Predicting Trial-to-Paid User Conversion in Video Streaming Platforms Using Session-Based Behavioral Data

Anastasija Kostovska and Hristijan Gjoreski
Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University in Skopje, Rugjer Boshkovik Str. 18, 1000 Skopje, N.Macedonia
nikolovska.anastasija@gmail.com, hristijang@feit.ukim.edu.mk

Keywords: Trial-to-Paid Conversion, Video Streaming Analytics, Marketing Optimization, Machine Learning.

Abstract: Predicting user conversion from free trial to paid subscription is an important task for video streaming platforms. It helps to improve marketing efforts and reduce unnecessary costs. Traditional marketing methods, such as sending many emails, often give low results and can cause users to lose interest. This study is based on behavioral data from 60,000 trial users of a video streaming platform, of whom about 10% converted to paid subscriptions. The dataset includes user metadata, session activity, and conversion outcomes, providing rich signals of viewing behavior. In this paper, we present a machine learning approach to predict the chance of conversion by analyzing user behavior during the trial period. We test both single-model and ensemble-model classifiers, using features based on session activity, device usage, and content variety. Since only 10% of users convert, we tested different sampling strategies, including SMOTE, undersampling on 10,000 samples, and a combined SMOTE + undersampling approach. The results show that ensemble methods, especially Random Forest, perform better than simpler models in terms of F1-score and recall for the minority class. Furthermore, the feature importance analysis shows that session duration, number of used devices, and diversity of content are strong indicators of conversion. These findings show that machine learning can support more effective and targeted marketing, with less unwanted communication.

1 INTRODUCTION

Online video content has become a regular part of people's lives and plays an important role in different industries. The growth of social networks and various streaming platforms has strongly supported this trend. As the video streaming environment continues to change, it is influenced by user behavior, market developments, strong competition, and cultural factors [1]. From user perceptions and platform adoption to multicultural engagement, live sports, gaming, and the rise of ad-supported models, a range of powerful forces are redefining the future of streaming.

Video streaming platforms frequently offer trial periods to attract new users, hoping a portion will convert into paying subscribers. However, in most cases, only a small number of users decide to pay. Prior work has shown that trial design and user behavior during trials strongly affect conversion outcomes [2]. Despite this, marketing strategies often rely on mass email campaigns, which not only have low effectiveness but can also reduce user interest and trust.

This paper explores a data-driven alternative by analyzing behavioral session data to predict which users are likely to convert to paid subscriptions. By examining patterns such as session duration, device usage, and viewing diversity, we aim to identify behavioral signals that distinguish potential subscribers from non-converting users.

From a scientific perspective, this study contributes in several ways. First, it formalizes the trial-to-paid conversion problem as a supervised classification task based on session-level behavioral dynamics- an area that remains underexplored in existing literature. Second, it provides an empirical comparison of multiple imbalance-handling strategies (SMOTE, undersampling, and hybrid methods) and quantifies their effects on model generalization and minority-class recall, both of which are critical for reliable prediction in imbalanced business datasets. Third, by combining interpretable models with feature importance analysis, the study identifies measurable behavioral indicators of subscription intent, bridging machine learning methods with actionable marketing insight.

2 RELATED WORK

User conversion prediction is a well-studied problem in domains such as e-commerce, mobile applications, and online media services. In the context of video streaming, predicting trial-to-paid conversion is particularly relevant due to the high customer acquisition costs and the competitive nature of the subscription-based content industry.

Many studies have focused on using user behavior to understand engagement and to predict whether a user will stay or leave. For example, Wang et al. [3] studied how users browse in e-commerce websites to predict whether a session will lead to a purchase, while recent work on mobile platforms has focused on churn and subscription renewal prediction using behavioral features [4]. These works show the importance of user interaction data, including frequency, recency, and type of usage, as predictors of intent.

In the streaming field, platforms like Netflix and Hulu have used methods like collaborative filtering and sequence modeling to recommend content and guess future user actions. However, these systems mainly focus on what content to recommend, not on predicting who will pay for a subscription. Gupta et al. [5] showed the value of session-level metrics, including session frequency and time spent on content, in understanding user satisfaction and predicting retention.

Different machine learning models have been used in conversion prediction, such as logistic regression, decision trees, and ensemble models like random forest and gradient boosting. Some newer methods include deep learning and models like LSTM that can work with time-based data. However, in business, simpler models are often preferred, because they are easier to understand and give clear results that can be used in planning.

Class imbalance is a common issue in conversion prediction because the number of users who convert is often much smaller than the number who do not. To address this problem, several studies have proposed resampling methods. Recent surveys [6] confirm that resampling remains an effective approach, while Wongvorachan and Bollen [7] showed that hybrid strategies combining SMOTE with undersampling often achieve better recall and F1-score for the minority class. Recent theoretical work has also analyzed the properties of SMOTE, providing insight into why it improves minority-class prediction [13].

In practical marketing situations, where it is important to identify as many potential converters as possible, SMOTE is usually preferred. It helps balance the dataset by generating synthetic samples of the minority class without losing any information from the majority group.

Our approach aligns with this stream of research by focusing on interpretability and business impact. We use real session data from a video streaming service to build classification models that can predict conversion. Compared to other studies, we give more attention to extracting practical conversion insights, such as typical time to conversion and device-based engagement, which can directly guide marketing strategy. Unlike prior studies that primarily focused on churn prediction or content recommendation, our work directly targets the trial-to-paid conversion stage in video streaming, which is a critical but less explored business challenge. The novelty of this paper lies in combining behavioral session features with imbalance-aware modeling to produce interpretable insights that can directly guide marketing strategies.

3 DATASET AND PREPROCESSING

This study uses behavioral data collected from a video streaming platform that offers free trial periods to new users. The dataset includes information from 60,000 trial users, and around 10% of them (5,500 users) converted to paid subscriptions. The data is grouped into three main sources: user metadata, session activity, and conversion records.

User information included identifiers, subscription outcomes, and the timing of conversions, along with additional attributes such as email-related metadata. Session-level behavior was captured from detailed activity logs, covering aspects such as session duration (start and end times), device characteristics (operating system, device type, application version, and user agent), and content engagement through channel identifiers. A clear conversion marker was added to distinguish users who subscribed.

Before analysis, sensitive or irrelevant attributes - such as regional information, technical edge data, and unique identifiers - were removed. Sessions with incomplete device information were also excluded to ensure data consistency and completeness.

To ensure the dataset only contained meaningful behavioral signals, all users (both converted and non-converted) who did not generate any session activity were excluded. Since these users provide no interaction data, they bring no value for the algorithms, and the absence of sessions makes it impossible to learn any useful patterns about their behavior. This filtering step reduced noise in the

dataset and improved the quality of subsequent analyses.

3.1 Feature Construction

To prepare the data for machine learning, we aggregated the session-level records into user-level features representing engagement, device usage, browsing diversity, and temporal activity. These behavioral variables capture different aspects of how users interact with the platform during the trial period and serve as predictors of conversion. A detailed description of all constructed features and their assessment criteria is provided in Table 1.

3.2 Data Preparation

Categorical fields representing primary device operating system and device type were encoded into numerical categories using LabelEncoder to facilitate model training. Users without recorded sessions were retained in the dataset, with all behavioral fields initialized to zero, preserving the original trial population distribution.

The resulting dataset consists of one row per user, with aggregated behavior features and a binary label showing if the user converted. This structured data serves as input for the next predictive modeling phase.

4 METHODS AND CLASS IMBALANCE

4.1 Model Selection and Training Setup

The proposed framework for predicting trial-to-paid user conversion is summarized in Figure 1, which outlines the complete data flow, from data preprocessing and feature construction to resampling, model training, and evaluation. Each user is represented by a feature vector containing aggregated behavioral metrics (e.g. unique_channels, total_duration_min, unique_user_agents), paired with a binary label indicating conversion (1 = paid, 0 = not paid).

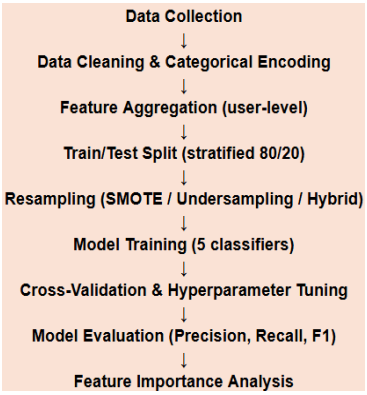

Figure 1: Training and evaluation pipeline for trial-to-paid user conversion prediction.

The task is formulated as a supervised classification problem, where each user is denoted as a feature vector x_i and a corresponding label $y_i \epsilon \{0,1\}$, where $y_i = 1$ represents a converting user. The goal is to learn a classification function $f(x_i)$ that estimates the probability of conversion $P(y_i = 1 \mid x_i)$, where probabilistic estimates are obtained directly or through calibrated model outputs, depending on the algorithm.

Since our main aim is to find users with high potential to convert during the trial period, we selected machine learning models that are both well-known and easy to interpret.

We employed five classification algorithms: Logistic Regression, Support Vector Machine (SVM) with both linear and RBF kernels, Random Forest, k-Nearest Neighbors (kNN), and Extreme Gradient Boosting (XGBoost). Logistic Regression serves as a strong baseline, providing interpretable feature weights and allowing easy understanding of each predictor's contribution to the model [8].

Table 1: Variables used for behavioral assessment.

Variable	Description	Assessment Criterion
total_session_count	Total number of sessions during	Indicate frequent platform usage
total_duration_min	Total minutes watched during	Indicates stronger engagement
avg_duration_min	Average session duration	Measures consistency
max_duration_min	Longest single session	Captures peak engagement
unique_device_type_count	Number of device types used	Indicates cross-device behavior
unique_os_count	Number of OS used	Indicates cross-device behavior
unique_channels	Unique content channels watched	Reflects content exploration
unique_user_agents	Distinct user agents per user	Indicates multi -environment use
active_days	Days between first and last session	Measures sustained activity

SVM is effective for high-dimensional problems, and its RBF kernel captures non-linear relationships between features [9]. Random Forest, an ensemble of decision trees, performs robustly even when features are noisy or correlated [10]. kNN is a simple yet effective algorithm that classifies users based on their similarity to others [11]. XGBoost is a powerful gradient boosting algorithm designed for structured data, offering regularization and efficient handling of missing values, which often improves predictive performance [12].

Hyperparameter optimization was performed for the Random Forest and Gradient Boosting models using grid search combined with 3-fold cross-validation.

4.2 Handling Class Imbalance

One of the main challenges in our modeling task is the imbalance between the two classes. Only about 10% of users convert to paid subscriptions, while 90% do not. This large imbalance can cause classification models to focus too much on predicting the majority class, which reduces their ability to detect converting users, the group that is most important for the business.

To address this issue, we used three different resampling strategies aimed at improving model sensitivity to the minority class: SMOTE (Synthetic Minority Over-sampling Technique) [13], random undersampling [14], and a combined approach involving both methods [15].

SMOTE is a popular oversampling technique that creates synthetic samples of the minority class by interpolating between existing examples. This helps to balance the dataset without removing any data from the majority class and improves the model's ability to learn from minority patterns.

In contrast, random undersampling reduces the number of samples in the majority class to match the size of the minority class. While this method may result in the loss of potentially useful information, it reduces model bias toward the dominant class and often speeds up training.

The undersampling size of 10,000 samples was chosen after preliminary experiments that balanced computational efficiency and class representation. This number ensured that all 5,500 positive (converted) users were retained while reducing the majority class to a comparable scale, thus maintaining a 1:1.8 ratio that improved model recall without excessive information loss.

To further balance the trade-off between sample diversity and information preservation, we also applied a hybrid method that combines SMOTE and random undersampling. This approach enhances minority class representation while retaining a meaningful portion of majority class examples.

By applying and comparing these resampling strategies and focusing on imbalance-aware evaluation, we aimed to improve the model's ability to detect converting users while preserving stable overall performance.

5 EXPERIMENTAL SETUP

In order to evaluate the predictive performance of various classification algorithms for trial-to-paid user conversion, we adopted a holdout validation approach, splitting the dataset into 80% training and 20% testing using stratified sampling to preserve the ratio of converted and non-converted users in both subsets. This way, the evaluation simulates real-world conditions, with the same class imbalance that exists in production.

Each machine learning model was trained on the same training data and tested on the same test set to ensure fair comparison.

To optimize performance, we conducted hyperparameter tuning using 3-fold cross-validation on the training data. For Random Forest, we varied the number of trees (n_estimators), tree depth (max_depth), and minimum samples required to split a node (min_samples_split). For Gradient Boosting, we tuned the number of estimators, maximum depth, and learning rate. The tuning was guided by the F1-score, which is more appropriate for imbalanced data because it balances precision and recall for the minority class (converted users). After tuning, we selected the best settings for each model and evaluated their performance on the test set.

The evaluation metrics - precision, recall, F1-score, and accuracy - follow the standard definitions provided by Powers (2011) [16].

In this study, these metrics were used to assess both overall model correctness and the ability to identify likely paying users (the minority class). Because the dataset is highly imbalanced, recall and F1-score were given higher priority than accuracy. Recall indicates how many actual converters were correctly identified, while precision reflects the reliability of these predictions. The F1-score, as the harmonic mean of precision and recall, provided a single measure balancing both aspects. Accuracy was used only as a general indicator of stability across models.

6 EXPERIMENTAL RESULTS

The experimental results for the baseline classification models are summarized in Tables 2-4. All models were trained using three sampling strategies to address class imbalance: SMOTE, undersampling on 10,000 samples, and a combined SMOTE + undersampling approach, and were evaluated on the original test set. This decision avoids data leakage and gives a more realistic view of how the models would perform on new, unseen users in real deployment scenarios.

Focusing first on the undersampling approach, Table 2 presents the baseline performance of all five models. Among the simpler models, Logistic Regression achieved an accuracy of 0.9596 but showed moderate F1-score (0.78) and recall (0.77) for the minority class, highlighting its limitations in capturing complex, non-linear user behavior. k-Nearest Neighbors provided slightly better recall for the minority class (0.81) but lower precision (0.76), reflecting a trade-off between detecting positive converters and avoiding false positives.

Table 2: Baseline model performance with undersampling.

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.9596	0.78	0.77	0.78
Decision Tree	0.9699	0.83	0.84	0.84
k-Nearest Neighbors	0.9592	0.76	0.81	0.78
Random Forest	0.9775	0.90	0.85	0.87
Gradient Boosting	0.9795	0.92	0.85	0.88

Tree-based models performed better than simpler approaches. The Decision Tree classifier reached an F1-score of 0.84 for the positive class, showing balanced precision and recall. Random Forest and Gradient Boosting achieved the best results, with Gradient Boosting reaching the highest accuracy (0.9795) and F1-score (0.88), while Random Forest delivered a strong F1-score (0.87) and high precision (0.90), highlighting its ability to identify likely converters while limiting false positives.

Building on these results, we further analyzed Random Forest and Gradient Boosting. Random Forest consistently delivered the strongest overall performance, as shown in Table 3. Among the sampling strategies, undersampling provided the highest F1-score for the minority class (0.87), while recall was slightly higher for the combined SMOTE + undersampling approach (0.86). SMOTE alone achieved slightly lower performance across all metrics.

Table 3: Random Forest performance across sampling methods.

Sampling Method	Accuracy	Precision	Recall	F1-score
SMOTE + Undersampling	0.9747	0.86	0.86	0.86
Undersampling	0.9775	0.90	0.85	0.87
SMOTE	0.9759	0.89	0.84	0.86

Gradient Boosting achieved the highest F1-score for class 1 with undersampling (0.88), though recall was slightly lower (0.85) than with SMOTE or the combined method (0.86) (Table 4). This clearly shows that undersampling provides the best overall balance between precision and F1-score for accurately identifying likely paying converters.

Table 4: Gradient Boosting performance across sampling methods.

Sampling Method	F1-score	Recall
SMOTE + Undersampling	0.87	0.86
Undersampling	0.88	0.85
SMOTE	0.87	0.86

The confusion matrices for Random Forest and Gradient Boosting provide further insight into performance with the undersampling approach. Random Forest correctly classified 939 of 1,108 positive cases, with 169 misclassified as negative and 104 negatives as positive (Figure 2). Gradient Boosting correctly identified 943 positives, misclassifying 165 as negative and 83 as positive (Figure 3). These results show that both ensemble methods achieve high accuracy for the minority class while keeping false positives low, confirming their suitability for conversion prediction.

Overall, these results demonstrate that ensemble methods, particularly Random Forest and Gradient Boosting, are the most effective models for trial-to-paid conversion prediction, and that undersampling on 10,000 samples consistently provides slightly better performance than SMOTE or the combined approach, making it the preferred strategy for this task.

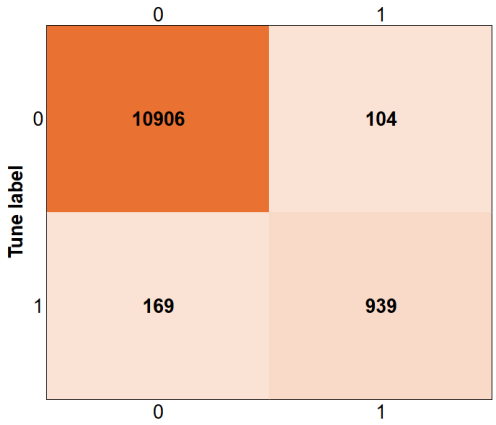

Figure 2: Confusion matrix for Random Forest using undersampling.

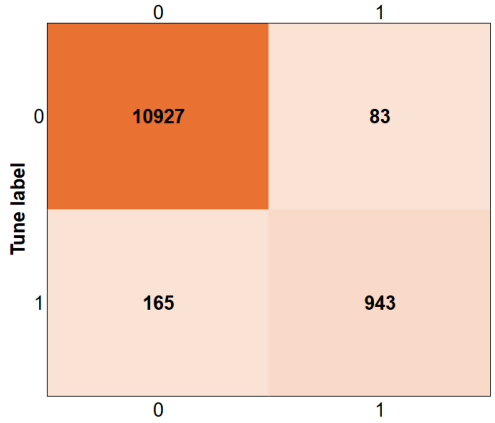

Figure 3: Confusion matrix for Gradient Boosting using undersampling.

7 FEATURE IMPORTANCE ANALYSIS

Understanding which user behaviors influence trial-to-paid conversion is critical for improving prediction and guiding marketing strategies. To investigate this, we analyzed feature importance using the Random Forest classifier on the undersampled dataset. The results are shown in Figure 4, which presents the relative importance of the top features.

The Random Forest model distributed importance across multiple features, with max_duration_min (0.204) and unique_user_agents (0.192) emerging as the strongest predictors of conversion. These features capture both the intensity and diversity of user engagement: users who spend longer maximum time on content and access the service from multiple user agents are more likely to convert. Total_duration_min (0.169) and unique_device_type_count (0.103) also contributed significantly, emphasizing that both overall engagement and device diversity play an important role. Features such as avg_duration_min (0.095) and unique_channels (0.088) had lower, yet meaningful contributions, suggesting that while average session duration and content variety matter, they are secondary to maximum engagement and user agent diversity.

Figure 4 clearly shows that engagement intensity and device diversity are the most influential factors in predicting trial-to-paid conversion, which aligns with expectations: users who interact more deeply and flexibly with the platform during the trial period are more likely to convert. This insight can help inform targeted marketing efforts, such as recommending content to highly engaged users or providing incentives for multi-device usage.

8 CONCLUSIONS

This study explored the prediction of trial-to-paid user conversion in video streaming platforms using behavioral session data and machine learning. Five classification models were evaluated - Logistic Regression, Decision Tree, k-Nearest Neighbors, Random Forest, and Gradient Boosting - under three imbalance-handling strategies: SMOTE, random undersampling, and a hybrid combination of both.

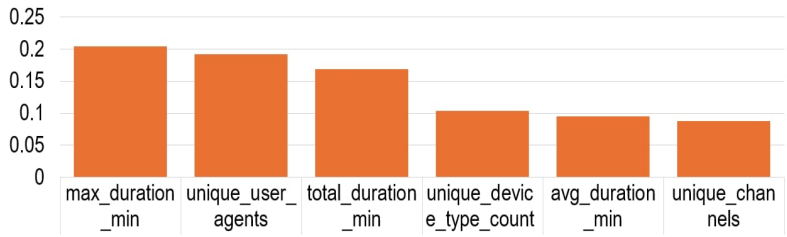

Figure 4 Feature importance for predicting trial-to-paid conversion derived from the Random Forest classifier (undersampling).

The experiments demonstrated that ensemble models, particularly Random Forest and Gradient Boosting, deliver the most consistent and accurate results for identifying likely subscribers. Among the tested strategies, undersampling on 10,000 samples provided the best trade-off between precision and recall, confirming its suitability for highly imbalanced behavioral data.

Feature importance analysis revealed that engagement intensity and device diversity are the strongest behavioral indicators of conversion. Specifically, maximum session duration and the number of unique user agents emerged as dominant predictors, followed by total viewing time and content diversity. These findings suggest that users who engage deeply and interact with the platform across multiple devices show a higher intention to subscribe.

From a scientific perspective, this work contributes by framing conversion prediction as a quantifiable behavioral modeling problem rather than a purely marketing task. It establishes an interpretable methodology linking user engagement metrics with conversion probability and extends existing research on imbalanced learning by comparing the impact of different resampling strategies on real-world behavioral data.

While the results are promising, this study has certain limitations, such as relying on data from a single platform and focusing on aggregated behavioral features rather than real-time session dynamics. Future research could address these aspects to further improve generalizability and predictive power.

REFERENCES

[1] A. Rajesh, V. G. Menon, S. Joseph, A. Paul, S. H. Ahmed, and Y. Zhang, "User behavior analysis in video streaming using ensemble learning," IEEE Transactions on Consumer Electronics, vol. 68, no. 3, pp. 233–241, Aug. 2022, doi: 10.1109/TCE.2022.3181340.

[2] S. Wang, B. Fu, K. Yang, and S. Li, "The impact of free trial duration on subscription conversion in online video platforms: Evidence from a randomized experiment," Electronic Commerce Research and Applications, vol. 47, p. 101057, Jan. 2021, doi: 10.1016/j.elerap.2020.101057.

[3] Y. Wang, T. Zhuang, R. Zhang, X. Lin, and J. Han, "Will this online shopping session succeed? A dynamic conversion prediction framework," in Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM), Atlanta, GA, USA, 2022, pp. 4230–4234, doi: 10.1145/3511808.3557575.

[4] J. Zhang, H. Chen, Y. Li, and W. Xu, "Loyalty and churn prediction on mobile social network platforms using machine learning," Scientific Reports, vol. 14, no. 1, p. 40704, 2024, doi: 10.1038/s41598-024-40704-7.

[5] A. Dedieu, R. Mazumder, Z. Zhu, and H. Vahabi, "Hierarchical modeling and shrinkage for user session length prediction in media streaming," arXiv preprint arXiv:1803.01440, 2018, doi: 10.48550/arXiv.1803.01440.

[6] W. Chen, J. Zhu, and L. Qu, "A survey on imbalanced learning," Artificial Intelligence Review, vol. 57, no. 4, pp. 1–34, 2024, doi: 10.1007/s10462-024-10931-9.

[7] T. Wongvorachan and K. A. Bollen, "A comparison of undersampling, oversampling, and hybrid methods for imbalanced data classification," Information, vol. 14, no. 1, p. 54, Jan. 2023, doi: 10.3390/info14010054.

[8] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed. Hoboken, NJ: Wiley, 2013.

[9] J. Cervantes, F. García-Lamont, L. Rodríguez-Mazahua, and A. López, "A comprehensive survey on support vector machine classification: Applications and challenges," Neurocomputing, vol. 408, pp. 189–215, 2020, doi: 10.1016/j.neucom.2019.10.118.

[10] P. Probst, M. N. Wright, and A.-L. Boulesteix, "Hyperparameters and tuning strategies for random forest," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 9, no. 3, e1301, 2019, doi: 10.1002/widm.1301.

[11] S. Syriopoulos and A. Kalampalikis, "kNN classification: A review," Annals of Mathematics and Artificial Intelligence, 2024, doi: 10.1007/s10472-023-09882-x.

[12] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), San Francisco, CA, USA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[13] D. Elreedy and A. F. Atiya, "A theoretical distribution analysis of the synthetic minority oversampling technique (SMOTE)," Machine Learning, vol. 113, pp. 1613–1639, 2024, doi: 10.1007/s10994-022-06296-4.

[14] T. Wongvorachan, S. He, and O. Bulut, "A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining," Information, vol. 14, no. 1, p. 54, 2023, doi: 10.3390/info14010054.

[15] Y. Lu, Y.-M. Cheung, and Y. Y. Tang, "Hybrid sampling with bagging for class imbalance learning," in Proceedings of the 20th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2016), Taipei, Taiwan, 2016, pp. 14–26, doi: 10.1007/978-3-319-31753-3_2.

[16] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," J. Mach. Learn. Technol., vol. 2, pp. 37–63, 2011.