Ying Tan
Yuhui Shi (Eds.)

Communications in Computer and Information Science 2017

Data Mining and Big Data

8th International Conference, DMBD 2023
Sanya, China, December 9–12, 2023
Proceedings, Part I

Part 1

Springer

Communications in Computer and Information Science 2017

Rationale

The CCIS series is devoted to the publication of proceedings of computer science conferences. Its aim is to efficiently disseminate original research results in informatics in printed and electronic form. While the focus is on publication of peer-reviewed full papers presenting mature work, inclusion of reviewed short papers reporting on work in progress is welcome, too. Besides globally relevant meetings with internationally representative program committees guaranteeing a strict peer-reviewing and paper selection process, conferences run by societies or of high regional or national relevance are also considered for publication.

Topics

The topical scope of CCIS spans the entire spectrum of informatics ranging from foundational topics in the theory of computing to information and communications science and technology and a broad variety of interdisciplinary application fields.

Information for Volume Editors and Authors

Publication in CCIS is free of charge. No royalties are paid, however, we offer registered conference participants temporary free access to the online version of the conference proceedings on SpringerLink (http://link.springer.com) by means of an http referrer from the conference website and/or a number of complimentary printed copies, as specified in the official acceptance email of the event.

CCIS proceedings can be published in time for distribution at conferences or as post-proceedings, and delivered in the form of printed books and/or electronically as USBs and/or e-content licenses for accessing proceedings at SpringerLink. Furthermore, CCIS proceedings are included in the CCIS electronic book series hosted in the SpringerLink digital library at http://link.springer.com/bookseries/7899. Conferences publishing in CCIS are allowed to use Online Conference Service (OCS) for managing the whole proceedings lifecycle (from submission and reviewing to preparing for publication) free of charge.

Publication process

The language of publication is exclusively English. Authors publishing in CCIS have to sign the Springer CCIS copyright transfer form, however, they are free to use their material published in CCIS for substantially changed, more elaborate subsequent publications elsewhere. For the preparation of the camera-ready papers/files, authors have to strictly adhere to the Springer CCIS Authors' Instructions and are strongly encouraged to use the CCIS LaTeX style files or templates.

Abstracting/Indexing

CCIS is abstracted/indexed in DBLP, Google Scholar, EI-Compendex, Mathematical Reviews, SCImago, Scopus. CCIS volumes are also submitted for the inclusion in ISI Proceedings.

How to start

To start the evaluation of your proposal for inclusion in the CCIS series, please send an e-mail to ccis@springer.com.

Ying Tan · Yuhui Shi
Editors

Data Mining and Big Data

8th International Conference, DMBD 2023
Sanya, China, December 9–12, 2023
Proceedings, Part I

Springer

Editors
Ying Tan
Peking University
Beijing, China

Yuhui Shi
Southern University of Science and Techn
Shenzhen, China

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-981-97-0836-9 ISBN 978-981-97-0837-6 (eBook)
https://doi.org/10.1007/978-981-97-0837-6

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Paper in this product is recyclable.

Preface

The Eighth International Conference on Data Mining and Big Data (DMBD 2023) was held on December 9–12, 2023 in Sanya, China. DMBD 2023 served as an international forum for researchers to exchange the latest advances in theories, models, and applications of data mining and big data as well as artificial intelligence techniques. DMBD 2023 was the eighth event after the successful first event (DMBD 2016) at Bali Island of Indonesia, second event (DMBD 2017) at Fukuoka City of Japan, third event (DMBD 2018) at Shanghai of China, fourth event (DMBD 2019) at Chiang Mai of Thailand, fifth event (DMBD 2020) at Belgrade of Serbia, sixth event (DMBD 2021) at Guangzhou of China and seventh event (DMBD 2022) at Beijing of China virtually.

These two volumes (CCIS vol. 2017 and vol. 2018) contain papers presented at DMBD 2023. The contents of those papers cover some major topics of data mining and big data. The conference received 79 submissions, at least three reviewers per submission in a double-blind review. The committee accepted 38 regular papers to be included in the conference program with an acceptance rate of 48.1%. The proceedings contain revised versions of the accepted papers. While revisions are expected to take the referee's comments into account, this was not enforced and the authors bear full responsibility for the content of their papers.

DMBD 2023 was organized by the International Association of Swarm and Evolutionary Intelligence (IASEI), and co-organized by Peking University and Southern University of Science and Technology, Computational Intelligence Laboratory of Peking University (CIL@PKU), Advanced Institute of Big Data, Beijing, Key Lab of Information System Requirement, Science and Technology on Information Systems Engineering Laboratory, and technically co-sponsored by City Brain Technical Committee, Chinese Institute of Command and Control (CICC), International Neural Network Society, and also supported by Nanjing Kangbo Intelligent Health Academy, Springer-Nature, and Beijing Xinghui High-Tech Co. The conference would not have been such a success without the support of these organizations, and we sincerely thank them for their continued assistance and support.

We would also like to thank the authors who submitted their papers to DMBD 2023, and the conference attendees for their interest and support. We thank the Organizing Committee for their time and effort dedicated to arranging the conference. This allowed us to focus on the paper selection and deal with the scientific program. We thank the Program Committee members and the external reviewers for their hard work in reviewing the submissions; the conference would not have been possible without their expert reviews. Furthermore, this work is partially supported by the National Natural Science Foundation of China (Grant No. 62250037, 62276008, and 62076010), and also partially supported by the National Key R&D Program of China (Grant No. 2022YFF0800601).

Finally, we thank the EasyChair system and its operators for making the entire process of managing the conference convenient.

December 2023

Ying Tan
Yuhui Shi

Organization

General Chair

Ying Tan	Peking University, China

Programme Committee Chairs

Yuhui Shi	Southern University of Science and Technology, China
Wenbin Zhang	Michigan Technological University, USA

Advisory Committee Chairs

Xingui He	Peking University, China
Gary G. Yen	Oklahoma State University, USA

Technical Committee Co-chairs

Benjamin W. Wah	Chinese University of Hong Kong, China
Guoying Wang	Chongqing University of Posts and Telecommunications, China
Enhong Chen	University of Science and Technology of China, China
Fernando Buarque	Universidade of Pernambuco, Brazil
Haibo He	University of Rhode Island Kingston, USA
Jihong Zhu	Tsinghua University, China
Jin Li	Guangzhou University, China
Kay Chen Tan	Hong Kong Polytechnic University, China
Nikola Kasabov	Auckland University of Technology, New Zealand
Qirong Tang	Tongji University, China
Yew-Soon Ong	Nanyang Technological University, Singapore
Yi Zhang	Sichuan University, China

Invited Speakers Session Co-chairs

Andres Iglesias	University of Cantabria, Spain
Shaoqiu Zheng	28th Research Institute of China Electronics Technology Group Corporation, China

Special Session Co-chairs

Ben Niu	Shenzhen University, China
Kun Liu	Advanced Institute of Big Data, China

Publications Co-chairs

Radu-Emil Precup	Politehnica University of Timisoara, Romania
Weiwei Hu	Tencent Corporation, China

Publicity Co-chairs

Eugene Semenkin	Siberian Aerospace University, Russia
Junqi Zhang	Tongji University, China

Finance and Registration Chairs

Andreas Janecek	University of Vienna, Austria
Suicheng Gu	Google Corporation, USA

Conference Secretariat

Wenbo Yan	Peking University, China

Program Committee

Muhammad Abulaish	South Asian University, India
Abdelmalek Amine	Tahar Moulay University of Saida, Algeria
Sabri Arik	Istanbul University, Turkey
Nebojsa Bacanin	Singidunum University, Serbia
Carmelo J. A. Bastos Filho	University of Pernambuco, Brazil

Chenyang Bu	Hefei University of Technology, China
Bin Cao	Tsinghua University, China
Junfeng Chen	Hohai University, China
Walter Chen	National Taipei University of Technology, Taiwan, China
Shi Cheng	Shaanxi Normal University, China
Prithviraj Dasgupta	U. S. Naval Research Laboratory, USA
Khaldoon Dhou	Texas A&M University Central Texas, USA
Hongyuan Gao	Harbin Engineering University, China
Weifeng Gao	Xidian University, China
Ke Gu	Changsha University of Science and Technology, China
Roshni Iyer	UCLA, USA
Ziyu Jia	Beijing Jiaotong University, China
Mingyan Jiang	Shandong University, China
Colin Johnson	University of Nottingham, UK
Liangjun Ke	Xi’an Jiaotong University, China
Lov Kumar	National Institute of Technology, Kurukshetra, India
Germano Lambert-Torres	PS Solutions, Brazil
Tai Le Quy	Leibniz University Hannover, Germany
Ju Liu	Shandong University, China
Jun Liu	Carnegie Mellon University, USA
Kun Liu	Advanced Institute of Big Data, China
Qunfeng Liu	Dongguan University of Technology, China
Yi Liu	Advanced Institute of Big Data, China
Hui Lu	Beihang University, China
Wenjian Luo	Harbin Institute of Technology (Shenzhen), China
Haoyang Ma	National University of Defense Technology, China
Jinwen Ma	Peking University, China
Chengying Mao	Jiangxi University of Finance and Economics, China
Mengjun Ming	National University of Defense Technology, China
Seyedfakhredin Musavishavazi	BAuA, Federal Institute for Occupational Safety and Health, Germany
Sreeja N. K.	PSG College of Technology, USA
Qingjian Ni	Southeast University, China
Neelamadhab Padhy	GIET University, India
Mario Pavone	University of Catania, Spain
Yan Pei	University of Aizu, Japan
Xin Peng	Hainan University, China

Mukesh Prasad	University of Technology, Sydney, Australia
Radu-Emil Precup	Politehnica University of Timisoara, Romania
Aniket Shahade	SSGMCE, India
Min Shi	Hunan University of Science and Technology, China
Zhongzhi Shi	Institute of Computing Technology, Chinese Academy of Sciences, China
Jiten Sidhpura	Sardar Patel Institute of Technology, India
Adam Slowik	Koszalin University of Technology, Porland
Ying Tan	Peking University, China
Eva Tuba	University of Belgrade, Serbia
Mladen Veinović	Singidunum University, Serbia
Guoyin Wang	Chongqing University of Posts and Telecommunications, China
Hong Wang	Shenzhen University, China
Hui Wang	Nanchang Institute of Technology, China
Yuping Wang	Xidian University, China
Ka-Chun Wong	City University of Hong Kong, China
Shuyin Xia	Chongqing University of Posts and Telecommunications, China
Jianhua Xu	Nanjing Normal University, China
Rui Xu	Hohai University, China
Yu Xue	Nanjing University of Information Science and Technology, China
Yingjie Yang	De Montfort University, UK
Peng-Yeng Yin	National Chi Nan University, China
Ling Yu	Jinan University, China
Hui Zhang	Southwest University of Science and Technology, China
Jie Zhang	Newcastle University, UK
Jiwei Zhang	Beijing University of Posts and Telecommunications, China
Xiaosong Zhang	Tangshan University, China
Yong Zhang	China University of Mining and Technology, China
Yuchen Zhang	Northwest A&F University, USA
Xinchao Zhao	Beijing University of Posts and Telecommunications, China
Shaoqiu Zheng	28th Research Institute of China Electronics Technology Group Corporation, China
Jiang Zhou	Texas Tech University, USA

Additional Reviewers

Cai, Long
Dhou, Khaldoon
Hu, Zhongyuan
Jin, Feihu
Lei, Jiaqi
Lian, Xiaoyu
Weinan, Tong
Wenbo, Yan
Zhang, Yixia
Zhang, Yixuan
Zhang, Zhenman

Contents – Part I

Data Mining Methods

Pretrained Language Models and Applications

Reinforcement Learning Approaches

Combinatorial Optimization Approaches

Contents – Part II

Information Security Approaches

Data Query and Image Clustering

Digital Simulation Methods

Machine Learning for Medical Applications

Comparison of Prediction Methods on Large-Scale and Long-Term Online Live Streaming Data

Huan Chen, Shuhui Guo, Siyu Lai, and Xin Lu(✉)

College of Systems Engineering, National University of Defense Technology, Changsha 410073, China
xin.lu.lab@outlook.com

Abstract. Effective prediction of online live streaming traffic plays a crucial role not only in optimizing network resource allocation for enhancing viewer experience but also in assessing factors impacting audience retention and the overall sustainability of streaming platforms. This study conducts a comprehensive evaluation of machine learning methods for online live streaming traffic prediction using extensive hourly traffic data. The dataset comprises 1,385,444,808 live streaming entries and encompasses 30,690,841 unique streamers from the Douyu platform, spanning December 2020 to April 2023. Various experimental settings are employed to compare the performance of these methods. Our findings reveal that among ten methodologies considered, the Bidirectional Long Short-Term Memory, Extra Tree (ET), and Random Forest models demonstrate consistent and robust performance. Particularly, the ET model exhibits outstanding accuracy and precision in predicting daily viewer counts when incorporating pertinent features. In the domain of large-scale and long-term live streaming data prediction, machine learning approaches surpass traditional time series forecasting methods. Moreover, our analysis underscores the significance of incorporating streamer count in enhancing the accuracy of network traffic prediction. Interestingly, while hourly features show limited impact, in certain scenarios, their inclusion may even diminish the predictive efficacy of the models.

Keywords: Online live streaming · Machine learning · Time series prediction · Extra Tree · Bi-LSTM

1 Introduction

Online live streaming, among the myriad Internet applications, has experienced rapid evolution propelled by its strong interactivity and the liberation from temporal and spatial constraints [1]. Since 2016, the surge in online live streaming has led to the emergence of numerous platforms, including Twitch, Douyu TV, TikTok, and others, captivating millions of users globally. The content spectrum of online live streaming has expanded from entertainment-focused gaming to diverse applications encompassing education, culture, sports, tourism, and beyond [2]. The onset of the COVID-19 pandemic further

Y. Tan and Y. Shi (Eds.): DMBD 2023, CCIS 2017, pp. 28–47, 2024.
https://doi.org/10.1007/978-981-97-0837-6_3

accelerated the momentum of online live streaming [3], facilitating the emergence of novel streaming methods such as "remote learning" [4], "e-commerce live streaming" [5], and "social live streaming" [6]. This evolution triggered a fervor for live streaming, exemplified by the staggering statistics: by December 2022, China recorded a soaring 751 million live streaming users, marking an increase of 47.28 million from December 2021, constituting a significant 70.3% of the total Internet user base [7].

The increasing coverage and popularity has brought numerous challenges to online live streaming platforms [8]. For example, game live streaming emphasizes on clarity and smoothness, while business live streaming requires stability. Faced with the differences in technology requirements, as well as the surge in traffic, platform operators need to use various technologies to maintain the stability and fluency of the platform. Another challenge is the high volume and frequency of "Danmu" (a subtitle system in online video platforms that allows users to overlay moving comments onto a playing video that are synchronized to the video timeline [9]), which puts great pressure on the system and imposes excessive technical requirements on the platform. Network traffic prediction can help platform operators optimize network management, and provide users with a smooth and stable live streaming experience. Furthermore, it can also evaluate the impact of factors such as viewers' behaviors, content preferences, and platform performance on the number of viewers, helping to understand the dynamic relationship between streamers and viewers, and further revealing factors that contribute to audience retention and overall platform sustainability.

In recent years, some scholars have been devoted to the research of live streaming prediction. In the related research on predicting the popularity of live streaming rooms, Kaytoue et al. [10] observed a strong correlation between the initial popularity of live streams and their future popularity. Based on this finding, they developed a linear regression model that utilizes the historical viewer count to predict the future viewer count. Furthermore, Jia et al. [11] demonstrated a strong correlation between the popularity of a live streaming room and the frequency at which the streamer conducts live streams. Arnett et al. [12] found through analysing live streaming data from Twitter, YouTube, and Instagram that the timing of account creation by streamers does not directly affect their popularity (measured by the number of viewers and fans). How, having a social media account is crucial for the growth of popularity. Netzorg et al. [13] propose a temporal analysis method that utilizes all relevant information available at time t to predict the eventual absolute popularity (measured by the number of fans) at time $t + \delta$. The predictive results indicate that the behaviors of streamers play a significant role in predicting their popularity.

In other related research on live streaming prediction, Nascimento et al. [14] created a linear regression model to predict the amount of chat based on the number of viewers logged into a channel. In an effort to infer the future income of streamers with users' attributes as features, Tu et al. [15] identified GBDT as one of the most effective/accurate decision tree algorithms compared to Cart, AdaBoost, and RF. They also find that streamers' income is most affected by the number of fans, and that streamers with more fans tend to receive more virtual gifts. Chen et al. [16] collected over 9.5 million Danmu data from 500 live streaming rooms on Douyu platform and proposed a novel model that integrates multiple types of semantic information from Danmu, including sentiment, topics, as well as information on the viewers, and used these information to predict the value of virtual gifts sent by the viewers. The results showed that the gifting behaviours of viewers can be well predicted with features extracted from the Danmu data.

We can see that the majority of online live streaming studies focused on statistical analysis of live streaming data, while only a few attempted to study on traffic prediction. What's more, most of these prediction studies were limited with small sample size, short observation periods, and low generalizability of prediction results, etc. The accuracy and effectiveness of predictions still need to be further verified and improved in practical applications. To overcome these limitations, this study aims to conduct a thorough comparative analysis of various methods and models employed for traffic prediction based on large-scale, long-term live streaming data. Additionally, we will integrate different features to enhance the accuracy and interpretability of predictions. We utilize large-scale and long-term online live streaming data to employ different methods for prediction. Then, through analyzing factors which affect the traffic of live streaming platforms, we extract different features for prediction and evaluate their contribution in improving prediction accuracy with various algorithms.

2 Dataset and Methods

2.1 Live Streaming Data

Douyu, founded in 2016, is the leading online live streaming platform in China. As a typical live streaming platform, it has a large number of active users and provides rich live streaming services. According to Douyu's 2022 annual financial report [17], in the fourth quarter, the Monthly Active Users (MAU) on its mobile platform reached 57.4 million, while the number of paying users remained stable at 5.6 million.

In this study, continuous live streaming data from Douyu, including room ID, room name, live streaming category ID, streamer ID, nickname of streamer, start time of streaming, time of data retrieval, number of live viewers, and number of fans, were obtained with a deliberately developed crawler system by scraping all streaming rooms using Douyu's open data interface (API). The original dataset used in this study covers a time span of 839 days, from December 25, 2020, to April 12, 2023, with a time interval of 10 min. The dataset consists of 1,385,444,808 tuples, involving 30,690,841 unique streamers. As the focus of this study is primarily on the number of streamers and the number of viewers on the platform, the dataset is aggregated into daily and hourly basis, respectively. A summary of the original dataset is provided in Table 1.

Table 1. Overview of the original dataset.

Category	Description
Period of analysis	2020.12.25–2023.04.12
Duration	839 days
Time interval	10 min
Tuples	1,385,444,808
Number of unique streamers	30,690,841
Data attribute fields	room_id, room_name, cate_id, owerner_id, nick_name, show_time, now_time, online, fans

2.2 Prediction Methods

To assess the efficacy of live streaming prediction methodologies, we conduct a systematic comparison of prediction performance between two categories of methods: traditional time series prediction techniques and machine learning algorithms. Specifically, we evaluate ARIMA, SARIMA, BP, RNN, LSTM, GRU, Bi-LSTM, Random Forest, XGBoost, and Extra Tree models.

ARIMA. As the most widely used traditional time series prediction method [18]. $ARIMA(p, d, q)$ integrates the main features of autoregression (AR), differencing (I), and moving average (MA) models to address issues such as non-stationarity of time series, correlation between observations, and residual errors [19]. Equation (1) demonstrates the $ARIMA(p, d, q)$ model using the lag polynomial L [20, 21].

$$\left(1-\sum_{i=1}^{p}\varphi_i L^i\right)(1-L)^d=\left(1-\sum_{j=1}^{q}\theta_j L^i\right)\varepsilon_t. \tag{1}$$

In the equation, L^i denotes the lag operator, φ_i represents the autoregressive model parameter, θ_j represents the moving average parameter, and ε_t is the error term.

SARIMA. SARIMA extends ARIMA by including the seasonal terms (P, D, Q) to capture repetitive patterns within the data's seasonal cycles. Assuming that y_t is a non-stationary time series, w_t represents a Gaussian white noise process, $E_p(B^m)$ represents a seasonal moving average polynomial, $\Theta_Q(B^m)$ demonstrates a seasonal moving average polynomial, and B is a backshift operator. Equation (2) presents the SARIMA model [22].

$$E_p(B^m)\phi_p(B)(1-B^m)^D(1-B)^d y_t=\Theta_Q(B^m)\theta_q(B)w_t. \tag{2}$$

Backpropagation (BP). BP is a key algorithm in neural networks that enables the optimization of the network's parameters, specifically the connection weights between neurons. It works by propagating the error from the network's output back to its inputs, allowing the weights to be adjusted in a way that reduces the error [23]. This iterative process gradually brings the network's output closer to the desired target output, and achieves better prediction performance.

Recurrent Neural Network (RNN). RNN is a robust ANN that can store and utilize information from previous time steps as input for the current time step, and use existing time series data to predict future data over a specific length of time [22]. This architecture enables RNN to effectively handle sequential data with temporal dependencies and remember important features of the input sequential data, making it widely used for time series prediction tasks. The single RNN cell is represented mathematically by Eq. (3) [22].

$$h_t = \tanh(W[h_{t-1}, x_t] + b). \tag{3}$$

where b represents the bias matrix, W denotes the weight matrix, and h_t and h_{t-1} are the hidden states at the current and previous time steps, respectively.

Long Short Term Memory (LSTM) [24]. LSTM differs from the basic structure of traditional RNNs, it introduces a long-term memory cell state and utilizes "gates" (forget gate, input gate, output gate [20]) to regulate the state and output at different time steps. By employing this approach, LSTM addresses issues such as gradient vanishing, gradient exploding and insufficient long-term memory capacity commonly encountered in RNNs [25], demonstrating significant effectiveness in handling sequential problems. The LSTM unit structure is shown in Fig. 1.

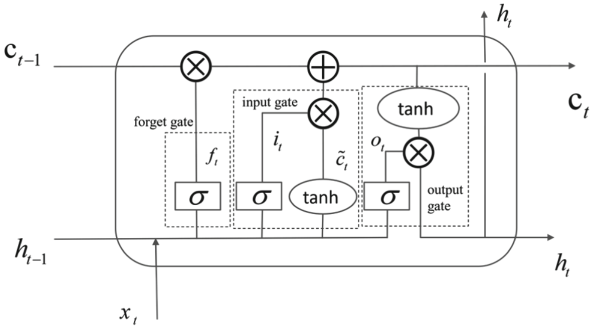


Fig. 1. The LSTM unit structure.

Gated Recurrent Unit (GRU). GRU is another variation of the RNN model. Similar to LSTM, GRU introduces a long-term memory cell state and utilizes "gates" (the reset gate and update gate) to control information and regulate the states and outputs at different time steps [26, 27]. Compared to LSTM, the GRU model has fewer parameters and is computationally more efficient but may exhibit slightly weaker modeling capabilities in certain tasks.

Bidirectional Long Short Term Memory (Bi-LSTM). Bi-LSTM is a variant of the LSTM model, proposed by Graves et al. [28] in 2005. The hidden layer of Bi-LSTM consists of both forward LSTM cell states and backward LSTM cell states [29]. One LSTM cell state considers the forward input and past information, while the other considers the backward input and future information. This structure enables simultaneous consideration of past and future information, resulting in improved predictive performance [30]. The Bi-LSTM network structure is shown in Fig. 2.

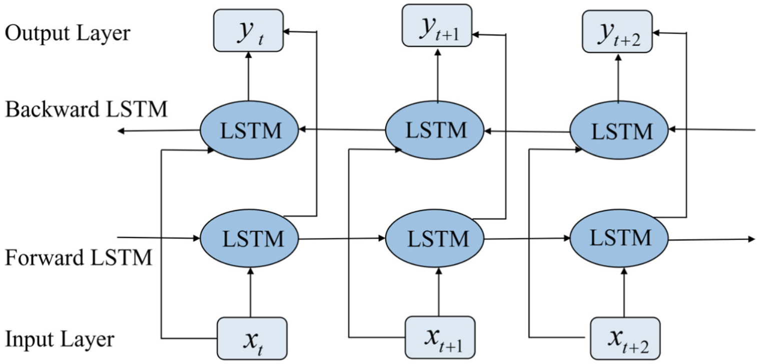


Fig. 2. The Bi-LSTM network structure.

Random Forest (RF). The core idea of Random Forest is to make predictions and classifications by constructing a set of decision trees [31]. Each decision tree is built independently, based on different random samples and feature subsets [32]. And final prediction result of the random forest is determined by voting or averaging the predictions from each decision tree [33].

Extreme Gradient Boosting (XGBoost) [34]. XGBoost is an ensemble learning algorithm based on gradient boosting algorithm and decision tree [34, 35]. It iteratively trains multiple weak learners (typically decision trees) and combines them to create a strong classifier for predicting and classifying complex data. In each iteration, XGBoost fits the residuals of the previous model to gradually improve the prediction performance. Additionally, XGBoost employs innovative techniques such as regularization [36], automatic handling of missing values, and parallel computing to enhance both the accuracy and efficiency of the model.

Extra Tree (ET). ET builds an ensemble of unpruned decision or regression trees according to the classical top-down procedure [37–40]. Its two main differences with other tree-based ensemble methods are that it splits nodes by choosing cut-points fully at random and that it uses the whole learning sample (rather than a bootstrap replica) to grow the trees [41]. Each individual tree within ET is trained on the original dataset, and during the construction process, ET randomly selects a feature value to split the tree. By combining the predictions from multiple trees, typically through voting or averaging, ET can make accurate predictions and handle complex datasets with high dimensional feature spaces. A schematic diagram of an extra tree algorithm is shown in Fig. 3.

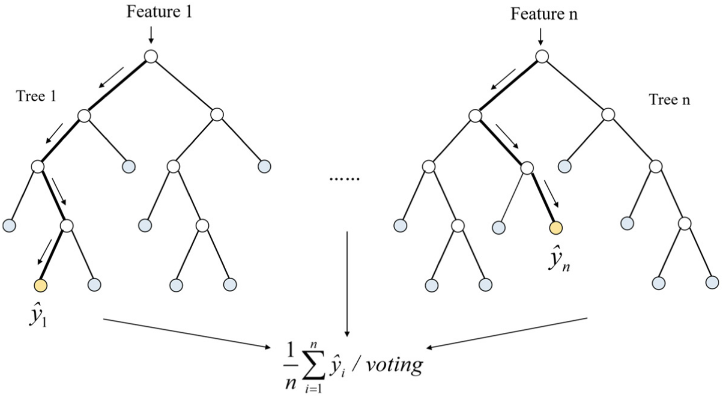


Fig. 3. Extra Tree Algorithm

2.3 Experimental Design

We evaluate the importance of different variables and incorporate different features into the models for prediction and fitting online streaming data. Therefore, the experimental has two parts: the first part does not involve feature inputs, while the second part includes the addition of features for prediction.

We select data from December 25, 2020, 00:00 to April 12, 2023, 24:00 using a sliding window approach to construct training and validation datasets for prediction. The prediction target is the viewer count for each hour of the next day, totaling 24 h. Specifically, the sliding window size is 168, including past 168 time steps, with a pre-diction time window size of 24.

The univariate live streaming data is used as both input and output in the first part, while multivariate data (including streamer count, hours of the day, and the combination of both) is used as input with univariate output in the second part. Evaluation metrics such as MAPE, MAE, MSE, and RMSE are calculated to assess the prediction performance.

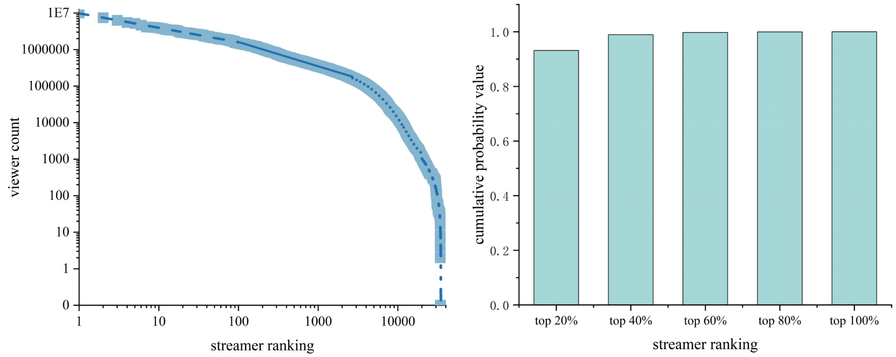
3 Results

3.1 Data Overview and Temporal Analysis

Basic Statistics. In this section, we investigate the distribution of viewer count for all live rooms on the platform at a specific moment. For example, on April 12, 2023, at 21:00 (as shown in Fig. 4 (a)), the number of viewers in different live rooms follow a power-law distribution with an exponential cutoff [29]. In other words, the distribution of viewer count in the top-ranked (less than 1%) live rooms follows a power-law distribution, represented by the Zipf distribution (as shown in Eq. (4)). However, beyond that, the distribution of viewer count follows an exponential distribution, represented by Eq. (5). The specific forms of the distributions and the values of their parameters can be found in Table 2.

$$y = cx^{\beta}. \tag{4}$$

$$y = ax^{\left(-\frac{x}{t}\right)} + y_0. \tag{5}$$



(a) (b)

Fig. 4. (a) displays the distribution of viewer count in the live rooms on the platform at 21:00 on April 12, 2023. The horizontal axis represents the ranking of the streamers, and the vertical axis represents viewer count. (b) depicts a cumulative probability bar chart of the viewer count in the top 20% to 100% rankings of streamers.

Table 2. Fitting results of distributions of viewer count.

	range of x	range of parameters	R^2
Zipf Distribution	[1, 100]	$c = 9.83 \times 10^6$ $\beta = -0.40$	0.99684
	[100, 2700]	$c = 3.41 \times 10^7$ $\beta = -0.66$	0.99799
Exponential Distribution	[2700, 20000]	$a = 4.33 \times 10^5$ $t = 2794$ $y_0 = 1118$	0.99982
	[20000, 35000]	$a = 1.53 \times 10^4$ $t = 7961$ $y_0 = -196$	0.98439

The analysis of the viewer count variations reveals a strong heterogeneity across live rooms, with a few highly popular top-ranked live rooms attracting the majority of the viewer, while the tail-end live rooms have very few viewers. Furthermore, a cumulative probability bar chart of the viewer count was plotted (as shown in Fig. 4 (b)), confirming that the viewer distribution on the platform follows the Pareto Principle, also known as the 80/20 rule. This highly heterogeneous distribution pattern has resulted in a few streamers becoming internet celebrities, as they possess greater attractiveness and influence over the viewers compared to ordinary streamers. This further validates the rationale behind the increased live streaming load caused by internet celebrities' live shows or official live events.

Major Events. To gauge the influence of significant events on streaming metrics, we analyze a segmented subset of streaming data spanning from September 22, 2022, to January 22, 2023, with a specific focus on major esports tournaments within this timeframe. The graphical representation of these pivotal events, showcased in Fig. 5, illustrates the fluctuations in viewer count (depicted by the blue line) juxtaposed with the timelines of the esports tournaments (highlighted in colored regions).

Observing the trends delineated by the blue line, it becomes evident that esports tournaments generally coincide with an uptick in viewer count. However, notable exceptions warrant attention. For instance, the yellow and purple regions in the visual correspond to the League of Legends World Championships. Notably, the yellow segment, encapsulating the group stage, exhibits an upward trajectory in traffic. Conversely, the purple segment, encapsulating the quarter-finals, semi-finals, and finals, displays a downward trend. This divergence could be attributed to the limited advancement of Chinese teams beyond the quarter-finals during the 2022 League of Legends World Championships.

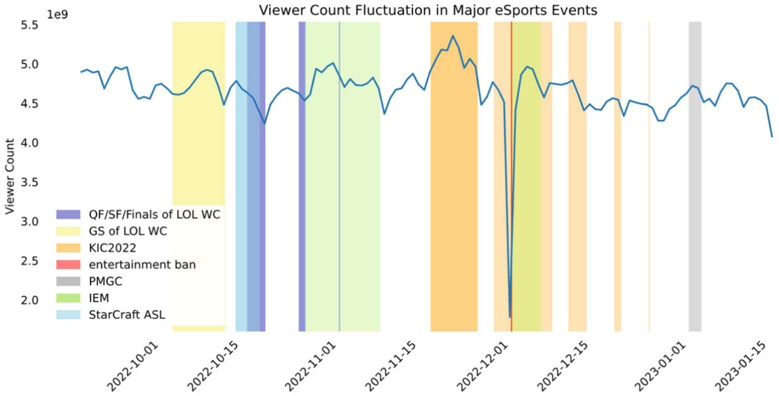
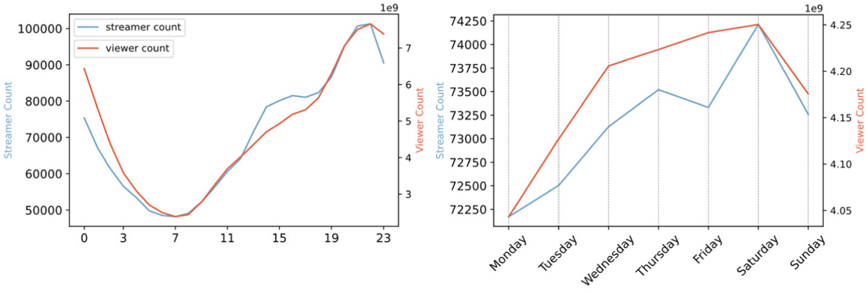


Fig. 5. Changes in the number of viewers associated with major esports events.

Furthermore, the red segment represents an aberrant dip in traffic on December 6, 2022, attributable to a nationwide entertainment suspension in remembrance of former president Jiang Zemin. This suspension led to the cessation of public entertainment activities across the country, profoundly affecting streaming traffic on that day.

General Viewer Characteristics. Line graphs illustrating the trends in the number of streamers and viewers over time are generated from the online live streaming data, as shown in Fig. 6. Studies have shown that the domestic live streaming platform load showed a significant intra-day effect, showing inverted N-type [12]. Our data also exhibits this pattern. As seen in Fig. 6 (a), the load variation on the live streaming platform follows a clear pattern within a day: viewer count in the live rooms decreases starting from midnight and gradually rises after reaching a minimum value at around 6–8 am (typically 7 am). Thereafter, it reaches the peak of the day between 9–11 pm, followed by a decline. Moreover, it appears that the fluctuations in both the number of viewers and the number of streamers are synchronized.

Figure 6 (b) represents the average traffic variation trend from Monday to Sunday. It demonstrates that the number of streamers and viewers remain relatively stable throughout the week, ranging from approximately 72,000 to 74,000 streamers and 404 million to 425 million viewers. Additionally, the changes in the average number of streamers and viewers are not perfectly synchronized but exhibit a similar trend. Specifically, the quantities gradually increase from Monday to Saturday, reaching a peak on Saturday, and then decrease (although there may be slight fluctuations on certain days).



(a)

(b)

Fig. 6. Daily traffic data and weekly traffic Data. In (a), the traffic changes within a day are plotted in hours, while in (b), the average traffic changes from Monday to Sunday are plotted.

Stationarity Test and Data Autocorrelation Analysis. In this case, We employ the Augmented Dickey-Fuller (ADF) test to assess the stationarity of the data. According to the ADF test, the test statistic is calculated as −6.35, and the p-value is 2.685557e−08, indicating that the series is stationary. This implies that the fluctuations in the time series data are predictable in the long run and not influenced by long-term trends. Furthermore, autocorrelation analysis conducted on the series of viewer count revealed two significant peaks at lag 24 and lag 168, as shown in Fig. 7. This indicates that the live streaming traffic data exhibits a cyclic pattern of autocorrelation with a period of one day (24 h) and one week (7 days, 168 h), which aligns with typical characteristics of network traffic data. This finding suggests that time can be considered as an important feature in designing prediction algorithms for live streaming traffic data. By incorporating features such as hours of the day and days of the week, we can capture the cyclic patterns and enhance the accuracy of predictions.

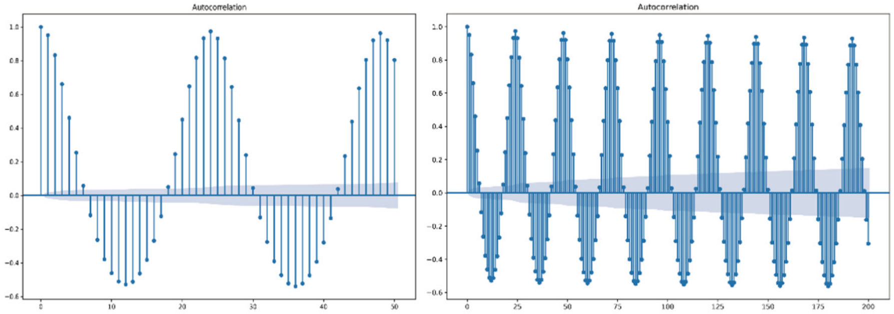


Fig. 7. Autocorrelation test result plot for different lag periods. The autocorrelation plot illustrates the strength of the correlation between lag periods and observed values. The horizontal axis represents the lag periods, with the left plot showing a lag of 50 and the right plot showing a lag of 200. The vertical axis represents the autocorrelation coefficient.

3.2 Univariate Prediction

This section showcases the prediction performance and comparative analysis of ten different models used in this study for forecasting the viewer_count for the next 24 h in a univariate forecasting task. The specific results are presented in Table 3 and Fig. 8, which display error levels, comparisons between models, and the fit be-tween predicted and actual values in the validation set.

The accuracy of final predictions depends on the selection of model parameters. We employed grid search to find optimal hyperparameters and further improved performance through fine-tuning. For instance, the Bi-LSTM model was tested with various hidden layer sizes, learning rates, and iterations, determining the best parameter combination. The final model parameters are as follows:

① ARIMA: the order values for p, d, and q are set as 6, 0, and 5 respectively.
② SARIMA: the order is set as (2,0,2), and the seasonal order is set as (2,0,2,12).
③ BP: the input layer size is 168, the output layer size is 24, the hidden layer size is 150, the learning rate is 0.001, and it iterates for 150 rounds.
④ RNN: the first layer size of 256, with a tanh activation function and a hidden state sequence at each time step; the second layer size is 100, with a tanh activation function and does not return a sequence. The final layer is a fully connected layer with 24 neurons. The loss function used is mean squared error, the optimizer is Adam with a learning rate of 0.001, and the training batch size is 64.
⑤ LSTM: only one hidden layer with 20 neurons and uses the ReLU activation function.
⑥ GRU: a hidden layer size of 256, an output layer size of 24, a batch size of 32, and a learning rate of 0.0003.
⑦ Bi-LSTM: the input layer, hidden layer, and output layer have 168, 128, and 24 neurons respectively. The training iterations, batch size, activation function, loss function, and optimization function are 150, 128, LeakyReLU, MAE, and Nadam. The learning rate is dynamically adjusted based on epochs (reduced to 1/10 every 50 epochs).
⑧ RF: 100 regression trees.

⑨ XGBoost: 1000 estimators, each with a maximum depth of 3 and a learning rate of 0.01. Additionally, training will stop early if there is no improvement in the validation set error for 50 consecutive rounds.
⑩ ET: 100 regression trees with a random state of 0.

Among the evaluated models, the Bi-LSTM model demonstrates the most outstanding performance. Specifically, the Bi-LSTM model achieves the lowest MAPE value of 0.00994, as well as the smallest MAE, MSE, and RMSE values. Compared to the worst-performing AMIRA model, the Bi-LSTM model has an MAPE value that is only 4.75% of its MAPE value, resulting in a 95.25% improvement in accuracy. Compared to the relatively better-performing RF model, the Bi-LSTM model shows a 2.55% improvement in accuracy.

These results indicate that the Bi-LSTM model exhibits minimal errors and high accuracy in predicting online live streaming traffic. In contrast, other neural network models such as RNN, LSTM, and GRU, as well as the traditional time series model SARIMA, perform relatively poorer. This suggests that the Bi-LSTM model, based on the bidirectional long short-term memory network, possesses an advantage in network traffic forecasting tasks. However, it is worth noting that apart from the Bi-LSTM model, several other models also display commendable performance. For instance, the Random Forest (RF), XGBoost, and Extra Tree (ET) models yield relatively good results across multiple evaluation metrics.

Table 3. Comparison of evaluation metric values for each model (without features).

Model	MAPE	MAE	MSE	RMSE
AMIRA	0.2093	1.055×10^9	1.52×10^{18}	1.23×10^9
SARIMA	0.0144	7.35×10^7	9.45×10^{15}	9.72×10^7
BP	0.031	1.48×10^8	3.696×10^{16}	1.92×10^8
RNN	0.031	1.42×10^8	2.94×10^{16}	1.71×10^8
LSTM	0.0857	4.12×10^8	2.53×10^{17}	5.03×10^8
GRU	0.021	9.94×10^7	1.41×10^{16}	1.19×10^8
Bi-LSTM	**0.00994**	$\mathbf{5.11 \times 10^7}$	$\mathbf{5.409 \times 10^{15}}$	$\mathbf{7.35 \times 10^7}$
RF	0.0102	5.17×10^7	6.11×10^{15}	7.82×10^7
XGBoost	0.0175	8.56×10^7	1.01×10^{16}	1.01×10^8
ET	0.012	6.34×10^7	9.26×10^{15}	9.62×10^7

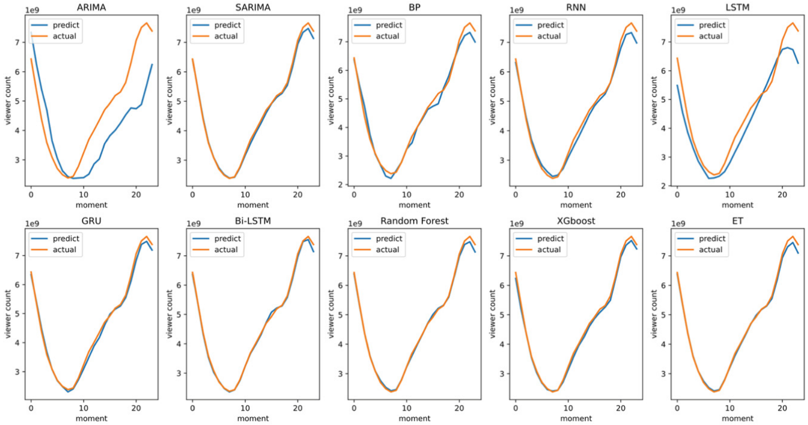
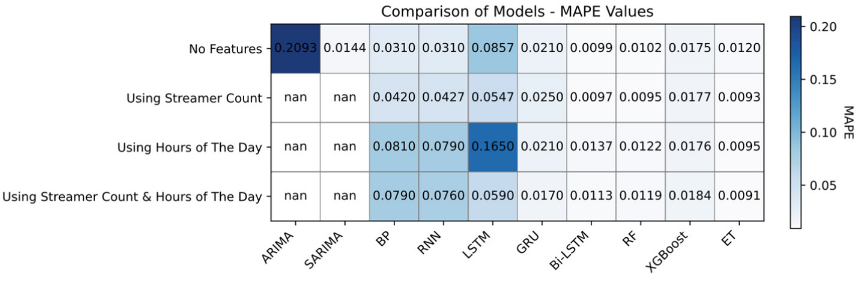


Fig. 8. The fitting performance of models (without features).

3.3 Feature Importance

In this partition, we incorporate additional features to predict viewer count. Specifically, we have chosen three features: streamer count, hours of the day, and the combination of both features. To emphasize the impact of features on the results, the model parameters in this section are set to be the same as those in the univariate prediction section. The selection of these features is based on their potential impact on the online live streaming platform's viewer count.

Figure 9 and Fig. 10 respectively illustrate the comparison of MAPE values under different conditions and the variation of four evaluation metrics. In Fig. 9, lighter colors indicate lower MAPE values and smaller model errors.



	ARIMA	SARIMA	BP	RNN	LSTM	GRU	Bi-LSTM	RF	XGBoost	ET
No Features	0.2093	0.0144	0.0310	0.0310	0.0857	0.0210	0.0099	0.0102	0.0175	0.0120
Using Streamer Count	nan	nan	0.0420	0.0427	0.0547	0.0250	0.0097	0.0095	0.0177	0.0093
Using Hours of The Day	nan	nan	0.0810	0.0790	0.1650	0.0210	0.0137	0.0122	0.0176	0.0095
Using Streamer Count & Hours of The Day	nan	nan	0.0790	0.0760	0.0590	0.0170	0.0113	0.0119	0.0184	0.0091

Fig. 9. The comparison of MAPE values with and without features.

Streamer Count. The error levels and comparison between different models after incorporating streamer count as an input feature for prediction are shown in Table 4. ARIMA and SARIMA models are excluded as they can only perform univariate forecasting. The results show that Bi-LSTM, RF and ET have similar performance. Among them, ET has the lowest MAPE value. However in terms of MAE, MSE, and RMSE, Bi-LSTM slightly outperforms ET and RF.

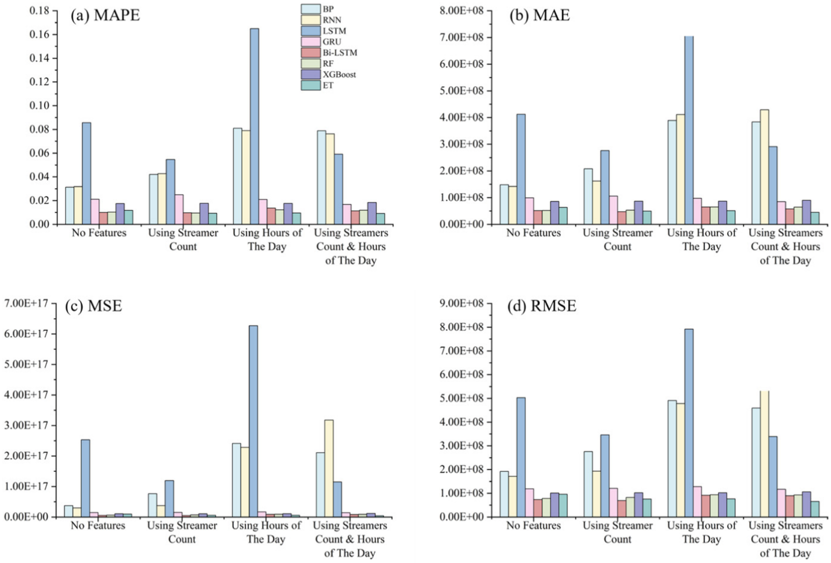


Fig. 10. The changes in evaluation metric values. The figure above illustrates the changes in four evaluation metrics - MAPE, MAE, MSE, and RMSE - under different scenarios: without features, with streamer count, with hours of the day, and with the combination of streamer count and hours of the day. These changes reflect the performance and variations of each model under different conditions.

Furthermore, BP, RNN, GRU, and XGBoost models perform between Bi-LSTM and LSTM models. While their MAPE values are slightly higher than the Bi-LSTM model, they remain relatively low. However, their MAE, MSE, and RMSE values are significantly higher than those of the Bi-LSTM model, indicating a larger gap between their predicted and observed values. The LSTM model performs the worst, with the highest MAPE value (0.0547), as well as highest MAE, MSE, and RMSE values, suggesting relatively large prediction errors.

Comparing these results with the previous univariate forecasting, after incorporating streamer count as a feature, the Bi-LSTM model's MAPE decreased from 0.00994 to 0.00967, indicating an improvement in prediction accuracy. Other metrics such as MAE, MSE, and RMSE also show improvements. LSTM, RF, and ET models are no exception to this. However, the BP model experiences a significant decline in performance after incorporating streamer count as a feature. Overall, there is an enhancement in overall prediction accuracy and a clear improvement in other metrics when considering the entire dataset. Nevertheless, incorporating this feature may not always lead to improvements and can potentially result in a decrease in prediction accuracy.

Hours of the Day. The error levels and comparison between different models after incorporating hours of the day for prediction are shown in Table 5. Incorporating hours of the day into the model results in varying levels of performance. In comparison to other models, Bi-LSTM, RF, and ET models exhibit better accuracy and lower percentage errors. Specifically, in this task, Bi-LSTM and RF models perform similarly. And ET

Table 4. Comparison of evaluation metric values for each model (with streamer count).

Model	MAPE	MAE	MSE	RMSE
BP	0.042	2.08×10^{8}	7.64×10^{16}	2.76×10^{8}
RNN	0.0427	1.62×10^{8}	3.73×10^{16}	1.93×10^{8}
LSTM	0.0547	2.76×10^{8}	1.19×10^{17}	3.46×10^{8}
GRU	0.025	1.06×10^{8}	1.47×10^{16}	1.21×10^{8}
Bi-LSTM	**0.00967**	$\mathbf{4.72 \times 10^{7}}$	$\mathbf{4.78 \times 10^{15}}$	$\mathbf{6.92 \times 10^{7}}$
RF	**0.00947**	$\mathbf{5.29 \times 10^{7}}$	$\mathbf{6.79 \times 10^{15}}$	$\mathbf{8.24 \times 10^{7}}$
XGBoost	0.0177	8.67×10^{7}	1.03×10^{16}	1.02×10^{8}
ET	**0.00927**	$\mathbf{4.93 \times 10^{7}}$	$\mathbf{5.70 \times 10^{15}}$	$\mathbf{7.55 \times 10^{7}}$

demonstrates the best performance, with an MAPE value of 0.00951, and the lowest values for MAE, MSE, and RMSE (4.93×10^{7}, 5.70×10^{15}, and 7.55×10^{7}, respectively). Compared to the worst-performing LSTM model, the ET model has an MAPE value that is only 5.76% of its MAPE value, resulting in a 94.24% improvement in accuracy. Compared to the relatively better-performing RF and Bi-LSTM model, the ET model shows a 22.05% and 30.58% improvement in accuracy, respectively. It is evident that when incorporating hours of the day as a feature, the ET model has the advantage.

Table 5. Comparison of evaluation metric values for each model (with hours of the day).

Model	MAPE	MAE	MSE	RMSE
BP	0.081	3.89×10^{8}	2.41×10^{17}	4.91×10^{8}
RNN	0.079	4.11×10^{8}	2.28×10^{17}	4.78×10^{8}
LSTM	0.165	7.18×10^{8}	6.27×10^{17}	7.92×10^{8}
GRU	0.021	9.76×10^{7}	1.63×10^{16}	1.28×10^{8}
Bi-LSTM	0.0137	6.48×10^{7}	8.35×10^{15}	9.14×10^{7}
RF	0.0122	6.48×10^{7}	8.80×10^{15}	9.38×10^{7}
XGBoost	0.0176	8.67×10^{7}	1.04×10^{16}	1.02×10^{8}
ET	**0.00951**	$\mathbf{5.10 \times 10^{7}}$	$\mathbf{5.82 \times 10^{15}}$	$\mathbf{7.63 \times 10^{7}}$

Considering the three tasks, the Bi-LSTM, RF, and ET models consistently display stable prediction performance, consistently ranking among the top three in terms of prediction accuracy. Specifically, before incorporating features, the Bi-LSTM model exhibits highest prediction accuracy, and the inclusion of the number of streamers as a feature further lowers its MAPE, indicating improved accuracy. However, the performance of the Bi-LSTM model does not significantly improve after incorporating hours of the day as a feature, whereas the ET model demonstrate superior performance and predictive accuracy. In contrast, the LSTM model consistently performs relatively poorly, with lower prediction accuracy before and after incorporating features, as evidenced by higher error metrics. Additionally, the BP and RNN models show relatively poor prediction performance. The GRU and XGBoost models demonstrate good prediction accuracy, ranking in the middle range. It can be observed that in this task, the ET model is more suitable for utilizing hours of the day for prediction.

Streamer Count and Hours of the Day. Table 6 displays the comparative error levels and results of different models upon integrating streamer count and hours of the day for prediction. Notably, upon inclusion of streamer count as a feature alongside hours of the day, all models showcased varying degrees of performance enhancement compared to solely incorporating hours of the day.

Several plausible explanations account for this observed phenomenon: First, the addition of streamer count furnishes the models with deeper insights into network live streaming load, establishing a correlation between streamer count and viewer count. This infusion of data grants the models enhanced understanding of data patterns and trends, consequently bolstering prediction accuracy. Second, the amalgamation of streamer count and hours of the day offers a more holistic and interconnected dataset. This amalgamated input empowers the models to decipher interactions among multiple features, facilitating a more comprehensive comprehension of the data and, consequently, more precise predictions.

Table 6. Comparison of evaluation metric values for each model (with streamer count and hours of the day).

Model	MAPE	MAE	MSE	RMSE
BP	0.079	3.84×10^8	2.11×10^{17}	4.59×10^8
RNN	0.076	4.29×10^8	3.18×10^{17}	5.64×10^8
LSTM	0.059	2.91×10^8	1.15×10^{17}	3.39×10^8
GRU	0.017	8.50×10^7	1.37×10^{16}	1.17×10^8
Bi-LSTM	0.0113	5.76×10^7	7.99×10^{15}	8.94×10^7
RF	0.0119	6.43×10^7	8.63×10^{15}	9.29×10^7
XGBoost	0.0184	9.00×10^7	1.12×10^{16}	1.06×10^8
ET	**0.00913**	$\mathbf{4.45\times10^7}$	$\mathbf{4.27\times10^{15}}$	$\mathbf{6.54\times10^7}$

Furthermore, in this task, the ET model continues to perform the best compared to the other models, with a MAPE value of 0.00913. It also achieves the lowest values in terms of MAE, MSE, and RMSE (4.45×10^7, 4.27×10^{15} and 6.54×10^7, respectively). The Bi-LSTM and RF models follow in performance, while the BP and RNN models exhibit relatively higher MAPE values, indicating larger errors in their predictions. Compared to the worst-performing BP model, the ET model has an MAPE value that is only 11.56%

of its MAPE value, resulting in a 88.44% improvement in accuracy. Compared to the relatively better-performing RF and Bi-LSTM model, the ET model shows a 23.28% and 19.20% improvement in accuracy, respectively.

When comparing these results with the previous three experiments, some trends and changes can be observed. Bi-LSTM, RF, and ET consistently demonstrate top performance among the models, with low MAPE, MAE, MSE, and RMSE values. And Bi-LSTM and ET outperform RF, indicating their stability and accuracy in dealing with large-scale and long-term online live streaming data prediction problems. The ET model, while slightly inferior to the Bi-LSTM model in the experiment without features (though still performing well), it consistently exhibits excellent or even the best performance in other experiments, making it the preferred model for predicting large-scale and long-term online live streaming load when features are incorporated.

Different models exhibit variations in handling different features. For instance, Bi-LSTM shows a stronger capability to leverage information from streamer count compared to hours of the day. Similarly, the ET model also demonstrates a similar pattern. However, the performance of the ET model is less influenced by the features and exhibits minimal fluctuations, demonstrating its relative stability.

4 Conclusions and Discussion

Our study centers on the analysis of extensive, prolonged online live streaming data. We conduct a thorough comparative assessment of diverse methods and models utilized for traffic prediction. Moreover, our approach involves the integration of various features aimed at augmenting prediction accuracy and refining the interpretability of prediction outcomes.

The findings highlight the performance stability of the Bi-LSTM, ET, and RF models in the realm of large-scale and long-term live streaming data prediction. Notably, the ET and Bi-LSTM models exhibit exceptional accuracy and precision, facilitating more precise forecasts of network traffic fluctuations while maintaining lower Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). Particularly, the ET model stands out as the most effective upon incorporating diverse features.

Our comparative analysis underscores the superiority of machine learning and deep learning models over traditional time series forecasting methods in predicting online live streaming traffic. The Bi-LSTM and ET models emerge as preferred choices, followed by RF, due to their superior performance.

In terms of feature integration, the inclusion of streamer count significantly enhances the performance of specific models. Conversely, the inclusion of hours of the day yields marginal improvements in predictive outcomes. Interestingly, experiments integrating both streamer count and hours of the day outperform experiments solely focused on hours of the day. However, some models exhibit slightly reduced performance in the combined approach compared to experiments solely reliant on streamer count. This reiterates the influential role of streamer count in improving prediction accuracy, while the impact of hours of the day remains minimal and, in some cases, may even diminish the predictive performance of the models.

Several potential avenues for optimization and future research merit consideration. First, while this study draws conclusions based on the provided dataset and specific problem parameters, selecting the optimal model should extend beyond individual model performance metrics. Factors like model complexity, training duration, interpretability, and scalability are crucial considerations that should influence the ultimate model selection. Expanding the scope of factors influencing viewer count for comprehensive analysis and prediction would be beneficial for future research. Incorporating additional variables that potentially impact live streaming traffic could offer a more holistic understanding of prediction dynamics. Furthermore, future investigations might delve into a more detailed examination of online live streaming traffic prediction. This could involve a targeted analysis, focusing on aspects intricately tied to live streaming operations. For instance, analyzing and predicting viewer counts for multiple or multiple-category live streaming rooms could provide valuable insights into granular operational aspects. Alternatively, exploring innovative ensemble models presents an opportunity to enhance operational efficiency. Novel approaches in ensemble modeling could specifically address concerns such as the training speed of the Bi-LSTM model, thus improving overall model efficiency.

References

1. Shu-Hui, G., Xin, L.: Live streaming: data mining and behavior analysis. Acta Phys. Sinica **69**(83) (2020)
2. Sharma, S., Gupta, V.: Role of twitter user profile features in retweet prediction for big data streams. Multimedia Tools Appl. **81**, 27309–27338 (2022)
3. Liu, X.: The market changes and causes of game live streaming industry from 2019 to 2020 by case study of HUYA. In: The 2022 International Conference on Economics, Smart Finance and Contemporary Trade (2022)
4. Heim, A.B., Patel, R.J.: Remote learning options. Science **377**(6601), 22–24 (2022)
5. Chen, H., Dou, Y., Xiao, Y.: Understanding the role of live streamers in live-streaming e-commerce. Electron. Commer. Res. Appl. **59**(C), 101266 (2023)
6. Qian, T.Y., Seifried, C.: Virtual interactions and sports viewing on social live streaming platforms: the role of co-creation experiences, platform involvement, and follow status. J. Bus. Res. **162**, 113884 (2023)
7. (CNNIC)ew, t.C.I.N.I.C.: The 51st edition of the "statistical report on internet development in china". Report 1009-3125 (2023)
8. Mengxuan, K., Junping, S., Pengfei, F.A.N.: Survey of network traffic forecast based on deep learning. Comput. Eng Appl. **57**(10), 1–9 (2021)
9. Yan, Z., Yang, Z., Griffiths, M.D.: "Danmu" preference, problematic online video watching, loneliness and personality: an eye-tracking study and survey study. BMC Psychiatry **23**(1), 523 (2023)
10. Kaytoue, M., Silva, A., Cerf, L., Meira Jr, W., Raıssi, C.: Watch me playing, i am a professional: a first study on video game live streaming. In: Proceedings of the 21st International Conference on World Wide Web, pp. 1181–1188 (2012)
11. Jia, A.L., Shen, S., Epema, D.H., Iosup, A.: When game becomes life: the creators and spectators of online game replays and live streaming. ACM Trans. Multimedia Comput. Commun. Appl. (TOMM) **12**(4), 1–24 (2016)
12. Arnett, L., Netzorg, R., Chaintreau, A., Wu, E.: Cross-platform interactions and popularity in the live-streaming community. In: The 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–6 (2019)

13. Netzorg, R., Arnett, L., Chaintreau, A., Wu, E.: PopFactor: live-streamer behavior and popularity. In: International Conference on Web and Social Media (2021)
14. Nascimento, G., et al.: Modeling and analyzing the video game live-streaming community. In: 2014 9th Latin American Web Congress, pp. 1–9 (2014)
15. Tu, W., Yan, C., Yan, Y., Ding, X., Sun, L.: Who is earning? Understanding and modeling the virtual gifts behavior of users in live streaming economy (2018)
16. Chen, Z., Shen, J., Zhu, M., Hu, B., Liu, A.: Predicting virtual gifting behaviors in live streaming using Danmaku information. In: 2022 8th International Conference on Big Data Computing and Communications (BigCom), pp. 190–198 (2022)
17. Douyu reports fourth quarter 2022 unaudited financial results (2023/03/20 2023)
18. Zhang, Y., Meng, G.: Simulation of an adaptive model based on AIC and BIC ARIMA predictions. J. Phys: Conf. Ser. **2449**, 012027 (2023)
19. Siami-Namini, S., Tavakoli, N., Namin, A.S.: A comparison of ARIMA and LSTM in forecasting time series (2018)
20. Pierre, A.A., Akim, S.A., Semenyo, A.K., Babiga, B.: Peak electrical energy consumption prediction by ARIMA, LSTM, GRU, ARIMA-LSTM and ARIMA-GRU approaches. Energies **16**, 4739 (2023)
21. Guenoupkati, A., Salami, A.A., Kodjo, M.K., Napo, K.: Short-term electricity generation forecasting using machine learning algorithms: a case study of the Benin electricity community (C.E.B). In: TH Wildau Engineering and Natural Sciences Proceedings, vol.1 (2021)
22. ArunKumar, K., Kalaga, D.V., Kumar, C.M.S., Kawaji, M., Brenza, T.M.: Comparative analysis of gated recurrent units (GRU), long short-term memory (LSTM) cells, autoregressive integrated moving average (ARIMA), seasonal autoregressive integrated moving average (SARIMA) for forecasting covid-19 trends. Alexandria Eng. J. **61**(10), 7585–7603 (2022)
23. Sadeq, J.M., Qadir, B.A., Abbas, H.H.: Cars logo recognition by using of backpropagation neural networks. Measure. Sens. **26**, 100702 (2023)
24. Li, Y.F., Cao, H.: Prediction for tourism flow based on LSTM neural network. In: 6th International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI). Procedia Computer Science, vol. 129, pp. 277–283 (2018)
25. Amalou, I., Mouhni, N., Abdali, A.: Multivariate time series prediction by RNN architectures for energy consumption forecasting. Energy Rep. **8**, 1084–1091 (2022)
26. Cho, K., Merrienboer, B.V., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: encoder-decoder approaches (2014)
27. Fu, R., Zhang, Z., Li, L.: Using LSTM and GRU neural network methods for traffic flow prediction (2016)
28. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. **18**(5), 602–610 (2005)
29. Doulamis, A.D., et al.: A convolutional neural network face recognition method based on BILSTM and attention mechanism. Comput. Intell. Neurosci. **2023**, 2501022 (2023)
30. Li, Z.Y., Ge, H.X., Cheng, R.J.: Traffic flow prediction based on BILSTM model and data denoising scheme. Chin. Phys. B **31**(4), 214–223 (2022)
31. Alakus, C., Larocque, D., Labbe, A.: Covariance regression with random forests. BMC Bioinform. **24**(1), 258 (2023)
32. Lin, Y., Jeon, Y.: Random forests and adaptive nearest neighbors. J. Am. Stat. Assoc. **101**(474), 578–590 (2006)
33. Moon, J., Kim, Y., Son, M., Hwang, E.: Hybrid short-term load forecasting scheme using random forest and multilayer perceptron. Energies **11**(12), 3283 (2018)
34. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system (2016)
35. Lei, T.M.T., Ng, S.C.W., Siu, S.W.I.: Application of ANN, XGBoost, and other ml methods to forecast air quality in Macau. Sustainability **15**(6), 5341 (2023)

36. Amjad, M., Ahmad, I., Ahmad, M., Wróblewski, P., Kamiński, P., Amjad, U.: Prediction of pile bearing capacity using XGBoost algorithm: modeling and performance evaluation. Appl. Sci. **12**(4), 2126 (2022)
37. Xia, B., Zhang, H., Li, Q., Li, T.: Pets: A stable and accurate pre dictor of protein-protein interacting sites based on extremely-randomized trees. IEEE Trans. NanoBioscience **14**(8), 882–893 (2015)
38. Zhou, Q., Ning, Y., Zhou, Q., Luo, L., Lei, J.: Structural damage detection method based on random forests and data fusion. Struct. Health Monit. **12**(1), 48–58 (2013)
39. Zhou, Q., Zhou, H., Ning, Y., Yang, F., Li, T.: Two approaches for novelty detection using random forest. Expert Syst. Appl. **42**(10), 4840–4850 (2015)
40. Xu, Y., Zhao, X., Chen, Y.: Research on a mixed gas classification algorithm based on extreme random tree. Appl. Sci.-Basel **9**(9), 1728 (2019)
41. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Mach. Learn. **36**(1), 3–42 (2006)