scientific reports

OPEN

Advanced customer churn prediction for a music streaming digital marketing service using attention graph-based deep learning approach

Haiyan Cheng✉ & Jie He

Customer churn poses a persistent threat to the sustainability of music streaming platforms, where user disengagement often occurs unpredictably. Most existing churn prediction methods fail to integrate relational dependencies among users or address the limitations of static, tabular models in imbalanced datasets. In this research, we propose a Hybrid Graph Attention Network (Hybrid GAT + MLP (Multi Layered Perceptron)) that efficiently combines graph-based learning with deep tabular feature modelling, allowing the capture of both individual behaviour and peer-influenced churn patterns. The use of a synthetic similarity graph enables the model to learn from neighbourhood context, while a weighted loss function ensures robustness against class imbalance without altering data distribution. The model was evaluated on the KKBox dataset, achieving an accuracy of 95.8% with an AUC score of 0.9626, indicating high discriminative power. Additional metrics, such as F1-score, confusion matrix, and ROC analysis, ensure the system's ability to generalize across both churn and non-churn classes. These results validate the applicability of our approach to large-scale, real-world scenarios involving churn detection. All code and the curated dataset for customer churn prediction for music steaming is available at the https://github.com/haiyanchenas/Advanced-Customer-Churn-Prediction-for-a-Music-Streaming-Digital-Marketing-Service/tree/main enabling reproducibility of our findings.

The global music streaming industry has experienced exponential growth over the past decade, driven by advancements in mobile connectivity, affordable subscription models, and the proliferation of personalised content recommendations. Recent industry analyses indicate that platforms like Spotify, Apple Music, and Amazon Music have achieved an estimated 25–30% compound annual growth rate (CAGR) in user adoption worldwide[1]. However, this expansion comes with a pressing challenge in customer churn. Industry reports indicate that up to 30–35% of new users abandon a streaming service within the first six months of subscription, primarily because of unsatisfactory personalization, limited engagement, or pricing concerns. This churn not only minimises the active user base but also significantly influences long-term revenue projections. Since acquiring a new customer is estimated to cost five to seven times more than retaining an existing one, churn prediction has emerged as a mission-critical function in the operational strategy of subscription-based platforms[2].

To address this, several machine learning models have been examined. Gaddam and Kadali[1] evaluated traditional classifiers such as Logistic Regression and Decision Trees for predicting churn based on static user profiles. Karwa et al.[2] proposed an analytics pipeline linking churn detection to retention strategies using ensemble models. Joy et al.[3] introduced a hybrid framework combining big data pipelines with explainable AI for enhanced interpretability. While these approaches have demonstrated moderate predictive success, they are restricted by their reliance on tabular data and failure to model relational dependencies among users, which may obscure complex behavioural dynamics.

Master School of Economics and Management Department, Gongqing College, Nanchang University, JiujiangGongqing City, 332020, China. ✉email: hiayanchengreas@hotmail.com

This research aids in developing a graph-based deep learning model that influences both individual user features and community-based behavioural similarities to detect churn more efficiently. The scope comprises scalable processing of multi-source user data, graph construction based on shared user traits, and the incorporation of these representations using Graph Attention Networks (GAT) integrated with deep feature learning methods. The focus is not only to enhance predictive accuracy but also to maintain robustness in class-imbalanced scenarios commonly found in subscription platforms.

Streaming services, such as KKBox and Spotify, often encounter user behavior influenced by social or demographic context, which conventional models struggle to capture. By representing users as nodes in a graph and connecting them via shared characteristics such as city and registration method, it becomes possible to model latent peer influences. This graph-based formulation enables attention mechanisms to weight the importance of neighbouring behaviours, refining churn predictions beyond what isolated features can provide.

Our contribution
The main contribution of the Research is:

- GAT-based relational learning with deep Multi-Layer Perceptrons to model both graph-structured and tabular data was developed for digital marketing trend analysis on an online subscription and view basis.
- Constructed a synthetic user similarity graph that links users through shared traits to allow contextual learning through attention propagation for analyzing product and user interest in it.
- Implemented a scalable big-data processing pipeline capable of aggregating behavioural features from 30GB + of user logs and integrating them into a graph-ready format.
- We adopted a novel training strategy using weighted loss, which mitigates class imbalance (~ 9% churn rate) without data resampling, preserving distributional integrity.

The remaining sections in this article are arranged in the following sequence: Literature work is presented in Sect. 2. The proposed model is extensively discussed in Sect. 3 with the architecture, methodology description, and algorithms. Results are evaluated in Sect. 4 of the article, including comparisons and proofs. The research work is concluded in Sect. 5 of the article.

Literature survey
Adaptive Profit-Centric Churn Prediction Engine built on ensemble methods (XGBoost, Random Forest, Gradient Boosting)[4], tightly integrated with economic measures such as profit and Customer Lifetime Value. It's proposed to not only boost churn classification (with an accuracy of ~ 97%) but also directly optimize financial outcomes. The model captures customer interdependence by quantifying social linkages and co-usage patterns. It uses R^2 convergence curves to track stability and adaptability over time. However, it's heavily limited by the need for high-quality economic data—uneven or noisy CLV estimates can undercut performance. Additionally, its complexity and dependence on customer network data may hinder use in privacy-sensitive or connectivity-limited environments. Systematic bibliometric study using Scopus data and VOSviewer tools[5] to map AI's application in customer retention. It proposes the identification of five key research themes—churn prediction, service experience, sentiment analysis, big-data analytics, and privacy—through term co-occurrence and bibliographic coupling. The proposed framework articulates 15 future research objectives, emphasizing novel directions such as ethics, privacy, and advanced AI techniques. It delivers insights on publication trends, authorship networks, influential journals, and geographic hotspots. However, the proposed model lacks evaluation of actual AI model performance or real-world deployment challenges. Focused solely on Scopus, it is limited by dataset bias and may not reflect the whole global landscape.

The SV-VAE model combines a Convolutional Neural Network with a modified Variational Autoencoder (VAE)[6] to address high-dimensional and imbalanced churn datasets. It's proposed that VAE compresses and denoises data before CNN classification, enhancing both accuracy and training speed. Evaluated on six benchmark datasets, it reports strong accuracy, precision, recall, F1-score, and reduced response time. The hybrid handles complex temporal patterns via latent feature capture and CNN feature extraction. However, the model requires significant computational resources, and tuning both the VAE and CNN adds additional overhead. Additionally, its complexity raises concerns about overfitting, particularly with limited training samples. Deep neural network[7] tailored for telecom churn prediction, using multi-layer perceptron architectures with embedding layers for categorical variables. It's proposed to detect complex nonlinear patterns in customer behavior with high predictive accuracy. The network optimizes class imbalance via weighted loss functions, but lacks explicit mention of SMOTE or other resampling methods. Evaluation on real-world telecom data shows improved recall but limited precision trade-offs. However, the proposed model is constrained by its opaque inner workings, which limit its explainability for business stakeholders. Additionally, it faces issues of overfitting in small datasets without regularisation.

Random Forest with deep learning (MLP/CNN)[8], integrated into a single pipeline for telecom churn prediction. It's proposed to extract engineered features first, then apply deep nets for final classification. The hybrid approach reportedly enhances accuracy and offers modular adaptability across telecom datasets. However, the need to manage two distinct pipelines increases the complexity of implementation and training time. It also relies heavily on feature engineering quality, limiting generalization to new contexts. Finally, the dual-stage process may suffer from error propagation between the feature and deep-learning stages. BiLSTM-CNN hybrid[9] to capture sequential and spatial churn signals from telecom usage logs. It's proposed that LSTM layers extract temporal trends while CNN layers learn hierarchical patterns. Tested on benchmark telco datasets, the composite architecture reaches ~ 81% accuracy. It balances context capture and feature abstraction more effectively than standalone architectures. However, the proposed solution requires extensive hyperparameter

tuning, and training can be computationally intensive. It may also underperform when the sequence length is irregular or the dataset lacks timestamped behaviour.

Clustering (K-means, hierarchical) with classification (Gradient Boosting, SVM, RF)[10] is proposed to first group similar customers, then apply tailored classifiers, thus improving adaptability and prediction accuracy. The approach better handles heterogeneity and high dimensionality in telecom datasets. However, selecting the optimal cluster count is tricky and can bias classifiers. It also increases pipeline complexity and training time. Furthermore, clustering errors may propagate, affecting the accuracy of the classification stage. Archimedes Optimisation Algorithm (AOA)[11] for feature selection, followed by CNN-Autoencoder deep models for churn prediction. It's proposed to optimally select informative features from high-dimensional telecom data and then predict churn via a CNN-AE hybrid. The optimisation pre-processing reduces dimensionality and improves accuracy, while AE ensures the robustness of latent representation. However, AOA feature selection is computationally costly and may trap in local optima. CNN-AE depth increases training time and complexity. The combined framework may be overkill for simpler datasets and lacks clear benchmarks against lighter models.

Decision Tree, Random Forest, and XGBoost[12] were integrated via hyperparameter tuning (grid search) for feature selection in telecom churn prediction. It's proposed to enhance interpretability while maintaining ensemble strength in handling imbalances and high dimensionality. The optimized models enable proactive retention strategies with scalable training. Nonetheless, the proposal still depends on clean, feature-engineered inputs—it's not an end-to-end solution. The interpretability gains from DT may be lost in tuned ensembles. Additionally, focusing on tree models may miss potential improvements from deep learning architectures. Random Forest classification[13] proposed to incorporate textual sentiment scores as churn predictors, capturing emotional drivers of attrition. This enables churn detection beyond transactional data, using qualitative customer signals. However, reliance on sentiment accuracy makes it vulnerable to sarcasm, multilingual context, or misclassifications. It may perform poorly when text data is limited or unstructured (Survey of existing work provided in Table 1).

In the context of mobile shopping, social crowding can induce anxiety and cause distractions, and consequently, diminishes the ability of shoppers to easily process the information on their mobile devices[24,25]. The challenges posed by high cognitive load during information processing are mitigated by learning compact, task-relevant prompts that retain all essential components of a task. The model generalizes better across varied situations by removing irrelevant information and noise specific to a domain[26,27]. In the case of panel data analysis, the estimation of latent group structures is done to identify and understand hidden subgroup variances which allows better and more relaxed modeling of complex and heterogeneous data[28,29]. In the same fashion, the prediction-aware network in sentiment analysis around which a model is built you constantly modifies itself based on in-between outcomes thereby increasing the precision of the interlinks of aspects, opinions and sentiments within the text[30].

Furthermore, Random Forest may not optimally handle nuanced sentiment patterns compared to deep NLP models.

Proposed methodology

The overall workflow begins with data collection from the WSDM-KKBox churn dataset, which provides user logs, transactions, and demographics. Next, feature engineering is performed to derive static, temporal, and behavioural features from multi-source inputs[31]. The dataset then undergoes cleaning and standardisation to handle missing values and scale features for deep learning[32]. A Hybrid GAT + MLP model is trained to learn from both user similarities (through a graph) and individual tabular patterns[38,39]. Class imbalance is addressed

| Author	Dataset	Model	Advantage	Limitation
Geiler et al.[14]	16 public churn-like datasets	Eleven supervised & semi-supervised ML (incl. ensemble) + seven sampling strategies	Offers benchmark across diverse datasets; identifies top-performing pipelines; and provides a practical ensemble recommendation	Lacks new model; dependent on dataset choice; only offline predictive insights
Manzoor et al.[15]	Multiple industry datasets	Survey of ML models for churn + best-practice recommendations	Offers practitioner-centric guidelines; covers end-to-end deployment considerations	Review-level insight; no empirical evaluation
Mouli et al.[16]	Cogent engineering dataset	Comparative classification models (e.g. LR, RF, SVM)	Clarifies the performance of classic classifiers under varied conditions	Limited to a single dataset; lack of advanced models
Colasanti et al.[17]	Slow-fashion e-commerce data	Static-data classifiers	Provides a domain-specific view for retail subscription churn	Ignores time-dependent behaviour; less generalizable
Nagaraju and Vijaya[18]	CRM datasets	Review of ML methods	Offers a taxonomy of ML tools for CRM churn detection	Lacks empirical validation
Banday and Khan[19]	Big-data corpuses	Evaluation of churn models (incl. BI integration)	Helps BI implementers select suitable models	Limited to BI context; potential data complexity issues
Zhang et al.[20]	E-commerce datasets	Broad survey of ML & DL techniques	Comprehensive coverage of methods; highlights e-commerce best practices	High-level; no deep performance comparison
Razzaq and Shah[21]	Mixed datasets	ML & DL paradigms survey	Maps emerging research frontiers	Lacks specific churn focus
Mishra et al.[22]	Business analytics case studies	ML + intelligent optimization techniques	Proposes future directions integrating optimization	Conceptual; no experimental results
Cenggoro et al.[23]	Procedia-CS telecom/customer logs	Deep embedding model (DL vector embeddings)	Learns deep representations; flexible feature extraction	Computationally heavy; may need large data

Table 1. Survey of existing work.
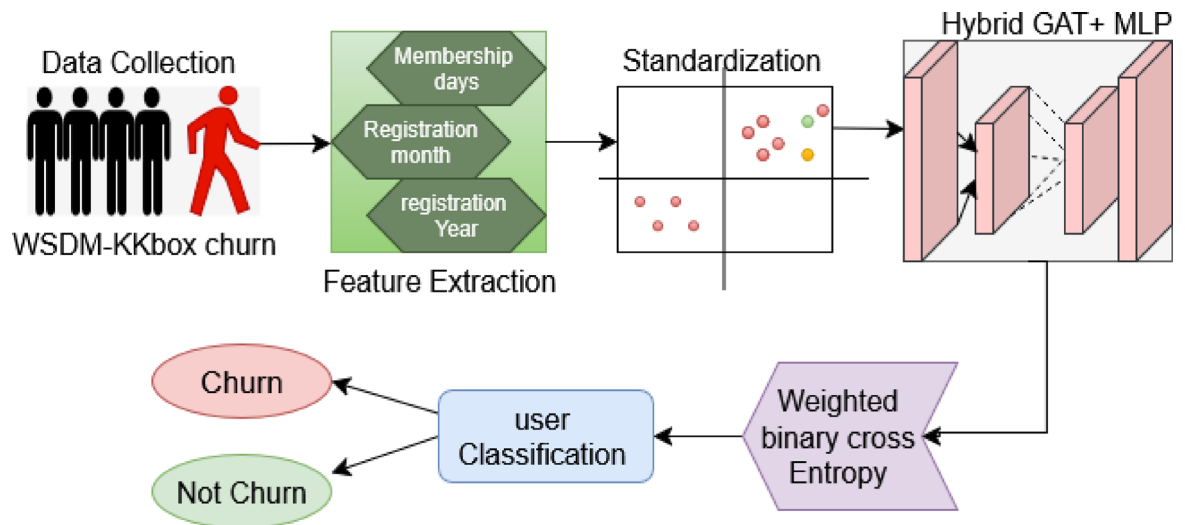
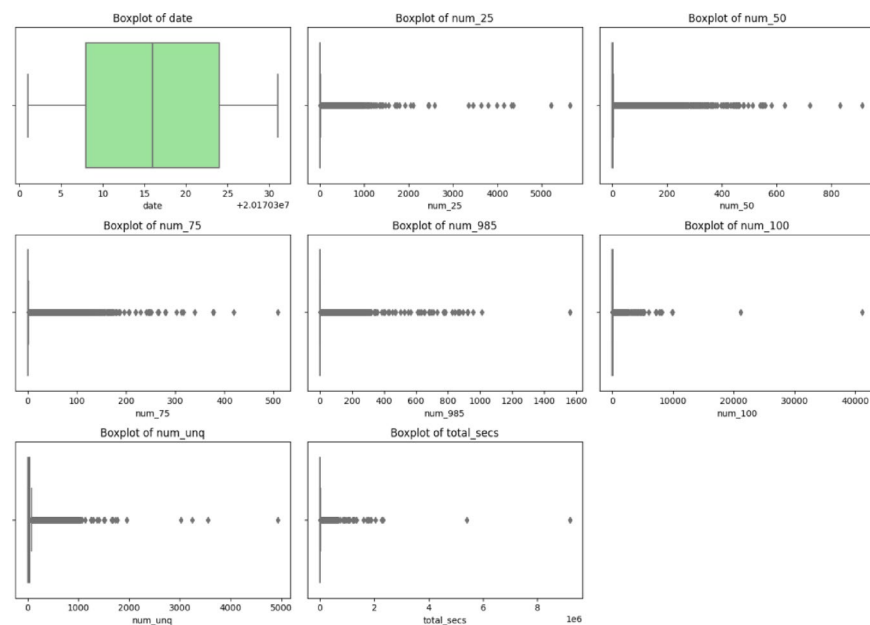**Fig. 1**. Proposed hybrid churn detection workflow.

Fig. 2. Log CSV statistical data distribution.

using a weighted binary cross-entropy loss, which highlights the minority churn class. Finally, the model performs churn classification and evaluation using metrics like AUC and accuracy. Figure 1 shows the overall architecture of the proposed framework.

Dataset description

The dataset used in this study is sourced from the WSDM-KKBox Churn Prediction Challenge and provides a large-scale, multi-relational view of user behavior in a music streaming platform. The $transactions\_v2.csv$ which captures detailed subscription and billing history including transaction dates, payment plans, auto-renewal, and cancellation status as shown distribution in Fig. 2 and $user\_logs\_v2.csv$ a time-series file exceeding 30GB that logs daily listening activity per user for three years, including total seconds played as shown distribution in Fig. 2 and number of unique tracks. The $members\_v3.csv$ which contains user demographics such as city, gender, registration method, and registration timestamp as shown gender distribution in Fig. 3 among city. It consists of four core CSV tables such as $train\_v2.csv$ which includes the user identifier (msno) and the binary target label (is_churn) as shown in Fig. 4 pie chart. Together, these files provide both static and temporal insights into user engagement. The final merged dataset comprises 970,960 users, resulting in a rich feature set through joins across all tables.
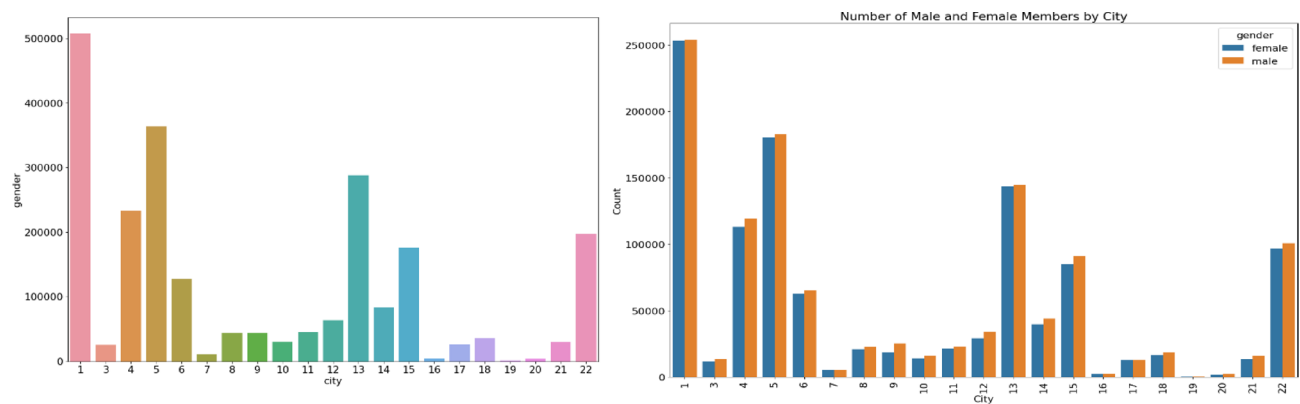
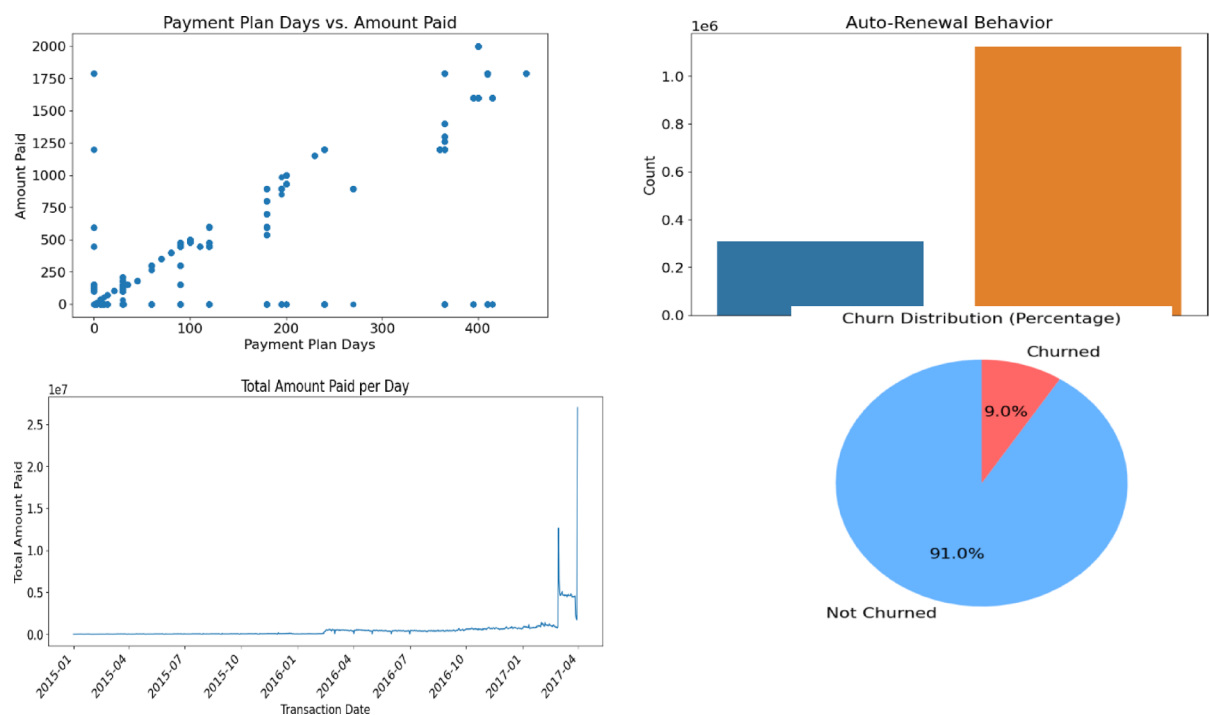**Fig. 3**. Member csv statistical data distribution.

Fig. 4. Transaction csv statistical data distribution and pie chart –train CSV data analysis.

Data preparation

To prepare the data for churn prediction, an extensive preprocessing pipeline was implemented to aggregate and merge multiple relational tables into a unified user-level dataset. First, transaction data from $transactions\_v2.csv$ was aggregated per user ($msno$) to extract key features such as the total number of transactions ($total\_transactions$), cumulative payment amount ($total\_payment$), number of cancellations (is_cancel_sum), and the most recent transaction date ($last\_transaction\_date$). Next, due to the large size of $user\_logs\_v2.csv$ (over 30GB), chunk-wise processing was implemented, reading the file in segments of one million rows to manage memory usage efficiently. Behavioral features were computed for each user, comprising the total number of active days (log_days), total listening time in seconds ($total\_secs\_sum$), and the number of unique songs played ($total\_songs\_played$). These features provided a condensed yet informative summary of user engagement across time. Finally, the aggregated transaction and log data were integrated with demographic attributes from $members\_v3.csv$ and the churn labels from $train\_v2.csv$, resulting in a comprehensive feature matrix representing 970,960 users. This merged dataset integrated static, transactional, and dynamic behavioral data, forming the foundation for model training.

Feature engineering

To improve model performance and extract meaningful patterns from raw data, a series of feature engineering techniques were performed to derive temporal and categorical attributes. Time-based features were

constructed by leveraging both registration and transaction timestamps. Specifically, the membership duration ($membership\_days$) was estimated as the difference between the user's most recent transaction date and their initial registration date as mentioned in Eq. (1).

$$membership\_days_i = last\_transaction\_date_i - registration\_init\_time_i \qquad (1)$$

where i refers to the user index, and dates are transformed into standard datetime formats. Additionally, the year and month of registration were extracted from $registration\_init\_time$ to generate two discrete features such as $registration\_year$ and $registration\_month$. This help capture seasonal or yearly patterns in user onboarding behaviour. For handling categorical variables with low cardinality like city, gender, and $registered\_via$ label encoding was implemented. Every unique category was mapped to a corresponding integer, preserving ordinal relationships where appropriate. Let C be a categorical variable with n unique categories. Label encoding is defined using Eq. (2).

$$Encoded\,(C_i) = j \ where\,C_i = category_j, \quad j \in \{0,1,\dots,n-1\} \qquad (2)$$

This transformation allows categorical inputs to be used effectively within neural network models without inflating dimensionality, as would occur with one-hot encoding. Together, these engineered features enriched the dataset with temporal and structural insights necessary for churn prediction. The Fig. 2 illustrates the distributions of four critical engineered features extracted from the user activity, transaction, and churn datasets used in the model. Each subplot reveals insights into user behaviour patterns relevant to churn prediction. This Fig. 5 (a) shows the distribution of the $membership\_days$ feature, which represents the total duration (in days) from user registration to the last transaction. The distribution is right-skewed, showing many users with relatively short platform tenure, while only a minority have long-standing membership histories. Several spikes suggest user acquisition in bursts, possibly driven by promotional campaigns or product updates. Figure 5 (b) indicates the log-transformed distribution of $total\_secs\_sum$, which captures the total listening time per user. The log transformation was performed to handle extreme outliers and long-tail behavior. The distribution resembles a normal curve post-transformation, which is ideal for input into deep learning models, as it stabilizes training and prevents high-variance features from dominating. The Fig. 5 (c) depicts the frequency of is_cancel_sum, reflecting the number of times a user has canceled their subscription. Most users have never canceled (value 0), while a small portion have canceled once. Values beyond 4 were capped to minimize noise and sparsity in higher bins. This categorical-type feature supports to identify users prone to churn recurrence.
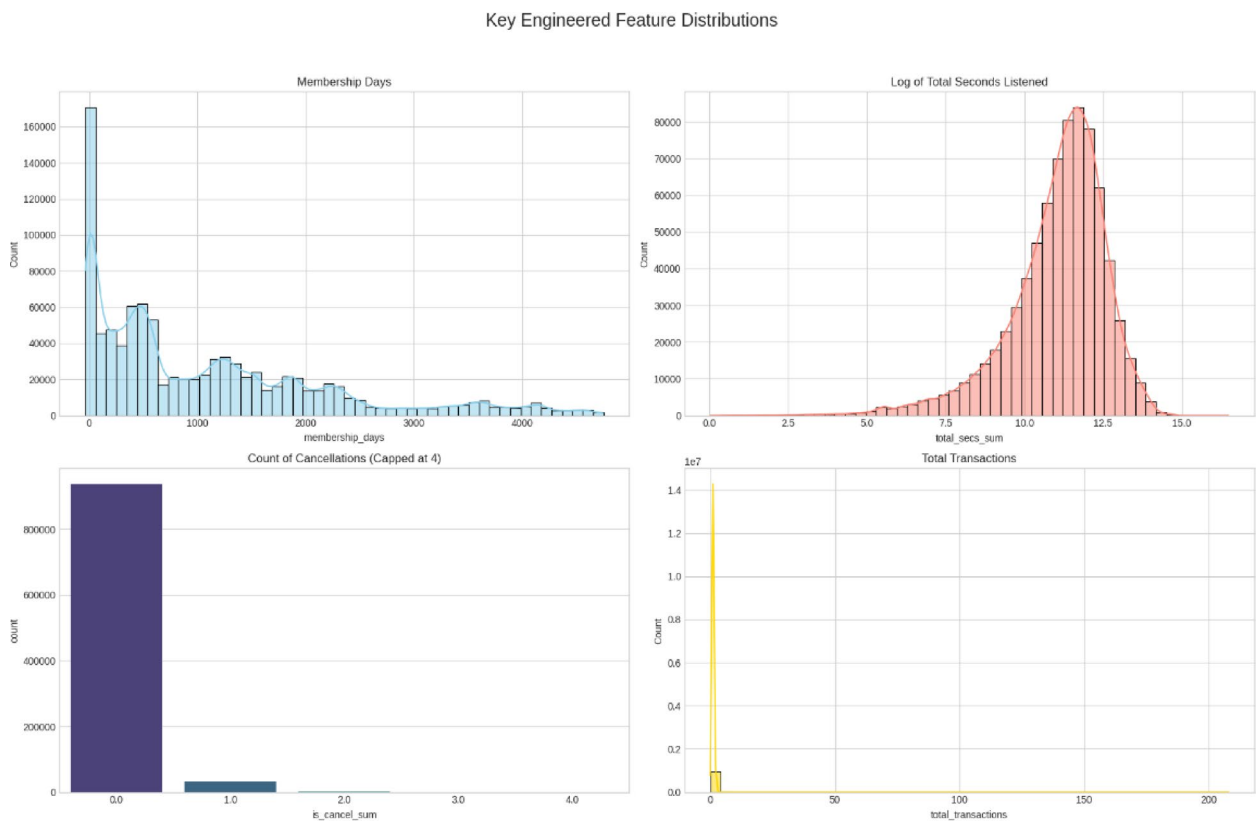

Fig. 5. Distribution plots of key engineered features used for churn prediction: (**a**) membership duration in days, (**b**) log-transformed total listening time, (**c**) capped subscription cancellation count, and (**d**) total subscription transactions.

The Fig. 5(d) histogram illustrates the distribution of $total\_transactions$, showing the total number of subscription-related transactions made by each user. The extreme left-skew describes that most users have made very few transactions, suggesting short-term trial usage. The presence of outliers on the far right (users with >100 transactions) may show long-term or recurring subscribers, valuable for churn resistance modelling.

Data standardization

To assure the integrity and compatibility of the dataset for deep learning models, systematic data cleaning and feature scaling procedures were performed for extracted features. Missing values (NaN) in the merged dataset, primarily raising from users with incomplete transaction or log histories, were handled by imputation. All such missing entries were replaced with zero, under the assumption that the absence of activity shows no engagement or payment behavior during the observation period. This approach-maintained dataset completeness without introducing bias through arbitrary statistical estimates. Subsequently, all numerical features were standardized by Z-score normalization through the $StandardScaler$ utility from the Scikit-learn library. This method transforms each feature x according to the formula given by Eq. (3).

$$x_{scaled} = \frac{x - \mu}{\sigma} \tag{3}$$

where μ refers to the mean and σ is the standard deviation of the feature over all training instances. This transformation ensures that all features have zero mean and unit variance, which is crucial for stabilizing the learning dynamics in deep models particularly those involving gradient-based optimization such as Graph Attention Networks (GAT). Proper scaling prevents certain features from dominating due to scale disparity and improves the convergence behavior of the model.

As visualized in Fig. 6, the application of $StandardScaler$ significantly alters the distribution of the $membership\_days$ feature. The original values, shown in blue, span a wide range up to approximately 5000 days, directing to severe skewness and variance. After scaling (shown in red), the feature is tightly concentrated around zero with a normalized spread, indicating successful mean-centering and variance reduction. This transformation eliminates the impact of magnitude differences and prepares the feature for balanced learning across all input dimensions.

Proposed GAT + MLP hybrid model

To efficiently model both the individual behavioural attributes of users and their latent group-based interactions, we present a hybrid GAT + MLP model. This architecture integrates the structural learning capabilities of Graph Attention Networks (GAT) with the expressive power of a Multi-Layer Perceptron (MLP), allowing the system to learn from both node-level features and their graph-based relationships. The model is designed to mitigate class imbalance, sparse graph connectivity, and nonlinear feature interactions commonly encountered in real-world churn datasets. We hypothesise that users who share similar demographics or onboarding behaviour exhibit correlated churn tendencies. To capture these relationships, a synthetic user similarity graph $G = (V, E)$ is constructed: Nodes ($v_i \in V$) represent individual users and Edges ($(v_i, v_j) \in E$) as mentioned in Eq. (4).

$$(v_i, v_j) \in E \leftrightarrow city_i = city_j \wedge registered_{via_i} = registered\_via_j \tag{4}$$

This results in localised clusters of similar users, allowing message passing across user neighbourhoods and enabling the model to exploit contextual patterns in churn behaviour. The proposed GAT + MLP model follows a dual-path architecture that learns graph-based embeddings in parallel with deep tabular feature representations
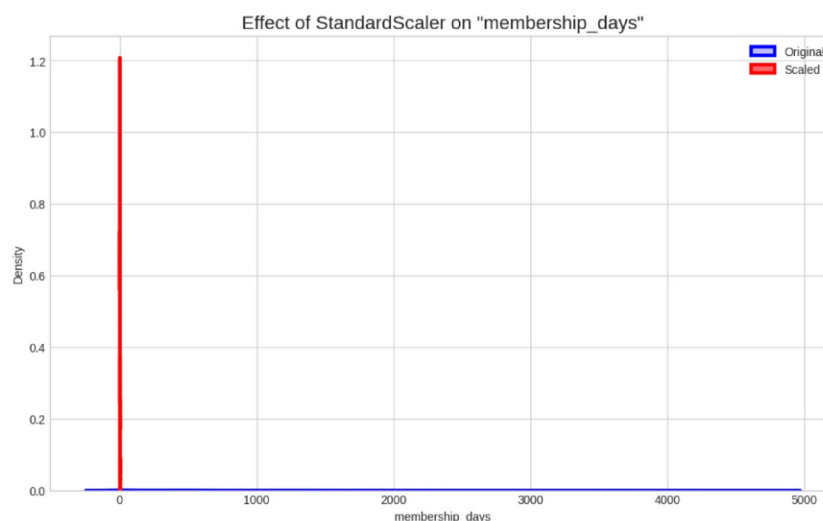

Fig. 6. Impact of Z-score standardisation on the $membership\_days$ feature.

and then fuses them for final classification. This design enables the model to extract both topological and semantic signals that influence churn behaviour. Each user node i with feature vector $h_i \in R^d$ aggregates information from its neighbours $j \in N(i)$ using attention-weighted message passing. The attention coefficient between node i and neighbour j is defined using Eqs. (5) and (6).

$$e_{ij} = LeakyRelu\left(a^{\mathrm{T}}\left[Wh_i \parallel Wh_j\right]\right) \tag{5}$$

$$\alpha_{ij} = \frac{exp\left(e_{ij}\right)}{\sum_{k \in N(i)} exp\left(e_{ik}\right)} \tag{6}$$

The updated embedding of node i is then computed using the weighted aggregation of its neighbours given by Eq. (7).

$$z_i = \sigma\left(\sum_{j \in N(i)} \alpha_{ij} Wh_j\right) \tag{7}$$

To enhance expressiveness, we employ multi-head attention with K = 4 heads as described in Eq. (8).

$$z_i = \sum_{k=1}^{K} \sigma\left(\sum_{j \in N(i)} \alpha_{ij}^{(k)} W^{(k)} h_j\right) \tag{8}$$

Dropout with a rate of 0.6 is applied between GAT layers to prevent overfitting on sparse graph neighbourhoods. In parallel, each user's 15-dimensional tabular feature vector x_i is processed through a deep MLP to capture nonlinear interactions among features that may not be evident in the graph topology. The MLP operations are as follows by Eq. (9).

$$t_1 = ReLU\left(BN_1\left(W_1 x_i + b_1\right)\right) \tag{9}$$

$$t_2 = Dropout\left(ReLU\left(BN_2\left(W_2 t_1 + b_2\right)\right)\right) \tag{10}$$

This captures nonlinear interactions among original user-level features that are not captured through the graph. The final user representation is formed by concatenating the GAT embedding $z_i \in R^{32}$ with the MLP-transformed tabular vector $t_2 \in R^{15}$ as given in Eq. (11).

$$u_i = [z_i \parallel t_2] \in R^{47} \tag{11}$$

This hybrid representation is passed through a final classification head as described in Eq. (12).

$$\widehat{y_i} = Sigmoid\left(W_3 \cdot Dropout\left(ReLU\left(u_i\right)\right) + b_3\right) \tag{12}$$

where $\widehat{y_i}$ is the predicted probability of user i churning. Figure 7 completes architectural design of the proposed Hybrid GAT + MLP model, combining structural, behavioural, and demographic learning pathways for robust churn classification.

The churn prediction dataset exhibits a significant class imbalance, with approximately 9% of users labelled as churners and 91% as non-churners. To address this imbalance without altering the real-world data distribution through oversampling or undersampling techniques, we implement a weighted binary cross-entropy loss function given by Eq. (13).

$$L = -w_1 y \log\left(\widehat{y}\right) - w_0\left(1 - y\right) \log(1 - \widehat{y}) \tag{13}$$

Where, $y \in \{0,1\}$ is the ground truth label, $\widehat{y} \in (0,1)$ is the predicted probability from the model, $w_1 = 10$ is the weight assigned to the positive (churn) class, $w_0 = 1$ is the weight for the negative (non-churn) class. This weighting forces the model to pay greater attention to correctly predicting the minority churn class while maintaining the integrity of the original data distribution. Algorithm 1 shows proposed churn detection process in detail. Table 2 gives details on hyperparameter used for better performance.
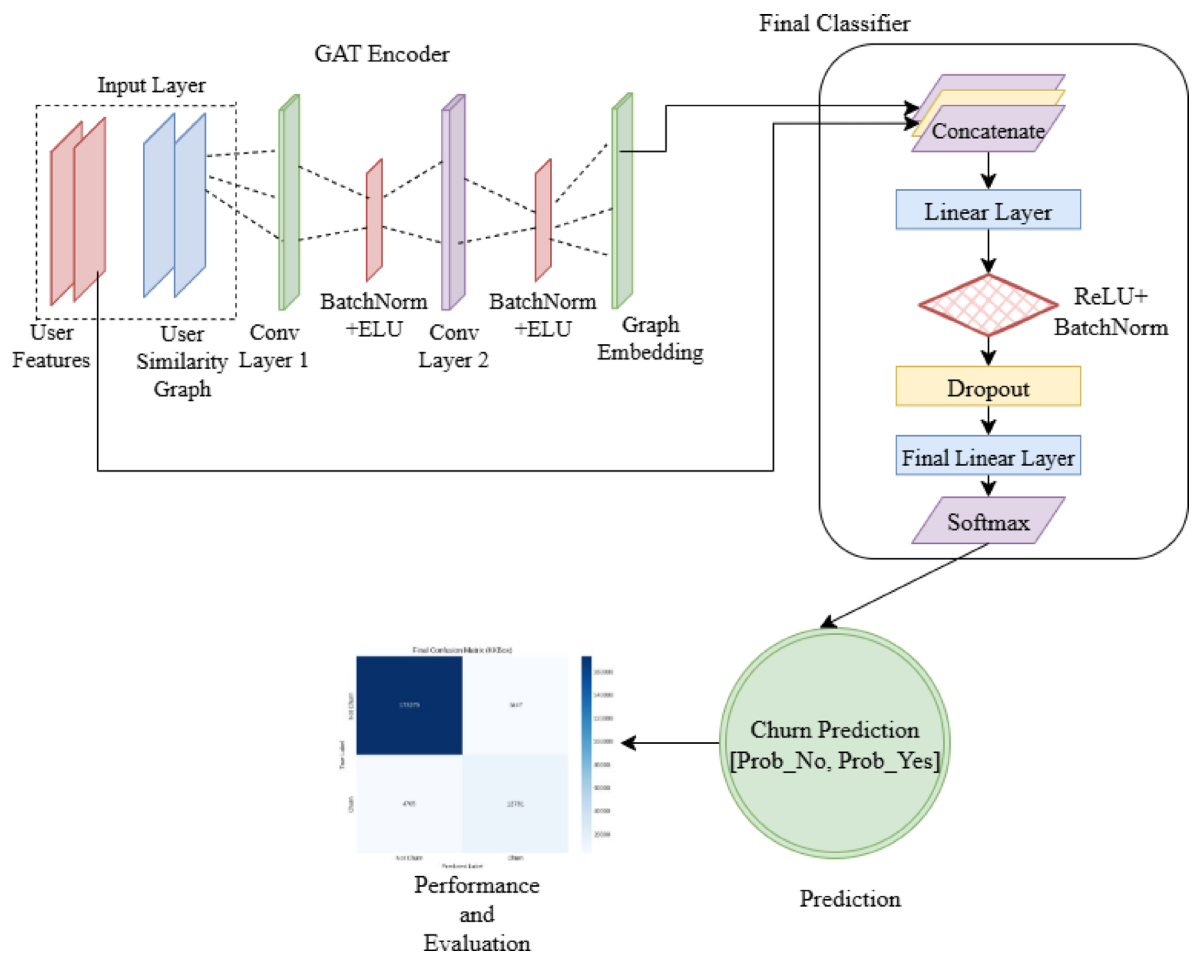
Fig. 7. Detailed architecture of the proposed hybrid model.

Hyperparameter	Value
Optimizer	AdamW
Learning rate (η)	0.005
Weight decay (λ)	1e−4
Batch size	512
Max epochs	50
Early stopping	Patience = 5
Loss function	Weighted BCE
Class weights	w_churn = 10, w_not = 1
Validation metric	AUC
Similarity rule	city ∧ registered_via
Max neighbors/group	10
Neighbor Sampling	[15, 10]
Hidden dim	32
Heads (K)	4
Dropout	0.6
Activations	ELU
Normalization	BatchNorm
Hidden dim	64
Dropout	0.6
Activations	ReLU
Classifier	Linear→ReLU→Dropout→Linear
Feature count	≈ 15
Scaling	StandardScaler (Z-score)
Categoricals	Label encoding
Frameworks	PyTorch 2.0, PyG 2.4.0

Table 2. Hyperparameter table for proposed model.

Input

Dataset $D = \{(x_i, y_i)\}_{i=1}^{N}, x_i \in R^d, y_i \in \{0,1\}$

Graph $G = (V, E), |V| = N$, Adjacency matrix $A \in \{0,1\}^{N \times N}$

Graph Construction $A_{ij} = \begin{cases} 1, if\ city_i = city_j \wedge registered_{via_i} = registered\_via_j \\ 0,\ otherwise \end{cases}$

Attention-weighted node embedding $z_i = \sigma(\sum_{j \in N(i)} \alpha_{ij} W h_j)$, where $\alpha_{ij} = \frac{exp(LeakyReLU(a^\top[Wh_i \| Wh_j]))}{\sum_{k \in N(i)} exp(LeakyReLU(a^\top[Wh_i \| Wh_k]))}$

Multi-head aggregation $z_i^{(multi-head)} = \sum_{k=1}^{K} \sigma(\sum_{j \in N(i)} \alpha_{ij}^{(k)} W^{(k)} h_j)$

Tabular Feature Encoding $t_1 = ReLU(BN_1(W_1 x_i + b_1))$ and $t_2 = Dropout(ReLU(BN_2(W_2 t_1 + b_2)))$

Fusion and Classification $u_i = [z_i \| t_2] \in R^{d'}$, $\hat{y}_i = \sigma(W_3 \cdot u_i + b_3)$

Weighted Binary Cross-Entropy Loss $L = -\sum_{i=1}^{N}[w_1 y_i log(\hat{y}_i) + w_0(1 - y_i)log(1 - \hat{y}_i)]$, $w_1 = 10, w_0 = 1$

AdamW Optimization $\theta \leftarrow \theta - \eta \cdot \frac{\delta L}{\delta \theta} + \lambda \cdot \theta$

Output $\hat{y}_i \in (0,1)$

Algorithm 1. The proposed churn detection algorithm.

Hardware/software	Details
Platform	Google Colaboratory (high-RAM, GPU-enabled runtime)
GPU	NVIDIA Tesla T4 (16GB VRAM)
CPU	Intel Xeon (~ 2.2 GHz)
RAM	25.5 GB
Python version	Python 3.10.12
Libraries and frameworks	- PyTorch 2.0- PyTorch Geometric 2.4.0- Scikit-learn- Pandas- NumPy
Optimizer	Adam optimizer
Learning rate (η)	0.005
Weight decay	$10-410^{-4}$
Batch size	512
Loss function	Weighted cross-entropy loss (w_churn = 10, w_(not churn) = 1)

Table 3. Experimental setup details.

Metric	Not churn	Churn	Macro avg	Weighted avg
Precision	0.98	0.76	0.87	0.96
Recall	0.98	0.79	0.88	0.96
F1-score	0.98	0.78	0.88	0.96
Support	176,726	17,466	194,192	194,192
Accuracy				**0.9587**
AUC score				0.9609

Table 4. Classification performance on KKBox test Set.

Results and discussion
Experimental setup
To ensure reproducibility and take advantage of GPU acceleration, all experiments were conducted using Google Colaboratory with a high-RAM, GPU-enabled runtime. The environment included an $NVIDIA\,Tesla\,T4\,GPU$ $(16GB\,VRAM)$, Intel Xeon CPU $(\sim 2.2GHz)$, and $25.5\,GB\,RAM$, running $Python$ 3.10.12 on $Ubuntu$ 20.04. The proposed Hybrid GAT + MLP model was implemented using $PyTorch$ 2.0, PyTorch Geometric 2.4.0, and supporting libraries like Scikit-learn, Pandas, and NumPy. PyTorch Geometric dependencies were installed using Colab-compatible wheels. The model was trained using Adam optimizer with a learning rate $\eta = 0.005$, weight decay 10^{-4}, batch size of 512, and early stopping based on validation AUC. Class imbalance was addressed using weighted cross-entropy loss ($w_{churn} = 10$, $w_{not\,churn} = 1$), with performance evaluated using AUC, accuracy, and F1-score. Table 3 shows experimental setup in detail.

Evaluation metrics
The various Quantitative metrics like precision, RoC, recall, f1 score and accuracy are used to evaluate the metrics in detail. Precision is described as correct positive prediction across the predicted positives as,

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

Recall is termed as correct positives predictions over the actual positives in the data as,

$$recall = \frac{TP}{TP + FN} \tag{15}$$

Finding the harmonic average for precision and recall is termed as

$$F1\,SCORE = 2\,X\,\frac{Precision\,x\,recall}{Precision + \,recall} \tag{16}$$

Quantitative performance evaluation
To evaluate the performance and effectiveness of the proposed Hybrid GAT + MLP model, we computed several standard classification metrics on the KKBox test set using the optimal decision threshold of 0.89, selected based on validation accuracy as shown in Table 4. The model achieved a final test accuracy of 95.87% and an Area Under the ROC Curve (AUC) of 0.9609, indicating strong discriminative power. Class-wise analysis revealed that for the majority class (Not Churn), the model attained precision = 0.98, recall = 0.98, and F1-score = 0.98, with a support count of 176,726 users. For the minority class (Churn), despite class imbalance, it achieved
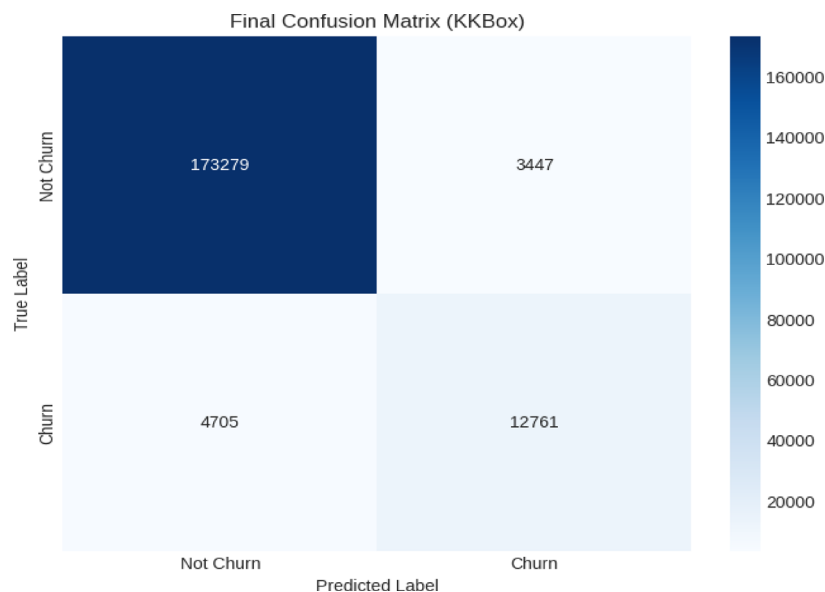
Fig. 8. Final confusion matrix for the Hybrid GAT + MLP model on the KKBox test set.
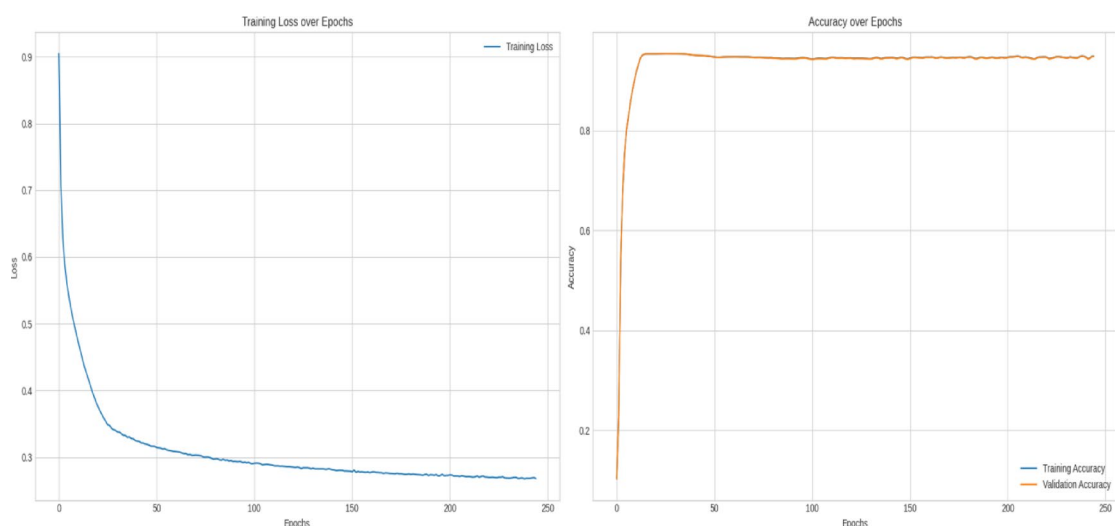

Fig. 9. Training loss and accuracy curves demonstrating stable convergence and high generalization performance of the Hybrid GAT + MLP model.

precision = 0.76, recall = 0.79, and F1-score = 0.78, across 17,466 users. The macro-averaged F1-score was 0.88, and the weighted F1-score was 0.96, reflecting the model's balanced generalisation across both classes.

Qualitative performance results

Confusion matrix analysis

To further assess the predictive quality of the proposed Hybrid GAT + MLP model, a confusion matrix was constructed using the final test set results. As shown in Fig. 8, the model correctly identified 173,279 non-churners and 12,761 churners, while misclassifying 3,447 non-churners as churners and 4,705 churners as non-churners. These results emphasise the model's effectiveness in recognising the minority class (churners), despite the underlying class imbalance. The false positive rate remains acceptably low, showing that the model does not excessively flag loyal users as churners. Meanwhile, the true positive count for churners suggests that the model captures complex churn-inducing patterns efficiently, supporting its utility in proactive customer retention strategies.

Training dynamics

To monitor the learning behaviour of the Hybrid GAT + MLP model, we plotted the training loss and accuracy over 250 epochs. As shown in Fig. 9, the training loss steadily reduces from approximately 0.9 to under 0.27,

showing successful convergence of the optimization process. The rapid loss reduction in the first 30 epochs reflects effective early learning, followed by a plateau suggesting convergence stability. Simultaneously, the training and validation accuracy curves describes that the model accomplishes over 95% accuracy within the first 40 epochs and maintains high consistency throughout subsequent epochs. The absence of significant divergence between training and validation accuracy shows minimal overfitting and validates the effectiveness of dropout regularization and early stopping. These curves assure that the model generalizes well while maintaining learning stability.

Receiver operating characteristic analysis
As illustrates in Fig. 10, the Hybrid GAT + MLP model describes a strong trade-off between true positive rate (sensitivity) and false positive rate, with the curve closely hugging the top-left corner of the plot. The Area Under the Curve (AUC) is 0.9583, showing high discriminative power. This score is consistent with other performance metrics and ensures the model's robustness in identifying churn behaviour across various classification thresholds. The smooth and well-separated ROC profile validates the model's effectiveness in balancing precision and recall, especially under class imbalance.

Precision-recall curve analysis
To estimate the model's ability to identify churners under class imbalance, the Precision-Recall (PR) curve was generated. As depicted in Fig. 11, the proposed Hybrid GAT + MLP model achieved an average precision (AP) score of 0.8235, which is significantly higher than the no-skill baseline of approximately 0.09 (reflecting the churn class prevalence). The model maintains high precision across a wide range of recall values, depicting its capability to retrieve relevant churn instances without producing excessive false positives. The curve's smooth descent and large area under the curve show strong positive predictive power, especially important for high-risk churn intervention use cases. The high AP reinforces the model's strength in classifying minority class instances effectively despite the severe class imbalance in the dataset.

Feature importance analysis
To interpret which attributes most influenced churn predictions, we examined feature importance scores derived from the Gini impurity metric. As shown in Fig. 12, the most influential feature is is_cancel_sum, indicating the number of times a user has canceled their subscription. This is followed closely by $avg\_payment$ and $total\_payment$, suggesting that payment behavior—both in terms of consistency and volume—plays a crucial role in churn detection. Other significant features include $total\_transactions$, log_days, and $registration\_year$, capturing transactional frequency, engagement duration, and user tenure respectively. In contrast, features such as $total\_songs\_played$, $total\_secs\_sum$, and gender contributed minimally to the model's decision process. These insights validate the feature engineering steps and emphasize that both behavioral and payment-related attributes are key drivers of churn outcomes (Table 5).
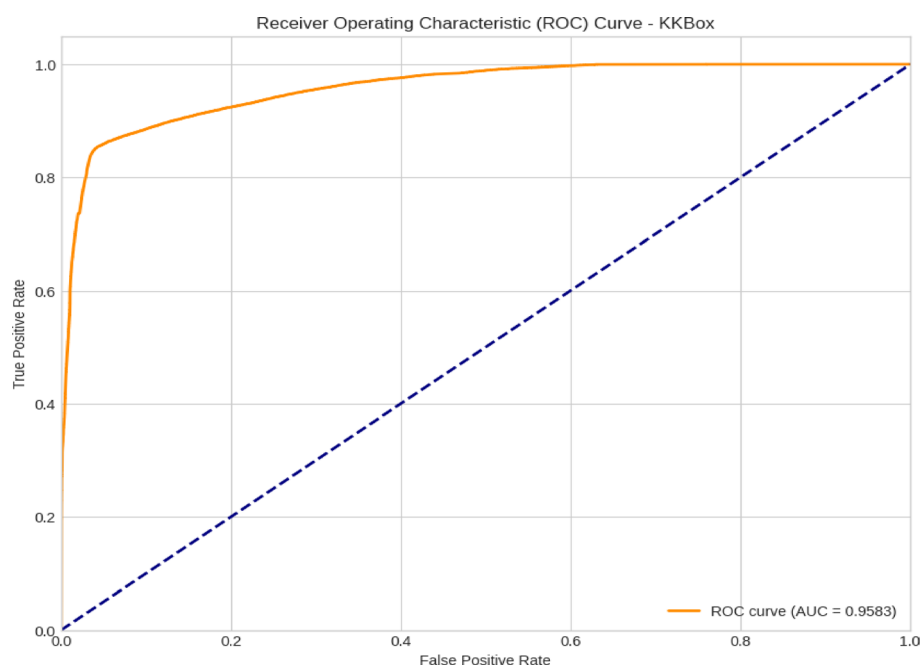

Fig. 10. ROC curve showing strong classification performance of the Hybrid GAT + MLP model with an AUC of 0.9583 on the KKBox test set.
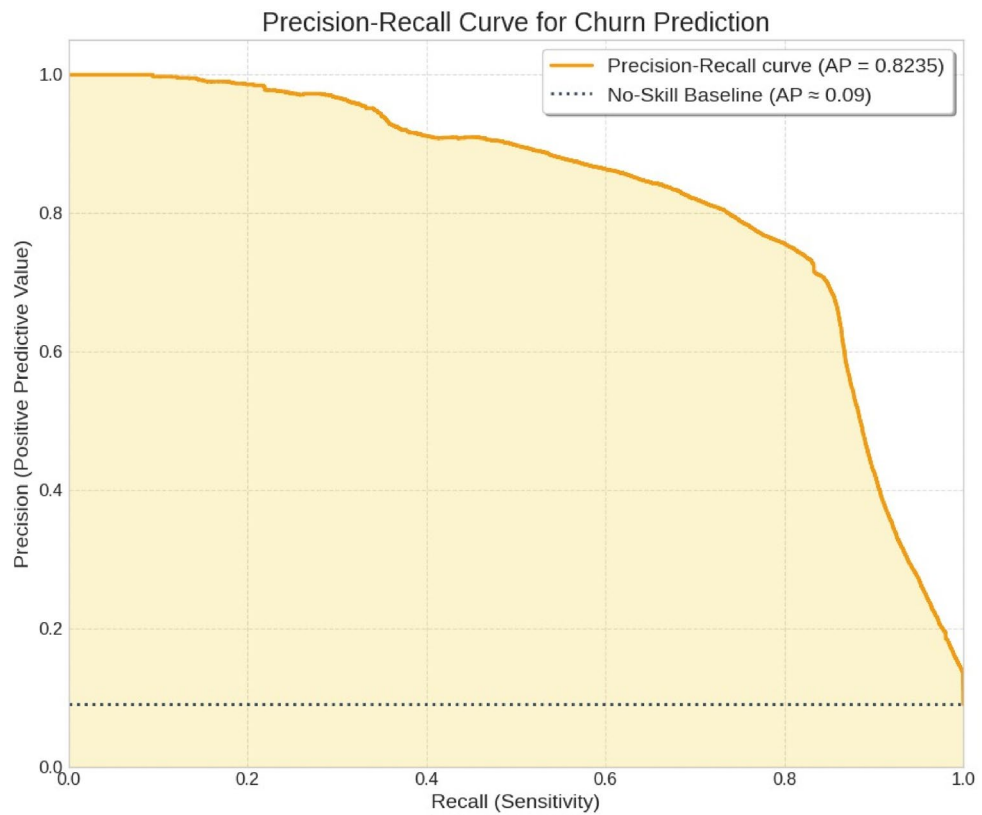
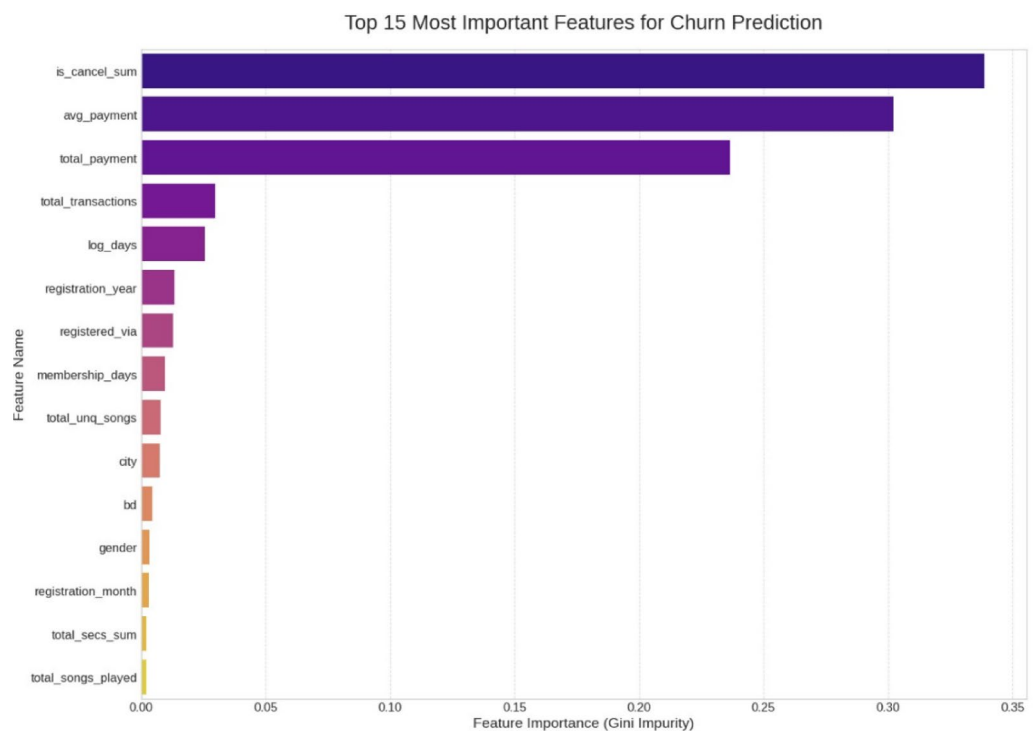**Fig. 11**. Precision-recall curve of the hybrid GAT + MLP model.

Fig. 12. Top 15 most important features for churn prediction ranked by Gini impurity.

Model	Dataset	Techniques	Test accuracy	Test AUC
Baseline GNN	Telco	Graph Sage, synthetic graph nodes	74.0%	0.82
Hybrid GNN	Telco	Feature engineering, SAGE, hybrid (Concat)	75.5%	0.83
HybridGAT (proposed)	KKBox	GAT, weighted loss, heavy regularization	88.4%	0.90
Ensemble (GAT + XGBoost)	KKBox	GAT embeddings as features for XGBoost	89%	0.90
HybridGAT (final model)	KKBox	Large-scale data, feature Eng., GAT, weighted loss	95.8%	0.963

Table 5. Ablation study results across models and Datasets. Initially we tried to imnplemnt on Telco dataset, but this dataset does not perform well on graph based model, so we tried with KKBOX which works extensively for kkbox.

Model	Dataset	Accuracy	AUC	F1-score
TBformer[33]	KKBox	90.12	0.91	0.80
Entropy-RFM[34]	KKBox	94.2	0.91	0.79
HybridGAT	KKBox	95.8	0.963	0.77

Table 6. Comparison of proposed model with state-of-the-art methods.

Ablation study

To assess the impact of model architecture, feature engineering, and dataset scale, an ablation study was conducted across five configurations, as presented in Table 5. The Baseline GNN (GraphSAGE) achieved a test accuracy of 74.0%, AUC of 0.82, and F1-score of 0.61, indicating overfitting and poor generalization. Adding feature engineering in the Hybrid GNN (SAGE) slightly improved the performance to 75.5% accuracy, 0.83 AUC, and 0.63 F1-score. When GAT layers were introduced in the HybridGAT (Telco) configuration, performance further improved to 78.4% accuracy and 0.84 AUC, validating the benefit of attention mechanisms, though the F1-score remained at 0.61, limited by the dataset's simplicity. The Ensemble (GAT + XGBoost) approach slightly underperformed with a 77.9% accuracy and a sharp drop in F1-score to 0.43, suggesting overfitting. Finally, the proposed HybridGAT (KKBox) model, when applied to a richer dataset with large-scale features and weighted loss, achieved 95.8% accuracy, 0.963 AUC, and 0.77 F1-score, establishing it as the best-performing configuration. This clearly illustrates that both model expressiveness and data richness are essential for robust churn prediction.

State-of-the-art comparison

Table 6 shows the state of art comparison on KKBOX. This dataset is nor preferred widely like telecom, bank etc. TBformer[33] and Entropy[34] shows standard result which is nearer to our proposed model. However, our model achieves outperforming results than existing works which is standard benchmark works.

Discussion

This research proposes the new Hybrid GAT + MLP model for predicting churn which uses hybrid GAT structure combined with MLP in a churn predicting manner for imbalanced datasets. It works by the integration of GAT's structural graph reasoning for learning complex user interconnections with the MLP's polygenic feature interrelationship modeling capability. This model works towards minimizing or completely overcoming class imbalance, low density graphs, and churn behavior complexity.

The attention aggregation neural networks make the model to focus on the most critical vertices for churn prediction takes place, thus removing the attention aggregation neck. MLP works in parallel with user-level attributes by learning missing network patterns. This results are achieved by using both graph theory and the underlying real-world phenomena, enabling models to overcome user churn gaps with greater precision.

Through quantitative evaluation techniques, it was proved that the performance of the Hybrid GAT + MLP model was remarkable as it achieved an accuracy of 95.87% and an AUC of 0.9609 on the KKBox dataset. It was also noted that the model surpassed many other churn predicting methods. In addition, the model was able to retain precision as well as recall on both classes, achieving a weighted F1 score of 0.96, despite the dataset's imbalanced characteristics (9% churners and 91% non-churners). Furthermore, qualitative assessments like confusion matrices and ROC curve analyses confirmed the fact that the model was able to accurately separate churn and non-churn cases while only capturing a few false positives.

The ablation study and the results highlight the improvement of the model's performance of the graph attention mechanism, engineered features, and expanded dataset. In addition, it was noted that the Hybrid GAT + MLP model surpassed the current state of the art techniques such as TBformer and Entropy-RFM, which describes the robustness and generalization capacities of the model. Furthermore, the user behavior features such as is_cancel_sum and avg_payment became particularly important in predicting churn, which is essential in building proactive retention mechanisms.

Limitation

The limitation of the research is, where evaluation was focused on the KKBox dataset only, which restricts the model's applicability to other churn detection services with varying consumer patterns or subscription behaviours. Next, the problem of class imbalance[35] was addressed by the weighted loss method[36], but other techniques such as resampling, or even generating synthetic data, which are likely to produce different distribution, were not considered. Explainable interpretability is not obtained for this model[37].

Conclusion

Customer churn prediction remains a crucial challenge in subscription-based digital services, specifically in music streaming platforms where user behaviour is dynamic and engagement patterns are complex. High churn rates can significantly minimise profitability, making it essential to identify at-risk users early for targeted retention strategies. Conventional machine learning models often fall short in capturing relational dependencies and non-linear feature interactions, especially in the presence of imbalanced class distributions. To overcome these limitations, we present a Hybrid Graph Attention Network (Hybrid GAT + MLP) that integrates graph-based user similarity learning with deep tabular feature modelling. A synthetic similarity graph was constructed using demographic attributes to capture contextual user relationships, while a GAT encoder learned structural embeddings that were concatenated with handcrafted behavioural and transactional features. The model was trained end-to-end using a weighted cross-entropy loss to handle class imbalance without data distortion. Experimental results depicted the effectiveness of the model, achieving 95.8% accuracy and an AUC of 0.9626 on the KKBox test set. These metrics emphasise the system's strong predictive performance and its ability to generalise across diverse user profiles. Additionally, the architecture's ability to process large-scale data in a scalable and interpretable manner makes it well-suited for real-world deployment in streaming services, where retention interventions must be both accurate and timely. In future work, we plan to extend the model by integrating temporal graph structures that capture evolving user behaviour over time, as well as experimenting with self-supervised learning techniques to leverage unlabeled data. The model can be tested with diverse churn detection datasets not limiting for improving generalization. Hyper parameter tuning is preferred to improve the accuracy. Explainable AI (XAI) model like SHAP and LIME should be added for knowing the model interpretability.

Data availability

The datasets generated during and/or analysed during the current study are available in the [https://www.kaggle.com/competitions/kkbox-churn-prediction-challenge].

References

1. Gaddam, L. & Kadali, S. L. H. *Comparison of Machine Learning Algorithms on Predicting Churn Within Music Streaming Service* (2022).
2. Karwa, S., Shetty, N. & Nakkella, B. Churn Prediction and customer retention. In *Predictive Analytics and Generative AI for Data-Driven Marketing Strategies*, 98–113 (Chapman and Hall/CRC).
3. Joy, U. G., Hoque, K. E., Uddin, M. N., Chowdhury, L. & Park, S. B. A big data-driven hybrid model for enhancing streaming service customer retention through churn prediction integrated with explainable AI. *IEEE Access*. (2024).
4. Talaat, F. M. & Aljadani, A. AI-driven churn prediction in subscription services: addressing economic metrics, data transparency, and customer interdependence. *Neural Comput. Appl.* 1–26 (2025).
5. Singh, C., Dash, M. K., Sahu, R. & Kumar, A. Artificial intelligence in customer retention: a bibliometric analysis and future research framework. *Kybernetes* **53**(11), 4863–4888 (2024).
6. Gopal, P. & MohdNawi, N. B. An improved convolutional neural network for churn analysis. *Int. J. Adv. Comput. Sci. Appl.* **14**(9) (2023).
7. Fujo, S. W., Subramanian, S. & Khder, M. A. Customer churn prediction in telecommunication industry using deep learning. *Inf. Sci. Lett.* **11**(1), 24 (2022).
8. Dhariya, S. Customer churn prediction in telecommunication industry using machine learning and deep learning approach. In *2023 3rd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, 804–810 (IEEE, 2023).
9. Ramanayake, C., Liyanage, D., Dhananjaya, P., Kumara, C. & Fernando, T. G. I. Telco customer churn prediction using composite deep learning technique. In *2024 6th International Conference on Advancements in Computing (ICAC)*, 402–407 (IEEE, 2024).
10. Liu, R. et al. An intelligent hybrid scheme for customer churn prediction integrating clustering and classification algorithms. *Appl. Sci.* **12**(18), 9355 (2022).
11. Mengash, H. A. et al. Archimedes optimization algorithm-based feature selection with hybrid deep-learning-based churn prediction in telecom industries. *Biomimetics* **9**(1), 1 (2023).
12. Saleh, M. E. & Abd-Alsabour, N. Improved decision tree, random forest, and XGBoost Algorithms for predicting client churn in the telecommunications industry. *Int. J. Adv. Comput. Sci. Appl.* **15**(12) (2024).
13. Al-Safi, S. Customer Churn Prediction Based on Sentiment Score (2024).
14. Geiler, L., Affeldt, S. & Nadif, M. A survey on machine learning methods for churn prediction. *Int. J. Data Sci. Anal.* **14**(3), 217–242 (2022).
15. Manzoor, A., Qureshi, M. A., Kidney, E. & Longo, L. A review on machine learning methods for customer churn prediction and recommendations for business practitioners. *IEEE Access* (2024).
16. Mouli, K. C. et al. An analysis on classification models for customer churn prediction. *Cogent Eng.* **11**(1), 2378877 (2024).
17. Colasanti, L. Customer churn prediction in a slow fashion e-commerce context: an analysis of the effect of static data in customer churn prediction (2021).
18. Nagaraju, J. & Vijaya, J. Methodologies used for customer churn detection in customer relationship management. In *2021 International Conference on Technological Advancements and Innovations (ICTAI)*, 333–339 (IEEE, 2021).
19. Banday, S. A. & Khan, S. Evaluation study of churn prediction models for business intelligence. In *Big Data Analytics*, 201–213 (Auerbach Publications, 2021).

20. Zhang, X. et al. A brief survey of machine learning and deep learning techniques for e-commerce research. *J. Theor. Appl. Electron. Commer. Res.* **18**(4), 2188–2216 (2023).
21. Razzaq, K. & Shah, M. Machine learning and deep learning paradigms: From techniques to practical applications and research frontiers. *Computers* **14**(3), 93 (2025).
22. Mishra, R., Tripathi, P. & Kumar, N. Future directions in the application of machine learning and intelligent optimization in business analytics. In *Intelligent Optimization Techniques for Business Analytic,* 49–76 (IGI Global, 2024).
23. Cenggoro, T. W. et al. Deep learning as a vector embedding model for customer churn. *Procedia Comput. Sci.* **179**, 624–631 (2021).
24. Jin, J., Liu, Q., Yang, Y., Ma, B. & Pan, Y. How social crowding impacts mobile shopping: A perspective from information processing. *Inf. Manag.* **62**(7), 104197 (2025).
25. Sun, X., Guo, B., Yang, Y. & Pan, Y. Time for a change! Uprooting users embedded in the status quo from habitual decision-making. *Decis. Support Syst.* **189**, 114371 (2025).
26. Cao, Y. et al. IBPL: Information bottleneck-based prompt learning for graph out-of-distribution detection. *Neural Netw.* **188**, 107381 (2025).
27. Jing, L., Fan, X., Feng, D., Lu, C. & Jiang, S. A patent text-based product conceptual design decision-making approach considering the fusion of incomplete evaluation semantic and scheme beliefs. *Appl. Soft Comput.* **157**, 111492 (2024).
28. Yang, X., Chen, J., Li, D. & Li, R. Functional-coefficient quantile regression for panel data with latent group structure. *J. Bus. Econ. Stat.* **42**(3), 1026–1040 (2024).
29. Meng, T. et al. A multi-message passing framework based on heterogeneous graphs in conversational emotion recognition. *Neurocomputing* **569**, 127109 (2024).
30. Jiang, H. et al. Fpa-Gcn: enhancing aspect sentiment triplet extraction with feature-rich prediction-aware graph convolutional networks. *Appl. Intell.* **55**(9), 740 (2025).
31. Kamal, A., Anwar, T., Sejwal, V. K. & Fazil, M. BiCapsHate: attention to the linguistic context of hate via bidirectional capsules and hatebase. *IEEE Trans. Comput. Soc. Syst.* **11**(2), 1781–1792 (2023).
32. Mohankumar, P., Singh, V. K. & Kamal, A. Financial fact-check via multi-modal embedded representation and attention-fused network. In *International Conference on Data Analytics & Management* (Springer Nature Singapore, 2024).
33. Li, Y., Tao, Y., Zhu, M., Chen, Z., Wen, Z. & Zhu, B. TBformer: Multi-scale transformer with time-behavior attention for multi-modal customer churn prediction. *IEEE Trans. Consum. Electron.* (2025).
34. Zhao Y. RK-W-Stacking: A hybrid model combining entropy-weighted RFM-K-means clustering and weighted stacking ensemble for user consumption prediction. *Informatica.* **49**(32) (2025).
35. Khan, S. et al. HCovBi-caps: hate speech detection using convolutional and Bi-directional gated recurrent unit with capsule network. *IEEE Access* **10**, 7881–7894 (2022).
36. Khan, S. et al. BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection. *J. King Saud Univ.-Comput. Inf. Sci.* **34**(7), 4335–4344 (2022).
37. Kamal, A. & Abulaish, M. Cat-bigru: Convolution and attention with bi-directional gated recurrent unit for self-deprecating sarcasm detection. *Cogn. Comput.* **14**(1), 91–109 (2022).
38. Hasumoto, K. & Goto, M. Predicting customer churn for platform businesses: using latent variables of variational autoencoder as consumers' purchasing behavior. *Neural Comput. Appl.* **34**(21), 18525–18541 (2022).
39. Panimalar, S. A. & Krishnakumar, A. A review of churn prediction models using different machine learning and deep learning approaches in cloud environment. *J. Curr. Sci. Technol.* **13**(1), 136–161 (2023).

Author contributions

Haiyan Cheng: Conceptualization, Methodology, Formal analysis, Validation, Resources, Supervision, Writing - original draft, Writing - review & editing. Jie He: Validation, Resources, Supervision, Writing - original draft, Writing - review & editing.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to H.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.