Journal of Advances in Information Technology, Vol. 16, No. 6, 2025

Event-Driven Crowd Forecasting: Wi-Fi Sensing, Event Schedules, and Weather Data for Foot Traffic Prediction

Kurumi Muto[1,*], Akihisa Kodate[1], Noboru Sonehara[2], and Nobuharu Hiruma[3]
[1] College of Policy Studies, Tsuda University, Tokyo, Japan
[2] Research Institute for Policy Studies, Tsuda University, Tokyo, Japan
[3]AREAPORTAL, INC, Tokyo, Japan
Email: p20080mk@gm.tsuda.ac.jp (K.M.); kodate@tsuda.ac.jp (A.K.); sonehara@tsuda.ac.jp (N.S.); hiruma@areaportal.co.jp (N.H.)
*Corresponding author

***Abstract*—Large stadium events, such as sports games and concerts, have a significant impact on pedestrian flow, potentially resulting in missed business opportunities and threats to public safety. This study addresses the issue of sudden crowd surges caused by large-venue events by developing a predictive Big Data framework that integrates Wi-Fi sensing data, event schedules, and meteorological variables to forecast foot traffic. The study focuses on Sendagaya, Tokyo, an area that hosts multiple large-capacity venues. Using Random Forest regression, we evaluated our model that incorporates event and weather data and compared it with our baseline model, which uses only temporal features. Our results indicate that while temporal data alone yield stable predictions, incorporating event-specific and weather-related factors enhances the prediction accuracy of foot traffic peaks during large events. This research offers a scalable approach to urban crowd management, combining practical data collection using web scraping and multivariate modeling. These findings contribute both to Big Data modeling techniques and applications in local government, societal infrastructure planning, and marketing.**

***Keywords*—urban computing, foot traffic, Wi-Fi sensing, crowd prediction, cross-domain data**

I. Introduction

Urban districts with clusters of large-scale venues face unique challenges in managing pedestrian flow, especially during events that trigger sudden crowd surges. Inadequate preparation for surges of high-volume traffic poses risks to public safety and challenges to urban infrastructure. Sendagaya, a district in Tokyo with a significant concentration of high-capacity venues and neighboring major downtown areas of Shinjuku and Harajuku, exemplifies such a setting. This study specifically focuses on multiple large-scale event venues around Sendagaya Station, including National Stadium, Meiji Jingu Baseball Field, Chichibunomiya Rugby Stadium, Tokyo Metropolitan Gymnasium, and the National Noh Theater. Figure 1 illustrates these exact locations. Table I provides details on the capacity of each facility, with the National Stadium being the largest, accommodating up to 67,750 attendees. The high capacity and number of venues indicate the importance of event schedules in predicting foot traffic in the area, which is why they are incorporated into our models.

Despite the concentration of large-scale event venues, Sendagaya remains relatively quiet during non-event periods. Although Sendagaya Station is only two stops from Shinjuku Station, one of the busiest rail hubs worldwide, with an average of 650,602 daily passengers in 2023, Sendagaya Station recorded a modest daily average of only 16,567 passengers in the same year [1]. Besides the major venues, Sendagaya hosts the "Sendagaya Odori Shopping Street," featuring local stores and franchise shops, as well as colleges and offices. These businesses contribute to regular pedestrian flow during non-event times. Evidently, the foot traffic volume at Sendagaya exhibits a stark contrast between regular operation and periods of peak traffic, prompting a challenge to urban planning.

Peak operation in the Sendagaya area imposes significant stress on nearby commuter infrastructure. For example, one author has observed that, after a soccer match at the National Stadium, the commute time to Sendagaya Station increased to more than 20 minutes, compared with the usual 10 minutes. This is because the streets were packed, causing people to become stationary along the route. Such a sharp increase in the volume of pedestrians can pose substantial challenges for urban management and public safety. Furthermore, a previous study [2] has shown the drastically varying pedestrian volume in Sendagaya during the Tokyo 2020 Olympics and Paralympics, confirming the potential for fluctuation of foot traffic in the area. Historical incidents, such as the 2022 Halloween crowd crush in Itaewon, Korea, which resulted in more than 150 deaths [3], and the 2014 new year's eve stampede in Shanghai, in which 36 people lost their lives [4], underscore the gravity of these risks when pedestrian surges are not properly anticipated and managed.

Despite increasing attention to crowd management, existing studies have often focused on univariate prediction or short-term prediction. They have focused only on historical foot traffic without incorporating external factors such as event schedules and weather conditions, which may significantly influence foot traffic.

For example, previous research by Singh *et al.* [5] has predominantly explored univariate predictions based solely on

Manuscript received January 2, 2025; revised January 14, 2025; accepted February 26, 2025; published June 26, 2025.

doi: 10.12720/jait.16.6.916-926

historical foot traffic data. Abrishami *et al.* [6] similarly developed models using univariate Wi-Fi sensor data to forecast foot traffic in retail environments. Other studies, such as Tsukada *et al.* [7], primarily emphasized short-term predictions, often limited to a few hours or a couple of days in advance. Jiang *et al.* [8] also focused on real-time forecasts using GPS data and LSTM models for event-related crowd dynamics. By incorporating both event-specific and meteorological data into the predictive model, this research addresses the limitations of existing crowd prediction models that primarily rely on short-term or univariate analyses. This approach aims to create a more practical framework for foot traffic forecasting and risk management. The reason that event-specific variables are incorporated is grounded in a general assumption that event schedules influence pedestrian flow in multivenue districts such as Sendagaya. Similarly, weather data are incorporated, as routine foot traffic from offices, schools, and commercial areas is generally influenced by weather conditions.

The research follows three key steps: (1) collect foot traffic data using Wi-Fi sensors installed in Sendagaya, (2) extract event schedule data from official venue websites through web scraping, and (3) develop machine-learning models to assess the effectiveness of event-specific and weather-related variables in foot traffic. This approach aims to determine the utility of event-specific and meteorological information to predict foot traffic in an event-dense area, both for short-term and long-term predictions. Furthermore, the resulting model is highly applicable for predicting foot traffic in mixed-use areas that experience varying pedestrian volumes. The findings of this research demonstrate that while traditional time-based models offer a baseline accuracy (R^2 score of 0.74), incorporating event-specific and weather-related data yields a more nuanced prediction of pedestrian surges. These results establish a predictive framework that is not only relevant for Sendagaya but can be adapted to other event-dense urban areas to improve public safety.

TABLE I: List of facilities in Sendagaya and their capacities [9]–[13].

Facility Name	Capacity(Persons)
National Stadium	67,750
Meiji Jingu Baseball Field	30,703
Chichibunomiya Rugby Field	24,871
Tokyo Metropolitan Gymnasium	10,000
National Noh Theater	627

II. Literature Review

A. Foot Traffic Prediction using Wi-Fi-sensing Data

Wi-Fi sensing data have been widely used to analyze and predict pedestrian behavior in urban environments. Singh *et al.* [5] developed a model based on long-short-term memory (LSTM) to predict crowd densities during large-scale events in Brussels, Belgium. Their approach utilized univariate time-series data from Wi-Fi probe requests, experimenting with multiple LSTM variants. Similarly, Abrishami *et al.* [6] developed a foot traffic prediction system for retail stores using

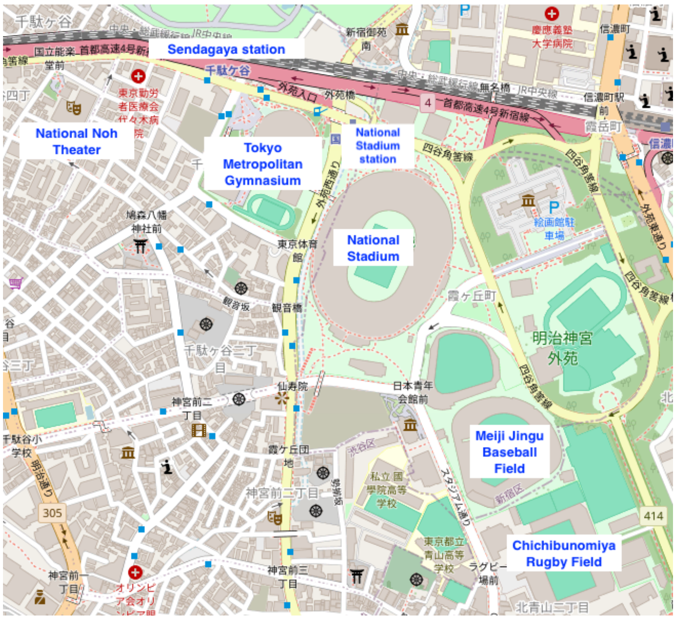

Fig. 1. Map of Sendagaya with stadiums and venues. Map data from OpenStreetMap (openstreetmap.org/copyright).

Wi-Fi sensor data collected from over 50 locations across the United States.

These models are primarily univariate and focus on short-term foot traffic, limiting their applicability in event-heavy areas such as Sendagaya. This study extends these approaches by incorporating event and weather data for a more nuanced long-term predictive model.

B. Crowd Forecast in Event-Based Contexts

Predicting crowd dynamics during events presents unique challenges owing to sudden and irregular surges in the volume of pedestrians. Tsukada *et al.* [7] developed a predictive model for crowd dynamics around large venues in the Greater Tokyo Area by integrating historical GPS congestion data with microblog posts (e.g., X, formerly Twitter). Their research is grounded in the fact that there are always posts about large-scale events on social media, and their model demonstrates the effectiveness of using event-specific social media posts as indicators of crowd density. They achieved higher accuracy than traditional baselines did, indicating the importance of incorporating event-specific features in crowd prediction. However, the model fails to extrapolated for longer-term forecasts (48 and 72 hours), since the relevant posts are generally shared only a day before the event.

Similarly, Jiang *et al.* [8] proposed a deep learning framework for city-wide crowd forecasting during large-scale events. The system, employing a multitask convolutional long short-term memory (LSTM), simultaneously predicts crowd density and flow based on GPS location data, using time-lag variables. However, the system's purpose is to predict real-time crowd dynamics up to 60 minutes in advance, and it is limited to the short-term prediction usage.

Inspired by [7], this study leverages event schedule data, which are often available weeks or months in advance on official venue websites. In addition, similar to [8], this current study employs time-lag variables from the previous foot traffic counts, and determines the effect of event schedule data by comparing the result when only time-lag variables are used versus when event schedule data are also used.

C. Crowd Analysis in Sendagaya

Traganmaturapot *et al.* [2] analyzed pedestrian patterns in Sendagaya using Wi-Fi sensor data during the Tokyo 2020 Olympics and Paralympics. The study revealed significant fluctuations in crowd movement across different times and events. For instance, a sharp increase in pedestrian volume was observed during the Olympics' opening ceremony, while the Paralympics' opening ceremony experienced only a modest rise.

Although the study effectively captured the temporal dynamics of foot traffic and revealed the great variance in Sendagaya's foot traffic, it did not focus on developing predictive models. The present research expands on this research by developing a predictive model that incorporates historical Wi-Fi data.

While previous studies demonstrate the potential of Wi-Fi and other data sources for foot traffic and crowd prediction, existing models are often limited to short-term, univariate approaches or specific scenarios. This study expands the scope by integrating multiple data sources - Wi-Fi, event, and weather data - within a multivariate long-term framework, offering a more comprehensive approach to prediction of foot traffic in urban areas with mixed use and heavy events.

D. Crowd Analysis in Sendagaya

Traganmaturapot *et al.* [2] analyzed pedestrian patterns in Sendagaya using Wi-Fi sensor data during the Tokyo 2020 Olympics and Paralympics. The study revealed significant fluctuations in crowd movement across different times and events. For instance, a sharp increase in pedestrian volume was observed during the Olympics' opening ceremony, while the Paralympics' opening ceremony experienced only a modest rise.

In another line of work, Traganmaturapot *et al.* [14] compared normal-period foot traffic against that on a major fireworks festival day in the Sendagaya area, demonstrating that crowd levels spiked most sharply just before the official start time of the fireworks (19:30). They also observed that while height of surrounding buildings did not affect foot traffic patterns, the proximity to event venues and the precise event scheduling significantly influenced visitor surges. These findings underscore the potential for sudden but temporary crowd densities during large-scale festivities. However, the authors primarily examined only one major festival and did not develop an extended predictive model spanning multiple event factors or time horizons. The present study addresses these gaps by integrating diverse event schedules (e.g., baseball games, concerts) and weather indicators in a predictive framework suitable for both short- and long-term foot traffic forecasting.

III. Data Collection

This study utilizes foot traffic data collected from Wi-Fi sensors, event schedule data obtained from venues' official websites, and weather data. This section elaborates on the methodologies used to collect data.

A. Foot Traffic Data

1) Wi-Fi Sensors: This study utilizes foot traffic data collected through multiple Wi-Fi sensors deployed in Sendagaya. These sensors capture probe requests emitted by Wi-Fi-enabled devices, such as smartphones, which are periodically broadcast to detect nearby Wi-Fi access points. The sensors identify individual devices using their unique media access control (MAC) addresses. To address privacy concerns, the collected MAC addresses are hashed, thereby the identifiers are anonymized. In other words, the original device cannot be inferred from the stored identifier.

It is important to note that Wi-Fi sensors do not detect all Wi-Fi-enabled devices; results in a gap between the actual foot traffic count and the detected count. However, previous research has reported a strong correlation between the number of detected Wi-Fi signals and the actual number of pedestrians. Fujii *et al.* [15] investigated the correlation between the number of Wi-Fi detections and the number of passengers passing through the gates at Kintetsu Nara Station. The research yielded a high correlation coefficient of $R^2 = 0.7179$, suggesting that although Wi-Fi detection does not capture all pedestrians, it provides a reliable measure to estimate actual foot traffic. The research further developed a model to estimate the actual number of passengers at Kintetsu Nara Station from Wi-Fi detection data using a carrier network. However, this modeling is beyond the scope of this study.

2) Locations: The geographic distribution of the Wi-Fi sensors is depicted in Figure 2. The coverage of each sensor is represented by the size of circles with a radius of 100 meters. This corresponds to the typical range for capturing Wi-Fi probe requests. Each sensor is assigned a unique identifier, such as "s01". This research focuses specifically on the access point labeled "s13", as it is located at a key intersection frequently traversed by individuals heading to National Stadium, Chichibunomiya Rugby Field, Meiji Jingu Baseball Field, and Tokyo Metropolitan Gymnasium. Figure 3 provides a street view of the s13 access point, which illustrates its urban environment.

3) API: The foot traffic data collected are made accessible via an API endpoint, which provides records in a format with columns as defined in Table II. The dataset is updated at three-minute intervals, meaning that each entry represents the cumulative foot traffic count over the past three minutes.

B. National Noh Theater's Schedule

Historical event schedule data for the National Noh Theater were obtained from the official website [16]. The website requires manual interaction to load the entire event list (e.g., scrolling and clicking the "more events" button). Thus, automated web scraping tools– specifically, WebDriver and ChromeDriver–, were employed to simulate user behavior.

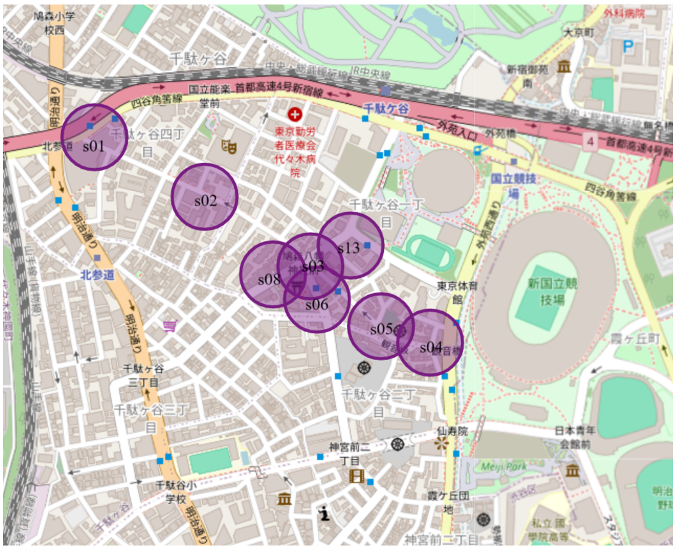

Fig. 2. Locations of Wi-Fi sensors.

Fig. 3. Street View of Access Point 13, Taken by the Authors.

TABLE II: Description of columns in foot traffic data.

Column	Meaning
Access Point	Unique identifier for each Wi-Fi sensor
Date	Date
Time	Time
Latitude	Latitude of the sensor
Longitude	Longitude of the sensor
Count	Cumulative foot traffic count from the previous entry

WebDriver is an open-source tool to automate web browser interactions, and ChromeDriver is a specific implementation that enables WebDriver to control Google Chrome [17]. The final dataset is organized into columns listed in Table III.

To maintain ethical standards during data collection, the automated system was designed to send requests at three-second intervals to avoid overwhelming the server. Furthermore, the website's terms of use were reviewed, confirming that there were no restrictions against automated browsing or data extraction.

TABLE III: Description of each column in the National Noh Theater event dataset.

Column	Description
Year	Year of the event
Month	Month of the event
Day	Day of the event
Genre	Type of performance (mostly "kyogen/noh", but sometimes "others")
Event Name	Name of the event
Start Time	Scheduled time for event starting if provided

C. Meiji Jingu Baseball Field's Schedule

Historical event data for Meiji Jingu Baseball Field were acquired from the official website [18]. Similar to the approach used for the National Noh Theater's data collection, WebDriver and ChromeDriver were utilized to automate interactions with the website, given that it required manual navigation (e.g., clicking through different event listings). A review of the website's terms of service confirmed that no restrictions were imposed on automated browsing activities. The resulting cleaned dataset comprises the features described in Table IV.

It should also be noted that the Meiji Jingu Baseball Field is slated for reconstruction as part of the Urban Redevelopment Project for the Jingu Gaien area [19]. Although detailed plans for the redevelopment have yet to be disclosed, the project received formal approval from the Tokyo Metropolitan Government in February 2023. Currently, the stadium remains operational; thus, this study does not address the future complexities introduced by the planned redevelopment.

D. National Stadium and Chichibunomiya Rugby Field

Both National Stadium and Chichibunomiya Rugby Field are managed by the Japan Sport Council, and their official sites share the same domain. Although only National Stadium has a publicly accessible archive of past events [20], the domain

TABLE IV: Description of columns in the Meiji Jingu Baseball Field event dataset.

Column	Description
Year	Year of the event
Month	Month of the event
Day	Day of the event
Event Name	Name of the event
Start Time	Scheduled time for event starting if provided
Is Professional Baseball	Indicator if the event is a professional baseball game (True/False)

has a public endpoint that offers historical event data for all its facilities, including the Yoyogi National Gymnasium. In this study, data were retrieved from this endpoint and systematically classified according to the venue.

The classification was performed as follows: (1) events that match entries in the National Stadium's historical archive were categorized under "National Stadium". (2) For the remaining events, if the textual description contained the keyword "rugby," they were assigned to "Chichibunomiya Rugby Field". This classification approach is based on the fact that the Chichibunomiya Rugby Field predominantly hosts rugby events, whereas other venues, such as Yoyogi National Gymnasium, typically do not host rugby matches. This enables relatively accurate categorization based on textual features.

Furthermore, the Chichibunomiya Rugby Field is also scheduled to undergo redevelopment as part of the Urban Redevelopment Project for the Jingu Gaien Area [19], and the newly reconstructed facility is expected to support a broader range of events beyond rugby [21]. This development may introduce additional complexities to future classification tasks; however, this study does not account for them as the field is still operational and detailed plans for reconstruction have yet to be announced.

The structure of the data retrieved is described in Tables V and VI, which detail the format of national stadium and Chichibunomiya Rugby Field data, respectively.

TABLE V: Description of columns in the National Stadium dataset.

Column	Meaning
Year	Year the event
Month	Month of the event
Day	Day of the event
Genre	Category of the event (e.g., soccer, rugby, field track, other)
Event Name	Name of the event
Open Time	Scheduled time for venue opening if provided.
Start Time	Scheduled time for event starting if provided.

E. Tokyo Metropolitan Gymnasium

Tokyo Metropolitan Gymnasium does not offer access to historical event data through its online platforms. However, the authors were able to obtain archival event records directly from the facility's administration, due to the partnership between the

TABLE VI: Description of columns in the Chichibunomiya Rugby Field dataset.

Column	Meaning
Year	Year the event
Month	Month of the event
Day	Day of the event
Event Name	Name of the event
Open Time	Scheduled time for venue opening if provided.
Start Time	Scheduled ending time for the event if provided.

authors' affiliated university and the gymnasium. The original data files were provided in Rich Text Format (RTF), a format that is not structured in a tabular manner and is therefore unsuitable for direct data analysis. To enable effective processing, the authors used ChatGPT (version o1-preview) within a temporary session, ensuring that data remained protected and would not be incorporated into model training or shared externally. ChatGPT was used to transform and reorganize the unstructured RTF files into a structured CSV format. The resulting dataset includes the columns specified in Table VII.

TABLE VII: Description of columns in the Tokyo Metropolitan Gymnasium dataset.

Column	Meaning
Year	Year the event
Month	Month of the event
Day	Day of the event
Event Name	Title of the event.
Open Time	Scheduled time for venue opening if provided.
Start Time	Scheduled time for event starting if provided.
End Time	Scheduled ending time for the event if provided.

F. Weather Data

Meteorological data for this research were obtained from the Japan Meteorological Agency through its official website [22]. The data correspond to the Tokyo observation point, specifically the Automated Meteorological Data Acquisition System (AMeDAS) station located in the Chiyoda Ward [23]. As the Chiyoda Ward is also located in Tokyo, it was reasonable to use the data from this AMeDAS station for Sendagaya. The weather variables incorporated in this study were temperature (°C), precipitation levels (mm), and snow accumulation (cm). The granularity of the data is by hour.

IV. Model Development

The collected data were used to train models to predict foot traffic at access point s13. This study involved two experiments for model building.

1) **A long-term prediction** comparing models trained with only time-related characteristics (baseline) versus models that incorporate events and weather-related variables (visually represented in Figure 4).
2) **A short-term prediction** comparing models trained with recent foot traffic delays, time-related characteristics, and weather-related and event-related variables (visually represented in Figure 5).

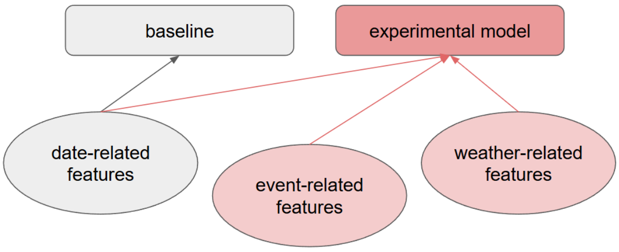

Fig. 4. Diagram to Explain the Long-term Prediction Models.

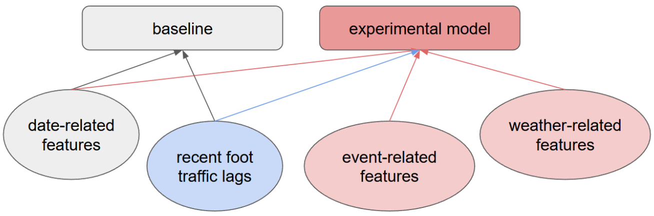

Fig. 5. Diagram to Explain the Short-term Prediction Models.

The decision to incorporate long- and short-term predictions was driven by the distinct practical needs that they address. Long-term predictions are particularly valuable for applications such as marketing strategy development, and infrastructure planning, in which sufficient lead time is essential for effective decision-making and implementation. Conversely, short-term predictions play a critical role in real-time optimization, enabling the dynamic adjustment and refinement of pre-established strategies in response to immediate conditions.

Both sections employ Random Forest models to determine the effectiveness of event-related and weather-related information in Sendagaya by comparing results when only time features (e.g., hour, minute, Boolean flags for whether it is lunchtime) are used versus when exogenous factors are included.

An ensemble learning method was chosen over neural network models like LSTM due to its emphasis on long-term prediction without a heavy dependence on recent foot traffic data. Recurrent neural network models such as LSTMs depend on recent time-series data for accurate prediction, which limits their applicability for long-term forecasts that require dependency on recent data. Therefore, Random Forest was chosen to better capture the relationships between exogenous factors (i.e., event schedules and weather) and foot traffic patterns, without being overly influenced by the short-term temporal dependencies that LSTMs typically prioritize.

Furthermore, Random Forest was chosen over other common ensemble methods, such as XGBoost, due to its stability in predicting foot traffic peaks. In our experiment, XGBoost's predicted peaks showed significant deviations from actual foot traffic peaks, as demonstrated in the comparison graph (Figure 6). In contrast, Random Forest provided more stable and moderate peak predictions that did not produce unrelated or spurious peaks. This stability is attributed to Random Forest's bagging mechanism, which averages predictions across

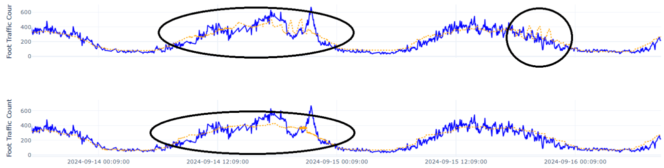

Fig. 6. Comparison of XGBoost and Random Forest predictions of foot traffic throughout the day.

Note: This date was chosen for demonstration purposes. The same trend was observed on other days as well.

multiple decision trees, inherently reducing the impact of extreme outliers and noisy fluctuations. This characteristic made Random Forest the optimal choice for our study.

A. Random Forest

Random Forest is an ensemble learning technique widely used for regression tasks. It builds multiple decision trees during the training process and aggregates their predictions by averaging them, thereby reducing overfitting and variance. The prediction from a Random Forest model can be mathematically expressed as:

$$\hat{f}(x) = \frac{1}{B} \sum_{b=1}^{B} T_b(x) \tag{1}$$

where B is the total number of trees, and $T_b(x)$ is the prediction of the b-th tree. Random Forest reduces overfitting by averaging multiple deep trees, resulting in a model with lower variance compared to individual decision trees.

B. Parameter Selection

The optimal hyperparameters for the Random Forest model were selected based on the configuration that produced the highest R^2 score. The hyperparameters were chosen from the following ranges:

- `n_estimators_list` = [100]
- `max_depth_list` = [10, 20]
- `min_samples_split_list` = [2, 5]
- `max_features_list` = [None]

C. Evaluation Metrics

To evaluate the performance of the model, we use the root mean squared error (RMSE). RMSE is calculated as:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{2}$$

where n is the total number of observations, y_i represents the actual values, and $\hat{y}_i$ represents the predicted values.

Additionally, we use RMSE percentage (RMSE%), which expresses RMSE as a percentage of the mean of the observed values, providing a relative error measure. RMSE% is calculated as:

$$\text{RMSE\%} = \left(\frac{\text{RMSE}}{\bar{y}} \right) \times 100 \tag{3}$$

where $\bar{y}$ is the mean of the actual values in the test data set.

D. Feature Importance Calculation

To understand the contribution of each feature to the prediction, we use the "gain" method to calculate the importance of the features. In Random Forest, the importance of a feature is measured by the average reduction in prediction error (or "gain") achieved when the feature is used for splitting.

E. Features

The features used to train the Random Forest models were carefully designed based on the observed trends in the dataset. Table VIII lists the features used. The features related to National Noh Theater were intentionally omitted, as they were found to degrade the performance of the model, likely because access point s13 is far from the theater and does not influence crowd behavior near s13.

F. Long-term Prediction

This section presents a comparison of two models:

1) **Baseline model**: This model is trained only with time-related variables, specifically those listed in the "Time" category in Table VIII (e.g., hour, day of the week). This model serves as a baseline to evaluate how well time-related information, excluding event-related and meteorological features, can explain foot traffic patterns.
2) **Extended model**: This model is trained with the full set of features listed in Table IV, including event-related and meteorological variables. If this model performs better, it suggests that event-related and meteorological information significantly enhances the prediction of foot traffic in a multi-venue area.

1) Metrics: Interestingly, incorporating event-related and meteorological features appears to degrade the model performance in terms of R^2 score and RMSE, as indicated in Table IX.

This result indicates that including additional features, such as event-related and meteorological data, might introduce noise or irrelevant information that complicates the model's ability to predict well. One potential reason for this observation is that the weather or event schedules may not always align closely with the foot traffic dynamics at s13, leading to overfitting or erroneous correlations.

2) Observation: Figures 7 and 8 reveal further insights into the model's predictive performance. Figure 8, which integrates event-related and meteorological variables, demonstrates more accurate predictions of pedestrian peaks during major events. In contrast, the model trained solely with time-related variables produces a smoother and more uniform prediction curve. This suggests that while event and meteorological data may not improve overall performance metrics, they contribute to capturing specific peak traffic patterns that are otherwise difficult to predict using only temporal variables. It is important to note that the model does not capture all peaks in foot traffic, and the underlying reasons for this will be addressed shortly.

Additionally, the visualizations of feature importance (Figures 9 and 10) highlight two observations:

1) **Temporal features, particularly the hour, have a strong influence.** As a result, the model trained solely on temporal features is able to deliver relatively stable predictions.
2) **Event-related and meteorological variables, such as temperature and event start times at nearby venues, play a key role.** While the impact of these exogenous variables varies, temperature and event start times are particularly influential in predicting peak events.

One critical limitation of the model is its inability to capture all pedestrian traffic peaks, likely due to missing data. A significant factor contributing to the performance is the large amount of missing start-time data for events at Tokyo Gymnasium, a factor that may have impaired the model's performance. As shown in Figure 10, the visualization of feature importance highlights the utility of event schedule data for Tokyo Gymnasium. However, as detailed in Table X, this dataset contains a substantial amount of missing entries, particularly missing event start times. The absence of such a critical feature likely impaired the model's ability to accurately capture peaks in pedestrian traffic.

This raises an important question about how to ensure the collection of high-quality, cross-domain data. For most venues, web scraping was successfully employed, as the HTML structure within the same domain is typically consistent, allowing for efficient data extraction. However, the files provided directly by Tokyo Gymnasium were an exception. These files contained event information presented primarily in natural language, making them unsuitable for direct data analysis without extensive preprocessing.

To address this, it is essential that nearby venues make event data publicly available on their websites. For cases in which internal datasets are shared, as with Tokyo Gymnasium, ensuring data quality and cleanliness becomes a critical task. Special attention should be paid to preprocessing and cleaning these datasets to minimize the risk of performance degradation in predictive models.

G. Short-term Prediction

This section presents a comparison of two models:

1) **Baseline model**: This model is trained using time-related variables and time-lag features that represent foot traffic counts from one to two hours prior. The choice to use these specific time lags aims to prevent the model from becoming overly reliant on recent foot traffic data. Dependence on very recent data (e.g., within three minutes) would result in predictions limited to an insufficiently short forecast horizon, restricting practical applications. By setting the time lag between one and two hours, the model can forecast foot traffic with a more effective lead time. The autocorrelation pattern in Figure 11 demonstrates that current foot traffic exhibits a significant correlation with foot traffic from 20 to 40 steps prior, which corresponds to data from one to two hours ago. This autocorrelation insight supports the selection of these time lags for model training.

TABLE VIII: Features and Explanation

Category	Feature Name	Explanation
Time	year	Year of the observation time
	month	Month
	day	Day
	hour	Hour
	minute	Minute (approximately 3 minute interval)
	day_of_week	Day of the week
	is_weekend	Indicator for weekend
	is_weekday_commute_morning	True if it's weekday and between 8 to 10 am, false otherwise
	is_weekday_commute_evening	True if it's weekday and between 5 to 7 am.
	is_weekday_lunch_time	True if it's weekday and between 11 am to 2 pm
	is_weekend_lunch_time	True if it's weekend and between 11 am to 2 pm
Weather	temperature(°C)	Temperature in Celsius
	precipitation(mm)	Precipitation in millimeters
	snowfall(cm)	Snowfall in centimeters
Meiji Jingu Baseball Field	jingu_start_until_min	Minutes until event in the field starts (max 6 hours)
	lag_jingu_start_min	Lag time after event in the field starts (max 2 hours)
	is_after_two_hours_from_jingu	Indicator if two hours have passed since the field event started
National Stadium	kokuritu_opening_until_min	Minutes until the stadium opens for an event (max 6 hours)
	lag_kokuritu_event_start_min	Lag time after event in the stadium starts (max 2 hours)
	is_after_two_hours_from_kokuritu_event	Indicator if two hours have passed since event in the stadium started
	has_kokuritu_concert	Indicator for concert
	kokuritu_concert_vs_opening_min	Interaction feature for concert and opening time
	kokuritu_concert_vs_after_two_hours	Binary interaction feature. True if it's concert and post-event time is 2 hours.
Chichibunomiya Rugby Field	chichibunomiya_opening_until_min	Minutes until the field opens for an event
	lag_chichibunomiya_start_min	Lag time after event starts
	is_after_two_hours_from_chichibunomiya	Indicator if two hours have passed since event started
Tokyo Metropolitan Gymnasium	tokyo_gym_opening_until_min	Minutes until the gym opens for an event
	tokyo_gym_start_until_min	Minutes until event starts (max)
	lag_tokyo_gym_end_time	Lag time after event ends (max 2 hours)
	has_tokyo_gym_event	True if the gym has an event in the day
Others	cumulative_stadium_events	Total number of events of the day in the venues

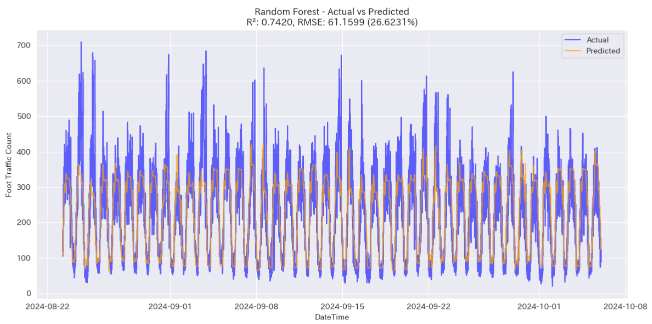

Fig. 7. Actual vs Predicted only with Time-related Variables (without recent lags).

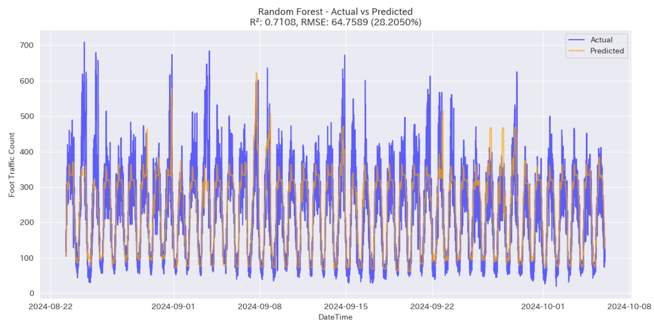

Fig. 8. Actual vs Predicted with Event-Related and Meteorological Variables (without recent lags).

TABLE IX: Comparison of Model Performance Metrics

Model	R^2 Score	RMSE	RMSE %
Time-related features only	0.7420	61.1599	26.6231
All features included	0.7108	64.7589	28.2050

2) **Extended model**: This model extends the baseline by incorporating event schedule data and weather information, along with the same time-lag variables used in the baseline model. Thus, this model uses not only historical traffic patterns, but also exogenous factors that may influence foot traffic patterns.

1) Metrics: The R^2 scores and the RMSE values for both models are similar, as shown in Table XI, indicating that

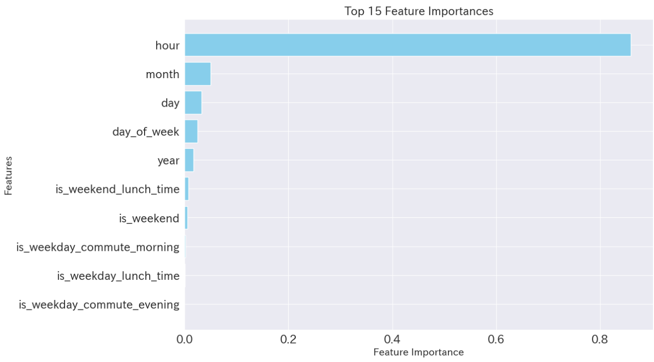

Fig. 9. Feature Importance: Time-Related Variables Only (without recent lags).

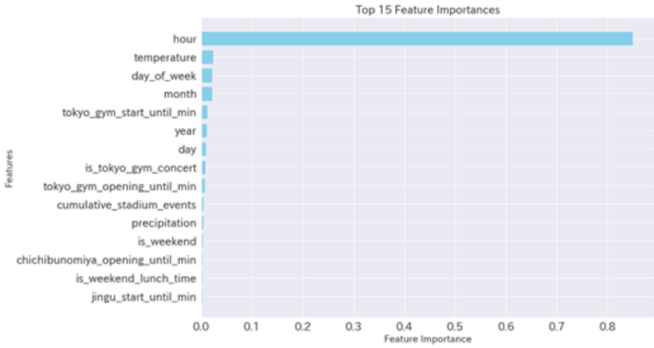

Fig. 10. Feature Importance: Event-Related and Meteorological Variables Included (without recent lags).

TABLE X: Percentage of Nan values in each dataset.

Facility	Percentage of NaN values (%)
National Stadium	13.67
Meiji Jingu Baseball Field	3.60
Chichibunomiya Rugby Field	0
Tokyo Gymnasium	94.62

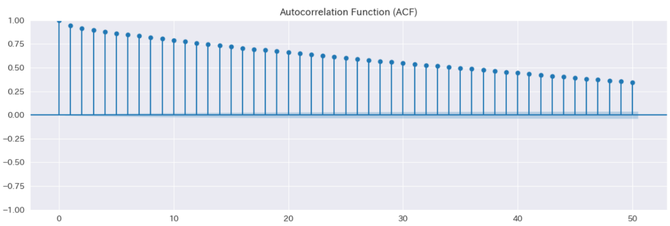

Fig. 11. Autocorrelation.

recent foot traffic data have a strong influence on short-term predictions.

TABLE XI: Comparison of Model Performance Metrics

Model	R^2 **Score**	**RMSE**	**RMSE %**
Time-related features only	0.7898	55.2070	24.0243
All features included	0.7595	59.0502	25.6967

2) Observation: In this experiment, the time-lag variables (representing foot traffic from one to two hours prior) effectively capture sudden surges in pedestrian volume, as illustrated in Figure 12. The baseline model successfully predicts peak periods without the need for supplementary event-related or meteorological variables. In contrast, Figure 13 indicates that the inclusion of these exogenous factors has only a marginal impact on the accuracy of the short-term prediction, reinforcing the importance of recent foot traffic data as the dominant predictor for short-term prediction.

The feature importance analysis (Figures 14 and 15) further confirms that the time-lag variables from recent foot traffic are primary contributors to the model's prediction. Interestingly, certain weather-related features (e.g., temperature, precipitation, and snowfall) also show moderate importance, suggesting their relevance for capturing fluctuations in short-term foot traffic patterns.

Furthermore, selected event-related features, such as *is_tokyo_gym_concert* and *tokyo_gym_opening_until_min*, exhibit notable impact. This implies that while not all exogenous variables enhance predictive accuracy, specific features contribute meaningfully to model performance, highlighting the potential value of targeted feature engineering in refining short-term predictions.

H. Summary of Long- and Short-Term Predictions

Comparing long- and short-term predictive performance reveals distinct dynamics in foot traffic forecasting. While short-term predictions are strongly influenced by recent time-lagged data, long-term forecasts benefit from integrating event-related and meteorological variables, particularly for modeling specific peak traffic.

V. Future Work

This study highlights several avenues for further research to improve foot traffic prediction in multi-venue urban areas.

First, optimizing features specifically for long- and short-term predictions can enhance the accuracy and applicability of the model. Future work will involve refining feature sets based on observed trends, as different types of data may hold varying predictive power over different time horizons.

Second, to make the predictive model actionable in real time, future efforts will focus on establishing a continuous deployment pipeline. This involves regularly scraping data from relevant event websites, such as stadium schedules, and integrating updated meteorological data. Automating data updates and model retraining will allow the system to adapt continuously to changing conditions, providing accurate, real-time forecasts of pedestrian volumes.

Third, expanding the model to include data from multiple access points across the Sendagaya area can improve the spatial accuracy of pedestrian predictions. Integrating data from additional Wi-Fi sensors will allow the model to capture broader movement patterns.

Moreover, beyond predicting crowd density, understanding the purpose of pedestrian movement offers valuable information for targeted public safety and commercial applications.

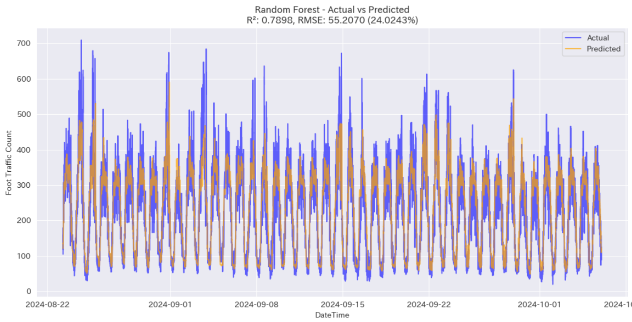

Fig. 12. Actual vs Predicted only with Time-Related Variables (with 1 hour to 2 hour foot traffic count lags).

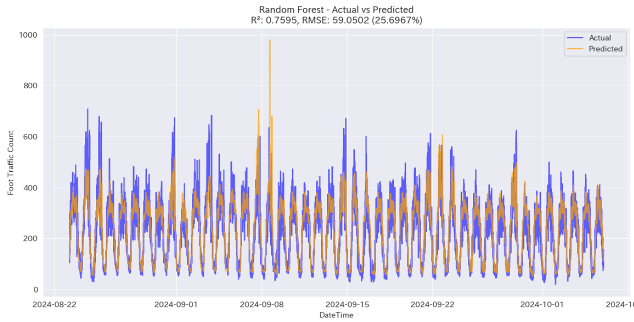

Fig. 13. Actual vs Predicted with Event-Related and Meteorological Variables (with 1 hour to 2 hour foot traffic countlags.

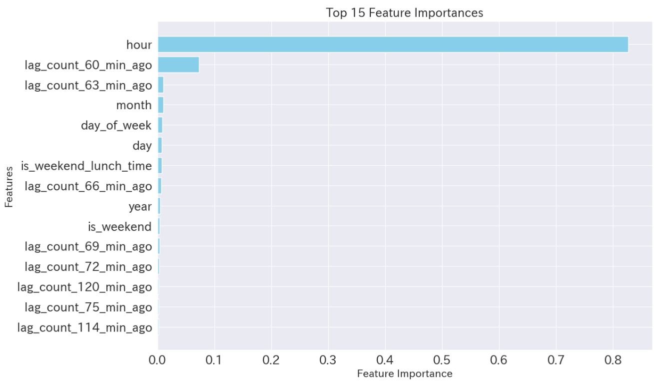

Fig. 14. Feature Importance: Time-Related Variables Only (with recent lags).

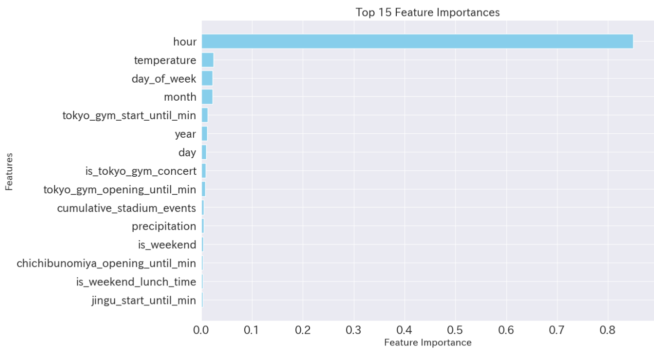

Fig. 15. Feature Importance: Event-Related and Meteorological Variables Included (with recent lags).

Future work will explore how metrics such as walking speed and direction can help distinguish between different intents of pedestrians, such as commuting, attending events, or shopping. Identifying purpose could enhance public safety management by allowing more targeted crowd control measures, while also allowing commercial uses, such as delivering location-based, targeted advertisements to relevant audiences.

VI. Conclusion

This study examined the use of event-specific and meteorological variables to enhance foot traffic prediction in Sendagaya, a district characterized by its concentration of high-capacity venues and fluctuating pedestrian volumes. By comparing baseline models—relying solely on time-related and time-lag features—with extended models incorporating event and weather data, the study evaluated the effectiveness of exogenous factors across both long-and short-term forecasting horizons.

The important findings from this study can be summarizes as following:

1) **In long-term prediction,** without dependency on recent foot traffic lags, **event- and weather-related information significantly improves the accuracy of prediction in foot traffic peaks.**
2) Particularly, **hour, event start times, and temperature** contribute significantly to the model.
3) The model struggled to capture all the peaks, likely because the dataset is missing start times of events in Tokyo Gymnasiums, presenting challenges in collecting cross-domain data.
4) In short-term prediction, with dependency on recent foot traffic lags, event- and weather-related information does not yield significant improvement on the baseline's performance, suggesting the substantial impact of recent foot traffic lags and exogenous features being noise.

Our results demonstrate that event- and weather-related variables significantly enhance long-term prediction accuracy by capturing pedestrian surges associated with major events. This finding is particularly valuable for urban management applications that require long-term foresight, such as public safety planning and resource allocation for large gatherings. Conversely, short-term predictions depend predominantly on recent foot traffic patterns captured through time-lag features, with event and weather data offering only marginal benefits. This underscores the dominant role of immediate historical data in short-term forecasts compared with exogenous factors.

This paper makes an important contribution to the field of crowd forecasting by demonstrating the value of incorporating cross-domain exogenous features—specifically event schedules and weather data—into long-term prediction models. Furthermore, it highlights web scraping as a practical and effective method for collecting diverse data sources in real time, a crucial aspect of comprehensive Big Data applications.

The findings emphasize the importance of tailoring prediction models to the forecasting horizon. While event- and weather-related variables are essential for long-term prediction accuracy, short-term forecasts rely heavily on the temporal dynamics of recent traffic data. This dual approach offers a comprehensive framework for managing pedestrian flow in mixed-use urban environments, providing actionable insights for urban planners, event organizers, and public safety officials aiming to optimize crowd movement and safety.

CONFLICT OF INTEREST

All authors declare that they have no conflicts of interest.

AUTHOR CONTRIBUTIONS

Noboru Sonehara and Nobuharu Hiruma deployed the Wi-Fi sensors in Sendagaya and set up data pipelines. Kurumi Muto conceived of the presented idea, carried out the experiments, and wrote the manuscript. Akihisa Kodate supervised the project, and revised the manuscript. All authors have approved the final version.

FUNDING

A part of this research was supported by the JSPS Kakenhi (JP23K11634).

REFERENCES

[1] East Japan Railway Company, “Passenger data,” [Online]. Available: https://www.jreast.co.jp/passenger/. [Accessed: Oct. 30, 2024].

[2] P. Traganmaturapot, N. Sonehara, N. Hiruma, N. Cooharojananone, J. Jirapongwanich, K. Yodsakuntip, and K. Atchariyachanvanich, “Foot traffic analysis using Wi-Fi sensor during the Tokyo 2020 Olympics and Paralympics,” in *2023 7th International Conference on Information Technology (InCIT)*, Bangkok, Thailand, Nov. 2023. doi: 10.1109/InCIT60207.2023.10413175.

[3] J. Mackenzie, “Itaewon crush: Survivors are still tormented a year on,” *BBC News*, Oct. 25, 2023. [Online]. Available: https://www.bbc.com/news/67203564. [Accessed: Sep. 28, 2024].

[4] BBC News, “Shanghai new year crush kills 36,” *BBC News*, Jan. 1, 2015. [Online]. Available: https://www.bbc.com/news/world-asia-china-30646918. [Accessed: Sep. 28, 2024].

[5] U. Singh, J.-F. Determe, F. Horlin, and P. De Doncker, “Crowd forecasting based on WiFi sensors and LSTM neural networks,” IEEE Transactions on Instrumentation and Measurement, vol. 69, no. 9, pp. 6121-6131, 2020. doi: 10.1109/TIM.2020.2969588.

[6] Abrishami, S., Kumar, P., and Nienaber, W. “Smart stores: A scalable foot traffic collection and prediction system”. In Advances in Data Mining. Applications and Theoretical Aspects: 17th Industrial Conference, ICDM 2017, New York, NY, USA, July 12-13, 2017, Proceedings 17 (pp. 107-121). Springer International Publishing. doi: http://dx.doi.org/10.1007/978-3-319-62701-4_9.

[7] R. Tsukada, H. Zhan, S. Ishiwatari, M. Toyoda, K. Umemoto, H. Shang, and K. Zettsu, “Crowd forecasting at venues with microblog posts referring to future events,” in *2020 IEEE International Conference on Big Data (Big Data)*, Atlanta, GA, USA, Dec. 2020, pp. 3147-3155. doi: https://doi.org/10.1145/3472300.

[8] R. Jiang, Z. Cai, Z. Wang, C. Yang, Z. Fan, Q. Chen, and R. Shibasaki, “Predicting citywide crowd dynamics at big events: A deep learning system,” *ACM Trans. Intell. Syst. Technol.*, vol. 13, no. 2, pp. 1-24, Feb. 2022. doi: 10.1145/3472300.

[9] Tokyo Metropolitan Government, “National Stadium,” n.d. [Online]. Available: https://www.sports-tokyo-info.metro.tokyo.lg.jp/english/tokyoSportsFacilities/facility/02.html. [Accessed: Sep. 28, 2024].

[10] Jingu Stadium, “Facilities information,” n.d. [Online]. Available: http://www.jingu-stadium.com/qa/qa_facilities.html. [Accessed: Sep. 28, 2024].

[11] Japan Sport Council, “Chichibunomiya Rugby Stadium,” n.d. [Online]. Available: https://www.jpnsport.go.jp/chichibunomiya/sisetu/tabid/61/Default.aspx. [Accessed: Sep. 28, 2024].

[12] Tokyo 2020 Organizing Committee, “Tokyo Metropolitan Gymnasium,” n.d. [Online]. Available: https://www.2020games.metro.tokyo.lg.jp/taikaijyunbi/taikai/kaijyou/kaijyou_02/index.html. [Accessed: Sep. 28, 2024].

[13] National Noh Theater, “Seating plan,” n.d. [Online]. Available: https://www.ntj.jac.go.jp/assets/files/nou/zaseki/zaseki_nou_2.pdf. [Accessed: Sep. 28, 2024].

[14] P. Traganmaturapot, N. Sonehara, N. Hiruma, N. Cooharojananone, A. Kodate, and K. Atchariyachanvanich, “Comparative foot traffic analysis during normal periods and firework events using Wi-Fi sensors,” *2024 Joint 13th International Conference on Soft Computing and Intelligent Systems and 25th International Symposium on Advanced Intelligent Systems (SCIS&ISIS)*, Himeji, Japan, Nov. 2024. doi: 10.1109/SCISISIS61014.2024.10760113.

[15] H. Fujii, Y. Mizoguchi, Y. Morie, H. Yoshii, and N. Sonehara, “On big data analysis integrating ticket gate counting data with Wi-Fi access log data,” Internet and Operation Technology Symposium 2018, IPSJ Symposium series, pp. 54-61, 2018. doi: http://id.nii.ac.jp/1001/00192582/.

[16] National Noh Theater, “Past Event Search,” [Online]. Available: https://www.ntj.jac.go.jp/search/past/. [Accessed: Sep. 29, 2024].

[17] Google, “ChromeDriver,” Chrome Developers Documentation, [Online]. Available: https://developer.chrome.com/docs/chromedriver. [Accessed: Sep. 29, 2024].

[18] Jingu Stadium, “Event schedule,” [Online]. Available: http://www.jingu-stadium.com/event/. [Accessed: Sep. 29, 2024].

[19] Minato City, “Redevelopment Project of the Jingu Gaien Area,” Minato City Website, [Online]. Available: https://www.city.minato.tokyo.jp/saikaihatsu/kankyo-machi/toshikekaku/shigaichi/saikaihatsu/zingugaien.html. [Accessed: Sep. 30, 2024].

[20] Japan Sport Council, “Event Schedule at National Stadium,” [Online]. Available: https://www.jpnsport.go.jp/kokuritu/event/tabid/444/Default.aspx. [Accessed: Sep. 29, 2024].

[21] Japan Sport Council, “Relocation and maintenance of Chichibunomiya Rugby Stadium,” [Online]. Available: https://www.jpnsport.go.jp/corp/Portals/0/corp/shisetsu/rugbypamph.pdf. [Accessed: Sep. 29, 2024].

[22] Japan Meteorological Agency, “Weather statistics and climate data,” [Online]. Available: https://www.data.jma.go.jp/stats/etrn/index.php. [Accessed: Sep. 29, 2024].

[23] Japan Meteorological Agency, “List of AMeDAS Station Locations,” [Online]. Available: https://www.jma.go.jp/jma/kishou/know/amedas/ame_master.pdf. [Accessed: Sep. 29, 2024].