Temporal Pattern Analysis of Baggage Impact on Flight Operations

Necip Gozuacik[a], Adem Tekinbas[b], Engin Sag[c], Onur Adiguzel[d] and Sibel Malkos[e]
Technology and Innovation Department, Siemens Kartal R&D Center, Istanbul, Turkey

Keywords: Baggage Analytics, Big Data, Exploratory Data Analysis, Pattern Analysis, Time-Based Baggage Records.

Abstract: Transportation hubs like airports are increasingly complex due to globalization, resulting in diverse operations and stakeholders with often conflicting objectives. Operational inefficiencies, exacerbated by unpredictable demand and external factors such as weather, can cause economic losses and reduce sustainability. This paper aims to apply holistic approach via exploratory data analysis, integration of datasets, advanced pre-processing techniques and reveal pattern analysis. This will be a prerequisite work for solving the inefficiencies by centralizing monitoring and tracking of all operations Additionally, given the large volume of structured and unstructured data, this study underscores the importance of big data processing. Utilizing NoSQL database technology, specifically Cassandra, enables scalable, high-performance handling of millions of rows of data. The conclusions offer insights about the importance of temporal pattern analysis regarding future AI developments.

1 INTRODUCTION

Transportation hubs, such as airports, are facing heightened complexity due to globalization, with an increasing variety of operations and stakeholders involved in the process chain. These stakeholders may have competing business objectives or internal conflicts (SOCFAI, 2024).

Operational inefficiencies are further intensified by unpredictable demand and external factors like weather, resulting in economic losses and reduced sustainability (Anupkumar, 2023). Measuring operational delays is essential to identify the inefficiencies, but integrating systems across operations and hubs is an overly complex process that often necessitates custom solutions. Enhancing efficiency requires centralized monitoring and tracking of all operations.

In this work, pattern analysis addressing a holistic approach is applied to flight and baggage datasets collected from several airports. This approach utilizes advanced analytical techniques to identify patterns, relationships, and anomalies within the datasets, ensuring a comprehensive understanding of the operational dynamics. By integrating domain knowledge with data-driven insights, the analysis aims to discover latent connections between baggage handling and flight operations, providing actionable intelligence for optimizing processes. Additionally, the holistic perspective ensures that the interplay of numerous factors, such as airline schedules and passenger behaviors, is thoroughly considered. The findings from this analysis can support decision-making processes, improve operational efficiency, and contribute to the development of predictive models for future scenarios.

Original research contribution and novelty here are to perform temporal pattern analysis of baggage records and reveal correlation between baggage analytics and flight operations with considering holistic approach. To achieve that analysis, various steps are applied such as data cleaning, irrelevant data removal, joint dataset generation linking flight and baggage datasets, feature extraction etc.

This paper consists of five sections. Section 2 gives information about activities and insights in literature. In Section 3, exploratory data analysis is

[a] https://orcid.org/0000-0003-0261-4404
[b] https://orcid.org/0000-0001-7935-7309
[c] https://orcid.org/0009-0001-7476-9803
[d] https://orcid.org/0000-0002-0195-4311
[e] https://orcid.org/0000-0002-2159-5766

Gozuacik, N., Tekinbas, A., Sag, E., Adiguzel, O. and Malkos, S.
Temporal Pattern Analysis of Baggage Impact on Flight Operations.
DOI: 10.5220/0013372800003905
In *Proceedings of the 14th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2025)*, pages 831-838
ISBN: 978-989-758-730-6; ISSN: 2184-4313

studied and outlined. Pattern analysis and interpreted results are included in Section 4. Finally, Section 5 concludes the paper.

2 LITERATURE REVIEW

In literature, there are several studies regarding applying pattern analysis and AI technology for airport and aircraft operations. From the overall perspective, majority of the publications focus on flight and passenger aspects. In contrast to this, there are less studies interest in baggage operations and their effects on airport operations and as well as flight operations.

In a study, existing research on AI applications in airports are synthesized and future research directions are identified (Amiri and Kusakci, 2024). The findings indicate a significant focus on airport administration and management, with AI enhancing security, traffic management, and passenger experience.

Developing predictive models for passenger arrival patterns at airport counters are studied specifically at İzmir Adnan Menderes Airport's international terminal, to enhance operational efficiency (Sayın et al., 2023). The research compares various predictive models, including linear regression and LSTM. This study contributes to the literature by providing insights into effective predictive modelling for airport operations, which can lead to improved resource allocation and passenger satisfaction

Another paper aims to improve the prediction of checked baggage volumes for flights, which is crucial for efficient aircraft load planning (Chen et al., 2024). The study utilizes a dataset from a major US airline, applying three machine learning algorithms.

Another study aims to provide a comprehensive review of demand analysis and forecasting algorithms for checked baggage flow at airports, addressing the inefficiencies in baggage handling (Jiang et al., 2024). The research employs a mixed-method approach, integrating quantitative data analysis with theoretical modelling.

A thesis study investigates to enhance the accuracy of baggage loading and un-loading duration predictions during KLM's turnaround process using data-driven and machine learning methods (Ochoa Barnuevo, 2023). The study contributes to both theoretical advancements in machine learning applications in aviation and practical implications for KLM's operational efficiency and customer satisfaction.

Real-time data sharing and collaborative decision-making process, particularly for transfer passengers are important to enhance operational efficiency (Guo et al., 2020). The mentioning study contributes theoretically by advancing the understanding of predictive analytics in airport operations and by providing a model that can be adopted by other airports to optimize their operations.

A standard set of measures is proposed to assess the expected performance of a baggage handling system through discrete event simulation (Le et al., 2012). A stochastic, periodic input modelling methodology from real and simulated data sources is proposed.

Dimensionality reduction methodology is also important especially for large volumes of data to enhance performance of pattern analysis and recognition (Petersen et al., 2024).

The state-of-the-art in the current approaches are often limited to a single perspective, focusing exclusively on specific domains such as flight-centric, passenger-centric, or baggage-centric views. In our study, the original contribution here is to evaluate baggage dataset from several aspects and outline various pattern analysis with combining flight dataset as a holistic view.

The advantage and novelty of this paper are to perform deep-dive analysis about time-based baggage records and reveal their impact on flight operations from the point of pattern analysis view. The highlighted points are going to be used as new feature candidates for predicting baggage operation volume and impacted flight operations in real world scenarios. This will be a baseline study from the point of data pre-processing and feature extraction for further AI model developments on this subject.

3 EXPLORATORY DATA ANALYSIS

In this section, provided dataset regarding the study is explored prior to pattern analysis. Content of the dataset is investigated from the point of several aspects such as descriptive analysis, pre-processing operation, storage strategy and visualization.

3.1 Dataset Description

Two datasets regarding flight and baggage information are analyzed in the study. The datasets are collected from several airports in Turkiye as part of the SOCFAI project and have been shared in this

publication in anonymized form. The authors do not have permission to share data. Initially, the data is collected for one month period. Details about the dataset are described in the following paragraphs with emphasizing major properties.:

- Flight Dataset: This dataset contains a substantial amount of data, with around 30K rows, related to flight operations over a month. This dataset includes detailed information about flight properties, aircraft features, arrival and departure times, airport codes, flight status, and various timestamps. The large volume of data enables extensive analysis of flight performance, on-time operations, and airport/airline efficiency. It can support decision-making processes, identify trends, and detect anomalies in the complex world of flight management.
- Baggage Dataset: This comprehensive dataset contains detailed information about baggage handling and tracking within an airport or airline system. With approximately 1M rows, it provides a complete record of baggage events, statuses, arrival details, security-related information, passenger profiles, and timestamps for various handling stages throughout the entire month. The scale of this dataset allows for in-depth analysis of baggage management processes, identification of trends and patterns, and optimization of operational efficiency across the entire baggage handling system.

Major features regarding flight and baggage dataset are summarized in Table 1 with feature name and feature description pair details. These features are interpreted in the upcoming sections.

Table 1: Major features of dataset.

Feature Name	Feature Description
ACTUAL_TIME	Actual time of flight.
BAG_NUMBER	Number of baggage in a single baggage record.
BAGGAGE_EVENT	Information about baggage operation itself.
BAGGAGE_STATUS_ALL	Status of baggage.
BIM_CREATE_DATE	First timestamp when baggage is entered to the system.
BIM_ID	ID of a baggage record.
CATEGORY	Identifies flight type (Domestic, International).
ESTIMATED_TIME	Estimated time of flight
FLIGHT_CODE	Combination of Airline Code and Flight Number.
FLIGHT_STATUS	Status of flight.
ID	Flight ID (also connection with Flight Dataset).
SCHEDULED_TIME	Scheduled time of flight.
SYSTEM_AIRPORT	Identify airport from the point of flight direction view.

3.2 Pre-Processing

Raw data is transformed into clean data via several pre-processing steps. These steps are divided into three major parts as data pre-processing, outlier analysis and repetition check. Data pre-processing operations are done through data simplification and data compression phases.

For each baggage data, feature column occupancy rate is calculated. Columns with occupancy rate below 100% are identified. . Columns with empty cell are joined to achieve 100% occupancy rate.

- ATD (Actual Time of Departure) and ATA (Actual Time of Arrival) are combined into ACTUAL_TIME; STD (Scheduled Time of Departure) and STA (Scheduled Time of Arrival) are combined into SCHEDULED_TIME; ETD (Estimated Time of Departure) and ETA (Estimated Time of Arrival) are combined into ESTIMATED_TIME regarding arrival or departure flight.
- SEAT_NUMBER, SECURITY_NUMBER, SEQUENCE_NUMBER, FILE_NAME is combined into PASSENGER_ID. i.e., 04B-33-33.
- BAG_NUMBER with empty cell is filled with the expected average bag number.
- The rest of the columns with empty cell are excluded.

After completing pre-processing steps, the following data analysis is conducted to help with the rest of the pre-processing steps. For each flight code with departure status, hourly baggage arrival pattern before flight time is extracted in Table 2. For example, flight code 1 is repeated 13 times for one month period and 100 baggage has arrived when there is more than 1 hour and less than 2 hours before flight time.

Table 2: Hourly Baggage Arrival Pattern per Flight Code.

Flight Code	Total Baggage Count	Flight Count	Hourly Flight Pattern (1 Hour, 2 Hours, 3 Hours, ...)
1	308	13	[0,100,0, …]
2	2656	22	[0,2,867, …]
3	5325	23	[0,27,33, …]

For each flight code, average baggage arrival versus time before flight is plotted as in Figure 1. Each color represents a distinct flight code. Excessive baggage count is observed as 1750. Also, it is found that there can be baggage arrivals even 12 hours before the flight. In this initial representation before cleaning and pre-processing steps, there are many outliers/anomalies regarding arrival pattern of baggage records in Figure 1. Many baggage records seem processed before too early hours than flight time.

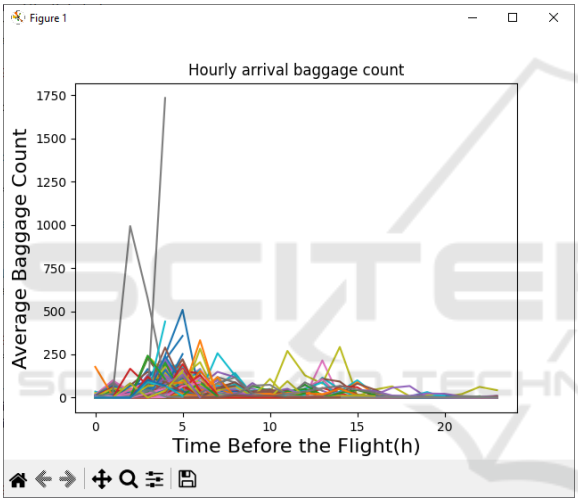

Figure 1: Hourly Baggage Arrival Pattern per Flight Code (Initial).

Outlier baggage records are excluded based on rules such as empty PASSENGER_ID, BIM_CREATE_DATE > ACTUAL_TIME and earlier timestamps do not exist in pre-determined date range. Repetition records are eliminated based on evaluating cases of BIM_CREATE_DATE, BIM_ID, FILE_NAME, ID, STATUS_INDICATOR and PASSENGER_ID feature relations. After pre-processing, hourly baggage arrival patterns for each flight code are again visualized in Figure 2. Furthermore, arrival pattern of baggage records is distributed in a reasonable way. Figure 2 indicates most activity concentrated in the few hours leading up to the flight time. Finally, because of data cleaning process, feature selection and feature merge, dimensionality reduction is also obtained like from (858K, 82) to (685K, 23).

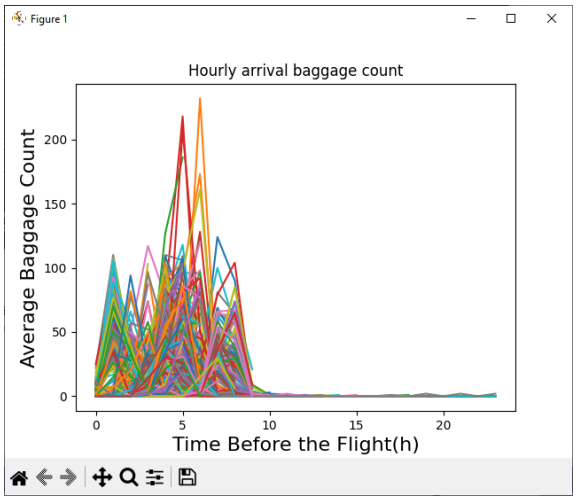

Figure 2: Hourly Baggage Arrival Pattern per Flight Code (Final).

3.3 Storage Strategy

Data can be classified as structured or unstructured. Structured data fits neatly into tables, while unstructured data cannot be easily mapped. Traditional Relational Database Management Systems (RDBMS) excel at handling structured data, but struggle with large volumes of structured or unstructured data. To address this, NoSQL databases have emerged as a more scalable, flexible, and distributed solution capable of efficiently managing both structured and unstructured big data. These advantages make NoSQL systems preferable for big data projects compared to traditional RDBMS technologies.

This study area deals with big data regarding millions of flights and baggage data with also future projection. Additionally, it requires to be expanded with real-time updates. Thousands of new rows will be added daily from the airport system with each flight, and new columns may be added as well. This will result in a high volume of data and transactions. Given these requirements, a NoSQL data management system is the most suitable solution for this big data study.

There are many NoSQL database technologies used in the software field. After careful research and requirement analysis process, it is decided that Cassandra will be the best option for this study (Khan et al., 2023). It is a wide-column store database, which is a type of NoSQL database that can be used for relational types of big data. Cassandra is an open-source database that can be distributed across multiple machines. It also provides a scalable, maintainable system with high performance and fault tolerance.

3.4 Visualization

To be able to interpret the dataset in a better way, data visualization is one of the most powerful methods used in practice. Regarding this, some major components of baggage dataset are visualized.

Distribution of flight categories can be seen in Figure 3. There are 2 categories here: INTERNATONAL and DOMESTIC. Their percentage are quite close. Count of INTERNATONAL flights is a bit more than DOMESTIC flights in that snapshot.

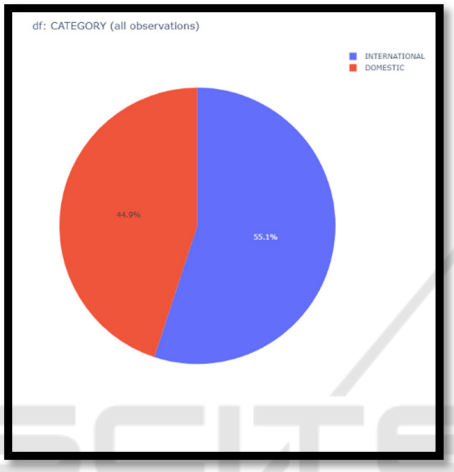

Figure 3: Distribution of Flight Categories.

Distribution of flight status can be seen in Figure 4. There are 5 categories here: ARRIVED, DEPARTED, OPEN, CANCELED, CLOSED. Majority of the flights belong to DEPARTED category. There are also numerous ARRIVED and OPEN flights. However, CANCELED and CLOSED flights are in negligible range. Focus on the study is on DEPARTED flights from the point of baggage operations.

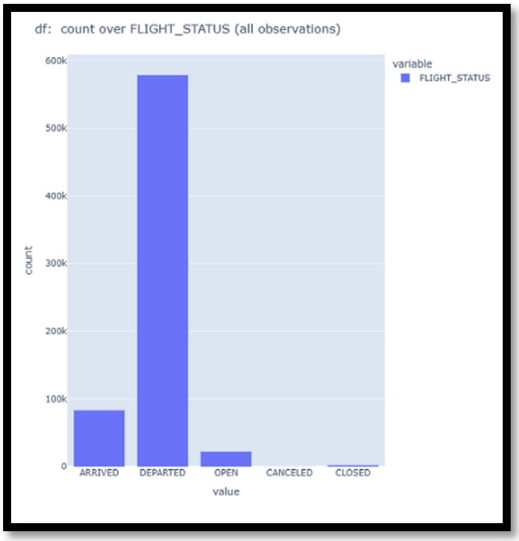

Figure 4: Distribution of Flight Status.

The distribution of system airports can be seen in Figure 5. There are 13 different system airport categories. The top 5 system airports look like ALA, ESB, ADB, ECN and TBS. ADB system airport is going to be focused primarily regarding request of data provider.

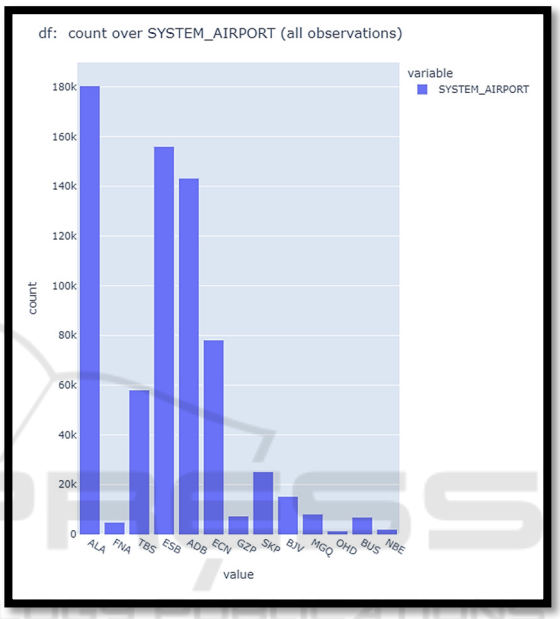

Figure 5: Distribution of Flight System Airports.

4 PATTERN ANALYSIS

Regarding pattern analysis of baggage records, two major approaches are detailed here. Firstly, average hourly baggage arrival pattern for each flight code is calculated for several flights in distinct days. Pattern Dissimilarity (PD) for a single flight here is defined as shown in formula 1. Numerator b_{ij} is the absolute difference for flight i at hour j and equals to |Hourly Baggage for Flight i at Hour j−Mean Hourly Baggage for Flight Code at Hour j|. Denominator a_j is the mean hourly baggage for the flight code at hour j. Please note that hour j here refers to delta before flight time.

$$PD_i = \frac{\sum_{j=0}^{T}(b_{ij})}{\sum_{j=0}^{T}(a_j)} \quad (1)$$

Let us there are 3 distinct flights for a flight code. Baggage count for 1 hour and 2 hours before flight time are examined. Flight 1: 10 baggage at hour 1, 30 baggage at hour 2; Flight 2: 5 baggage at hour 1, 20 baggage at hour 2; Flight 3: 15 baggage at hour 1, 10 baggage at hour 2. Mean baggage count: (10 + 5 + 15) / 3 at hour 1 equals to 10; (30 + 20 + 10) at hour 2 equals to 20. PD for Flight 1 is calculated as (|10 – 10| + |30 -20|) / (10 + 20) and equals to 0.33 (33%).

Here, the measure of PD is needed to demonstrate correlation with flight delay possibility. If there is an increase in PD of a flight, it directly impacts flight delay. For each calculated PD information, relation between flights is analyzed based on a threshold approach. A flight is marked delayed if the time difference between ACTUAL_TIME and SCHEDULTED_TIME is greater than 15 minutes. Overall number of delayed flights are counted and the delay ratio for each threshold is calculated. So that the correlation between PD ratio and flight delay is shown in Table 3.

Table 3: Correlation Between Pattern Dissimilarity Ratio and Flight Delay.

Pattern Dissimilarity Ratio Threshold	Flight Delay Ratio	Delayed Flight Count	Total Flight Count
30	57.30%	2199	3838
40	60.36%	1457	2414
50	61.75%	980	1587
60	66.98%	720	1075
70	68.61%	518	755
80	69.96%	396	566
90	71.87%	327	455
100	73.68%	266	361

Flight delay ratio over PD ratio threshold in given sample data in Table 3 is displayed in Figure 6. Results indicate that percentage of unexpected baggage records affect flight operations negatively such as delay in our case.

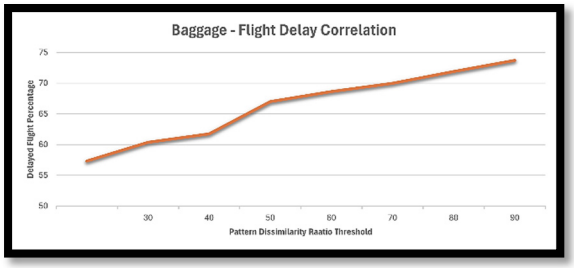

Figure 6: Pattern Dissimilarity Ratio – Flight Delay Correlation.

As a second pattern analysis of baggage records, a new feature is generated with name of MULTI_BIM_CREATE. It is calculated for each baggage record as the number of repetition baggage records with same ID, BIM_CREATE_DATE, PASSENGER_ID having STATUS_INDICATOR changed or deleted. Repetition records may appear in baggage dataset due to several reasons such as baggage rehandling, integration from multiple sources. For each flight, the percentage of baggage records that has repetition (MULTI_BIM_CREATE is greater than 1) over total baggage records is computed as shown in Table 4. For example, Flight ID 4 has total 154 baggage records. 42 records have repetition. So, MULTI_BIM_CREATE_RATIO is calculated as 27.27%.

Table 4: Multi BIM Create Ratio per Flight ID.

ID	MULTI_BIM_CREATE_RATIO
1	0.5 (1/199)
2	3.33 (2/60)
3	13.7 (10/73)
4	27.27 (42/154)

Correlation between MULTI_BIM CREATE_RATIO and flight delays is displayed in Table 5. A flight is marked as delayed if the time difference between ACTUAL_TIME and SCHEDULTED_TIME is greater than 15 minutes. For example, there are 464 flight records that have MULTI_BIM_CREATE_RATIO >= 15% and 66.16% of flights have delay in the first row.

Table 5: Correlation Between Multi BIM Create Ratio and Flight Delay.

MULTI_BIM_CREATE_RATIO	Delay Ratio	Delayed Flight Count	Total Flight Count
15	66.16%	307	464
16	66.83%	268	401
17	69.01%	236	342
18	70.3%	213	303
…	…	…	…
59	90.0%	9	10
60	100.0%	7	7

Flight delay percentage over MULTI_BIM CREATE_RATIO threshold is displayed in Figure 7. It indicates that percentage of repetitive baggage records affect flight operations negatively such as delay.

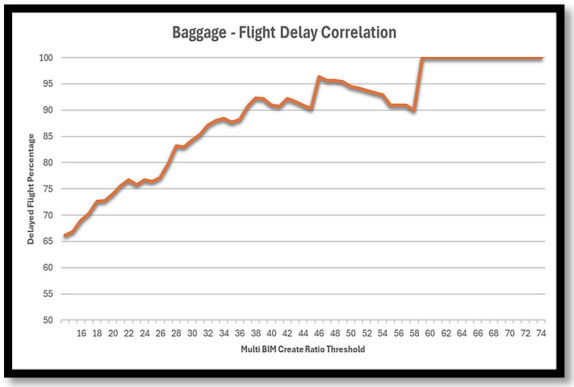

Figure 7: MULTI_BIM CREATE_RATIO – Flight Delay Correlation.

5 CONCLUSIONS

Maintaining baggage requests of passengers in an airport may have direct or indirect effect on other operations such as ground services, flight. In the study, temporal pattern analysis of baggage operations is investigated to determine if there is a correlation for flight delays in a holistic perspective.

Several pre-processing and cleaning strategies are applied on the given dataset with also considering cross relation between baggage features and flight features. Gathered results reveal two major founding. Firstly, creating multiple baggage records per passenger has a negative impact on the related departed flight operation. Secondly, increase in pattern dissimilarity ratio for baggage arrival correlates with flight delay possibility.

In future, extended version of dataset is going to be analyzed with the current systematic approaches and founding. Only statistical analysis may not be sufficient or flexible enough to manage the growing volume of data and the increasing number of features. To address this, the focus will shift towards incorporating AI-powered solutions to enhance the understanding of the effects of baggage records and operations on flight delays.

Developing AI-driven models can accurately predict baggage counts with daily, hourly, and even minute-level precision, considering both airport-specific and flight-specific factors. For this purpose, machine learning (Random Forest, Gradient-Boosting etc.), deep learning (Neural Network, CNN, LSTM etc.) and time-series (ARIMA, SARIMA etc.) approaches are planned to be utilized. This will enable airport operators to proactively manage baggage handling resources and optimize their operations, reducing the impact of baggage-related issues on flight delays.

ACKNOWLEDGEMENTS

This study was supported by Eureka-ITEA Project "SOCFAI" (Project Number: ITEA-21020). We extend our gratitude to TUBITAK for funding this project. Our special thanks go to our project partners, TAV Technologies, Siemens A.S for their invaluable contributions and collaboration. Additionally, thanks to SOCFAI Project Team for their technical contributions during the initial phase of the project.

REFERENCES

SOCFAI. (2024, October 15). Secure Open Collaboration Framework powered by Artificial Intelligence Homepage. https://www.socfai.com/

Anupkumar, A. (2023). Investigating the Costs and Economic Impact Of Flight Delays In The Aviation Industry and The Potential Strategies for Reduction.

Amiri, M. A., & Kusakci, A. O. (2024). A Scoping Review of Artificial Intelligence Applications in Airports. *Comput. Res. Prog. Appl. Sci. Eng. Trans. Ind. Eng, 10, 2900*.

Sayın, M. G., Aktaş, D. Y., Bolat, M., Çelenli, M. K., Dursun, B., Koç, G., & Üçkardeş, K. S. (2023). A Study Of Predicting Arrival Patterns Of Airport Passengers To The Counters On The Basis Of International Terminal. *Avrupa Bilim ve Teknoloji Dergisi*, (51), 63-74.

Chen, S., Park, C., Guo, Q., & Sun, Y. (2024). Advancing a major US airline's practice in flight-level checked baggage prediction. *Intelligent Transportation Infrastructure*, 3, liae001.

Jiang, B., Ding, G., Fu, J., Zhang, J., & Zhang, Y. (2024). An Overview of Demand Analysis and Forecasting Algorithms for the Flow of Checked Baggage among Departing Passengers. *Algorithms*, 17(5), 173.

Ochoa Barnuevo, M. L. (2023). Enhancing Baggage Handling Duration Predictions for KLM: A Data-driven and Machine Learning Approach Using Camera and Sensor Data (Bachelor's thesis, University of Twente).

Guo, X., Grushka-Cockayne, Y., & De Reyck, B. (2020). London heathrow airport uses real-time analytics for improving operations. *INFORMS Journal on Applied Analytics*, 50(5), 325-339.

Le, V. T., Zhang, J., Johnstone, M., Nahavandi, S., & Creighton, D. (2012, October). A generalised data analysis approach for baggage handling systems simulation. In *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 1681-1687). IEEE.

Petersen, P., Stage, H., Reis, P., Rauch, J., & Sax, E. (2024). Comparison of Dimension Reduction Methods for Multivariate Time Series Pattern Recognition. In *ICPRAM* (pp. 809-816).

Khan, W., Kumar, T., Zhang, C., Raj, K., Roy, A. M., & Luo, B. (2023). SQL and NoSQL database software architecture performance analysis and assessments—a systematic literature review. *Big Data and Cognitive Computing*, 7(2), 97.