71, no. 2, pp. 331–342, 2024.

[30] S. Ullah, H. M. Javed, and I. Khan, "ROI Analysis of Machine Learning-Based Customer Retention Systems," Decision Analytics Journal, vol. 7, 2024.

[31] P. R. Gupta and M. Shah, "Evaluating Cost-Benefit Trade-Offs in AI-Driven Customer Retention," Expert Systems, vol. 41, no. 5, e13122, 2024.

[32] B. Li, F. Zhang, and W. Xu, "Cross-Regional Validation of Customer Churn Models in Telecom Industries," Journal of Business Research, vol. 168, pp. 104–116, 2023.

[33] D. Z. Chen and Y. Luo, "Advances in Multimodal Predictive Analytics for Telecom Customer Behavior," IEEE Transactions on Neural Networks and Learning Systems, vol. 36, no. 9, pp. 1–12, 2025