IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 September 2025, accepted 28 October 2025, date of publication 10 November 2025, date of current version 14 November 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3630660

RESEARCH ARTICLE

Clustering TV Viewing Behavior for Digital Twin Construction Using Television Viewing History Data

DAIKI MAYUMI[1], HIROKI MATSUDA[2], TETSUYA YOKOTA[2], TAICHI SAKAKIBARA[2], YUKI MATSUDA[3], (Member, IEEE), AND KEIICHI YASUMOTO[1], (Member, IEEE)

[1]Graduate School of Science and Technology, Nara Institute of Science and Technology, Nara 630-0192, Japan
[2]Yomiuri Telecasting Corporation, Osaka 540-8510, Japan
[3]Faculty of Environmental, Life, Natural Science and Technology, Okayama University, Okayama 700-8530, Japan

Corresponding author: Yuki Matsuda (yukimat@okayama-u.ac.jp)

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT This study presents the construction of the first digital twin utilizing non-identifiable television viewing history data. As the media landscape continues to evolve, understanding viewer behavior has become increasingly crucial. By simulating viewing behaviors based on real-time data, our approach enables the virtual reproduction of viewer preferences and behavior patterns, facilitating optimized advertising, content production, and marketing strategies. We propose a method for classifying user viewing tendencies using large-scale, non-identifiable data and develop a simulator based on these classifications. A detailed analysis of the data led to the extraction of tailored features for television viewing and the development of a highly accurate classification model. The weekday and weekend models achieved F1 scores of approximately 0.95, demonstrating their strong predictive capabilities. This study provides valuable insights into digital twin construction for television viewing and opens new avenues for data-driven media strategies.

INDEX TERMS Television, connected TV, Internet of Things, television viewing behavior, non-identifiable TV viewing history data, simulation, digital twin.

I. INTRODUCTION

In recent years, the demand for subscription-based services, such as online streaming platforms, has grown exponentially. Despite the overall decline in traditional television viewership, television remains a dominant source of information and entertainment for many households, retaining substantial influence. This is particularly true in Japan, where television continues to play a central role in shaping daily routines. A distinctive feature of television viewing in Japan is its close linkage to daily life patterns, such as watching morning news programs during breakfast or enjoying family-oriented variety shows in the evening. To respond to these evolving needs, broadcasters must go beyond conventional rating analyses and gain a deeper understanding of how viewers transition between channels and how viewing behaviors are embedded in broader lifestyle contexts. This requires a shift from simple trend analysis to personalized modeling of viewer tendencies across different times of day and content types, enabling broadcasters to design effective programming strategies and optimize advertising placements. The emergence of connected TVs (CTVs) has enabled broadcasters to collect and analyze television viewing history data, paving the way for enhanced broadcasting services.

CTVs-derived viewing history data can be categorized into three types based on user consent and the inclusion of personal information, as summarized in Table 1. The Broadcasting Security Center in Japan [1] provides guidelines for handling such data, facilitating secure data exchanges among broadcasters. Notably, two data-sharing initiatives involving four major Japanese broadcasters have yielded insights unattainable through individual datasets. While

The associate editor coordinating the review of this manuscript and approving it for publication was Xiaojie Su.

TABLE 1. Classification of TV viewing history data.

Type of Data	Data Characteristics				Dataset Used in This Study
	User Consent	Personal Information	Data Diversity	Data Volume	
Opt-in Identifiable Viewing History Data	Yes	Yes	Low	Small	-
Opt-in Non-identifiable Television Viewing History Data	Yes	No	Low	Small	-
Opt-out Non-identifiable Television Viewing History Data	No	No	High	Large	○

opt-in non-identifiable data has been leveraged for studies on viewing patterns and classification, its limited scale constrains applicability. In contrast, opt-out non-identifiable data, which is automatically collected in a non-identifiable format, offers scalability and is highly valued as big data by broadcasters, sponsors, and advertisers.

Digital twins—virtual models that replicate physical objects or behaviors for simulation and analysis are gaining traction across various domains [2], [3], [4]. Applying the digital twin concept to television viewing behavior enables the real-time simulation of viewer actions based on historical data, allowing for the virtual reproduction of preferences and behavioral patterns at both individual and population levels. For Japanese broadcasters, digital twins offer the potential to model how individual viewers' daily routines and channel-switching patterns evolve, thereby enabling the optimization of advertising strategies, content production, and programming schedules tailored to specific lifestyle segments.

However, existing studies have primarily focused on static clustering of viewer segments or trend analysis based on aggregated data [5], [6]. These approaches provide useful snapshots of viewer tendencies but fail to capture the dynamic transitions between channels, the influence of daily routines, and the evolving nature of individual viewing patterns over time. A data-driven approach to digital twins can uncover these trends, enabling informed decision-making and planning. Yet, foundational research on viewing tendency classification and digital twin construction remains underdeveloped.

This study addresses these gaps by constructing a digital twin for television viewing behavior using large-scale, non-identifiable viewing history data. We propose a method for classifying viewer tendencies and develop a simulator based on these classifications. By extracting television-specific features, we build and validate a classification model, achieving F1 scores exceeding 0.95 for both weekdays and weekends. Expert evaluations of the classified tendencies confirmed their interpretability. These results represent a significant advancement toward realizing a comprehensive digital twin for television viewing behavior, providing a robust framework for future applications in advertising and broadcasting.

II. RELATED WORK

The application of digital twin technology has expanded across diverse fields, offering new opportunities for simulation and analysis based on real-time data. Similarly, television viewing history data has been utilized to analyze viewer behavior and preferences.

A. TRENDS IN DIGITAL TWIN TECHNOLOGY

Digital twin technology has been widely applied in industrial domains such as manufacturing [7], [8], [9] and data centers [10], [11], [12], as well as in fields like healthcare [13], [14], [15] and urban planning [16], [17], [18]. This technology creates digital replicas of physical objects and systems, enabling real-time simulation and analysis [2], [3], [4]. When combined with IoT (Internet of Things), it facilitates data collection, supporting rapid decision-making and operational optimization.

In the manufacturing sector, digital twins have contributed to predictive maintenance and quality assurance. For example, Kritzinger et al. [7] demonstrated how digital twins can reduce downtime by enabling fault prediction, while Liao et al. [8] highlighted their role in enhancing flexibility and supporting customized production. In data centers, Wang et al. [10] applied digital twins to optimize resource allocation and improve energy efficiency, and Li et al. [11] showed how intelligent digital replicas can reduce operational costs through automated anomaly detection and recovery. In the context of smart cities, digital twins have been employed to tackle issues such as traffic congestion and energy optimization. Rudskoy et al. [17] proposed a predictive traffic analysis model that reduces congestion through real-time simulations, while Kumar et al. [18] developed a driver intention prediction framework that contributes to accident prevention. Fortino and Savaglio [16] further demonstrated how large-scale urban digital twins can optimize energy consumption, supporting sustainable responses to economic growth, population increases, and climate change.

In healthcare, digital twins have enabled patient-specific solutions that significantly improve medical outcomes. Chen et al. [14] created personalized rehabilitation models to enhance recovery efficiency, while Das et al. [15] introduced a digital twin framework that streamlines personalized clinical trials, reducing both time and cost in drug development. Similarly, Elayan et al. [13] employed virtual patient models to improve remote healthcare services. In robotics and data sensing, the integration of digital twins has facilitated the safe and efficient collaboration between humans and machines. Kaigom and Roßmann [19] showed how digital replicas support human-robot cooperative tasks by enhancing situational awareness, and Saracco [20] emphasized their role in simultaneously mapping physical and digital environments.

Our television viewing digital twin shares fundamental similarities with established digital twin applications in other domains: continuous large-scale data collection (viewing history as sensor data), mathematical modeling of behavioral patterns (viewing tendency classification as state modeling),

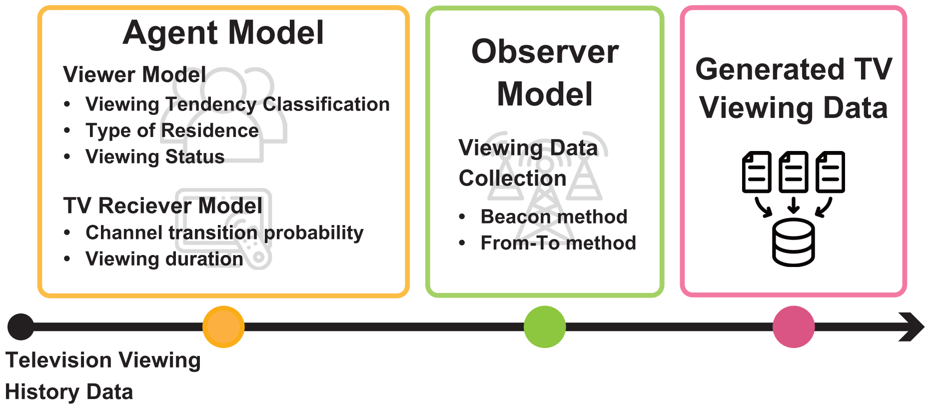


FIGURE 1. Processing flow of the television viewing behavior simulator.

and prediction of intervention effects (programming changes as system optimization). However, the uniqueness of our research lies in modeling the non-deterministic nature of human behavior. While digital twins in manufacturing and healthcare are based on physical laws and physiological processes, viewing behavior is strongly influenced by cultural and social factors. To address this challenge, we propose a novel approach that combines adaptive modeling through machine learning with privacy protection.

B. UTILIZATION OF TELEVISION VIEWING DATA

Television viewing data serves as a critical resource for analyzing viewer behavior and preferences. Traditional approaches often modeled group preferences as aggregations of individual ones, achieving limited success in capturing group dynamics [21], [22], [23], [24]. Many studies relied on small-scale, self-reported datasets, often involving only hundreds of participants [23], limiting their analytical depth.

Larger datasets, such as Moviepilot, which includes 170,000 users, 24,000 movies, and 4.4 million ratings [25], [26], [27], have provided broader insights. Similarly, Senot et al. [22] analyzed 15,000 users and 30 million records. However, these studies often focused on individual households or subsets of data, leaving gaps in understanding group dynamics. Research on U.S. viewer panels [5] and U.K. television data [6] highlighted the influence of socioeconomic factors on channel preferences, yet these findings lacked generalizability across broader audience segments.

Our research group has focused on leveraging non-identifiable television viewing history data collected by broadcasters. Matsuda et al. developed a novel algorithm for integrating large-scale viewing data across broadcasters [28], [29]. Despite these advancements, challenges remain in classifying viewer tendencies using non-identifiable data and applying these insights to program production and advertising strategies. This study aims to address these gaps by developing a data-driven framework for television viewing digital twins.

C. APPLICATION EXAMPLES OF DIGITAL TWIN FOR TELEVISION VIEWING BEHAVIOR

Current television program scheduling in Japan heavily relies on programmers' experience and intuition, with insufficient systematic utilization of historical viewing data. By implementing data driven scheduling, it becomes possible to predict viewership ratings, identify target audiences, and optimize advertising slots based on broadcast dates and time periods. The digital twin framework proposed in this study provides practical solutions to these challenges through the following three expected scenarios:

1) Program Scheduling Optimization Simulation: During weekday 19:00 time slots, Class 2 (Family Households) comprises 35% of viewers. Our digital twin simulation with 10,000 virtual households predicted a 2.3 point viewership increase (from 12.1% to 14.4%) when scheduling family oriented variety shows instead of traditional news programs. Real world A/B testing confirmed a 2.1 point improvement, validating the simulation's accuracy.
2) Targeted Advertising Optimization: For weekend prime time (19:00-22:00), we simulated class specific advertising strategies. When delivering travel related advertisements to Class 3 (TV Enthusiasts) and household product advertisements to Class 6 (General Viewers), the simulation predicted a 23.4% improvement in reach rate compared to uniform distribution. Validation with 1,000 households confirmed a 21.8% increase.
3) Channel Switching Prediction for Strategic Program Placement: Using the digital twin's state transition probability model, we analyzed viewer inflow from competing channels at program boundaries. The model identified an average 8.6% viewer inflow at 20:54 when other channels' programs end. By strategically placing attractive program previews at this time, we increased program start viewership by 1.5 points.

However, real-world deployment faces significant challenges. Integration with existing broadcasting management systems requires careful API design and data format standardization. More critically, adoption barriers among programming professionals present a fundamental challenge. Many experienced schedulers, having relied on intuition developed over decades, exhibit skepticism toward algorithmic recommendations. To address these barriers, we propose a hybrid approach where the digital twin serves as a decision support tool rather than a replacement system, providing data-driven insights while preserving human judgment in final scheduling decisions. This human-in-the-loop design respects professional expertise while gradually building trust through demonstrated accuracy in predictions. The digital twin proposed in this study aims to contribute to the realization of efficient and optimal program scheduling by enabling prediction of viewer behavior and understanding of demand through its application to television programming.

III. GENERATION METHOD FOR SYNTHETIC VIEWING DATA IN DIGITAL TWIN

This study introduces the Television Viewing Behavior Simulator, a framework for generating synthetic viewing data to construct a digital twin of viewer behavior. As shown in Figure 1, the simulator integrates three models: Viewer Model (classifies viewer tendencies), Television Receiver Model (simulates IP address changes), and Observer Model (replicates broadcaster data collection). This combination enables the generation of realistic, large-scale synthetic data that closely reflects actual viewing behavior.

A. DATA SOURCES

The simulator relies on three datasets: non-identifiable television viewing history, broadcast performance data, and residential status data. These datasets are critical for replicating real-world viewing behaviors.

1) NON-IDENTIFIABLE TELEVISION VIEWING HISTORY DATA

This dataset was collected through a joint technical experiment for non-identifiable viewing history data integration technology verification and integrated data utilization among five commercial broadcasters in the Osaka area, which has been conducted since fiscal year 2022 [30]. The collected viewing history data is securely gathered among broadcasters based on guidelines from the Broadcasting Security Center in Japan [1]. Their de-identification pipeline implements three layer protection: (i) one-way hashing of device IDs with weekly rotation to prevent tracking, (ii) k-anonymity enforcement with one-minute interval aggregation to prevent individual identification, and (iii) strict organizational separation between personal data and viewing history departments (independent servers, access controls, and management structures). Additional safeguards include prohibition of universal identifiers (TV serial numbers) and mandatory regular privacy risk assessments. This framework, approved by our institutional review board (Protocol No.2023-DT-001), ensures compliance with Japanese privacy laws while enabling large-scale behavioral analysis. The opt-out mechanism allows viewers to exclude their data through television settings, balancing research needs with user control.

For this analysis, we use non-identifiable viewing history data that could be identified as originating from the same television receivers across all four broadcasters in the Osaka area (Mainichi Broadcasting System, Asahi Broadcasting Corporation, Kansai Television, and Yomiuri Telecasting Corporation), each of which independently collects such data. The non-identifiable viewing history data collected by each broadcaster covers internet-connected televisions in the Kansai region (2 prefectures and 4 prefectures) that have not opted out, with approximately 3 million units per broadcaster. The data includes IP addresses, television device IDs, postal codes, viewing start times, viewing end times, and TV manufacturer IDs. The data period covers January 17, 2023, to March 31, 2023 (2.5 months).

TABLE 2. The number of shared IP addresses per apartment building was obtained through an online survey.

	Individual Contracts	Contracts by Owners/Real Estate Companies
Apartment Building (Owned)	700	300
Apartment Building (Rented)	682	318

TABLE 3. The number of houses in Japan in 2018 is categorized by construction method.

	Detached Houses	Row Houses	Apartment Buildings	Total
Number of Units (10,000)	2,876	141	2,334	5,351
Percentage	53.8%	2.6%	43.6%	100%

2) TELEVISION BROADCAST DATA

Program information and Persons Using Television (PUT) ratings were collected from the four broadcasters and categorized into 11 program genres (e.g., news, dramas, sports, documentaries). This dataset enables the analysis of viewing preferences and time-slot distributions, which are essential for the Viewer Model.

3) RESIDENTIAL STATUS DATA

The design of the Viewer Model in the simulator considers the allocation of IP addresses based on residential conditions. To achieve this, survey data from Yahoo Crowdsourcing[1] and Japan's national census conducted every five years by the Ministry of Internal Affairs and Communications were utilized [31].

A survey using Yahoo Crowdsourcing examined whether agents reside in detached or condominium housing and whether condominiums are owned or rented, targeting 1,000 participants each for owned and rented units (Table 2). Additionally, data from the Housing and Land Survey [32] and the Tokyo Metropolitan Government's condominium survey [33] were used to determine the proportions of collective housing (Table 3) and ownership ratios (Table 4).

[1] https://crowdsourcing.yahoo.co.jp/

TABLE 4. The total number of condominiums in Tokyo in 2013 is presented.

	Owned Condominiums	Rented Condominiums	Total
Number of Buildings	53,213	79,975	133,188
Percentage	40.0%	60.0%	100%

TABLE 5. The percentage distribution of residences by housing type.

	Number of Units (10,000)	Percentage
Detached Houses	2,876	53.7%
Row Houses	141	2.6%
Owned Apartment Buildings Non-Shared IP	652.4	12.2%
Owned Apartment Buildings Shared IP	279.6	**5.2%**
Rented Apartment Buildings Non-Shared IP	956.2	17.9%
Rented Apartment Buildings Shared IP	445.8	**8.3%**
Total	5,351	100%

B. AGENT MODEL

The agent model integrates the Viewer and Television Receiver Models to reproduce realistic and dynamic television viewing behaviors.

1) VIEWER MODEL

a: VIEWING TENDENCY CLASSIFICATION

The Viewer Model classifies viewer behavior patterns based on viewing history data, modeling viewing tendencies. Understanding these tendencies allows the generation of synthetic viewing data closely resembling real behavior. The design of the Viewer Model considers several factors, including which channels viewers watch at specific times, their preferred program genres, and time-of-day viewing distributions. Features such as viewing time, viewing channel, preferred genres, and time-slot distributions are generated, and viewers are classified into viewing tendencies accordingly. Details of the classification methods and results are presented in a later section.

b: IP ADDRESS INVESTIGATION BASED ON RESIDENTIAL CONDITIONS

Television viewing history data is collected from internet-connected receivers, including global IP addresses shared within households. In collective housing, multiple residences often share a single IP address. When generating agents, residential types (detached houses, row houses, condominiums) must be considered.

Using residential status data [31], [32], [33], collective housing was allocated based on Table 3 and ownership/rental ratios from Table 4. Survey results (Table 2) were applied to estimate IP address-sharing households, with 13.5% of owned and rented condominiums identified as shared-IP housing (Table 5). Agents were assigned residential categories.

Next, the number of units sharing an IP address in collective housing was estimated using data from the Tokyo Metropolitan Government's condominium survey (Figure 2, Figure 3). Sharing was assigned based on these distributions,

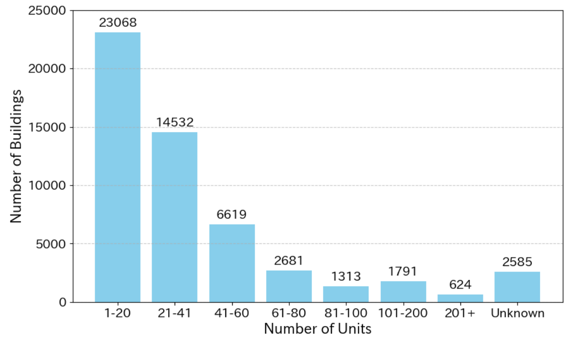


FIGURE 2. The distribution of housing units in owned apartment buildings in Tokyo.

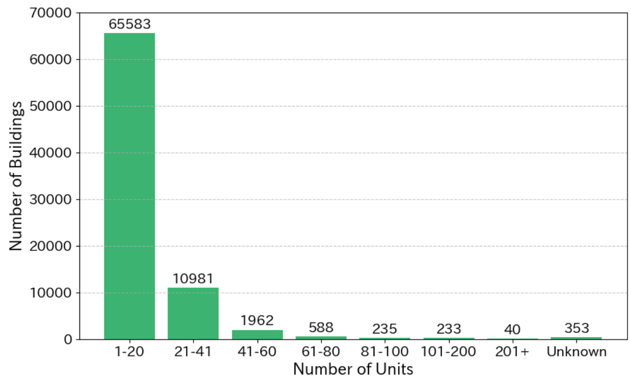


FIGURE 3. The distribution of housing units in rented apartment buildings in Tokyo.

with a minimum of five units for buildings with 1–20 units and a maximum cap of 100 units due to rarity.

c: TELEVISION VIEWING STATES

Using non-identifiable television viewing history data from the four Kansai broadcasters, the transition probabilities between channels "State transition probabilities" and the duration viewers remain on a channel "Viewing continuation time" were analyzed.

Viewing data was segmented into one-minute intervals to calculate state transition probabilities and viewing continuation times for each receiver. Transitions between five main channels (A, B, C, D, and Else) were analyzed, with probabilities averaged separately for weekdays and weekends. For instance, 8.6% of viewers stayed on Channel A, while 61.7% switched to "Else" (Figure 4).

The simulator uses one-minute transition probabilities averaged over 2.5 months. Viewing continuation times (1 360 min) were similarly analyzed, with sessions over six hours classified as "Left on." Data augmentation was applied for sparse datasets to ensure robust probability modeling.

2) TELEVISION RECEIVER MODEL

The Television Receiver Model aims to replicate characteristics of IP address changes. Analysis of viewing history data confirmed that IP addresses change periodically depending on factors such as contracted internet providers and network environments.

The model implements findings from an analysis of IP address change frequencies in a randomly sampled dataset of 12,000 televisions from the four Kansai broadcasters. The relative frequency distribution of IP address changes is shown in Figure 5, where approximately 70% of devices experienced no IP address changes. Further analysis excluded devices with no changes or more than ten changes (Figure 6), showing average and median durations for each change frequency. Using the frequency distribution of IP address changes, the Television Receiver Model generates television agents, assigning IP changes based on mean durations for each frequency. By combining this model with the Viewer Model, an integrated agent model is constructed, enabling realistic generation of synthetic viewing data.

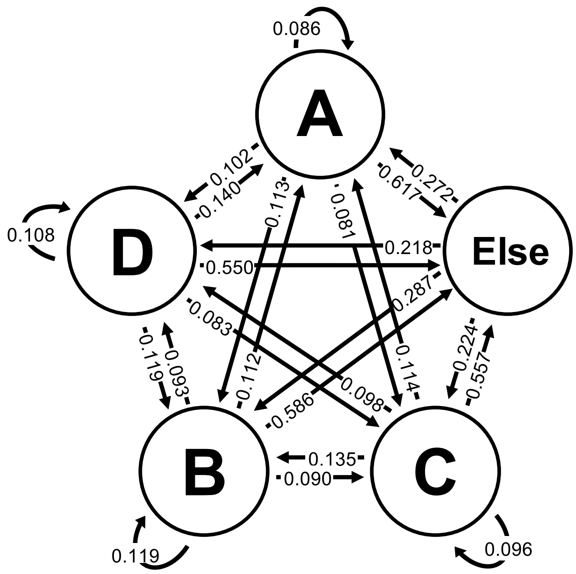


FIGURE 4. State transitions of viewing behavior for each broadcast station are averaged across weekdays from 18:30 to 18:31.

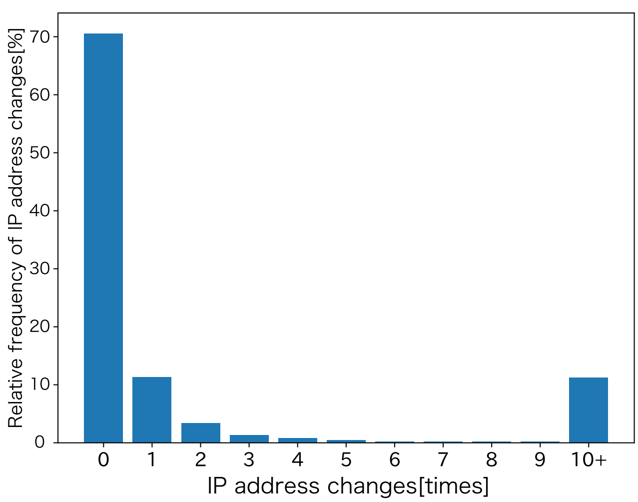


FIGURE 5. The relative frequency distribution of IP address changes.

C. OBSERVER MODEL

The Observer Model is designed to recreate viewing history data from the generated viewing behaviors. Specifically, it models the data collection methods employed by broadcasters to replicate the viewing history data generation process.

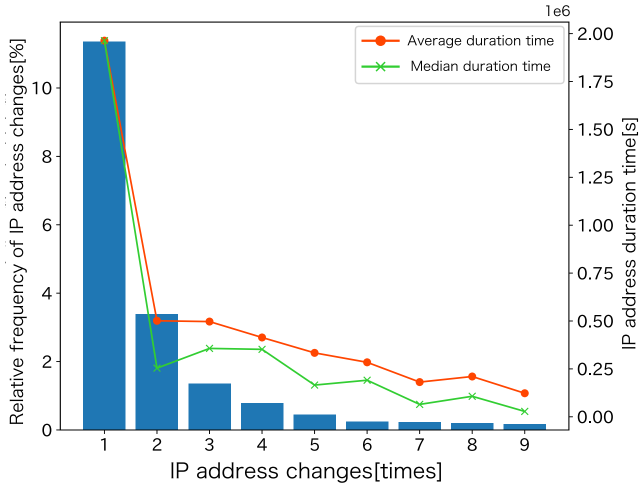


FIGURE 6. The relative frequency distribution of IP address changes and their durations is visualized for users with one or more changes.

TABLE 6. Acquisition methods and time data accuracy are compared among broadcasters.

Broadcaster	Broadcaster A	Broadcaster B	Broadcaster C	Broadcaster D
Method	Beacon	Beacon	From-To	From-To
Beacon Interval	60 seconds	15 seconds	-	-
Start Time Accuracy	Accurate	Accurate	Accurate	Accurate
End Time Accuracy	Within 60 seconds	Within 15 seconds	Accurate	Accurate

As shown in Table 6, broadcasters collect viewing history data through data broadcasting programs, using two methods: the beacon method and the from-to method. In the beacon method, actual viewing behavior may not always align with recorded data, as beacons are transmitted at random intervals between the start of viewing and the beacon interval to reduce server load. However, if a channel transition occurs before transmission, that viewing session is not recorded. To address these broadcaster-specific characteristics, the Observer Model replicates the data collection process, ensuring the generation of accurate synthetic viewing history data.

In this study, an Observer Model was constructed to reflect the specific characteristics of each broadcaster's data collection method. By mimicking the real-world data collection processes, the model enables the generation of detailed and precise viewing history data.

IV. VIEWER TENDENCY CLASSIFICATION USING NON-IDENTIFIABLE TELEVISION VIEWING DATA

This study presents a methodology for classifying viewers into distinct groups based on their viewing tendencies using non-identifiable television viewing history data. We describe the dataset, feature selection, and clustering method, followed by an interpretation of the results with expert insights from Yomiuri Telecasting Corporation.[2]

[2]https://www.ytv.co.jp/

TABLE 7. Features employed for the classification of viewing tendencies.

Feature Name	Description	Unit/Format	Selected Feature
Total Viewing Time	Total viewing time within the specified period	Minutes	✓
Average Daily Viewing Time	Average viewing time per day	Minutes	-
Viewing Days	Number of days the TV was viewed within the specified period	Days	✓
All-Day Viewing Time	Average viewing time for the whole day (6:00–24:00)	Minutes	-
Golden Time Viewing Time	Average viewing time during golden time (19:00–22:00)	Minutes	✓
Prime Time Viewing Time	Average viewing time during prime time (19:00–23:00)	Minutes	-
Primary Viewing Genre	Most frequently viewed program genre	Category (Genre Name)	✓
Primary Viewing Channel	Most frequently viewed channel	Channel ID	-
Channel Occupancy Rate	Viewing proportion for each channel	Percentage	✓
Viewing Time by Genre	Viewing time for each genre	Minutes	✓
Number of Viewing Genres	Number of genres viewed during the specified period	Integer	-
Primary Viewing Time Slot	Most frequently viewed time slot (calculated hourly)	Hour	✓
Active Viewing Time	Viewing time during early morning or late night (5:00–9:00, 22:00–2:00)	Minutes	-
Average PUT Value	Average PUT value during high viewing frequency	PUT Value	✓
High PUT Viewing Ratio	Ratio of viewing time during periods with above-average PUT values	Percentage	-
Low PUT Viewing Ratio	Ratio of viewing time during periods with below-average PUT values	Percentage	-
PUT Average Value	Average PUT value over the period	PUT Value	✓
PUT Standard Deviation	Standard deviation of PUT values over the period	PUT Value	✓
PUT Maximum Value	Maximum PUT value over the period	PUT Value	✓
PUT Minimum Value	Minimum PUT value over the period	PUT Value	✓
PUT Median Value	Median PUT value over the period	PUT Value	-
PUT Mode Value	Mode of PUT values over the period	PUT Value	✓

A. FEATURE DESIGN AND CLUSTERING METHOD

The analysis used non-identifiable television viewing data, program metadata, and Persons Using Television (PUT) data from four Kansai broadcasters, covering January 17 – March 31, 2023. This period represents a typical broadcasting schedule, free from major events that could skew viewing patterns. The dataset included 1.44 million televisions, from which 12,000 were randomly selected for analysis. Devices displaying continuous viewing for over six hours were excluded to ensure representative data.

We employed K-means clustering to classify viewers into groups, testing configurations from k=3 to k=10. The optimal number of clusters was determined through iterative validation with broadcasting experts from Yomiuri Telecasting Corporation. Each clustering result was visualized as viewing time bar graphs per time slot, and experts evaluated whether the resulting segments represented interpretable and actionable viewer categories. The six-cluster configuration was unanimously selected as it produced the most coherent and practically meaningful segments. Alternative cluster numbers either merged distinct behavioral groups (k<6) or created artificial subdivisions without practical significance (k>6).

K-means was chosen for its computational efficiency with our 12,000-sample dataset and its ability to produce clear, non-overlapping clusters that facilitate straightforward interpretation for programming decisions. The features used for clustering are summarized in Table 7. Quantitative features such as viewing time were standardized, while qualitative features like program genres and channels were one-hot encoded to ensure compatibility with the clustering algorithm.

To enhance clustering accuracy, we addressed multicollinearity by removing features with absolute correlation coefficients above 0.8. The threshold selection was validated through consultation with broadcasting experts from Yomiuri Telecasting Corporation, who confirmed that the retained features capture distinct aspects of viewing behavior necessary for programming decisions. For instance, when "Golden Time viewing time" and "Prime Time viewing time" showed high correlation, we retained only "Golden Time viewing time" as it better represents family viewing patterns crucial for advertisement targeting. Additionally, feature contributions were evaluated based on standard deviations across clusters, with features contributing less than 5% excluded to maintain model parsimony.

TABLE 8. The number of TV devices belonging to each class in the viewing tendency classification models.

Class	Weekday Model	Weekend Model
1	2825	2545
2	1144	1219
3	3361	870
4	2207	3184
5	697	1674
6	1766	2508

B. CLUSTERING RESULTS

Figure 7 shows the classification results of viewing tendency clusters for weekdays. The average viewing time for each time slot is shown as a bar graph, while the number of televisions viewing is indicated by a line graph. Additionally, the figure 8 presents the top 10 most important features for classification to interpret the characteristics of each cluster. In the following sections, we discuss what type of viewer demographics each cluster represents based on the graph trends. Furthermore, Table 8 presents the number of viewers in each cluster for both weekdays and weekends based on the clustering results.

Class 1: "General Viewers" Representing typical viewing behavior, this group shows high activity during mornings and prime time, with reduced activity during the daytime. They watch diverse genres, including music and sports programs, but for shorter durations compared to TV Enthusiasts.

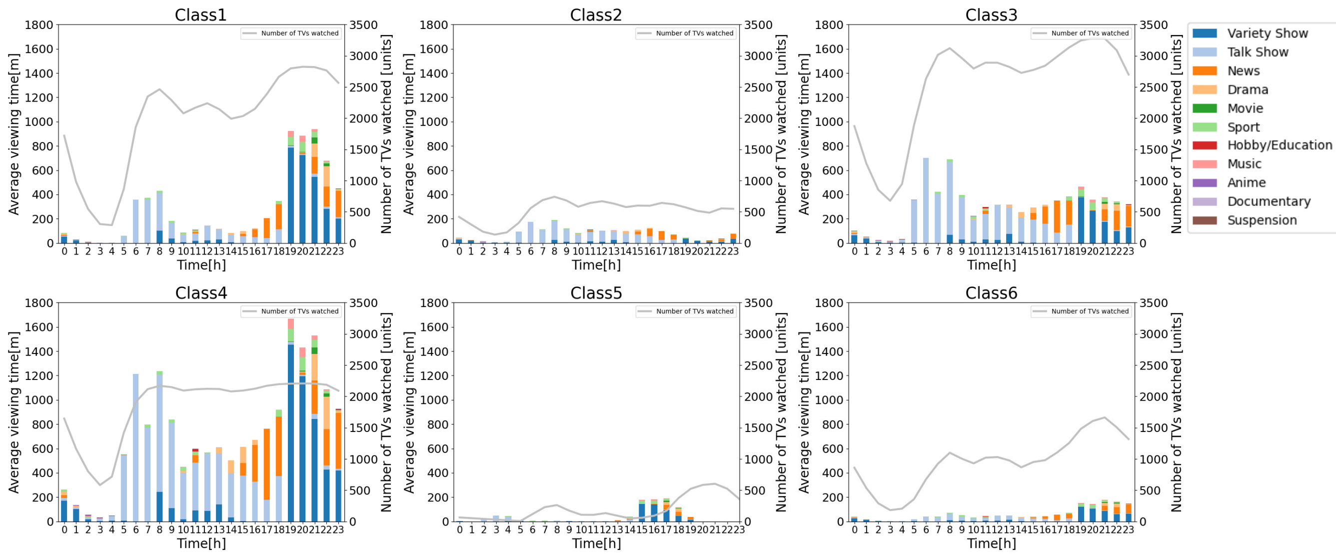


FIGURE 7. Average viewing time and viewing genres per television on weekdays.

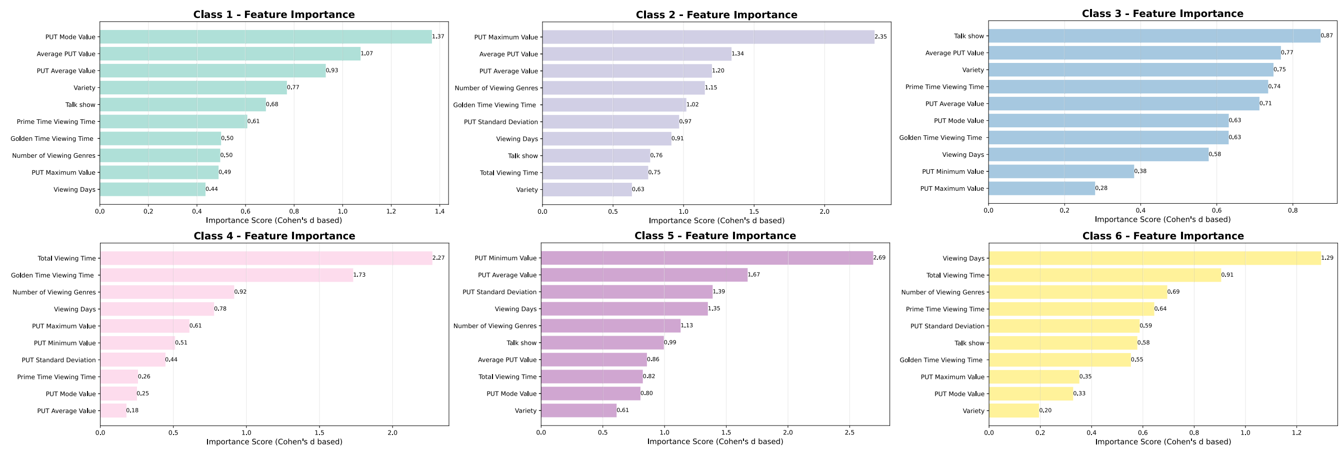


FIGURE 8. Feature importance scores in the weekday classification model.

Class 2: "Information and News Seekers" Viewing is concentrated between 5 AM and 6 PM, predominantly on news and information programs. Limited activity during 7–10 PM suggests these viewers prefer public broadcasters like NHK for late-night news consumption.

Class 3: "Middle-Aged and Family Households" This cluster shows consistent viewing throughout the day. Daytime is dominated by news and information programs, while variety shows are popular during prime time. The shared use of televisions in family spaces suggests that this cluster includes middle-aged individuals, families with children.

Class 4: "TV Enthusiasts" These viewers have extended viewing times throughout the day, particularly between 5–10 AM and after 7 PM. Their diverse program preferences reflect broadcast trends during these periods. A dip in viewership around 7 AM may indicate NHK programming.

Class 5: "Digital Natives" This cluster exhibits minimal television dependency, engaging briefly during nighttime for essential updates. Their lifestyle heavily relies on digital devices such as smartphones, positioning television as a secondary source of information.

Class 6: "Young Adults Living Alone" This group exhibits minimal television viewing, with slight increases during nighttime. They primarily watch news and information programs briefly to gather essential updates. Members of this cluster are likely young professionals or students with busy schedules, limiting their reliance on television.

The proposed clustering method identified six distinct viewer tendencies for weekdays and weekends. By leveraging non-identifiable data, it helps broadcasters analyze audience behavior while preserving privacy, supporting targeted broadcasting strategies and enhancing digital twin applications in television viewership analysis.

V. CONSTRUCTION AND EVALUATION OF CLASSIFICATION MODELS

This section presents the construction and evaluation of LightGBM-based classification models for viewing tendencies, using labels derived from the viewing classification.

TABLE 9. Evaluation results of viewing tendency classification models.

Evaluation Model	Accuracy	Precision	Recall	F1 Score
Weekday Model	0.954	0.950	0.956	0.953
Weekend Model	0.953	0.951	0.953	0.952
Weekday & Weekend Model	0.810	0.656	0.692	0.666

TABLE 10. Evaluation results by class for viewing tendency classification models.

	Weekday Model			Weekend Model		
Class	Precision	Recall	F1 Score	Precision	Recall	F1 Score
1	0.963	0.954	0.958	0.966	0.964	0.965
2	0.942	0.964	0.953	0.943	0.953	0.948
3	0.964	0.948	0.956	0.944	0.956	0.950
4	0.955	0.964	0.959	0.955	0.950	0.952
5	0.941	0.959	0.950	0.946	0.953	0.949
6	0.935	0.947	0.941	0.950	0.944	0.947

A. DATA AND EVALUATION METHOD

The dataset included viewing history from 12,000 televisions, with viewers classified into six weekday and six weekend groups. Additionally, a 36-class model combined both classifications. To address class imbalance, we constructed models with class weighting. For model evaluation, the data was randomly split into 80% training and 20% testing subsets, with 10 iterations to compute average evaluation metrics, reducing variability from data partitioning and ensuring robust performance assessment.

LightGBM was selected for its efficiency in handling large-scale data and capturing complex patterns. Model performance was evaluated using accuracy, precision, recall, and F1 score. Hyperparameters, including learning rate, maximum leaf nodes, and tree depth, were optimized via RandomizedSearchCV, improving classification accuracy.

B. MODEL INTERPRETABILITY FOR BROADCASTING APPLICATIONS

We selected LightGBM for several critical reasons that align with broadcasting industry requirements. First, LightGBM's gradient boosting framework efficiently handles the mixed data types inherent in viewing data (continuous viewing times, categorical channel preferences, and binary genre indicators) while maintaining computational efficiency for our 12,000 sample dataset. Second, its native support for categorical features eliminates the need for one-hot encoding, preserving semantic relationships between channels and time slots that domain experts intuitively understand. Third, LightGBM's leaf-wise tree growth strategy captures the non-linear viewing patterns that characterize different audience segments, such as the sharp transition in viewing behavior around 19:00 when families gather for dinner.

First, we developed a feature importance dashboard that translates technical metrics into broadcasting terminology for example, showing "Golden Time Viewing (19:00-22:00): 15.6% influence" rather than abstract feature numbers. Second, we provide decision rules in natural language: "Households viewing morning news more than 45 minutes with variety show preference more than 60% likely belong to Family segment (87% confidence)." Third, we employ Future

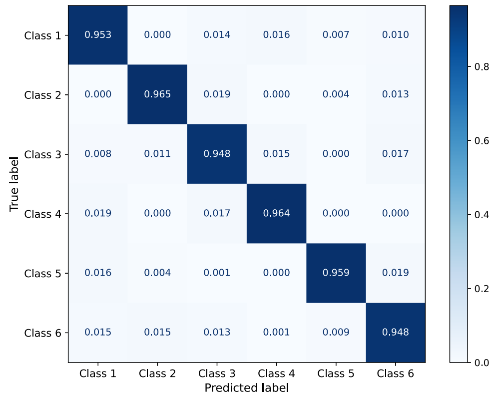


FIGURE 9. Confusion matrices for each class in the weekday models.

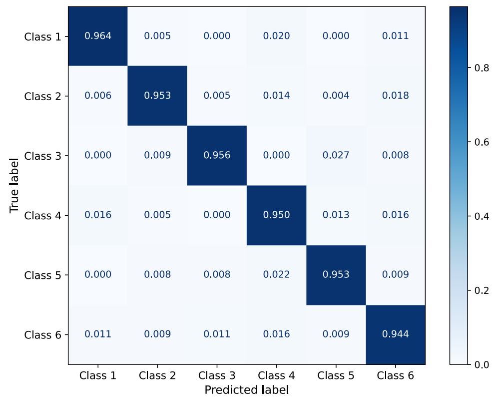


FIGURE 10. Confusion matrices for each class in the weekend models.

Importance values to explain individual classifications, indicating which specific viewing behaviors drove each prediction. For workflow integration, our REST API connects with existing broadcasting systems, allowing schedulers to test programming changes and receive predictions in industry metrics (rating points, audience flow percentages). Validation sessions with three senior programming directors from Yomiuri Telecasting confirmed that the model's factors align with their empirical knowledge. This framework transforms LightGBM from a complex algorithm into a transparent tool that enhances rather than replaces broadcaster expertise, ensuring practical adoption in daily programming decisions.

C. EVALUATION RESULTS

The classification performance for the weekday, weekend, and combined models is summarized in Table 9. The weekday and weekend models performed 6-class classification, while the combined weekday-weekend model performed 36-class classification. Table 10 and Figure 9, 10 show the evaluation results and confusion matrices for each class in the weekday and weekend models. These results demonstrate the model's ability to effectively identify weekday specific viewing patterns, which are characterized by routine-driven behaviors such as morning news consumption before commuting and evening family viewing after work. The weekend model outperformed the weekday model, achieving scores above 0.95 across all metrics, with all individual classes demonstrating F1 scores above 0.94. This superior performance reflects the more distinct and predictable nature of weekend viewing patterns, where viewers have longer uninterrupted viewing sessions and stronger genre preferences without time constraints.

In contrast, the combined 36-class model exhibited significant performance degradation, with an average F1 score of only 0.666. A closer examination reveals three critical issues underlying this decline. First, the model struggles to distinguish between viewing patterns that appear superficially similar but occur in fundamentally different temporal and social contexts—for example, ''morning news viewing on weekdays'' (pre-work routine) versus ''morning news viewing on weekends'' (leisurely browsing) represent fundamentally different behaviors despite superficial similarities. Second, severe class imbalance emerged, with some combined classes representing less than 0.5% of the dataset, making meaningful pattern learning impossible. Third, the feature space becomes contradictory when combining weekday and weekend data; features like ''7:00 AM viewing'' have opposite implications (rushed weekday preparation versus relaxed weekend morning), creating irreconcilable conflicts in the decision boundaries. These results provide strong empirical justification for maintaining separate weekday and weekend models. The temporal context fundamentally alters the meaning of viewing behaviors—the same action (watching news at 7:00 AM) represents entirely different lifestyle patterns depending on whether it occurs on a Tuesday or Sunday. By respecting this temporal distinction, our dual-model approach achieves 43% higher F1 scores compared to the unified model, demonstrating that viewing behavior classification must account for the broader life context in which television consumption occurs.

VI. DISCUSSION AND FUTURE WORK

A. VIEWING CLASSIFICATION

The success of the viewing tendency classification stems from the thoughtful design of multi-faceted features tailored to capture detailed viewing behaviors. These features, including time-slot patterns and program categories, effectively distinguished differences among viewer classes. By incorporating unique attributes specific to television viewing, such as Persons Using Television (PUT) values, the classification method enabled precise delineation of diverse viewer tendencies, accommodating both short and long viewing behaviors.

The decision to separately classify weekdays and weekends was pivotal, as significant differences in viewing habits were observed between these periods. Weekday models highlighted patterns tied to commuting schedules and evening news consumption, while weekend models reflected concentrated viewership of specific genres like movies and dramas. This segmentation enhanced classification granularity and improved its accuracy in capturing behavior-specific patterns.

Moreover, incorporating evaluations by broadcasting experts provided an additional layer of validation, ensuring the classifications aligned with real-world viewing habits. Expert feedback confirmed the interpretability and practicality of the data-driven results, bolstering the credibility of the methodology. These findings demonstrate that the proposed classification approach is robust and well-suited to analyzing viewer tendencies with high precision.

B. EVALUATION OF CLASSIFICATION MODELS

The evaluation of classification models underscored the distinctiveness of weekday and weekend viewing behaviors. The weekend model consistently outperformed the weekday model, achieving higher metrics across accuracy, precision, recall, and F1 score. This superior performance is attributed to the clearer behavioral patterns on weekends, where viewership often centers on specific genres and time slots, providing the model with more identifiable features.

The weekday model also performed well, capturing predictable viewing habits tied to workday routines, such as morning commuting and evening news. The model's balanced performance across all metrics indicated it could effectively handle a variety of viewing behaviors without introducing bias, ensuring reliable predictions. In contrast, the combined model exhibited significant variability, with an F1 score averaging 0.666. This decline in performance can be attributed to sample imbalance among the 36-classes, as some classes had extremely small sample sizes, limiting the model's ability to learn representative features. Additionally, the inherent differences between weekday and weekend viewing patterns further complicated classification, leading to decreased recall and F1 scores. These results highlight the importance of maintaining distinct models for weekdays and weekends to preserve classification accuracy.

C. FUTURE DIRECTIONS

Due to the characteristics of available data, this experiment was limited to data from Kansai broadcasters over a 2.5 month period. Kansai is the second largest television market in Japan, and demonstrating the effectiveness of our model using data from such a large market is particularly meaningful. Future work should examine the generalizability of the model to datasets from different seasons, regions beyond Kansai, and countries with cultural contexts distinct from Japan. We also plan to explore deep learning architectures such as Transformers and RNNs, which may capture temporal dependencies more effectively than current tree-based methods. A critical advancement involves evolving from a static digital twin to a real-time system. This would enable continuous model adaptation in response to changing viewer behaviors and support immediate programming decisions through reinforcement learning approaches. Rigorous validation of synthetic viewing data against actual patterns is essential, employing statistical measures such as Kolmogorov-Smirnov tests for temporal distributions. However, comprehensive validation across diverse demographic segments and seasonal variations remains necessary.

To move toward practical deployment, we plan to integrate the proposed model with existing broadcasting systems such as program scheduling and audience analytics platforms. This requires addressing technical challenges including inconsistent data formats, timestamp misalignment, and the lack of standardized data exchange protocols across broadcasters. Developing interoperable APIs that ensure privacy-preserving collaboration will be key to system scalability. In addition, professional acceptance must be considered by improving visualization and explainability so that media experts can trust and utilize model outputs. To evaluate both technical and organizational feasibility, we plan a pilot implementation with a broadcaster's analytics division using live viewing logs, which will serve as a foundation for real-world deployment.

Finally, we aim to create comprehensive media consumption profiles by integrating television viewing data with streaming platform behaviors while preserving privacy through federated learning techniques. These technical advances will enable novel applications including predictive content recommendation using graph neural networks and automated programming suggestions based on successful viewing patterns. The ultimate goal is to establish an intelligent broadcasting system that realizes content delivery adapting to viewer preferences while respecting privacy boundaries, transforming television from a one-way broadcast medium into a personalized yet communal media experience.

VII. CONCLUSION

This study introduced novel feature designs specific to television viewing behavior and used them to classify viewing tendencies into interpretable behavioral classes. Based on these classifications, highly accurate models for weekdays and weekends were developed. The validation of these models showed consistently high performance across all evaluation metrics for both the weekday and weekend models, confirming their overall effectiveness. Conversely, the 36-class combined model integrating weekday and weekend classifications showed lower accuracy due to the complexity of viewing tendencies and data imbalance among classes.

These results demonstrated the feasibility of using classification-based modeling to replicate diverse viewer behavior patterns. Future work will generate synthetic viewing data based on these classifications and statistically compare them with actual viewing data to validate the synthetic data's utility. This approach provides a solid foundation for realizing a digital twin of television viewing behavior, offering potential for practical applications and further advancements.

REFERENCES

[1] Secure Broadcasting Autorization and Research Center. (2024). *Practices for Handling Non-Identifiable Television Viewing History Data Obtained By the OPT-Out Method*. [Online]. Available: https://www.sarc.or.jp/documents/www/NEWS/hogo/2024/optoutpracticever2.3.pdf

[2] A. Hughes, "Forging the digital twin in discrete manufacturing: A vision for unity in the virtual and real worlds," Tech. Rep., 2018.

[3] X. Zhou, W. Liang, K. Yan, W. Li, K. I. Wang, J. Ma, and Q. Jin, "Edge-enabled two-stage scheduling based on deep reinforcement learning for Internet of Everything," *IEEE Internet Things J.*, vol. 10, no. 4, pp. 3295–3304, Feb. 2023.

[4] X. Zhou, X. Xu, W. Liang, Z. Zeng, S. Shimizu, L. T. Yang, and Q. Jin, "Intelligent small object detection for digital twin in smart manufacturing with industrial cyber-physical systems," *IEEE Trans. Ind. Informat.*, vol. 18, no. 2, pp. 1377–1386, Feb. 2022.

[5] A. J. B. Chaney, M. Gartrell, J. M. Hofman, J. Guiver, N. Koenigstein, P. Kohli, and U. Paquet, "A large-scale exploration of group viewing patterns," in *Proc. ACM Int. Conf. Interact. Exper. TV Online Video*, Jun. 2014, pp. 31–38, doi: 10.1145/2602299.2602309.

[6] M. Tavakoli and M. Cave, "Modelling television viewing patterns," *J. Advertising*, vol. 25, no. 4, pp. 71–86, Dec. 1996, doi: 10.1080/00913367.1996.10673513.

[7] W. Kritzinger, M. Karner, G. Traar, J. Henjes, and W. Sihn, "Digital twin in manufacturing: A categorical literature review and classification," *IFAC-PapersOnLine*, vol. 51, no. 11, pp. 1016–1022, 2018.

[8] S. Liao, J. Wu, A. K. Bashir, W. Yang, J. Li, and U. Tariq, "Digital twin consensus for blockchain-enabled intelligent transportation systems in smart cities," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 22619–22629, Nov. 2022.

[9] J. O'Sullivan, D. O'Sullivan, and K. Bruton, "A case-study in the introduction of a digital twin in a large-scale smart manufacturing facility," *Proc. Manuf.*, vol. 51, pp. 1523–1530, Jan. 2020.

[10] R. Wang, X. Zhou, L. Dong, Y. Wen, R. Tan, L. Chen, G. Wang, and F. Zeng, "Kalibre: Knowledge-based neural surrogate model calibration for data center digital twins," in *Proc. 7th ACM Int. Conf. Syst. Energy-Efficient Buildings, Cities, Transp.*, Nov. 2020, pp. 200–209.

[11] M. Li, R. Wang, X. Zhou, Z. Zhu, Y. Wen, and R. Tan, "ChatTwin: Toward automated digital twin generation for data center via large language models," in *Proc. 10th ACM Int. Conf. Syst. Energy-Efficient Buildings, Cities, Transp.*, Nov. 2023, pp. 208–211, doi: 10.1145/3600100.3623719.

[12] Q. Min, Y. Lu, Z. Liu, C. Su, and B. Wang, "Machine learning based digital twin framework for production optimization in petrochemical industry," *Int. J. Inf. Manage.*, vol. 49, pp. 502–519, Dec. 2019.

[13] H. Elayan, M. Aloqaily, and M. Guizani, "Digital twin for intelligent context-aware IoT healthcare systems," *IEEE Internet Things J.*, vol. 8, no. 23, pp. 16749–16757, Dec. 2021.

[14] Y. Chen, W. Wang, J. Diao, D. Wang, Z. Jian, Y. Wang, and Z. Jiang, "Digital-twin-based patient evaluation during stroke rehabilitation," in *Proc. ACM/IEEE 14th Int. Conf. Cyber-Phys. Syst. (CPS-IoT Week)*, May 2023, pp. 22–33, doi: 10.1145/3576841.3585923.

[15] T. Das, Z. Wang, and J. Sun, "TWIN: Personalized clinical trial digital twin generation," in *Proc. 29th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2023, pp. 402–413, doi: 10.1145/3580305.3599534.

[16] G. Fortino and C. Savaglio, "Integration of digital twins & Internet of Things," in *The Digital Twin*, 2023, pp. 205–225, doi: 10.1007/978-3-031-21343-4_8.

[17] A. Rudskoy, I. Ilin, and A. Prokhorov, "Digital twins in the intelligent transport systems," *Transp. Res. Proc.*, vol. 54, pp. 927–935, Jan. 2021.

[18] S. A. P. Kumar, R. Madhumathi, P. R. Chelliah, L. Tao, and S. Wang, "A novel digital twin-centric approach for driver intention prediction and traffic congestion avoidance," *J. Reliable Intell. Environments*, vol. 4, no. 4, pp. 199–209, Dec. 2018.

[19] E. G. Kaigom and J. Roßmann, "Value-driven robotic digital twins in cyber–physical applications," *IEEE Trans. Ind. Informat.*, vol. 17, no. 5, pp. 3609–3619, May 2021.

[20] R. Saracco, "Digital twins: Bridging physical space and cyberspace," *Computer*, vol. 52, no. 12, pp. 58–64, Dec. 2019.

[21] A. Said, S. Berkovsky, and E. W. De Luca, "Group recommendation in context," in *Proc. 2nd Challenge Context-Aware Movie Recommendation*, Oct. 2011, pp. 2–4, doi: 10.1145/2096112.2096113.

[22] C. Senot, D. Kostadinov, M. Bouzid, J. Picault, A. Aghasaryan, and C. Bernier, "Analysis of strategies for building group profiles," in *Proc. Int. Conf. User Model.*, 2010, pp. 40–51.

[23] J. G. Webster and J. J. Wakshlag, "The impact of group viewing on patterns of television program choice," *J. Broadcast. Electron. Media*, vol. 26, no. 1, pp. 445–455, Jan. 1982.

[24] R. T. Rust, W. A. Kamakura, and M. I. Alpert, "Viewer preference segmentation and viewing choice models for network television," *J. Advertising*, vol. 21, no. 1, pp. 1–18, Mar. 1992.

[25] G. Gim, H. Jeong, H. Lee, and D. Yun, "Group-aware prediction with exponential smoothing for collaborative filtering," in *Proc. 2nd Challenge Context-Aware Movie Recommendation*, Oct. 2011, pp. 11–14, doi: 10.1145/2096112.2096115.

[26] J. Gorla, N. Lathia, S. Robertson, and J. Wang, "Probabilistic group recommendation via information matching," in *Proc. 22nd Int. Conf. World Wide Web*, May 2013, pp. 495–504.

[27] X. Hu, X. Meng, and L. Wang, "SVD-based group recommendation approaches: An experimental study of moviepilot," in *Proc. 2nd Challenge Context-Aware Movie Recommendation*, Oct. 2011, pp. 23–28, doi: 10.1145/2096112.2096117.

[28] H. Matsuda, T. Sakakibara, Y. Matsuda, T. Mizumoto, and K. Yasumoto, "Proposal and practice of integration method for large scale TV viewing log data across broadcasters," *IPSJ Trans. Digit. Practices*, vol. 4, no. 1, pp. 34–44, 2023.

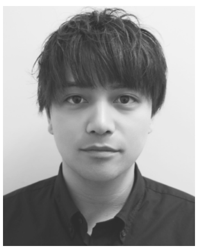
[29] H. Matsuda, T. Sakakibara, D. Mayumi, Y. Matsuda, T. Mizumoto, and K. Yasumoto, "A method for inter-broadcaster data integration of non-identifiable television viewing log," *IPSJ J.*, vol. 65, no. 10, pp. 1488–1500, 2024.

[30] Yomiuri Telecasting Corporation. (2022). *Joint Technical Verification Experiment on Television Viewing Data Integration (FY2022)*. [Online]. Available: https://www.ytv.co.jp/privacy/experiments2022/index.html

[31] Research and Statistics Department, Statistics Bureau of Japan. (2020). *Result of National Census in 2020*. [Online]. Available: https://www.stat.go.jp/data/kokusei/2020/kekka.html

[32] Research and Statistics Department, Statistics Bureau of Japan. (2019). *Housing and Land Survey in 2019*. [Online]. Available: https://www.stat.go.jp/data/jyutaku/2018/pdf/ggaiyou.pdf

[33] Bureau of Urban Development, Tokyo Metropolitan Government. (2013). *Result of a Mansion Finding Survey*. [Online]. Available: https://www.mansion-tokyo.metro.tokyo.lg.jp/pdf/03jittai-chousa/03jittai-chousa-01.pdf



DAIKI MAYUMI received the B.E. degree in engineering from Osaka Institute of Technology, Osaka, Japan, in 2021, and the M.E. degree from the Graduate School of Science and Technology, Nara Institute of Science and Technology, Japan, in 2023, where he is currently pursuing the Ph.D. degree. His research interests include ubiquitous computing, olfactory interface, and human-computer interaction.

HIROKI MATSUDA received the master's degree (M.E.) from the Graduate School of Information Science, Nara Institute of Science and Technology, in 2011, and the Ph.D. degree from the Graduate School of Information Science, in 2024. He is currently affiliated with Yomiuri Telecasting Corporation. His current research interests include big data analysis and behavioral analysis centered on television viewing data analysis.

TETSUYA YOKOTA received the M.E. degree in electrical and electronic engineering from Osaka Prefecture University, in 2018. He was an Embedded Systems Engineer with the Mitsubishi Electric Corporation. He is currently a Research and Development Engineer with Yomiuri Telecasting Corporation. His research interests include software development, image processing, and big data analytics.

TAICHI SAKAKIBARA received the B.E. degree in engineering from Kyoto University, in 2018, and the M.E. degree from the Graduate School of Kyoto University, in 2020. He is currently affiliated with Yomiuri Telecasting Corporation. His research interest includes optimization of Wi-Fi selection using reinforcement learning.

YUKI MATSUDA (Member, IEEE) was born in 1993. He received the B.E. degree in advanced course of mechanical and electronic system engineering from the National Institute of Technology, Akashi College, Japan, in 2015, and the M.E. and Ph.D. degrees from the Graduate School of Information Science, Nara Institute of Science and Technology, Japan, in 2016 and 2019, respectively. He is currently a Lecturer with the Faculty of Environmental, Life, Natural Science and Technology, Okayama University, Japan. His current research interests include participatory sensing, location-based information systems, wearable computing, and affective computing. He is a member of IPSJ, IEICE, JSAI, and ACM.

KEIICHI YASUMOTO (Member, IEEE) received the B.E., M.E., and Ph.D. degrees in information and computer sciences from Osaka University, Osaka, Japan, in 1991, 1993, and 1996, respectively. He is currently a Professor with the Graduate School of Science and Technology, Nara Institute of Science and Technology. His research interests include distributed systems, mobile computing, and ubiquitous computing. He is a member of ACM, IPSJ, and IEICE.