D2.1 State of the art

Deliverable ID:	**D2.1**
Project acronym:	**SynthAIr**
Grant:	**101114847**
Call:	**HORIZON-SESAR-2022-DES-ER-01**
Topic:	**HORIZON-SESAR-2022-DES-ER-01-WA1-7**
Consortium coordinator:	**Massimiliano Ruocco - SINTEF**
Edition date:	**14 May 2024**
Edition:	**00.02.00**
Status:	**Official**
Classification:	**PU**

Abstract

This deliverable discusses the state-of-the-art related to the use cases considered in the project, as well as relevant synthetic data modelling techniques to be used for elaboration of use cases. Based on multiple data-, modelling-, and stakeholder-related criteria, two promising use cases were selected for further elaboration in the project. The literature review serves as a starting point for the activities in WP3 (Synthetic Data Generation for Multivariate Time Series for ATM-automation) and WP4 (Universal Time Series Model for Prediction and Data Generation for ATM-automation), based on the selected use cases.

Authoring & Approval

Authors of the document

Name / Beneficiary	Position / Title	Date
TUD	Task lead	25/9/2023
SINTEF	Coordinator / WP1 Leader	17/11/2023
DEEPBLUE	WP2 contributor	26/01/2024
EUROCONTROL	WP2 contributor	26/01/2024

Reviewers internal to the project

Name / Beneficiary	Position / Title	Date
EUROCONTROL	WP2 contributor	30/01/2024

Approved for submission to the SJU By - Representatives of all beneficiaries involved in the project

Name / Beneficiary	Position / Title	Date
SINTEF	Coordinator / WP1 Leader	30/1/2024
DEEPBLUE	WP2 contributor	30/1/2024
EUROCONTROL	WP2 contributor	30/01/2024

Rejected By - Representatives of beneficiaries involved in the project

Name and/or Beneficiary	Position / Title	Date

Document History

Edition	Date	Status	Name / Beneficiary	Justification
00.00.01	25/9/2023	Initial structure	Alexei Sharpanskykh/TUD	
00.00.10	20/12/2023	Intermediate version	Alexei Sharpanskykh/TUD	
00.01.00	30/01/2024	Final draft	Alexei Sharpanskykh/TUD	
00.02.00	14/05/2024	SJU comments applied	Alexei Sharpanskykh/TUD	

SynthAIr

IMPROVED ATM AUTOMATION AND SIMULATION THROUGH AI-BASED UNIVERSAL MODELS FOR SYNTHETIC DATA GENERATION

This deliverable is part of a project that has received funding from the SESAR Joint Undertaking under grant agreement No 101114684 under European Union's Horizon 2020 research and innovation programme.

Table of Contents

List of Tables

Executive summary

This document describes the state of the art on the use cases considered in SynthAIr project, as well as on synthetic data modelling techniques to be applied for these cases. Five use cases were introduced in the SynthAIr Description of Action: prediction of turnaround time (UC1), flight delay prediction (UC2), passenger flow prediction (UC3), synthetic traffic generation (UC4), and flight diversion prediction (UC5). The related work on these use cases was reviewed both with respect to synthetic data generation approaches, as well as downstream tasks such as prediction and forecasting.

We can conclude from the literature review study that synthetic data generation in relation to the considered use cases was mostly done using classical statistical and simulation-based approaches. AI-based synthetic data generation approaches have been used only to a limited extent, mostly for generation of air traffic (UC4). On the other hand, many studies were performed for the downstream tasks related to the considered use cases, such as prediction and forecasting, which are however, not the central problems considered in SynthAIr. Therefore, although we provide a review of existing literature related to the downstream tasks in the context of the considered use cases, we do not aim at completeness. We also review open aviation databases that could be useful for training synthetic data generation models.

The state-of-the-art review was also done for synthetic data generation methods. Next to more traditional approaches, based on statistical analysis and simulation, we also reviewed more recent AI-based generative modelling techniques, such as autoencoders, Generative Adversarial Networks, and diffusion models, to be used in this project. Furthermore, we described some exemplary applications of generation modelling techniques in other domains.

Based on multiple selection criteria divided in 5 categories: Data, Analysis and Modelling, Integration, Stakeholders, and Validation, the most promising use cases were selected for further elaboration in the project. For this selection, an important contribution was obtained from interviews with several aviation stakeholders (airports, airlines, air navigation service providers, air transport researchers). For further elaboration we selected two use cases: flight delay prediction (UC2) and flight diversion prediction (UC5). UC4 might be considered in the future as an extension of UC2. UC3 is also considered as an option, because of its relevance to many stakeholders, however, it highly depends on the availability of data for model training, which is both confidential and privacy-sensitive. In addition, U-space-related cases might be considered in SynthAIr, which were not a part of the original proposal. We will apply state of the art generative modelling techniques to the selected use cases in WP3 (Synthetic Data Generation for Multivariate Time Series for ATM-automation) and WP4 (Universal Time Series Model for Prediction and Data Generation for ATM-automation).

1 Introduction

1.1 Purpose of this document

This deliverable introduces the problem of synthetic data generation using AI techniques and provides an overview of related work on the topics related to the project. More specifically, we review related work on the use cases considered in SynthAIr: prediction of turnaround time (UC1), flight delay prediction (UC2), passenger flow prediction (UC3), synthetic traffic generation (UC4) and flight diversion prediction (UC5), and open data that could be used for these cases. Moreover, we describe modelling and computational AI techniques for synthetic data generation, which could be used in the context of the considered use cases. We also reviewed other projects related to SynthAIr to identify possible links and learn from their results. Furthermore, based on a number of selection criteria and using findings from related work, we choose use cases which will be further elaborated in SynthAIr.

1.2 Scope

The state of the art review is done for both the use cases described in the SynthAIr project proposal, as well as AI modelling techniques for synthetic data generation, which could be used for these use cases. While reviewing the use cases in the existing literature, we considered both existing synthetic data generation approaches which were used for these use cases, as well as the downstream tasks, such as prediction and forecasting related to these cases. We also reviewed existing research on how synthetic data generation techniques were used in other application domains. We reviewed aviation open data sources that could be used for the use cases. Based on all this related work, using a number of selection criteria, which include modelling-related, data-related, and stakeholder-related categories, we chose use cases to be further elaborated in the SynthAIr project.

1.3 Structure of the document

In Section 2 basic principles of synthetic data generation are explained. The state of the art on the use cases considered in SynthAIr is discussed in Section 3. Section 4 considers open databases which could be used for modelling of the use cases. In Section 5 a state of-the art of existing AI-based synthetic data generation techniques is provided, also considering their application in other domains. Section 6 discusses the selection of use cases to be further considered in the project, based on the reviewed related work. Finally, related projects are considered in Section 7.

2 Motivation for synthetic data generation

In the area of Artificial Intelligence (AI) and Machine Learning (ML), data are fundamental for algorithm development and effectiveness. The surge in data-centric technologies has escalated the demand for extensive datasets that are not only important for algorithm training but also for ensuring their performance. This increasing need for large volumes of data is met with considerable challenges. Strict data privacy laws, such as the General Data Protection Regulation (GDPR), regulate the use of personal data stringently. In parallel, the difficult of data access in some domains, coupled with biases in existing datasets, can severely constrain the capacity and expansiveness of AI and ML applications. More in detail, the advancement and adoption of AI and ML is related to the **data access problem** that refer to the problems of **data quality**, **scarcity**, **privacy**, and **fairness**. More in detail:

- **Data Quality**: The assurance of high data quality remains a challenge. Models trained on data that is noisy, incomplete, or incorrect are prone to produce unreliable or inaccurate predictions. This misleads the training process, resulting in algorithms that are unable to generalize well to new data, thus compromising their utility in real-world applications.

- **Data Scarcity**: A substantial obstacle in AI development is the lack of sufficient data. Many domains suffer from a dearth of accessible datasets, either due to the nature of the field or prohibitive costs associated with manual data labelling. This scarcity hampers the ability to train models effectively, especially for tasks that require extensive data to capture the characteristics of complex patterns.

- **Data Privacy and Fairness**: Privacy concerns and the imperative for fairness restrict the availability of datasets in various sectors. Legal and ethical considerations often preclude the public release of data that could reveal sensitive information about individuals. This limitation has propelled the exploration of synthetic data as a feasible alternative. Synthetic data generation, when executed with rigor, can yield anonymized datasets that retain the statistical properties of the original data while adhering to differential privacy standards. These efforts are crucial for maintaining user privacy and ensuring that the resulting models do not perpetuate or amplify biases.

Addressing these challenges is necessary to harness the full potential of ML. Ensuring that data is of high quality, readily available, and used ethically is fundamental for the adoption and evolution of AI technologies and their successful deployment across industries.

3 State of the art on the use cases

In the Description of the Action, 5 use cases (UC) related to the operations of airports, ANSP and airlines were identified to be considered in the project: prediction of turnaround time (UC1), flight delay prediction (UC2), passenger flow prediction (UC3), synthetic traffic generation (UC4) and flight diversion prediction (UC5). Literature related to all these cases will be reviewed in the subsequent sections 3.1-3.5.

Furthermore, based on multiple selection criteria divided in 5 categories (refer to Table 8): Data, Analysis and Modelling, Integration, Stakeholders, and Validation, the most promising use cases will be selected in section 6 for further elaboration in the project.

So far, synthetic data generation using generative AI methods in application to air transport has been considered only to a limited extent. Therefore, most of the related work reviewed in the following subsections is on downstream tasks, such as prediction and forecasting, which is however, not the central problem considered in SynthAIr. Therefore, although we provide a review of important examples of downstream tasks in the context of the considered use cases, we do not aim at completeness.

3.1 UC1: Prediction of turnaround time

Turnaround time is the time between the aircraft is on-chocks until it is off-chocks, and is affected by several factors including the aircraft, the airline, the airport, weather and airport infrastructure. Since the last decade, according to yearly EUROCONTROL's Performance Review Reports, more than 40% of primary delay at airports is generated by the turnaround process. Past ATM research projects under and beyond SESAR have considered aircraft rotation at airports in a relatively aggregated way, i.e., from in-block to off-block, but rarely scrutinised the details of ground handling and turnaround process. At the same time, it is recognised in the European ATM Master Plan that there is a need to provide the airport stakeholders with early warning indicators of possible delays along the critical paths during the turnaround process, under the Operational Improvement AO-0818 on Extended Turnaround monitoring. Also, recent validation exercises on A-CDM, such as at Alicante regional airport reinforced the urgent need to increase the accuracy of the information exchanged with the Network Manager (NM) on the progress of the turnaround process to improve NM's planning. For the accuracy improvement better prediction of the turnaround process is needed. Better turnaround time estimates allow smaller deviations from the actual turnaround times, resulting in reduced delays [1].

The schedule of an aircraft turnaround can be divided into 5 main steps: boarding, fuelling, catering, cleaning and deboarding. Some steps are being executed in a strict sequence due to technical, legal or operational constraints. Other steps may be executed in parallel, such as boarding and baggage loading. The shortest path, measured in units of time, of parallel and sequential events yields the shortest turnaround time and is called the critical path. When processes on the critical path are delayed, the whole process is delayed. Each step of the turnaround procedure is a stochastic process depending on multiple factors, making the prediction of turnaround time a challenging problem. Turnaround time can be modelled at different levels of granularity for a given scope. The granularity spans from airport level to the level of each individual flight. While the scope can be as limited as the flights of a single airline in a single airport, or as general as all flights for a given category of airports.

A microscopic simulation-based approach to study turnaround and its impact on flight delays is proposed in [2]. Many important random factors influencing the duration of the turnaround were identified and quantified in this work. However, a limited dataset with aggregated data was used to calibrate the model at the microscopic level, which may impact the accuracy of the model predictions. Furthermore, it is not clear how generalizable is the developed model.

Methods based on machine learning techniques, neural networks in particular, were used to predict turnaround time [3]. For the airport under study, the authors were able to achieve the prediction accuracy up to 10 minutes. The authors used Artificial Neural Networks (ANN) to predict the flight turnaround time of flights at a large international hub airport in China. In the ANN model only 7 features are considered, namely: (1) aircraft stand (to account for distance for support vehicles), (2) aircraft type, (3) type of flight (domestic or international, accounting for immigration and custom inspections), (4) aircraft ground handler, (5) flight arrival time (to account for high turnaround demand), (6) number of arriving passengers and (7) number of departing passengers. With this low number of variables, the model was able to find the general trend in the data. It was not investigated whether or not the model can be generalizable for other airports too.

In collaboration with an aircraft ground handler, Van Hassel [4] proposes a Process Structure Aware Prediction (PSAP) approach to predict the taxi-in and turnaround duration for Boeing 737 flights of a major European low-cost carrier at Eindhoven Airport in an interpretable manner. In the PSAP approach, the turnaround process is split into a set of activities of which the cycle time is predicted. For this, Van Hassel considers two algorithms: (1) Random Forest and (2) Multilayer Perceptron (MLP). In the PSAP framework, Van Hassel found that the performance can be considered equal between the models. In an aggregated approach, when the process structure is not explicitly defined, MLP proved to be more capable in estimating the turnaround duration. However, Van Hassel only considers a handful of processes and factors in his research. For example, transfer passengers are not included in the model, nor is aircraft catering and cleaning considered.

Fricke and Schultz [1] present a statistical approach to determine the turnaround process duration. In the model, the turnaround process is split into five sub-processes, namely: (1) de-boarding, (2) cabin cleaning, (3) catering, (4) fuelling and (5) boarding. For every process, a Weibull or Gamma distribution (whichever describes the process best) is fit using operational data from a regional airport in the United States. Next, the critical path method (i.e. considering the sequential and parallel dependencies between turnaround steps) is used to obtain total turnaround process duration. In such a simplified model, only a handful of processes are considered, neglecting the influence of key turnaround process such as baggage and cargo (un)loading.

Asadi et al. [5], based on research in [1], propose a novel analytical convolution method to predict the Target Off-block Time (TOBT) of a flight, taking into account uncertainties in the turnaround process. The turnaround process is split into various sub-processes. For each process, the authors use and, in some cases, (re)parameterize the Weibull or Gamma distributions obtained in [1]. Next, using analytical convolution the stochastic process times were obtained. Taking into account the parallel and sequential dependencies between the various processes, Asadi et al. apply analytical convolution to obtain the Estimated Off-Block Time (EOBT).

Zhou et al. [6] present a deep learning approach using Gated Recurrent Units (GRU) to predict the departure time of a flight. Zhou et al. [236] train the model on data from a spoke airport in eastern China, comparable in size and passenger numbers to London Luton. In the model, Zhou et al. take the following into account: basic flight information (e.g. actual landing time, actual departure time), airport parameters (e.g., number of flights arriving and departing), weather and airline parameters. This

indicates that variables directly influencing the turnaround process are not taken into account. With respect to standard ANNs and Long Short-Term Memory (LSTM) neural networks, Zhou et al. found that neural networks with GRUs show higher predictive performance on their data sets.

In collaboration with a small (non-hub) airport and a Scandinavian full service carrier, Halmesaari [7] presents an explainable aggregated machine learning approach to predict the aircraft ground handling process duration. The approach consists of two steps: (1) obtaining a prediction of the turnaround process duration using a gradient boosted tree-ensemble regression model, XGBoost, and (2) extracting explanations from the model using a post-hoc explanatory framework, SHAP. One of the limitations in Halmesaari's study is the lack of data. Although the author found that in most cases the turnaround duration can be described by only a few variables, in cases where the turnaround is significantly longer than the scheduled duration the available data is not able to describe this discrepancy. Furthermore, provided the size and (special) characteristics of the airport considered in Halmesaari's work, it is doubtful that the proposed model and findings could be well generalized to other airports.

In their work, Luo et al. [8] aimed to forecast the duration of the turnaround process and aircraft off-block adherence in two separate models. To this end, they used data on the duration of the turnaround sub-processes obtained using computer vision techniques applied on camera images. These data, which were partial and prone to errors, were extended with a synthetic dataset obtained using an agent-based simulation model, which simulated the turnaround (sub-)processes based on actual data and domain knowledge. Unfortunately, due to data confidentiality, no further details were provided on the performance of the model. To forecast the duration of the turnaround process and confirm off-block adherence, Luo et al. trained various tree-based models on the synthetic and available data. As a side result, Luo et al. found that representing the (sub-)process durations by separate features (e.g. first passenger in, last passenger in) yields higher model performance.

Asadi and Fricke [9] employ fuzzy logic to predict the turnaround time of a flight. First, they transformed the probability distributions of (sub-)processes into a cumulative density function, which is mathematically equivalent to the fuzzy membership function. Next, they combined fuzzy logic and the critical path method to make an estimate of the turnaround time of a flight, taking into account the main turnaround processes. However, the approach suffers from some significant limitations. For example, only triangular and trapezoidal fuzzy membership are considered in the model, yielding an inherent loss of accuracy, as not every cumulative density function can be accurately described by such functions.

There are only few airports and airlines which collect detailed quantitative data about the steps of the turnaround process. Among them is Schiphol airport with its Deep Turnaround project. Furthermore, in [10] an approach based on deep learning and computer vision is proposed for detecting and monitoring turnaround activities. However, datasets collected using such approaches are not openly available. An approach described in [11] merges OpenSky data with EUROCONTROL's CPR data to produce off-/on-block time among other estimates. In this case, the turnaround process is modelled at a higher aggregation level.

In the reviewed literature, synthetic data is generated using (agent-based) simulation models, calibrated using limited real datasets and expert opinion.

3.2 UC2: Flight Delay Prediction

Flight delay is widely recognized as a critical performance indicator of air transport systems in aviation industry [12]. In literature, primary and secondary delays are distinguished. There exist several definitions for primary delay.

Beatty et al. [13] described it as initial delay, the delay which is initially created by the aircraft itself or its conditions. Furthermore, AhmadBeygi et al. [14] referred to this delay as root delay - the source of propagation throughout the network. This root delay is independent of any other delay created earlier on.

Propagated delay can be caused by four main reasons, namely, aircraft rotation, aircraft equipment, crew rotation, and transferring passengers. Lan et al. [15] defines propagated delay as delay that occurs when the aircraft is delayed on a prior flight. However, this definition only covers aircraft rotation and ignores the effects of crew rotation or transferring passengers. Kafle and Zou [16] stated that propagated delay occurs if connected resources are delayed in a flight downstream. This definition is more broad and is able to cover, not only delay caused by aircraft rotation, but also crew, passenger and airport resources. In Europe, the term reactionary delay is commonly used, whereas in the U.S., the term propagated delay or delay propagation is generally used.

According to the Bureau of Transportation Statistics of the US, the causes of delay can be categorized in five main categories, namely, air carrier delay, extreme weather delay, National Aviation System (NAS) delay, security delay, or late-arriving aircraft [17].

In the category air carriers, delays arise due to circumstances which are accountable to the operating airline, such as maintenance or crew uncertainties, unloading of baggage, aircraft cleaning, or other day-to-day operations, which involves the preparation of next flight. According to the Bureau of Transportation Statistics of the US, the category air carrier is the largest category and accounts for 32% of delays in the year 2022.

The extreme weather category presents all delays caused by significant meteorological conditions, which prevent the aircraft from flying or could delay the flight, such as tornadoes, blizzards, or hurricanes. Extreme weather is accountable for approximately 3% of flight delays.

The NAS category covers a broader set of causes which are all related to the national aviation system, such as non-extreme weather, volume restrictions, equipment problems, closed runway, or other reasons. The National Aviation System delays are monitored by the FAA and are recorded in a different database, called the Operations Network (OPSNET). This category is accountable for approximately 21% of flight delays. All non-extreme weather conditions, which are included in this category, could be mitigated, if improvements would be made to the NAS's capacity.

The fourth category of delay is due to aircraft arriving late, causing the flight at hand to depart late. This category accounts for roughly 30% of all delays in the national aviation system.

The last and the smallest part of delays is triggered by a security breach and only account for 0.3% of all delay caused in a year.

Jacquillat and Odoni [18] identified that delay propagation models fit in three main categories, namely, microscopic, mesoscopic, and macroscopic models. Microscopic models treat aircraft separately and consider a detailed layout of the analysed airport and its movements on the ground. Due to this high level of complexity, microscopic models are not well-suited for analysing the dynamic behaviour of delays of the overall network.

Mesoscopic models predict airport operations by using operational data, such as the runway configuration in use, the short-term demand or the ground vehicle schedule. Many machine learning studies fit into this category. Mesoscopic models heavily rely on operational data per airport and are therefore less suitable when modelling delay propagation on a larger scale.

Lastly, macroscopic models are defined at an airport level and thus make it possible to capture the effects of the whole network on for example flight schedules, airports capacities and the propagation of delays. In the literature, these models have been used to model a network of airports [19], often using machine learning-based studies.

Comprehensive reviews on flight delay prediction methods can be found here [12], [20].

A flight delay prediction model, if employed appropriately, can help commercial airports to reduce negative impacts of undesirable congestion, without necessarily investing in logistics and airport capacity development. The resulting decision support system is expected to be embedded within the flight information systems of commercial airports and integrated with their existing delay prediction engine. This ultimately can enable connected airports to collectively alleviate flight delay propagation within their network through collaborative efforts, such as sharing relevant information and synchronizing their delay predictions at regular intervals.

In most of the papers reviewed, real data were used to train and validate flight delay prediction models from open databases such as Airline On-Time Performance Data at http://www.transtats.bts.gov and EUROCONTROL R&D archive. No generative AI methods in application to this use case were found in the reviewed literature.

3.3 UC3: Passenger Flow Prediction

Passengers are an important source of uncertainty in air traffic management and airport operations. More than 50% of delayed flights are caused by passengers boarding late or not at all. These delayed flights cause instability in the overall air traffic planning. In particular, recurring issues with management of passenger flows at Amsterdam Schiphol Airport in summer 2022 resulted in cancelation of thousands of flights. In recent years, with the rapid growth of airport passenger flow, airport terminal processes such as security inspection, emergency response, check-in, baggage handling are facing tremendous pressure. In the process of transforming airports into ‘smart’ and ‘digital’ operations, airports must accurately anticipate changes in the number of passengers and their flows in the terminal to improve the quality of service, achieve efficient business operations and rationalize the allocation of resources.

Airport terminals serve as a gateway between the landside and the airside, with security procedures forming the interface between these two areas. The landside is freely accessible to everyone, while the airside can only be accessed by passengers and employees after passing through the security checkpoint. Important airport terminal activities include:

- *Passenger arrival at the airport*

- *Check-in*: the flight ticket and passport/ID of a passenger are submitted at the counter. Nowadays only the passport/ID is sufficient to allow the passenger to be checked-in and to hand over the boarding pass. Additionally, the luggage that must be checked-in is weighed and, if the weight limit is not exceeded, transported away by the baggage handling system. When the weight limit is exceeded, an

additional fee needs to be paid. To speed up the process, automatic self service check-in kiosks are increasingly used. Check-in may create a bottleneck in the flow of passengers when insufficient resources are allocated to it.

- *Security check*: at a security checkpoint, passengers and their cabin luggage are being checked for illegal items, such as weapons and flammable fluids. The screening of passengers and their luggage is done by both detection machines (e.g., X-ray machines) and manual checks by security agents. In some cases, additional, manual check of cabin luggage is required. The security checkpoints often represent an important bottleneck in a passenger flow. Even before the 9/11, security checkpoints naturally created a bottleneck in passenger flow, as all passengers had to pass through them. However, post 9/11, due to significantly heightened scrutiny, throughput rates drastically decreased. Airports attempt to maintain the minimum number of open lanes necessary to meet throughput requirements; however, this often leads to long queues, with 70% of passengers reporting such a perception. Being able to predict queueing time of passengers and to allocate airport resources accordingly is important for the performance improvement of the airport terminal. The prediction accuracy depends on many factors, among which delayed flights, issues with other connected modes of transportation such as trains, insufficient airport capacity, passenger characteristics, uncertainty related to connected passengers.

- *Passport control*: during the passport control, the identity of the passenger is checked, and security is ensured. This activity may also create a significant bottleneck in the flow of passengers. Airports cannot control this activity directly, as it is under the responsibility of the national gendarmerie and national police forces, for which security, and not necessarily efficiency, is the main priority.

- *Discretionary activities* such as retail, food, and beverages generate a significant amount of non-aeronautical revenue within the aviation industry. Discretionary activities affect passenger flow and global airport terminal performance.

- *Passenger boarding/deboarding*: During boarding, the boarding pass and ID of the passenger is checked one final time and most of the time, the passenger can then board the aircraft. Sometimes however, the operator responsible for boarding can ask the passenger to check-in the cabin luggage such due to a lack of cabin space. When this occurs, the boarding process is disrupted which can cause problems in terms of delays.

To model and analyse airport terminal processes often two classes of models are used: simulation-based models [21], [22] and data-driven models [23], [24]. Some of these models aim at prediction of passenger flows. In [25] four types of passenger flow prediction were identified: time series models, causal models, artificial intelligence models, and hybrid models.

Time series forecasting techniques comprise a wide array of statistical methods designed to predict future values using historical data. These approaches span from basic moving average models to intricate ARIMA and GARCH models. In [26] Dynamic Tobit models and Generalised Autoregressive Conditional Heteroskedasticity (GARCH) were employed to forecast monthly arrivals of domestic and international passengers at Corfu Airport in Greece. The combined model utilised 20 years of time series data, incorporating variables such as the number of arrivals, European GDP per capita, Greek GDP per capita, and disposable income. In [27] a SARIMA model was used, which combines autoregressive, moving average, and seasonal components to predict arrivals at the security

checkpoint. However, this method is inherently incapable of utilising flight schedule information, which carries significant value for refining arrival rate estimates. In [28] the authors focused on forecasting monthly passenger arrivals on a longer time horizon by employing methods such as moving average, single exponential smoothing method, Holt method, and Holt-Winter method. The objective was to estimate the required daily shuttle service levels, using relatively simple time series techniques as input to enhance operational and strategic level resource allocation.

In [29] boarding card data are used to estimate individual passengers Time To Departure arrival distributions for individual flights, which are then combined to determine the overall short-term arrival rate at a checkpoint. The study found that the Weibull distribution provides the best fit to the TTD arrival distribution from among Gaussian, Poisson, Gamma, and Lognormal distributions. However, the goodness of fit for the Weibull distribution is not thoroughly investigated.

In [30] proposed a causal approach to forecasting arrival rates, based on system dynamics. This technique employs historical data to estimate dwell times in three primary airport areas: check-in hall, security area, and departure hall. A gamma distribution is fitted to represent the probability of the duration a passenger spends in each section of the airport. Using scheduled flight departures and the estimated number of passengers, it becomes possible to estimate the arrival rate in each airport area. However, this approach assumes homogeneity of all passengers. Furthermore, the dwell time distributions are assumed to be static, without any variation throughout the day.

In general, while causal models present opportunities for improving short-term forecasting by incorporating external information and capturing non-linear relationships, their limitations, such as the need for high-quality input variables, their selection and extensive fine-tuning, are often problematic.

Detailed real data on passenger flows at airports are rarely available in open access. In [31] timing data on the security checkpoint process was provided. The reported data are related to passenger arrival times, X-ray image inspection times, decision type (cleared or not-cleared), physical search service times, and explosives trace detection service times. However, they do not report data that specifies how long passengers take to drop or collect luggage. Furthermore, the passenger type is not included either in their dataset. The authors only provide summary figures that describe the data, but do not provide the raw data. To the best of our knowledge, no security checkpoint dataset with a large amount of details exists is available in literature. One of the exceptions is a limited dataset provided in [23]. Data was collected for 13 different security checkpoint lanes, in 11 blocks of time. In many of these time blocks, data for a single lane was collected, while multiple lanes were open. A total of 2277 passengers, flying to 16 different destinations with 48 flights were followed. Three types of lanes were considered: standard, normal and service lanes. Data for standard lanes was gathered between 23 February 2018 and 17 April 2018, while data for normal and service lanes was collected on the experimental days: 17 December 2018 and 18 December 2018. Their analysis showed important differences between six identified passenger types.

Obtaining real data for passenger flow prediction is often problematic, as many existing airports are not equipped with necessary sensor technology to track passenger flows. Furthermore, passenger privacy aspects may also create additional obstacles for data collection. Detailed data about passenger flows are not publicly available. Existing passenger flow prediction models are often based on simulators calibrated using limited sets of real data.

3.4 UC4: Synthetic traffic generator for (fast and/or real-time) simulations purposes

Real-time simulations, on the one hand, commonly consist of a sequence of exercises. Each exercise represents a particular combination of conditions referred to as organisation (e.g., todays vs. advanced ATC tool support) and each organisation is commonly repeated multiple times across the whole sequence of exercises.

Performance metrics are computed for each organisation by computing a statistic (e.g., mean or median workload scores) across all the exercises within an organisation for the participants operating the controller working positions of sectors of the measured airspace. Inferences regarding benefits/concerns related to the change elements under scrutiny are sought by comparing performance metrics between organisations. In experimental design terminology, a real-time simulation is most often executed as a repeated measures experimental design. This means that the same group of participants perform the exercises of all conditions. Inferences on the differences between conditions, here organisations are complicated by the presence of systematic sequential effects. For instance, controllers become more and more familiar with the simulated traffic situations. They recognise previously experienced traffic patterns; they more easily anticipate the evolvement of traffic and hence identify and resolve conflicts with less effort and a higher degree of situational awareness. A solution to counteract the invalidating impact of sequential effect consists of using adequate traffic samples. Generating new traffic conditions with similar complexity, however, is a very tedious and time-consuming tasks currently performed by human experts.

Fast-time simulations, on the other hand, are frequently employed to test new ATM concepts (e.g., a new airspace design) or train reinforcement learning algorithms (e.g., to develop decision-support tools for air traffic controllers).

In the former case, Monte Carlo experiments with variable traffic conditions are widely applied to produce statistically significant results, with the parameters that determine a traffic condition rigorously randomised according to pre-defined probability distributions. In the latter case, variability in traffic conditions is essential for achieving a policy that generalises across multiple scenarios (even those not seen during the learning process). However, generating random yet representative traffic conditions is difficult and requires a thorough understanding of the characteristics that describe traffic conditions including their distributions. As a result, the appropriateness of the parametrisation as defined by the humans who design the experiment limits the representativeness of the created traffic conditions.

One of the most commonly used approaches for synthetic air traffic generation is based on generating simplified flight trajectories from extrapolated flight plan databases [32] using statistical methods. It provides a good balance between constant aircraft sets and recorded real data. More specifically, the following approach is followed in [32]: For each hour of a reference day, the number of flights between any two European airports is used as basis for the calculation of the hourly aircraft generation rate for each pair of source and destination airports. The number of flights is extrapolated by applying growth factors to the aircraft generation rates. On the basis of the hourly aircraft generation rates and the per-country EUROCONTROL growth factors, a reference day of synthetic average European air traffic for a year was generated. Flights were simulated in an asynchronous distributed way. The simulation

of each flight trajectory was implemented in an asynchronous time-stepped manner. Similar approaches for traffic generation were proposed in [33], [34].

In [35] an air traffic generator is proposed, which uses traffic patterns from real data, and produces a set of synthetic flights consistent with these data. Traffic pattern data contain a list of deterministic and probabilistic patterns with respect to several aspects such as flight periodicity, route and significant points.

Generative AI methods have also recently been used for air traffic generation. In particular, variational autoencoders were used in [36] to generate 4-dimensional aircraft trajectories modelling using Temporal Convolutional Networks and a prior distribution composed of a Variational Mixture of Posteriors. The proposed model has been trained on trajectories in the Terminal Manoeuvre Area of Zurich airport. The model has demonstrated abilities to generate complex and realistic trajectories. However, the authors note that the approach has difficulties generating trajectories based on uncommon events (e.g., go-arounds). In particular, the approach can generate realistic tracks based on events which are common in training data like holding patterns, but not go-arounds.

3.5 UC5: Flight diversion prediction

When an aircraft is unable to land at its original destination airport, it is diverted to an alternate airport. This event has economic and operational implications for airspace users. The fleet and crew schedules may be severely disrupted, and passengers and/or cargo must be transported as soon as possible to their original destination.

Diversions are also undesirable from the standpoint of the airport. When a massive number of flights are diverted to airports that are operating at or near capacity, they risk becoming critically overloaded. Diversions are triggered by many reasons, including adverse weather (e.g., low visibility), medical emergencies, unruly passengers, and technical problems. Recent work proposed a model that could assist in determining flights that are likely to divert because of adverse weather conditions at the destination airport. The model was trained via confident learning on four years of historical data, with the goal of pruning flight diversions caused by events other than weather. The reasoning was that these diversions are likely to be unpredictable, so the model should not attempt to learn them. Each of the observations was labelled as positive if the corresponding flight was diverted and negative otherwise, i.e., the model solved a binary classification task. According to aggregated performance metrics, the model has a high precision with a moderate recall, indicating that it is conservative but could miss some of the diversions. These limited performance indicators could be attributed to an underrepresentation of positive observations, which account for only 0.2% of the whole training dataset.

Machine learning models were used for flight diversion prediction. In particular, in a tree-based model was proposed in [37] which learned which flights are more likely to be diverted due to adverse weather conditions at the destination airport using historical traffic and weather data. The proposed model demonstrated high precision and moderate recall.

In supervised clustering method is proposed in [38] which combines feature attribution, dimensionality reduction, and clustering algorithms to identify the most representative features for characterising flight diversions due to weather and highlighting situations where predictions require careful consideration.

In [39] an approach for automated detection and prediction of diverted flights is proposed based on Support Vector Machines using publicly available data. The technique is able to classify a flight as diverted with a high accuracy, when the aircraft displays anomalous behaviour for an extended period of time.

No related literature on using generative AI methods for this use case was found.

4 Available open data

In this section we review several open databases available online and discuss their relevance to the use cases considered in SynthAIr.

4.1 EUROCONTROL R&D Archive data

EUROCONTROL shares flight data for research and development purposes, but users must agree to some rules. These include using the ATM Dataset only for research, not sharing or distributing it, acknowledging EUROCONTROL as the source, and understanding that the data comes as-is without any guarantees. The dataset covers historic commercial flights in fixed sample months of specific years, with a two-year delay before release.

The information about flights and their paths over points and airspaces comes from the flight plans submitted by airlines and other aircraft operators to EUROCONTROL Network Manager (NM). The NM's ATFM systems also generate flight profiles. All instrument flight rules (IFR) flights within the NM Area are required to submit their flight plans to NM. However, the point and airspace profile data in the 'actual' version of the data includes some updates from radar observation of the flight's path. Flight data only includes flights of ICAO types 'S' (scheduled) and 'N' (non-scheduled flight), excluding ICAO types of General aviation, Military and Other.

This database is formed from data collected from commercial flights operating in and over Europe. Furthermore, the collected data is enriched with live data from air navigation service providers' flight data systems, radar, and datalink communications. Moreover, additional data sources are used such as information about the route network.

Among the available datasets are:

- Detailed flight information.
- Flight trajectories (planned and actual).
- Airspace structure.
- Route network information.

The content of the single datasets is described below.

ECTL_ID	Unique numeric identifier for each flight in EUROCONTROL PRISME DWH
ADEP	ICAO airport code for the departure airport of the flight. The ICAO airport code or location indicator is a four-letter alphanumeric code designating each airport around the world. These codes are defined by the International Civil Aviation Organization and published in ICAO Document 7910: Location Indicators.
ADEP Latitude	Latitude of departure airport in decimal degrees.
ADEP Longitude	Longitude of departure airport in decimal degrees.

ADES	ICAO airport code for the destination airport of the flight. The ICAO airport code or location indicator is a four-letter alphanumeric code designating each airport around the world. These codes are defined by the International Civil Aviation Organization and published in ICAO Document 7910: Location Indicators. If a flight is diverted, then the ADES will be the actual airport where it landed.
ADES Latitude	Latitude of destination airport in decimal degrees.
ADES Longitude	Longitude of destination airport in decimal degrees.
Filed Arrival Time	Time of arrival (UTC) based on the last filed flight plan. It is the time at which the aircraft will land at the aerodrome according to the planned profile calculated for the flight.
Actual Off-Block Time	Off-Block Time (UTC) based on the ATFM-updated flight plan. The time that an aircraft departs from its parking position. This time may be known from flight data updates received by NM, or in the absence of such updates may be calculated from the known take-off time minus a standard taxi time value for the airport.
Actual Arrival Time	Time of arrival (UTC) based on the ATFM-updated flight plan. It is the time at which the aircraft lands at the aerodrome, reflecting the best picture that NM has based on available radar updates, ATFM messages received etc.
AC Type	The ICAO aircraft type designator is a two-, three- or four-character alphanumeric code designating every aircraft type that may appear in flight planning. These codes are defined by the International Civil Aviation Organization and published in ICAO Document 8643 Aircraft Type Designators.
AC Operator	Three-letter ICAO operator code. Aircraft operator codes are defined by ICAO and published in Document 8585. If the operator is unknown, not provided in the flight plan or not present in Document 8585 the value will be "ZZZ".
AC Registration	Aircraft registration. In accordance with the Convention on International Civil Aviation, all civil aircraft must be registered with a national aviation authority (NAA) using procedures set by each country. Every country, even those not party to the Chicago Convention, has an NAA whose functions include the registration of civil aircraft. An aircraft can only be registered once, in one jurisdiction, at a time. The NAA allocates a unique alphanumeric string to identify the aircraft, which also indicates the nationality of the aircraft, and provides a legal document called a Certificate of Registration, one of the documents which must be carried when the aircraft is in operation.
ICAO Flight Type	ICAO Flight Type: S – Scheduled, N - Non-scheduled commercial operation
STATFOR Market Segment	Market segment definitions can be found in http://www.eurocontrol.int/sites/default/files/content/documents/official-documents/facts-and-figures/statfor/statfor-market-segments-rules-for-sid-2016-definition.xls
Requested FL	Requested cruising flight level from the flight plan.

Actual Distance Flown (nm)	Distance flown in nautical miles, corresponding to the 'actual' profile below.

Table 1: Flights File - flight details from EUROCONTROL Network Manager flight plans in PRISME Data Warehouse (DWH)

ECTL_ID	As in Flights file above
Sequence Number	Numeric sequence number of the points crossed by the flight in chronological order. (Points can be not only known named waypoints, navaids, etc. but also intermediate points inserted by NM profile-generation processes.)
Time Over	Time (UTC) at which the point was crossed
Flight Level	Altitude in flight levels at which the point was crossed
Latitude	Latitude in decimal degrees
Longitude	Longitude in decimal degrees

Table 2: Filed flight points. Actual flight points are also provided as for the filed ones

ECTL_ID	As in Flights file above
Sequence Number	Numeric sequence number of the airspace entered by the flight in chronological order
FIR ID	The identifier of the FIR
Entry Time	Time (UTC) the flight entered the airspace
Exit Time	Time (UTC) the flight exited the airspace

Table 3: Filed flight airspaces. Actual flight airspaces are also provided as for the filed ones

ECTL_ID	As in Flights file above
Sequence Number	Numeric sequence number of the airspace entered by the flight in chronological order
AUA ID	The identifier of the AUA
Entry Time	Time (UTC) the flight entered the airspace
Exit Time	Time (UTC) the flight exited the airspace

Table 4: Filed ATC unit airspaces. Actual ATC unit airspaces are also provided as for the filed ones

External ID	Unique ID of the AIRAC cycle. AIRAC stands for Aeronautical Information Regulation And Control and defines a series of common dates and an associated standard aeronautical information publication procedure for States. It consists of four digits. The first two digits represent the year in which the AIRAC was published. The third and the fourth digits represent the sequential number of the AIRAC cycle.
Date From	First date at which AIRAC data is valid.
Date To	Last date at which AIRAC data is valid.

Table 5: Information about AIRACs (Aeronautical Information Regulation And Control)

Route ID	Unique route identifier. According to ICAO Annex 11 basic designators for ATS routes shall consist of a maximum of five, in no case exceed six, alpha/numeric characters in order to be usable by both ground and airborne automation systems. The designator shall indicate the type of the route: high/low altitude, specific airborne navigation equipment requirements (RNAV), aircraft type using the route primarily or exclusively. A. The basic designator consists of one letter of the alphabet followed by a number from 1 to 999. The letters may be: 1. A, B, G, R — for routes which form part of the regional networks of ATS routes and are not area navigation routes; 2. L, M, N, P — for area navigation routes which form part of the regional networks of ATS routes; 3. H, J, V, W — for routes which do not form part of the regional networks of ATS routes and are not area navigation routes; 4. Q, T, Y, Z — for area navigation routes which do not form part of the regional networks of ATS routes. B. Where applicable, one supplementary letter shall be added as a prefix to the basic designator as follows: 1. K — to indicate a low level route established for use primarily by helicopters; 2. U — to indicate that the route or portion thereof is established in the upper airspace; 3. S — to indicate a route established exclusively for use by supersonic aircraft during acceleration/deceleration and while in supersonic flight. C. Where applicable, a supplementary letter may be added after the basic designator of the ATS route as a suffix as follows: 1. F — to indicate that on the route or portion thereof advisory service only is provided; 2. G — to indicate that on the route or portion thereof flight information service only is provided; 3. Y — for RNP1 routes at and above FL200 to indicate that all turns on the route between 30 and 90 degrees shall be made within the tolerance of a tangential arc between the straight leg segments defined with a radius of 22.5 NM; 4. Z — for RNP1 routes at and below FL190 to indicate that all turns on the route between 30 and 90 degrees shall be made within the tolerance of a tangential arc between the straight leg segments defined with a radius of 15 NM.
Sequence Number	Numeric sequence number of a point on the route
Latitude	Latitude in decimal degrees of a point on the route
Longitude	Longitude in decimal degrees of a point on the route

Table 6: Information about the routes

Airspace ID	Unique identifier of the FIR (could also be a UIR, Upper Information Region)

Min Flight Level	Minimum vertical boundary of the airspace volume expressed as a flight level, repeated for each point
Max Flight Level	Maximum vertical boundary of the airspace volume expressed as a flight level, repeated for each point
Sequence Number	Numeric sequence number of a boundary point of the FIR's shape
Latitude	Latitude in decimal degrees
Longitude	Longitude in decimal degrees

Table 7: Flight Information Regions (FIRs)

EUROCONTROL R&D Archive data OpenSky can be useful as one for modelling use cases UC2, UC4 and UC5.

4.2 OpenSky

The OpenSky Network is a community-based receiver network which continuously collects air traffic surveillance data (specifically, ADS-B and Mode S messages) and makes it available to researchers. OpenSky works with off-the-shelf sensors run by volunteers distributed over Central Europe. As noted in [40], more than 30 % of Europe's commercial air traffic is captured in OpenSky. In contrast to the existing services offering live visualization of air traffic on Internet (e.g., Flightradar24), OpenSky offers access to the historical raw data necessary for independent research.

OpenSky collects the following primary data of ADS-B-equipped aircraft: aircraft identification, its position and velocity. In addition to the aircraft state vector, some aircraft also broadcast status messages that contain information on emergencies, priority, capability, navigation accuracy category, and operational modes. Furthermore, OpenSky stores metadata for each message, including timestamp of the reception, the receiving sensor's ID, the ADS-B checksum, and the raw message as a hex string.

OpenSky provides an initial abstraction of the data by separating messages from any aircraft into flights. It can be used represent real movement of aircraft in the air space, and thus, to model aircraft routes and traffic density.

OpenSky database is not sufficient as the data source for any of the use cases considered in SynthAIr, however, it can be useful as one of the data sources for the use cases UC2, UC4 and UC5.

4.3 BTS database

Bureau of Transportation Statistics (BTS) of the US department of transportation collected diverse transport-related data, which can be openly downloaded from the website https://www.bts.gov/ and used.

In relation to SynthAIr, the following databases are of interest:

Air Carrier Statistics (Form 41 Traffic)- U.S. Carriers (All Carriers): Contains monthly data reported by certificated U.S. (U.S. and foreign) air carriers on passengers, freight and mail transported. Also includes aircraft type, service class, available capacity and seats, and aircraft hours ramp-to-ramp and airborne.

Airline On-Time Performance Data: Contains monthly data reported by US certified air carriers, specifically scheduled and actual arrival and departure times for flights.

Airline Origin and Destination Survey: Origin and Destination Survey is a 10% sample of airline tickets from reporting carriers. Data includes origin, destination and other itinerary details of passengers transported.

Intermodal Passenger Connectivity: The Intermodal Passenger Connectivity Database is a nationwide data table of passenger transportation terminals, with data on the availability of connections among the various scheduled public transportation modes at each facility. In addition to geographic data for each terminal, the data elements describe the availability of rail, air, bus, transit, and ferry services. This data has been collected from various public sources to provide the only nationwide measurement of the degree of connectivity available in the national passenger transportation system.

BTS database is a useful data source for modelling use cases UC2, UC4 and UC5. However, it covers predominantly the air traffic over the US.

5 State of the art on synthetic data generation

5.1 Synthetic Data: definition and type

Synthetic data has been identified as an approach to overcome these challenges. It involves the use of algorithms to create data that statistically resembles real-world data but does not correspond to any real individuals or events. This method holds promise for maintaining the statistical validity of datasets while avoiding issues related to privacy and data access. **Synthetic data generation (SDG)** refers to methods and strategies that can mitigate these challenges, providing a pathway to more robust and equitable ML solutions.

Synthetic data are algorithmically generated data that simulates the statistical characteristics of real-world phenomena without replicating specific events or individual records. This form of data is a synthetic artifact, engineered to emulate the statistical distributions of actual data while avoiding the issues tied to real data collection and usage.

Different kind of synthetic data can be generated, each tailored to specific applications in machine learning. Here a brief overview of the types of synthetic data that can be generated:

- **Tabular Synthetic Data**: This kind of data mimics the structure of traditional databases or Excel spreadsheets. Its generation involves creating rows and columns of categorical and numerical data, which is instrumental for tasks such as data anonymization, imbalance correction in training datasets, or database testing without exposing sensitive information [41], [42], [43], [44], [45], [46].

- **Image Synthetic Data**: By leveraging generative models, synthetic images can be created to augment datasets where collecting real images is impractical or privacy-sensitive. These images are pivotal for training robust computer vision models, especially in domains like medical imaging or autonomous driving, where real data can be scarce or highly regulated [47], [48], [49], [50], [51].

- **Text Synthetic Data**: Artificially generated text data can serve to enhance language models' understanding and generation capabilities. This synthetic data supports tasks like chatbot training, sentiment analysis, and other NLP applications, often addressing the shortage of labelled data or the need to protect privacy in sensitive text corpuses. One prominent approach in this domain is the use of Generative Pre-trained Transformers (GPT), which have significantly advanced natural language processing (NLP) tasks by generating high-quality synthetic text data [52], [53].

- **Audio Synthetic Data**: Synthetic audio samples, including speech, music, or ambient sounds, can be created to train models for applications like speech recognition, audio classification, and virtual assistant technologies. Such data is particularly useful when real audio data collection is challenged by noise, privacy issues, or environmental constraints [54], [55], [56], [57], [58].

- **Time Series Synthetic Data**: This type of data is structured to reflect temporal dynamics and is crucial for models that predict stock market movements, weather patterns, or energy consumption forecasting. Its generation helps in creating diverse scenarios for model training without the wait for real-time data accumulation [59], [60], [61], [62], [63].

- **Spatial Synthetic Data**: Artificial spatial data helps in geospatial analyses, urban planning, and environmental modelling. Generated to represent geographic locations and their attributes, it can be used in GIS systems and spatial data infrastructures where real data may be limited due to geographical constraints or sensitivity [64], [65], [66].

- **Video Synthetic Data**: This involves the creation of artificial footage that can train models to understand and interpret dynamic scenes, which is essential in security surveillance, sports analytics, and the development of interactive media. Generated videos can provide diverse and voluminous datasets necessary for complex model training, bypassing the lengthy and costly process of capturing real-world videos [67], [68], [69].

In conclusion, synthesizing data offers a regulated, expandable, and generally more ethical method for gathering data and training models across different fields, propelling advancements in machine learning while protecting privacy and diminishing the dependence on real-world data acquisition. Aligned with SynthAIr's mission and our specific field of interest, we have chosen to concentrate primarily on the synthesis of tabular and time series data. Hence, our forthcoming review of the state-of-the-art will specifically address the progress in synthetic data generation techniques pertaining to these two categories of synthetic data.

5.2 Applications and Use of Synthetic Data

Synthetic data is emerging as a transformative solution across various domains and industries where there are issues in data access and data quality. In this section we will consider as example three main application domains where the use of synthetic data has been started to be considered: **healthcare**, **finance** and **automotive and robotics**.

5.2.1 Synthetic Data in Healthcare

The integration of synthetic data within healthcare research and practice has recently become an important aspect of modern medical innovation. This progression has been defined by some recent different literature review and scientific studies highlighting different aspect of using synthetic generated data in healthcare [70], [71], [72], [73]. Among the others these works:

- delves into the creation and application of synthetic electronic health records (EHRs). This method provides a valuable solution for training diagnostic models, balancing the need for comprehensive data while safeguarding patient privacy.

- highlights the role of synthetic data in enhancing drug development processes, particularly in clinical trial simulations. This approach assists in evaluating new treatments' efficacy and safety before proceeding to human trials, thus optimizing the research process.

- reflects the application of synthetic data in medical imaging, a field where real data is often scarce or sensitive. This approach aids in developing and refining diagnostic tools, for example in areas like radiology.

- addresses the challenge of data imbalance in healthcare datasets. They explore the generation of synthetic data to represent rare conditions, thereby aiding in the development of accurate diagnostic models.

- underscores the enhancement of machine learning applications in medicine through synthetic data, especially in training models with limited real-world data.

5.2.2 Synthetic Data in Finance

The use of synthetic data in finance has become increasingly prominent, offering solutions for various challenges such as data scarcity, privacy concerns, and the need for robust testing environments. This trend is evident in several key studies and applications within the domain.

First of all, [74] offers a comprehensive overview of the potential and challenges in generating synthetic data for finance. This study provides insights into the opportunities synthetic data presents in finance, along with the practical challenges and considerations involved in its implementation.

More specifically, [75] delves into generating synthetic financial transactions to enhance anti-money laundering models. This approach aids in creating realistic transaction datasets for testing and improving the efficacy of financial monitoring systems. In [76], synthetic data is utilized to augment training datasets for deep reinforcement learning models in financial trading. This study exemplifies the enhancement of predictive models in trading by leveraging synthetic datasets to simulate various market conditions. Similarly [77] presents a methodology for generating synthetic financial time-series data. This research is pivotal in simulating realistic market scenarios, which are essential for testing financial strategies and models under various market conditions.

These studies together highlight the increasing relevance and varied uses of synthetic data within the financial sector. They demonstrate how synthetic data is being utilized to improve anti-money laundering processes and refine trading models. Additionally, the ability of synthetic data to emulate intricate financial scenarios showcases its essential role in the ongoing development of the finance industry.

5.2.3 Synthetic Data in Automotive and Robotics

In the automotive sector, the development of autonomous driving systems heavily relies on synthetic data. In [78], the authors discuss the importance of understanding the gap between synthetic and real-world data in autonomous driving applications. This research is crucial for improving the reliability and safety of autonomous vehicles.

In robotics, the integration of synthetic data for real-time applications is explored in [79]. This study addresses the challenge of detecting and interacting with humans in real-time using synthetic data, a critical aspect for developing responsive and safe robotic systems. In the same area, [80] presents a methodology for generating synthetic data to train and enhance machine learning models in robotics. This approach facilitates the development of more advanced and efficient robotic systems capable of complex tasks.

The study [81] illustrates the use of synthetic data in creating virtual environments for testing and validating automotive systems. This approach allows for extensive testing under various scenarios, which would be impractical or unsafe in real-world settings. In [82]the focus is on procedural modelling techniques to generate synthetic data for automotive applications. This method provides a scalable and efficient way to create diverse datasets for testing and validation purposes.

Lastly, [83] demonstrates the application of synthetic data in industrial settings, specifically for object recognition tasks. This is particularly relevant in robotics, where accurate object recognition is fundamental for various applications.

These studies collectively underscore the growing importance and diverse applications of synthetic data in automotive and robotics. From enhancing autonomous driving systems and robotic

interactions to facilitating rigorous testing and object recognition, synthetic data has become an important tool in advancing these fields.

5.3 Synthetic Data Generation methods

The development of synthetic data has become increasingly sophisticated, with several methodologies emerging to create datasets that cater to a variety of analytical needs. Broadly categorized, the methods for synthetic data generation include **statistical modelling**, **agent-based modelling** (as mentioned in section 3.1), and **generative modelling**, each grounded in different principles of simulation and data synthesis. In this section, we will provide an overview of the different categories of synthetic data generation methods with more focus on generative modelling, which is the main methods we will use in SynthAIr project.

5.3.1 Statistical Modelling

Statistical modelling techniques have long been employed in synthetic data generation, particularly in fields where the underlying data distributions are well understood or can be accurately modelled. These methods typically benefit from using probabilistic models to generate synthetic data that closely resembles the characteristics of the original dataset. These methods encompass a range of approaches, including parametric and non-parametric models, as well as classical statistical techniques such as Monte Carlo simulations, bootstrapping, and resampling methods in addition to data augmentation methods such as Synthetic Minority Over-sampling Technique.

Parametric models [84], such as linear regression [85], logistic regression [86] and Mixture Models or Copula [87] assume a specific functional form for the data distribution and estimate the parameters from observed data. These models are useful when the underlying distribution can be reasonably approximated by the chosen parametric form. Non-parametric models, on the other hand, make fewer assumptions about the data distribution and instead rely on flexible structures to capture complex patterns [84]. Techniques like kernel density estimation (KDE) [88] and nearest neighbour methods [89] fall into this category and are particularly advantageous in scenarios where the data distribution is unknown or highly non-linear.

Monte Carlo simulations represent a versatile approach to synthetic data generation, especially in scenarios where there are complex interactions between variables or stochastic processes. It uses random sampling techniques to estimate complex numerical results by generating numerous random samples [90]. In the context of synthetic data generation, Monte Carlo methods can be utilised to generate synthetic datasets that mimic the statistical properties of the observed data.

Resampling methods, such as bootstrapping, are another flavour of statistical techniques that involve sampling with replacement from the observed data to create synthetic datasets that mimic the properties of the original data [91]. These methods are particularly useful when dealing with small datasets or when there is a need to estimate sampling distributions or confidence intervals.

Synthetic Minority Over-sampling Technique (SMOTE) [92] is a popular statistical method for data augmentation, especially for handling imbalanced datasets. It generates synthetic data of the minority class by interpolating between existing samples. A similar approach is Adaptive Synthetic Sampling (ADASYN) [93], which adaptively generates synthetic samples of the minority class, focusing on the harder-to-classify samples. While these methods can help balance the class distribution, they are

limited to generating synthetic data within the convex hull of the original data, which may not always capture the true underlying data distribution.

Variants of SMOTE can improve the quality of the generated synthetic data. Borderline-SMOTE [94] adapts the SMOTE algorithm to focus on samples near the decision boundary, which are more likely to be misclassified. SVMSMOTE [95] uses a Support Vector Machine (SVM) to identify the borderline hyperplane separating the minority and majority classes and then generates synthetic minority samples along this hyperplane. However, especially in noisy datasets, generating synthetic samples near the decision boundary may introduce noise into the synthetic data. To address this issue, the Safe-Level SMOTE [96] introduces a safe-level parameter to control the generation of synthetic samples. Similarly, the KMeans-SMOTE [97] avoids noisy samples by using k-means clustering to identify safe clusters with a high ratio of minority observations and generate synthetic samples within these clusters.

While statistical modelling techniques offer interpretability and control over the generated data distribution, they may struggle to capture the full complexity of real-world datasets, particularly in high-dimensional or non-linear settings. Additionally, the performance of these methods heavily relies on the adequacy of the chosen model assumptions and the quality of parameter estimation from limited observed data. Despite these limitations, statistical modelling remains a valuable tool in synthetic data generation, especially when combined with domain knowledge and expert judgment to tailor models to specific applications.

5.3.2 Agent-Based Modelling

Agent-based modelling (ABM) is a computational model that has gained popularity in various fields, including economics, sociology, biology, and ecology [98], [99], [100]. ABM entails creating models of autonomous agents, such as individuals, organizations, or other entities, and simulating their interactions within a specified environment. By modelling the behaviours and interactions of individual agents, ABM can be used to generate synthetic datasets that capture emergent phenomena at the macroscopic level.

In ABM, agents typically follow a set of rules or algorithms that govern their behaviour and interactions with other agents and the environment [101]. These rules can be based on empirical data, theoretical principles, or a combination of both. By varying the parameters of the model or introducing random elements, ABM can generate diverse synthetic datasets that explore different scenarios or hypotheses.

One advantage of ABM is its ability to capture complex, nonlinear dynamics that arise from the interactions between individual agents. By simulating the behaviour of multiple agents over time, ABM can generate synthetic datasets that exhibit emergent properties, such as self-organization, adaptation, and evolution. This makes ABM particularly useful for studying phenomena that are difficult to observe or replicate in real-world settings, such as the dynamics of social networks, the spread of infectious diseases, or the evolution of ecosystems.

While agent-based modelling can potentially offer a powerful framework for generating synthetic data that provide valuable insights into the behaviour of complex systems and inform decision-making in various domains, it also presents challenges related to model complexity, computational resource requirements, and validation against real-world data. Additionally, ABM requires careful consideration of the underlying assumptions and simplifications made in modelling agents' behaviours and interactions, as these assumptions can influence the model's predictive accuracy and generalizability.

5.3.3 Generative Modelling

Generative models in AI, particularly those utilizing deep learning approaches, have significantly advanced synthetic data generation. Techniques like **Generative Adversarial Networks** (GANs) [102], **Variational Autoencoders** (VAEs) [103] and **Diffusion Models** [104] have been pivotal in generating high-fidelity data across domains, including images, text, and tabular data. The generative aspect of these models is rooted in their ability to learn the distribution of the input data and produce new instances that could have plausibly come from the same distribution. Such capabilities are not just impressive but also practically valuable in augmenting datasets, especially when dealing with privacy-sensitive or rare data scenarios. By bolstering datasets with synthetic yet realistic examples, these models help in training more robust and generalizable machine learning models. In the following subsections we will focus on the use of GAN, VAE and Diffusion model with focus on timeseries data.

In time-series data, synthetic data generation is notably intricate due to inherent temporal patterns. The generative process must adeptly capture both feature distributions and temporal relationships. Deep learning techniques are particularly adept at modelling these multifaceted relationships. However, real-world scenarios often present limited time-series data, either in sample numbers or historical length. For instance, predicting stock market trends for newly public companies or forecasting staffing needs for newly inaugurated retail outlets can be challenging due to data paucity. Such scenarios necessitate a data generation technique that is robust despite limited data and allows for the introduction of specific time-series patterns pertinent to the use-case.

5.3.3.1 Generative Adversarial Network (GAN)

Some of the recent advancements in synthetic data generation have hinged on Generative Adversarial Network GANs, introduced in 2014 [79] especially those employing recurrent neural networks (RNNs) for both generation and discrimination [106], [107]. However, the inherent complexity of temporal relationships means that the conventional approach of binary discrimination between real and synthetic data falls short in capturing temporal dependencies. This has led to the exploration of specialized mechanisms within GAN networks, such as the fusion of supervised training, typically used in autoregressive models, with the unsupervised training of GANs [106].

A comprehensive review [108] delves into the application of GANs for time-series data, highlighting the benefits of GANs as tools for data augmentation. These benefits range from addressing data shortage issues by augmenting smaller datasets, to data recovery, noise reduction, and the generation of differentially private datasets that safeguard sensitive information. The review also enumerates several state-of-the-art GAN models tailored for time-series data, such as C-RNN-GAN [109], RCGAN [107], TimeGAN [106], and SigCWGAN [110]. However, a persistent challenge with RNN-based GAN models is their inability to produce extended, realistic synthetic sequences. This limitation stems from RNNs processing time-steps sequentially, leading to more recent time-steps disproportionately influencing the generation of subsequent ones. This sequential processing makes it challenging for RNNs to establish relationships between distant time-steps within an extended sequence.

The transformer architecture, characterized by its multiple self-attention layers [111], has recently gained prominence. Demonstrating superior performance over other neural network architectures, such as CNNs for images and RNNs for sequential data [112], [113], transformers have showcased their versatility across various tasks. Recent endeavours have sought to integrate the transformer model within GAN architectures to enhance synthetic data quality or streamline the training process [114] for tasks like image and text generation. Given that transformers were originally designed to handle

extended text sequences and are immune to the vanishing gradient problem, they theoretically should outperform RNN-based models in time-series data generation.

Building on this premise, a recent study [115] introduced a transformer-based GAN model, TTS-GAN, for synthetic time-series data generation. This approach trained a distinct GAN model for each dataset class. However, a limitation of this approach was its struggle to train GAN models for classes with limited training instances, making it challenging to generate realistic sequences for such classes. To address this, the study proposed a conditional GAN for time-series generation, termed TTS-CGAN [116]. This model was trained on data from all classes concurrently, allowing for controlled data generation for specific classes by priming the model with the appropriate input. This holistic training approach enabled TTS-CGAN to benefit from transfer learning effects between classes, facilitating better low-level feature representation learning. The deeper network layers simultaneously fine-tuned high-level features for each class. The study showcased TTS-CGAN's efficacy using novel similarity metrics and experiments that highlighted the impact of synthetic data augmentation on classification tasks.

However, it has been demonstrated that GANs may capture less diversity compared to state-of-the-art likelihood-based models. The training process of GANs is often fraught with challenges, including a tendency to collapse without correct chosen of hyperparameters and loss functions [117]. Additionally, GANs are known for their instability during training [118] and are susceptible to the mode collapse issue, where the model fails to capture the variety in data [119]. These limitations have prompted a shift towards alternative models like Variational Autoencoders (VAEs) and diffusion models, which offer more stability and diversity in synthetic data generation.

5.3.3.2 Variational Autoencoder (VAE)

The current state of the art in Variational Autoencoders (VAEs) for time series generation is marked by significant advancements that cater to the diverse challenges posed by time series data.

One development is for example the integration of causal mechanisms within VAE frameworks, as seen in [120], which focuses on medical time series data. This approach not only enhances predictive accuracy but also adds an interpretive layer to the generated data, crucial in medical applications. Hybrid models that amalgamate VAEs with other forecasting techniques are also gaining prominence. For instance [121] combines VAEs with additional models to boost forecasting precision, indicating a trend towards leveraging the strengths of multiple machine learning techniques. Bidirectional priors in VAEs, as introduced in another study [122], represent a significant leap in generating complex time series patterns. This innovation opens avenues for generating more nuanced and accurate time series data, crucial for various applications. Addressing incomplete time series data, a challenge in real-world applications, has been tackled through models that integrate Neural Ordinary Differential Equations (ODEs) with VAEs, as demonstrated in [123]. This approach elegantly handles data gaps, preserving the integrity of time series analysis.

In data augmentation, the efficacy of Beta-VAE has been explored, particularly in comparison to models like WGAN-GP, as shown in [124]. This comparison is crucial for understanding the best practices in time series data augmentation, especially for enhancing classification performance.

The application of Koopman theory (Hamiltonian systems and transformation in Hilbert space) in VAEs, tailored for both regular and irregular time series data [125], marks a significant stride in bridging linear dynamical systems and nonlinear time series analysis. This method provides a new lens through which time series data can be understood and modelled. Finally, the generation of multivariate time series data, which is essential in handling complex, real-world scenarios, has been advanced through models

like TimeVAE [126]. This approach highlights the need for and effectiveness of specialized techniques in managing the intricacies of multivariate data.

In conclusion, the landscape of VAEs in time series generation is characterized by a rich tapestry of methodologies, each addressing specific facets of time series analysis. From enhancing predictive accuracy with causal mechanisms to innovating in data completion and multivariate analysis, these advancements collectively push the boundaries of what can be achieved in this evolving field.

5.3.3.3 Diffusion Models

Diffusion models is a subset of deep learning-based generative models that recently gained interest in various machine learning applications due to their generative capabilities. Recent examples are in image synthesis [127], video generation [128], and natural language processing [129].

In recent years, diffusion models have been extended to time series-related applications, such as forecasting [130], imputation [131], and synthesis [132]. Time series forecasting involves predicting future data points based on historical observations, while imputation deals with filling in missing values in incomplete series. Time series generation, or synthesis, diverges from these by aiming to create new time series samples that maintain the characteristics of the observed data.

In conclusion, the domain of synthetic data generation, especially for time-series data, is rapidly evolving. As the demand for data-driven solutions continues to grow, the development and refinement of generative models will play a pivotal role in shaping the future of various industries including the aviation domain.

6 Case selection and next steps

In this section, based on multiple selection criteria divided in 5 categories (refer to Table 8): Data, Analysis and Modelling, Integration, Stakeholders, and Validation, the most promising use cases are selected for further elaboration in the project. The filled elements of the evaluation table are based on the analysis of existing literature and interviews with stakeholders (airports, airlines, ANSPs, air transport researchers). This assessment will be further refined and matured in the coming months, also taking into account the feedback from the first upcoming advisory board meeting, and the final version will be presented in deliverable D2.2 “Definition of use cases”.

As an important input to this selection, we used feedback on the considered use cases from aviation stakeholders (airports, airliners, ANSPs, air transport researchers) obtained in the interviews we conducted in the first months of the project. In the following we discuss some important observations we obtained during the interviews in relation to the specific use cases.

UC1: Turnaround time prediction

Although turnaround time prediction is crucial for the optimization of air transport operations, the interviewed airport stakeholders did not consider this use case as particularly relevant to be considered in SynthAIr. In the past years more and more airports have been collecting (detailed) data about the durations of turnaround operations, which they use for prediction of their turnaround times either using statistical methods or machine learning models. These data are, however, usually confidential. Interviewed airlines confirmed the importance of this process for their planning. However, often they collect sufficient data about handling of their aircraft. They usually are able to make accurate turnaround time predictions on a short term, however long-term predictions are difficult, as many factors influence the turnaround processes, including propagated delays, passenger flows, ground handling processes. ANSPs recognized the relevance of this use case. Some airport stakeholders, as well as university researchers, identified the necessity of considering turnaround in relation to other related processes, such as passenger flow management in airport terminals. In such a way, the airport system could be considered and analysed in a holistic way. However, including all these processes substantially complicates the use case.

UC2: Flight delay prediction

The interviewed airport stakeholders identified the lack of real time data exchange between airports concerning their local traffic situation and delays. They pointed that if a flight was delayed at an origin airport, the lack of timely communication with the destination airport might cause issues with resource allocation (e.g., of ground service equipment, airport personnel) and suboptimal planning of operations at the corresponding destination airport. Currently, some airports develop own models to assess traffic situation and delays at other airports in their airport network.

The interviewed airlines identified the general difficulty of predicting the reactionary, accumulated delay of their aircraft in the end of a day and confirmed the relevance of this use case for airlines. One of the airlines pointed to a high uncertainty associated with flight delays which impedes decision-making of AOCC controllers, which makes this use also important for them.

ANSPs also recognized the importance and relevance of this use case. Despite the fact that many delay prediction models and tools exist, one ANSP interviewee argued that it would be still useful to generate synthetic data representing diverse delay propagation scenarios for different disruptions.

UC3: Passenger flow prediction

The airport stakeholders confirmed that passengers are an important source of uncertainty in the air transportation system. In particular, predicting flows of transfer passengers is problematic, because of the lack of data about them. It is also difficult to predict whether or not a passenger will be on time for boarding. Airlines collect information about passengers, such as the numbers of passengers on particular flight, demand data, check-in time, baggage check-in, in some cases, about the time when a passenger passed the security check. However, because of confidentiality and sensitivity of passenger-related data, exchange of information between airports and airlines is quite limited. Thus, synthetic data generation of passenger flows would be useful for airlines.

Airport passport control is another important bottleneck with a large uncertainty, since airports do not usually have control over this process. Many airports collect data about passenger flows, in particular, using camera's, mobile phone signals and new LiDAR technology. However, these data are usually confidential.

Interviewed university researchers working on (multimodal) passenger flow modelling indicated that they lack data on flight schedules, as well as passenger numbers and flows at the airport and individual flights (load factor). Synthetic data of these types would be useful for them.

UC4: Synthetic Traffic Generator

This use case is closely related to UC2. The interviewed ANSPs recognized that many traffic generators were developed in the past (many of them are not openly available), however, more synthetic traffic data could still be useful to generate, in particular for novel concepts of operations, such as UTM/UTS operations.

UC5: Flight diversion prediction

The airport and ANSP stakeholders confirmed that flight diversion prediction is a relevant case, which, however, to a lesser degree was considered in the related literature. This use case is however might be challenging to model, as flight diversions do not occur often, and there could be diverse causes, which are not always easy to understand from the available data. Furthermore, flight diversion often involves ad-hoc coordination among airlines, pilots, air traffic controllers and airports and is influenced by many factors such as airport capacity, airport departure/arrival sequencing, weather conditions, air traffic complexity. Thus, more data need to be collected or generated to better understand and represent flight diversion under different causes and conditions.

When flight is being diverted, airlines usually do not have information about which other flights were diverted too and are often making reactive decisions based on air traffic controllers' instructions and airport availability. Because of this, this use case is less interesting for airlines.

	Criterion description	UC1_Turnaround Time Prediction	UC2_Flight Delay Prediction	UC3_Passenger Flow Prediction	UC4_Synthetic Traffic Generator for Fast and Real-Time Simulations	UC5_Flight Diversion Prediction
Data	**Availability**	Many airports and airlines collect detailed turnaround data. High level data about off-block/on-block times of aircraft is openly available	Historical flight data, including delays, is available in open databases	Data is partially available (not openly); data about transfer passengers is limited; some areas of airport terminals are less represented in data; data exchange about passengers between airports and airlines is limited	Historical flight data is available in open databases	Historical flight data is available in open databases
	Confidentiality	Detailed data about turnaround steps is usually confidential	Historical data is openly available	This data is usually confidential	Historical data is openly available	Historical data is openly available
	Required and available data types and format - Determine the types of data and the format in which data will be collected and stored, whether structured (e.g., databases), semi-structured (e.g., JSON), or	Openly available data are usually in csv format. The types of data required for high level modelling usually include on- and off-block times of aircraft/flights. Data required for detailed modelling	Openly available data are usually in csv format. The available types of data usually include Flight date, Flight number, Carrier code, Tail number, Flight origin and destination. In addition, the following data is often available: Flight Latitude, Longitude,	(Aggregated) passenger flow data from camera's and LiDARs; passenger check-in time; passenger security check-time; passenger baggage data. More detailed data about transfer passengers is needed. More detailed data about passenger discretionary activities	Openly available data are usually in csv format. The available types of data usually include Flight date, Flight number, Carrier code, Tail number, Flight origin and destination. In addition, the following data is often available: Flight Latitude, Longitude,	Openly available data are usually in csv format. The available data include date, flight number, flight origin, flight original destination, the airport to which the flight was diverted. Useful information about the diversion

	unstructured (e.g., text documents)	include the duration of all turnaround steps, as well as about the resources used for these steps	Flight level data over time; flight on-block and off-block time; scheduled and actual arrival/departure time. The available data forms a good basis for this use case. In addition, airport- and weather-related data might be necessary for high quality delay prediction.	is needed. Interviewed researchers indicated that even flight schedules are difficult to obtain.	Flight level data over time; flight on-block and off-block time; scheduled and actual arrival/departure time.	causes is often not available. Weather information is highly relevant for this use case.
	Quality - Define the level of data accuracy, completeness, and reliability required (consider data cleansing and validation processes)	The data quality varies; detailed data often contain a significant amount of missing and incorrect entries; high level data are usually mostly correct	Available data is mostly correct; some fields might be missing/incorrect	Data are usually aggregated at the level of probability distributions of passenger numbers, waiting times, arrival times, throughput. Not all areas of airport terminals are represented in data. Data about transfer passengers is limited	Available data is mostly correct; some fields might be missing/incorrect	Available data is mostly correct; some fields might be missing/incorrect
	Privacy and security - Ensure that data requirements align	Detailed data are privacy-sensitive, as they reflect the performance	Information about the causes of delays might not be always available	Passenger-related data have high privacy sensitivity	No critical sensitivities	No critical sensitivities

	with privacy and security regulations, and establish measures to protect sensitive information	characteristics of individual ground operators				
	Frequency - how often data should be collected or updated, whether in real-time, daily, weekly, or at some other interval	High-level data are collected per flight on a daily basis; the frequency of detailed data collection varies, depending on the airport	Daily, on a flight basis	Frequently, e.g., with 15-minutes time intervals	On a flight basis	Available on a daily basis
	Documentation - Updated documentation that describes the data's source, structure, and usage to facilitate understanding and collaboration	Data is usually well documented.	Data in open databases is well documented.	Well described in the databases of airports and airlines.	Data in open databases is well documented.	Data in open databases is well documented.
Analysis and modelling	**Analytic capability** of performance variability qualitatively and quantitatively (i.e. amount of uncertainty in prediction tasks)	High level modelling is feasible; low level modelling highly depends on the availability of data. The amount of uncertainty for a particular airline and airport might be limited.	The systemic nature of delays, especially propagated ones make it difficult to make predictions	Passengers are one of the major sources of uncertainty in air transport systems	The effects of disruptions might be uncertain; airline flight scheduling is another source of uncertainty, as well as ATC decisions	Flight diversions are difficult to predict in advance; available data on flight diversions is also limited; the quality of flight diversion prediction largely depends on the quality of weather

						information/prediction.
	Complexity of modelling/prediction	High for complex hub airport, like Schiphol; low/medium for small- and medium-size airports	Prediction might be complex for propagated delays, which are influenced by many diverse airline-, airport-, and ATM-related factors. Synthetic generation of data for diverse (disrupted) air traffic scenarios appears to be feasible	Passenger behaviour is not easy to model. Furthermore, passenger flow management measures are airport- and situation-specific, and are difficult to model even when historical data is available	High, as many factors influence 4D-trajectories of aircraft	There are many causes of flight diversion, with weather as the most significant one; predicting flight diversion is a challenging task, which is influenced by many airport-, ATC- and weather-related factors, however, generating synthetic data on diverted flights might be feasible
	Formal **analysis capability**	Developed models can be well analysed	Developed models can be well analysed	It is expected that developed models will have high uncertainty with limited analysis capabilities	Developed models can be well analysed	Predictive power of the models might substantially depend on the type of the cause of diversion, airport, and air traffic complexity
	Level of **abstraction** (for modelling)	High, as well as low-level modelling is possible	At the level of individual flights, as well as aggregated airport delays	Per passenger or at the level of passenger flows	At the level of individual flights, as well as aircraft flows	At the level of individual flights
	Level of **Generalization**	High for small- and medium-size airports;	Possibly high (considering different	Probably, not high, and airport-specific	Expected to be high	Generalization might be difficult across

		might be problematic for large hub airports	airport network configurations and disruptions)	(however, largely depends on the level of modelling)		different causes of diversions and airports; however, for some types of causes and airports (such as weather) the level of generalization might be high
	Time **horizon** for predictive/forecasting models	Typically, on an hour basis (per flight) – tactical and operational planning; models might also be used for strategic planning	On an hour basis – tactical and operational planning; models might also be used for strategic planning	Short time horizon, e.g., 15 minutes; models might also be used for strategic planning	On an hour basis – tactical and operational planning; models might also be used for strategic planning	On an hour basis (per flight) – tactical and operational planning; models might also be used for strategic planning
	Expected quality of the model	High for small- and medium-size airports; might be problematic for large airports	Potentially, synthetic data could well represent real data; the downstream task of delay prediction is also often addressed well in the literature, in the context of specific cases	Synthetic data generation might be problematic. But it depends on the scope of modelling and the airport size. Prediction models might work well for specific airports/conditions	Probably high, however depends on the scope of modelling	Data generation models for particular types of causes (such as weather) might be of good quality

Stakeholders	**Interest, Relevance, Priority and Urgency**	Interviewed airports did not give a high priority to this case. Airports and airlines collect their own data on turnaround which they use for defining the critical path for the turnaround process. This use case is relevant to airlines; however, they didn't recognize its relevance for SynthAIr.	Interviewed airports pointed at the need of availability of data about the real time traffic situation at the connected airports; some airlines also pointed at a high uncertainty of delay-related information in their planning. ANSPs identified the need to generate synthetic date to better represent different scenarios of delay propagation through a network under different disruptions. This use case has a high priority for all aviation stakeholders.	There is much interest to this use case both from airports and airlines, as it is directly relevant to their operational planning and resource allocation. Available data if often scarce. The stakeholders would profit from synthetic data generation.		The airport and ANSP stakeholders recognized the relevance of this case, as well as limited data available. Better prediction of diversions would allow airports to better prepare and allocate their resources to handle aircraft. On the other hand, one interviewed airline indicated that this case is less relevant for them, because airlines are often making decisions concerning flight diversion reactively
	Novelty	There exist several studies on synthetic data generation using statistical and simulation models; also, the turnaround prediction task was addressed successfully	Use of generative AI for generating (disrupted) traffic scenarios is currently very limited. Simulations models were used in the past for this purpose.	Synthetic data generation for this case is usually done using statical and simulation models, using historical data. However, it is not clear how realistically such models would be able	Synthetic data generation using AI is done to a limited extend in the existing literature; simulators are also used to a limited extent for this purpose	To the best of our knowledge, synthetic data generation was not used for this use case

		in several previous studies.	There exist many studies on delay prediction (downstream task)	to represent new scenarios, not considered in the historical data. Prediction of passenger flows is considered in related literature, with different degree of accuracy, depending on the airport systems considered		
Validation	**Validation difficulty**	Validation of high-level models is not problematic; validation of low-level models depends on the availability of data on the individual steps of turnaround, which is confidential	Developed models both for synthetic data generation and delay prediction could be validated using historical data	Possible, if sufficient data available, which might be problematic, considering confidentiality, data gathering and privacy issues	Validation can be done using historical data and with operational experts	Validation can be done using historical data and with operational experts

Table 8: Matrix for the selection of the use cases

In the following, based on the case selection table above, we make the choice of use cases to be further elaborated in SynthAIr.

UC1 was not selected for multiple reasons: Many existing AI-based studies are already able to make high quality high-level predictions about turnaround time. Furthermore, statistical, and simulation-based methods exist for synthetic data generation for this case. Furthermore, the aviation stakeholders did not recognize the urgency of considering this use case in SynthAIr, although it is still relevant and important for their operations. Data for low level, detailed modelling are confidential, and are not easily accessible for research purposes. Modelling UC1 might be also too complex, since the use cases UC2, UC3, and UC5 could be also seen as its contributors.

UC2 was selected for further elaboration in SynthAIr for the following reasons: All interviewed aviation stakeholders recognized the relevance and importance of this case for their operational planning. In particular, the interviewed airports considered that data generation models to be developed for this use case could be used to produce data representing disruptions and their effects on the traffic originating at the connected airports. In such a way, the airport would be able to better anticipate and allocate their resources. Furthermore, the models developed in this use case can be used to fill in missing or correct erroneous data in the existing open databases with historical flight data. Air transport researchers would also profit from such models, as they could use them to generate traffic data under diverse disruptions. Our interviews also identified the need for this. **Data generation for UC2 is closely related to UC4 and could be seen as a part of it.** UC2 is also chosen for the reason of availability of open historical flight data which could be used for model training. UC2 can be considered at the level of flight schedules, without detailed modelling of aircraft trajectories. Synthetic data can be generated to represent propagation of delays through an airport network under different disruptions.

UC3 was identified as potentially interesting for further elaboration in SynthAIr. Although the interviewed airport and airline stakeholders recognized the relevance and importance of this use case, nevertheless obtaining real data for training synthetic data generation models is a major obstacle for this use case. Furthermore, airports apply diverse, often ad-hoc passenger flow management measures, adapted to particular situations, which are difficult to model and generalize. The final decision on whether or not this case will be chosen will depend on the availability of data and will be made in the coming month.

Disrupted air traffic scenarios considered in UC2 are selected to be modelled as a part of UC4. The experience to be gained and lessons learned with modelling of UC2 will provide further research directions with respect to UC4, e.g., detailed aircraft 4D-trajectory modelling.

UC5 was selected in SynthAIr for further elaboration for the following reasons: The airport and ANSP interviewed stakeholders recognized the relevance and importance of this case, as well as the lack of available data on diverted flights, for which synthetic data generation could be a solution. At the same time, open historical flight databases can be used for training of data generation models. AI-based synthetic data generation has not been considered for this use case. UC5 has also a relation with UC2, as delays could be precursors for flight diversion.

In addition, we also had an interview with U-space researchers, who identified two possible UAS-related use cases, which we consider in the following. A UAS mission usually defines a planned trajectory to be flown by a UAS. However, during the mission execution, the operator may decide to deviate from the planned trajectory. Furthermore, the trajectory may be influenced by environmental

factors. Synthetic data generation can be used to represent such trajectory deviations and further improve operational planning taking into account uncertainties. The other possible use case concerns synthetic data generation of images representing the dynamics of population flows in regions over which UASs are flying. The population density is reflected by pixel intensities in these images. Such synthetically generates images can be used to improve safety and efficiency of UAS path planning.

To conclude, use cases UC2 and UC5 have been chosen for further elaboration in SynthAIr. UC2 might potentially be further extended to UC4. UC3 is potentially interesting to consider, however, it will only be chosen if real data become available for training in the coming months. Furthermore, U-space-related cases considered above might be of interest for SynthAIr too. The synthetic data generation techniques reviewed in Section 5, in particular the ones based on GANs, transformers, and diffusion models appear to be promising to be applied for the selected use cases in WP3 (Synthetic Data Generation for Multivariate Time Series for ATM-automation) and WP4 (Universal Time Series Model for Prediction and Data Generation for ATM-automation).

The integration of advanced generative techniques reviewed in Section 5, such as Generative Adversarial Networks, transformer-based architectures, Variational Autoencoders, and diffusion models into the SynthAIr framework promises to revolutionize synthetic data generation for Air Traffic Management (ATM). These models excel in modelling and synthesizing complex time series data, which is common in aviation scenarios due to factors such as flight traffic variability from weather disruptions or operational anomalies. Models like Transformer are part of these generative approaches and offer advantages in processing timestamped and sequential data, making them ideal for capturing complex temporal correlations within aviation timestamped data, such as the progression of flight delays or the scheduling of aircraft routes. The ability of these AI models to handle sequential dependencies could be crucial for example for predicting the effects of irregular operations across the flight network.

Moreover, techniques like Diffusion Models and VAE applied to timestamped data, are able to train model to synthetize high quality synthetic data with high degree of fidelity (i.e., generation of realistic flight trajectories and passenger flows) with stability in the training process.

By exploring these sophisticated generative models, the SynthAIr project aims to produce synthetic datasets that are not only diverse and rich but also maintain high levels of accuracy and realism. This will enable more effective decision-making in ATM, fulfilling the needs encapsulated in the selected use cases UC2 and UC5, and potentially extending to UC4.

7 State of the art on related projects

AI is considered one of the main enablers to overcome the current limitations in the ATM system. A new field of opportunities arises from the general introduction of AI, enabling higher levels of automation and impacting the ATM system in different ways [133].

The Strategic Research and Innovation Agenda (SRIA) describes high-level R&I needs/challenges that AI should tackle in aviation:

1. **Trustworthy AI powered ATM environment** – Consider aviation/ATM AI infrastructure that can capture the current and future information required to support AI-enabled applications with the required software developments processes, using robust architectures for ATC systems to provide ATCOs and pilots with good level of confidence of automation and decision aiding tools.
2. **AI for prescriptive aviation** - AI will help aviation to move forward from a reactive (to act when a problem appears) to a predictive (anticipating a problem, enabling innovative preventive actions) and even a prescriptive paradigm (adding the capability to identify a set of measures to avoid the problem).
3. **Human – AI collaboration: digital assistants** - The interaction between humans and machines powered by AI, or other sub-branches such as reinforcement learning (RL), explainable AI (XAI) or natural language processing (NLP), will positively impact the way humans and AI interact. These advances aim to increase human capabilities during complex scenarios or reduce human workload in their tasks, not to define the role of the human or to replace the human, but to support him.
4. **AI Improved datasets for better airborne operations**- Datasets are essential to AI-based application development. R&I should be conducted to generate and in particular to enable the automation of such aviation-specific data sets from a large variety of on-board and ground communication across the network, which could then a enable broad range of AI-based applications for aviation (e.g. voice communications between ATC and pilots).

The following projects highlight the state-of-the-art in SESAR R&D Artificial Intelligence in Air Traffic Management.

7.1 SESAR 2020 Wave 1 and Wave 2

7.1.1 Exploratory research

ARTIMATION

Transparent Artificial Intelligence and Automation to Air Traffic Management Systems (ARTIMATION) project main aim was to introduce innovative AI methods to predict air transportation traffic and to optimise traffic flows based on the domain of explainable artificial intelligence. The project duration was from July 2020 to December 2022.

The project addressed two different use cases: a (1) Delay Prediction and Propagation tool, sub-task to optimise the use of runways, and the (2) Conflict Detection and Resolution to better understand

how ATCOs can be supported in terms of decision-making in the context of conflict resolution (different visualization techniques providing different explainability levels).

The Delay Prediction use case aimed at optimizing the runway use introducing explainability through the visualization of parameters influencing an aircraft delay.

For delay prediction, a comparison among the ML models shows XGBoost performed better than other models. From the algorithmic perspective, XGBoost is more scalable and better at handling spare trees and optimizing errors than RF and GBDT. XGBoost is also a much faster algorithm for learning with large datasets compared to other ML methods. Considering take-off time delay propagation, while comparing the three ML models, (i.e., RF, XGBoost and LSTM models), LSTM performed better than the other, although the overall accuracy was not so good. The LSTM was considered better at solving sequence or temporal dependency; however, it requires a large number of data than RF and XGBoost. The results may be due to insufficient samples in the dataset, and the sequences only depend on two previous flight information. The HMI and visual presentation as a way to improve explainability on the results was less explored, further research was considered necessary to better understand how to identify the most relevant parameters to be shown event by event to the Air Traffic Controllers.

The conflict detection uses case experiment consisted in a low fidelity human in the loop simulation with 21 participants (11 professional ATCOs and 10 ATCO students). The duration was one hour of conflict resolution tasks using three explanation conditions: (1) Black Box, where only the selected solution is presented, (2) Heat Map, where a corpus of potential solution is displayed thanks to a density map, (3) Story Telling, where data driven storytelling technique was applied to convey the explication of the proposed solution. The data was collected through debriefings at the end of the session, over-the-shoulder observations, questionnaires, and neurophysiological measurements. In general, expert ATCOs were less optimistic about the conflict resolution visualization in terms of performance improvement. Higher transparency was considered more useful for less timely critical or tasks or operational phases in which the ATCOs are subject to lower risk of cognitive workload, like planning tasks [134].

MAHALO

The Modern ATM via Human/Automation Learning Optimisation (MAHALO) project, which lasted from July 2020 to December 2022, aimed at exploring new avenues for human-AI teaming in Air Traffic Control (ATC) environments. It focused on two key concepts: strategic conformance and transparency. Strategic conformance refers to the alignment of Machine Learning (ML) models with the strategies and preferences of human controllers. Transparency pertains to the development of AI systems that convey their decision-making processes in a manner that is interpretable by human operators, utilizing clear textual and visual cues.

The project created a hybrid ML system that combined Supervised Learning and Reinforcement Learning techniques to perform Conflict Detection & Resolution (CD&R) tasks. This hybrid model was integrated into an enhanced prototype ATC display featuring transparency elements for visualizing and contextualizing the AI control behaviour. After several development trials, the project culminated in two field studies conducted in two European countries, involving a total of 34 ATCOs. During these studies, controllers' behaviour was recorded in a pre-test phase and used to train the strategic conformance ML system. The main experiment trials then manipulated the strategic conformance of the ML models (either personalized, group average, or optimized) and the transparency of the conflict resolution advisories (as either a basic vector depiction, an enhanced graphical diagram, or a diagram-

plus-text presentation). The results [135] were measured by objective performance and behavioural data, as well as self-reported workload and survey responses. The results revealed a significant impact of strategic conformance on controllers' response to advisories, with controllers responding more positively to advisories that matched their preferred separation distance. No main effects of advisory transparency were found, but transparency did interact with strategic conformance.

MAHALO concluded that increased transparency could benefit understanding of the system and/or situation but does not necessarily benefit acceptance of a system and agreement with its advisories. The effect might be the opposite, where increased transparency decreases acceptance and agreement simply because the system is offering an explanation that reveals that its reasoning is different from that of the operator [136].

7.1.2 Industrial research

As part of SESAR projects there was SESAR PJ04 project, known as Total Airport Management (TAM), this project proposed the evolution toward a “performance-driven” airport through holistic monitoring of demand and capacity and the decision making based on more reliable information with enhanced decision impact assessment. The duration of the project was between 2019 and 2023.

The project PJ04 - TAM delivered two (2) SESAR Solutions through which dedicated tasks to support validation activities have been performed: (1) Solution PJ.04-01 ‘Enhanced Collaborative Airport Performance Planning and Monitoring’ that builds on the work performed in SESAR1 specifically in relation to SESAR Solution #21 (’Airport Operations Plan and AOP-NOP Seamless Integration’) and (2) Solution PJ.04-02 ‘Enhanced Collaborative Airport Performance Management’ that focused on an enhanced collaborative airport performance management, facilitated by access to real-time information captured in the form of performance dashboards showing ‘what has happened’, ‘what is happening’ but importantly ‘what is predicted to happen’. Work has been performed in the specific context of environmental impact planning and monitoring in order to ensure that environmental performance is fully integrated into the airport operations management process. (Partial V2 Solution maturity achieved) [137].

For the frame of SynthAIr project solution PJ.04-02 ‘Enhanced Collaborative Airport Performance Management’ has addressed aspects that can be particularly important to consider, especially the work performed as part of the 'Digital Smart Airports' work package, Operational Improvement 29.1 - Airside/Landside Performance Management (targeting V3 maturity).

This operational improvement addressed Airport Airside/Landside Performance Management which can be enhanced through incorporation of a rationalised dashboard fed with all landside and airside. This is expected to lead key performance indicators covering TAM processes such as passenger, baggage or stand, and achieved thanks to:

- Awareness on potential airport performance degradation (through integrated models that forecast future performance).
- Impact assessment and evaluation of predefined solution scenarios trading-off KPIs (supported by previously-performed post-analysis activities, and possibly by machine learning capabilities, integrated models that permit stakeholders to model what-if scenarios).
- Collaborative Demand and Capacity Balancing (DCB) decision-making between airport stakeholders for potential re-evaluation of solution scenarios and selection of the one that

would consist in the best trade-off between key performance areas (KPAs) and best limit the overall airport performance deterioration.

As outcomes of the project the following aspects were considered important as R&D next steps:

• The enhancement of the airside processes with the inclusion of landside (passenger and baggage flow) process outputs (shared in the AOP via the TOBT update) that can affect ATM performance.

• The question of intermodality can cover the notion of an 'integrated' passenger experience for example a journey with a combined rail and flight ticket issued at a single point of sale. In addition, the passenger experience linked to a flight journey also encompasses the question of both access to and egress from the airport before and after the journey. Different modes of transport (road, rail, …) can be used for airport access and therefore the flight element of the journey is part of an overall multi-modal process. The work performed in this area will address how an increased knowledge of transport performance of covering airport access can be made available to airport stakeholders as a means of identifying potential access issues likely to have ramifications on the punctuality of operations.

• There is a need to enhance the information sharing and collaborative decision making between the airside and landside processes in an airport. These two processes have traditionally been managed in isolation, but in reality, there is a significant degree of 'coupling' between these two processes with the performance of one process having a potential for a significant impact on the other. For example, a landside process performance issue can have an impact on punctuality (passengers not being at the gate on time) which in turn can have ramifications on the parking stand use. The work performed in Wave 1, which focused on the construction of performance dashboards at both the individual process level, and holistic level will be further developed and possibly supported by local and/or network-based services/applications. Similarly, the support to decision-making will be enhanced by the further development of tools such as 'what-if' functionality as well as the use of enhanced predictions through techniques such as machine learning. Business intelligence/machine learning should help stakeholders to share the same vision and collaborate in root cause analyses incorporating real-time information presenting both "what has happened" and also "what is predicted to happen" through forecast or predicted future airport performance and what-if capabilities enabling the proactive management of situations.

• The management of the turnaround process is fundamental to the punctuality performance of an airport and to the predictability of its operations. Work will focus on a detailed monitoring of the different processes relating to the turnaround to provide an early warning indicator of process and infrastructure inefficiencies / issues / failures, resulting in possible delays.

Main outcomes:

At the end of wave 3 most guidelines offered indications of greater granularity within the "AI Explainability" building block, not at a technical level, i.e., algorithmic, but at a much more operational one and concerning the activities of frontline operators for whom the personalization, transparency, and human-machine interaction aspects are crucial. One of the main outtakes was that operators identify as more important to trust the system than getting explanations on the AI decisions in the tactical phase. Trust has its foundations on the certification and training phases, where respective actors should dig into the system to understand its behaviour, validate it and, eventually, build this fundamental trust. These aspects may be accompanied by regulatory changes in terms of liability. The projects developed principles and recommendations which may serve as a starting point to address explainability in future AI applications in ATM [138].

Another conclusion of the exploratory projects from SESAR wave 3 was related to the cost- and time-consuming aspect of data acquisition - and that the data quality assurance for AI purposes. One recommendation was the creation of an open access data lake/repository of the ATM/AI community with raw data, but also data that is pre-processed and cleaned, according to data quality standards might be an interesting approach to save time and costs, especially for ER projects which usually develop AI solutions on a proof-of-concept level [x], which is a topic, also highlighted in The Fly AI Report [138].

EASA itself recently released an expansion of the Trustworthy AI building blocks (EASA First usable guidance for Level 1 machine learning applications - Issue 01), in which it introduces a clearer differentiation between Development Explainability, to be achieved at an algorithmic phase, and Operational Explainability, which instead must be pursued within the front lines of operators. More generally, this concept fits into the broader impact of what Human Factors can bring to Artificial Intelligence and Machine Learning in safety critical organizations. Human AI Teaming will aspects become more prominent soon and, again, the constructs of personalization, transparency, and human-machine interaction will become more decisive [136].

7.2 SESAR 3 Digital European Sky 3 R&D

TRUSTY

Launched in September 2023, TRUSTY is an exploratory project that focuses on human-AI teaming for Remote Digital Tower (RDT) operations. RDT systems allow human operators to control airport airspace remotely through audiovisual and sensor-based aids. TRUSTY seeks to augment the monitoring and alerting capabilities of current RDT systems by introducing explainable AI (XAI) and Multimodal Machine Learning (MML) models that would aid human controllers in crucial tasks such as taxiway and runway monitoring -e.g., bird hazards, unauthorized airspace use by drones, aircraft-runway misalignment during approach [139].

Central to the project's vision is the implementation of a human-centred design methodology that fosters effective collaboration between humans and AI. This approach is multi-faceted, emphasizing: (1) A transparent design, where AI operation and decision-making are contextualized by visual and textual explanations. (2) AI adaptability where the model learns and adjusts its behaviour to individual. (3) User acceptance and trust by involving and gathering feedback from end users in the system design and implementation loops. The project is expected to build a conceptual framework that defines of trustworthy design of machine learning models design of AI-tools to support RDT tasks and validate the models and relevant interface prototypes through Field Operational Tests. The lessons learned will be consolidated into a set of guidelines that will delineate best practices for developing explainable AI-based systems for RDT operations.

MultimodX

Integrated Passenger-Centric Planning of Multimodal Transport Networks (MultimodX) [140] is an exploratory research project that aims at assuring a more efficient, predictable, and environmentally sustainable door-to-door passenger journey focusing on air and rail as natural multimodal partners. The project will develop a set of innovative solutions and decision-making tools to support the coordinated planning and management of multimodal transport networks. Specifically, the project will develop a modelling and evaluation framework, and a solution to enable the coordinated design of air and rail schedules according to expected demand behaviour. The modelling approach that will be used by this project has been identified as interesting to consider for some of SynthAIr use cases, namely the ones that deal with predictions.

ASTRA

The AI-enabled tactical FMP hotspot prediction and resolution (ASTRA) is an exploratory project aims to advance the capabilities of Air Traffic Flow and Capacity Management in predicting and resolving traffic hotspots at a pre-tactical stage. Currently, traffic planning is based on flight plans submitted to the Flow Management Position (FMP) several days before operation. However, the static nature of such data cannot accommodate for dynamic changes such as convective weather, ground delays or airspace closures, leading to inaccuracies in traffic hotspot forecasts. The project seeks to fill the operational gap between FMP and CWP by developing an AI-based decision support tool capable of predicting and resolving traffic hotspots with a longer look-ahead time. The tool aims at reducing last-minute interventions, decreasing the burden on CWP controllers, and ultimately enabling more efficient use of airspace [141].

ASTRA aspires to deliver a solution at TRL 2 by developing an AI-based FMP function for traffic hotspot prediction and a Human Machine Interface (HMI) concept that enables operators to interact with the function. The solution will be demonstrated and validated by human-in-the-loop Real-Time Simulations in a representative operational environment.

JARVIS

JARVIS is an Industrial research project that will develop three AI-based solutions (digital assistants) aimed to team with their human counterparts an (1) airborne digital assistant to support crew in single pilot operations; (2) an ATC digital assistant and (3) an airport digital assistant will increase the level of automation in airports, enhancing safety and efficiency. The digital assistants will be delivered at a TRL4 level of maturity, each solution (DA) will validate the different features of the assistant in different validation exercises [142].

JARVIS aims at delivering novel methods for AI trustworthiness, providing the relevant means of compliance to assure the robustness and stability of the AI/ML algorithms, starting from the evaluation of data quality.

A cloud-based AI infrastructure will be developed to gather a massive data lake with information coming from ANSP and airports. Its architecture will allow the data to be labelled, processed, and used for training different advanced ATM tools. A robust human-AI teaming framework based strategic

automation, decision support, shared situational awareness (e.g., joint cognition) in order to support the actor trust.

JARVIS foundational AI work package will collect lessons learned, best practices, and standard approaches/procedures from the different solutions AI capabilities, namely in terms of Assured AI, Human AI-teaming and Big Data & Cloud Infrastructures.

DARWIN

Digital Assistants for Reducing Workload & Increasing collaboration (DARWIN) is a Fast track for innovation project. AI-based automation for cockpit and flight operations are the key enabler for single pilot operations (SPO). The project aims to develop digital assistants to support SPO operations, assuring the same (or higher) level of safety and same (or lower) workload as operations with a full crew today. The project will deliver solutions that enable operational efficiency and route flexibility, considering the complexity of the future airspace. The results will support the commercial and operational viability of those new airspace users [143].

8 References

[1] H. Fricke and M. Schultz, 'Improving aircraft turn around reliability', in *Third International Conference on Research in Air Transportation*, 2008, pp. 335–343.

[2] H. Jin, 'A framework for the analysis of aircraft turnaround at congested airports', Apr. 2019, Accessed: Dec. 14, 2023. [Online]. Available: http://hdl.handle.net/1853/61278

[3] Y. Gao, Z. Huyan, and F. Ju, 'A Prediction Method Based on Neural Network for Flight Turnaround Time at Airport', *2015 8th International Symposium on Computational Intelligence and Design (ISCID)*, vol. 2, pp. 219–222, 2015.

[4] O. van Hassel, *Predicting the turnaround time of an aircraft: a process structure aware approach (Master Thesis)*. TU Eindhoven, 2019.

[5] E. Asadi, J. Evler, H. Preis, and H. Fricke, 'Coping with Uncertainties in Predicting the Aircraft Turnaround Time at Airports', in *Operations Research Proceedings 2019*, J. S. Neufeld, U. Buscher, R. Lasch, D. Möst, and J. Schönberger, Eds., in Operations Research Proceedings. , Cham: Springer International Publishing, 2020, pp. 773–780. doi: 10.1007/978-3-030-48439-2_94.

[6] H. Zhou, W. Li, Z. Jiang, F. Cai, and Y. Xue, 'Flight Departure Time Prediction Based on Deep Learning', *Aerospace*, vol. 9, no. 7, p. 394, 2022.

[7] E. Halmesaari, *Interpretable Machine Learning for Prediction of Aircraft Turnaround Times (MSc thesis)*. Aalto University, 2020.

[8] M. Luo, M. Schultz, H. Fricke, B. Desart, F. Herrema, and R. B. Montes, 'Agent-based simulation for aircraft stand operations to predict ground time using machine learning', in *2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC)*, IEEE, 2021, pp. 1–8. Accessed: Jan. 04, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9594325/

[9] E. Asadi and H. Fricke, 'Aircraft total turnaround time estimation using fuzzy critical path method', *Journal of Project Management*, vol. 7, no. 4, pp. 241–254, 2022.

[10] S. Yıldız, O. Aydemir, A. Memiş, and S. Varlı, 'A turnaround control system to automatically detect and monitor the time stamps of ground service actions in airports: A deep learning and computer vision based approach', *Engineering Applications of Artificial Intelligence*, vol. 114, p. 105032, Sep. 2022, doi: 10.1016/j.engappai.2022.105032.

[11] J. Fitzgerald, E. Spinielli, A. Tart, and R. Koelle, 'Reference Trajectories: The Dataset Enabling Gate-to-Gate Flight Analysis', *Engineering Proceedings*, vol. 13, no. 1, Art. no. 1, 2022, doi: 10.3390/engproc2021013014.

[12] A. Sternberg, J. Soares, D. Carvalho, and E. Ogasawara, 'A Review on Flight Delay Prediction', *Transport Reviews*, vol. 41, no. 4, pp. 499–528, Jul. 2021, doi: 10.1080/01441647.2020.1861123.

[13] R. Beatty, R. Hsu, L. Berry, and J. Rome, 'Preliminary Evaluation of Flight Delay Propagation through an Airline Schedule', *Air Traffic Control Quarterly*, vol. 7, no. 4, pp. 259–270, Oct. 1999, doi: 10.2514/atcq.7.4.259.

[14] S. AhmadBeygi, A. Cohn, Y. Guan, and P. Belobaba, 'Analysis of the potential for delay propagation in passenger airline networks', *Journal of Air Transport Management*, vol. 14, no. 5, pp. 221–236, Sep. 2008, doi: 10.1016/j.jairtraman.2008.04.010.

[15] S. Lan, J.-P. Clarke, and C. Barnhart, 'Planning for Robust Airline Operations: Optimizing Aircraft Routings and Flight Departure Times to Minimize Passenger Disruptions', *Transportation Science*, vol. 40, no. 1, pp. 15–28, Feb. 2006, doi: 10.1287/trsc.1050.0134.

[16] N. Kafle and B. Zou, 'Modeling flight delay propagation: A new analytical-econometric approach', *Transportation Research Part B: Methodological*, vol. 93, pp. 520–542, Nov. 2016, doi: 10.1016/j.trb.2016.08.012.

[17] United States. Department of Transportation. Bureau of Transportation Statistics, 'Transportation Statistics Annual Report 2022', doi: 10.21949/1528354.

[18] A. Jacquillat and A. R. Odoni, 'Endogenous control of service rates in stochastic and dynamic queuing models of airport congestion', *Transportation Research Part E: Logistics and Transportation Review*, vol. 73, pp. 133–151, Jan. 2015, doi: 10.1016/j.tre.2014.10.014.

[19] N. Pyrgiotis, K. M. Malone, and A. Odoni, 'Modelling delay propagation within an airport network', *Transportation Research Part C: Emerging Technologies*, vol. 27, pp. 60–75, Feb. 2013, doi: 10.1016/j.trc.2011.05.017.

[20] T. Wang, Y. Zheng, and H. Xu, 'A Review of Flight Delay Prediction Methods', in *2022 2nd International Conference on Big Data Engineering and Education (BDEE)*, Aug. 2022, pp. 135–141. doi: 10.1109/BDEE55929.2022.00029.

[21] S. Janssen, A. Sharpanskykh, and R. Curran, 'Agent-based modelling and analysis of security and efficiency in airport terminals', *Transportation Research Part C: Emerging Technologies*, vol. 100, pp. 142–160, Mar. 2019, doi: 10.1016/j.trc.2019.01.012.

[22] S. Janssen, A. van den Berg, and A. Sharpanskykh, 'Agent-based vulnerability assessment at airport security checkpoints: A case study on security operator behavior', *Transportation Research Interdisciplinary Perspectives*, vol. 5, p. 100139, May 2020, doi: 10.1016/j.trip.2020.100139.

[23] S. Janssen, R. van der Sommen, A. Dilweg, and A. Sharpanskykh, 'Data-Driven Analysis of Airport Security Checkpoint Operations', *Aerospace*, vol. 7, no. 6, Art. no. 6, Jun. 2020, doi: 10.3390/aerospace7060069.

[24] R. Félix Patrón, P. Scala, M. Mujica Mota, and A. Murrieta Mendoza, *Airport passenger flow prediction using simulation data farming and machine learning*. 2021.

[25] B. Chen, X. Zhao, and J. Wu, 'Evaluating Prediction Models for Airport Passenger Throughput Using a Hybrid Method', *Applied Sciences*, vol. 13, no. 4, Art. no. 4, Jan. 2023, doi: 10.3390/app13042384.

[26] M. Karlaftis, 'DEMAND FORECASTING IN REGIONAL AIRPORTS: DYNAMIC TOBIT MODELS WITH GARCH ERRORS', 2008, Accessed: Nov. 21, 2023. [Online]. Available: https://dspace.lib.ntua.gr/xmlui/handle/123456789/28204

[27] Z. Li, J. Bi, and Z. Li, 'Passenger Flow Forecasting Research for Airport Terminal Based on SARIMA Time Series Model', *IOP Conf. Ser.: Earth Environ. Sci.*, vol. 100, no. 1, p. 012146, Dec. 2017, doi: 10.1088/1755-1315/100/1/012146.

[28] A. Akincilar, 'A Methodology for Shuttle Scheduling in Airports That Ensures Mitigating Arriving Passenger Congestion Under Uncertain Demand', *IEEE Intelligent Transportation Systems Magazine*, vol. 14, no. 2, pp. 105–114, Mar. 2022, doi: 10.1109/MITS.2021.3049359.

[29] M. N. Postorino, L. Mantecchini, C. Malandri, and F. Paganelli, 'Airport Passenger Arrival Process: Estimation of Earliness Arrival Functions', *Transportation Research Procedia*, vol. 37, pp. 338–345, Jan. 2019, doi: 10.1016/j.trpro.2018.12.201.

[30] L. Lin, X. Liu, X. Liu, T. Zhang, and Y. Cao, 'A prediction model to forecast passenger flow based on flight arrangement in airport terminals', *Energy and Built Environment*, vol. 4, no. 6, pp. 680–688, Dec. 2023, doi: 10.1016/j.enbenv.2022.06.006.

[31] K. Leone and R. (Rachel) Liu, 'Improving airport security screening checkpoint operations in the US via paced system design', *Journal of Air Transport Management*, vol. 17, no. 2, pp. 62–67, Mar. 2011, doi: 10.1016/j.jairtraman.2010.05.002.

[32] T. Gräupl, 'Validation of Simulated Synthetic Air Traffic against Recorded Air Traffic in Germany', Jan. 2017.

[33] F. Hoffmann, U. Epple, M. Schnell, and U.-C. Fiebig, 'Feasibility of LDACS1 cell planning in European airspace', in *2012 IEEE/AIAA 31st Digital Avionics Systems Conference (DASC)*, Oct. 2012, pp. 1–17. doi: 10.1109/DASC.2012.6383053.

[34] C.-H. Rokitansky, M. Ehammer, and Th. Graupl, 'Newsky - building a simulation environment for an integrated aeronautical network architecture', in *2007 IEEE/AIAA 26th Digital Avionics Systems Conference*, Oct. 2007, p. 4.B.4-1-4.B.4-11. doi: 10.1109/DASC.2007.4391905.

[35] J. A. Besada, J. Portillo, G. de Miguel, R. de Andrea, and J. M. Canino, 'Traffic analysis and synthetic scenario generation for ATM operational concepts evaluation', in *2009 IEEE/AIAA 28th Digital Avionics Systems Conference*, Oct. 2009, p. 2.A.1-1-2.A.1-14. doi: 10.1109/DASC.2009.5347557.

[36] T. Krauth, A. Lafage, J. Morio, X. Olive, and M. Waltert, 'Deep generative modelling of aircraft trajectories in terminal maneuvering areas', *Machine Learning with Applications*, vol. 11, p. 100446, Mar. 2023, doi: 10.1016/j.mlwa.2022.100446.

[37] R. Dalmau and G. Gawinowski, 'Learning With Confidence the Likelihood of Flight Diversion Due to Adverse Weather at Destination', *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 5, pp. 5615–5624, May 2023, doi: 10.1109/TITS.2023.3235741.

[38] R. Dalmau and G. Gawinowski, 'The effectiveness of supervised clustering for characterising flight diversions due to weather', *Expert Systems with Applications*, vol. 237, p. 121652, Mar. 2024, doi: 10.1016/j.eswa.2023.121652.

[39] C. Di Ciccio, H. van der Aa, C. Cabanillas, J. Mendling, and J. Prescher, 'Detecting Flight Trajectory Anomalies and Predicting Diversions in Freight Transportation', *Decision Support Systems*, vol. 88, May 2016, doi: 10.1016/j.dss.2016.05.004.

[40] M. Schäfer, M. Strohmeier, V. Lenders, I. Martinovic, and M. Wilhelm, 'Bringing up OpenSky: A large-scale ADS-B sensor network for research', in *IPSN-14 Proceedings of the 13th International Symposium on Information Processing in Sensor Networks*, Apr. 2014, pp. 83–94. doi: 10.1109/IPSN.2014.6846743.

[41] H. Zhang *et al.*, 'Mixed-Type Tabular Data Synthesis with Score-based Diffusion in Latent Space'. arXiv, Oct. 14, 2023. doi: 10.48550/arXiv.2310.09656.

[42] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, 'Modeling Tabular data using Conditional GAN'. arXiv, Oct. 27, 2019. doi: 10.48550/arXiv.1907.00503.

[43] A. J. Ohrt, 'Probabilistic Tabular Diffusion for Counterfactual Explanation Synthesis', Master thesis, NTNU, 2023. Accessed: Dec. 20, 2023. [Online]. Available: https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/3097652

[44] L. Xu and K. Veeramachaneni, 'Synthesizing Tabular Data using Generative Adversarial Networks'. arXiv, Nov. 27, 2018. doi: 10.48550/arXiv.1811.11264.

[45] J. Fonseca and F. Bacao, 'Tabular and latent space synthetic data generation: a literature review', *Journal of Big Data*, vol. 10, no. 1, p. 115, Jul. 2023, doi: 10.1186/s40537-023-00792-7.

[46] M. Hernandez, G. Epelde, A. Alberdi, R. Cilla, and D. Rankin, 'Synthetic data generation for tabular health records: A systematic review', *Neurocomputing*, vol. 493, pp. 28–45, Jul. 2022, doi: 10.1016/j.neucom.2022.04.053.

[47] A. Brock, J. Donahue, and K. Simonyan, 'Large Scale GAN Training for High Fidelity Natural Image Synthesis', in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net, 2019. Accessed: Jan. 24, 2024. [Online]. Available: https://openreview.net/forum?id=B1xsqj09Fm

[48] T. Luhman and E. Luhman, 'High Fidelity Image Synthesis With Deep VAEs In Latent Space'. arXiv, Mar. 23, 2023. doi: 10.48550/arXiv.2303.13714.

[49] K. Pandey, A. Mukherjee, P. Rai, and A. Kumar, 'DiffuseVAE: Efficient, Controllable and High-Fidelity Generation from Low-Dimensional Latents'. arXiv, Nov. 29, 2022. doi: 10.48550/arXiv.2201.00308.

[50] Y. Leng, Q. Huang, Z. Wang, Y. Liu, and H. Zhang, 'DiffuseGAE: Controllable and High-fidelity Image Manipulation from Disentangled Representation', in *Proceedings of the 5th ACM International Conference on Multimedia in Asia*, in MMAsia '23. New York, NY, USA: Association for Computing Machinery, Jan. 2024, pp. 1–7. doi: 10.1145/3595916.3626402.

[51] A. Sinha, J. Song, C. Meng, and S. Ermon, 'D2C: Diffusion-Denoising Models for Few-shot Conditional Generation'. arXiv, Jun. 12, 2021. doi: 10.48550/arXiv.2106.06819.

[52] H. Wang, J. Li, H. Wu, E. Hovy, and Y. Sun, 'Pre-Trained Language Models and Their Applications', *Engineering*, vol. 25, pp. 51–65, Jun. 2023, doi: 10.1016/j.eng.2022.04.024.

[53] 'Introducing ChatGPT'. Accessed: Jan. 24, 2024. [Online]. Available: https://openai.com/blog/chatgpt

[54] A. Caillon and P. Esling, 'RAVE: A variational autoencoder for fast and high-quality neural audio synthesis'. arXiv, Dec. 15, 2021. doi: 10.48550/arXiv.2111.05011.

[55] H. H. Tan, Y.-J. Luo, and D. Herremans, 'Generative Modelling for Controllable Audio Synthesis of Expressive Piano Performance', *CoRR*, vol. abs/2006.09833, 2020, Accessed: Jan. 24, 2024. [Online]. Available: https://arxiv.org/abs/2006.09833

[56] M. Baas and H. Kamper, 'GAN You Hear Me? Reclaiming Unconditional Speech Synthesis from Diffusion Models', in *2022 IEEE Spoken Language Technology Workshop (SLT)*, Jan. 2023, pp. 906–911. doi: 10.1109/SLT54892.2023.10023153.

[57] S. Pascual, G. Bhattacharya, C. Yeh, J. Pons, and J. Serrà, 'Full-Band General Audio Synthesis with Score-Based Diffusion', in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10096760.

[58] F. Kreuk *et al.*, 'AudioGen: Textually Guided Audio Generation', in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, OpenReview.net, 2023. Accessed: Jan. 24, 2024. [Online]. Available: https://openreview.net/pdf?id=CYK7RfcOzQ4

[59] A. Madane, M. Dilmi, F. Forest, H. Azzag, M. Lebbah, and J. Lacaille, 'Transformer-based conditional generative adversarial network for multivariate time series generation'. arXiv, Oct. 05, 2022. doi: 10.48550/arXiv.2210.02089.

[60] C. Hu, Z. Sun, C. Li, Y. Zhang, and C. Xing, 'Survey of Time Series Data Generation in IoT', *Sensors*, vol. 23, no. 15, Art. no. 15, Jan. 2023, doi: 10.3390/s23156976.

[61] A. Desai, C. Freeman, Z. Wang, and I. Beaver, 'çç'. arXiv, Dec. 07, 2021. doi: 10.48550/arXiv.2111.08095.

[62] X. Li, A. H. H. Ngu, and V. Metsis, 'TTS-CGAN: A Transformer Time-Series Conditional GAN for Biosignal Data Augmentation'. arXiv, Jun. 27, 2022. doi: 10.48550/arXiv.2206.13676.

[63] D. Lee, S. Malacarne, and E. Aune, 'Vector Quantized Time Series Generation with a Bidirectional Prior Model'. arXiv, Apr. 01, 2023. doi: 10.48550/arXiv.2303.04743.

[64] C. Wu, Y. Chen, P. Chou, and C. Wang, 'Synthetic Traffic Generation with Wasserstein Generative Adversarial Networks', in *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, Dec. 2022, pp. 1503–1508. doi: 10.1109/GLOBECOM48099.2022.10001157.

[65] Y. Zhu, Y. Ye, Y. Wu, X. Zhao, and J. Yu, 'SynMob: Creating High-Fidelity Synthetic GPS Trajectory Dataset for Urban Mobility Analysis', presented at the Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track, Nov. 2023. Accessed: Jan. 24, 2024. [Online]. Available: https://openreview.net/forum?id=oz4AGs0phP

[66] T. Cunningham, K. Klemmer, H. Wen, and H. Ferhatosmanoglu, 'GeoPointGAN: Synthetic Spatial Data with Local Label Differential Privacy', *CoRR*, vol. abs/2205.08886, 2022, doi: 10.48550/ARXIV.2205.08886.

[67] C. Madarasingha, S. R. Muramudalige, G. Jourjon, A. Jayasumana, and K. Thilakarathna, 'VideoTrain++: GAN-based adaptive framework for synthetic video traffic generation', *Computer Networks*, vol. 206, p. 108785, Apr. 2022, doi: 10.1016/j.comnet.2022.108785.

[68] K. Mei and V. M. Patel, 'VIDM: Video Implicit Diffusion Models'. arXiv, Nov. 30, 2022. doi: 10.48550/arXiv.2212.00235.

[69] S. Palazzo, C. Spampinato, P. D'Oro, D. Giordano, and M. Shah, 'Generating Synthetic Video Sequences by Explicitly Modeling Object Motion', in *Computer Vision – ECCV 2018 Workshops: Munich, Germany, September 8-14, 2018, Proceedings, Part II*, Berlin, Heidelberg: Springer-Verlag, Jan. 2019, pp. 492–499. doi: 10.1007/978-3-030-11012-3_37.

[70] D. McDuff, T. Curran, and A. Kadambi, 'Synthetic Data in Healthcare', *CoRR*, vol. abs/2304.03243, 2023, doi: 10.48550/ARXIV.2304.03243.

[71] A. Gonzales, G. Guruswamy, and S. R. Smith, 'Synthetic data in health care: A narrative review', *PLOS Digital Health*, vol. 2, no. 1, p. e0000082, Jan. 2023, doi: 10.1371/journal.pdig.0000082.

[72] T. Kokosi and K. Harron, 'Synthetic data in medical research', *BMJ Medicine*, vol. 1, no. 1, 2022, doi: 10.1136/bmjmed-2022-000167.

[73] R. J. Chen, M. Y. Lu, T. Y. Chen, D. F. K. Williamson, and F. Mahmood, 'Synthetic data in machine learning for medicine and healthcare', *Nat Biomed Eng*, vol. 5, no. 6, pp. 493–497, Jun. 2021, doi: 10.1038/s41551-021-00751-8.

[74] S. A. Assefa, D. Dervovic, M. Mahfouz, R. E. Tillman, P. Reddy, and M. Veloso, 'Generating synthetic data in finance: opportunities, challenges and pitfalls', in *Proceedings of the First ACM International Conference on AI in Finance*, in ICAIF '20. New York, NY, USA: Association for Computing Machinery, Oct. 2021, pp. 1–8. doi: 10.1145/3383455.3422554.

[75] E. Altman, B. Egressy, J. Blanusa, and K. Atasu, 'Realistic Synthetic Financial Transactions for Anti-Money Laundering Models', *CoRR*, vol. abs/2306.16424, 2023, doi: 10.48550/ARXIV.2306.16424.

[76] C. Liu, C. Ventre, and M. Polukarov, 'Synthetic Data Augmentation for Deep Reinforcement Learning in Financial Trading', in *3rd ACM International Conference on AI in Finance, ICAIF 2022, New York, NY, USA, November 2-4, 2022*, D. Magazzeni, S. Kumar, R. Savani, R. Xu, C. Ventre, B. Horvath, R. Hu, T. Balch, and F. Toni, Eds., ACM, 2022, pp. 343–351. doi: 10.1145/3533271.3561704.

[77] M. Dogariu, L.-D. Ştefan, B. A. Boteanu, C. Lamba, B. Kim, and B. Ionescu, 'Generation of Realistic Synthetic Financial Time-series', *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 18, no. 4, p. 96:1-96:27, Mar. 2022, doi: 10.1145/3501305.

[78] N. Gadipudi *et al.*, 'Synthetic to Real Gap Estimation of Autonomous Driving Datasets using Feature Embedding', in *2022 IEEE 5th International Symposium in Robotics and Manufacturing Automation (ROMA)*, Aug. 2022, pp. 1–5. doi: 10.1109/ROMA55875.2022.9915679.

[79] M. Tzelepi, C. Symeonidis, N. Nikolaidis, and A. Tefas, 'Real-time synthetic-to-real human detection for robotics applications', in *2022 13th International Conference on Information, Intelligence, Systems & Applications (IISA)*, Jul. 2022, pp. 1–5. doi: 10.1109/IISA56318.2022.9904394.

[80] 'Learning from synthetic data generated with GRADE', *arXiv.org*, vol. abs/2305.04282, May 2023, doi: 10.48550/arXiv.2305.04282.

[81] P. N. Canas, J. D. Ortega, M. Nieto, and O. Otaegui, 'Virtual passengers for real car solutions: synthetic datasets'. arXiv, May 13, 2022. doi: 10.48550/arXiv.2205.06556.

[82] M. Methini and J. Priyadharshini, 'Procedural Modelling for Synthetic Data Generation in Automotive Applications', in *2022 International Conference on Communication, Computing and Internet of Things (IC3IoT)*, Mar. 2022, pp. 1–4. doi: 10.1109/IC3IOT53935.2022.9767965.

[83] C. A. Akar, J. Tekli, D. Jess, M. Khoury, M. Kamradt, and M. Guthe, 'Synthetic Object Recognition Dataset for Industries', in *2022 35th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, Oct. 2022, pp. 150–155. doi: 10.1109/SIBGRAPI55357.2022.9991784.

[84] 'Differences Between a Parametric and Non-parametric Model | Baeldung on Computer Science'. Accessed: May 13, 2024. [Online]. Available: https://www.baeldung.com/cs/ml-parametric-vs-non-parametric-models

[85] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*. in Springer Texts in Statistics. New York, NY: Springer US, 2021. doi: 10.1007/978-1-0716-1418-1.

[86] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. in Springer Series in Statistics. New York, NY: Springer, 2009. doi: 10.1007/978-0-387-84858-7.

[87] T. Schmidt, 'Coping with copulas', *Copulas-From theory to application in finance*, vol. 3, pp. 1–34, 2007.

[88] G. R. Terrell and D. W. Scott, 'Variable Kernel Density Estimation', *The Annals of Statistics*, vol. 20, no. 3, pp. 1236–1265, 1992.

[89] T. Cover and P. Hart, 'Nearest neighbor pattern classification', *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967, doi: 10.1109/TIT.1967.1053964.

[90] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*. in Springer Texts in Statistics. New York, NY: Springer, 2004. doi: 10.1007/978-1-4757-4145-2.

[91] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and Their Application*. Cambridge University Press, 1997. [Online]. Available: https://github.com/Johnnyboycurtis/TSI/blob/master/doc/Davison%2C%20Hinkley%20-Bootstrap%20Methods%20and%20their%20Application.pdf

[92] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, 'SMOTE: synthetic minority over-sampling technique', *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[93] H. He, Y. Bai, E. A. Garcia, and S. Li, 'ADASYN: Adaptive synthetic sampling approach for imbalanced learning', in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, Ieee, 2008, pp. 1322–1328. Accessed: May 16, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/4633969/

[94] H. Han, W.-Y. Wang, and B.-H. Mao, 'Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning', in *Advances in Intelligent Computing*, vol. 3644, D.-S. Huang, X.-P. Zhang, and G.-B. Huang, Eds., in Lecture Notes in Computer Science, vol. 3644. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 878–887. doi: 10.1007/11538059_91.

[95] H. M. Nguyen, E. W. Cooper, and K. Kamei, 'Borderline over-sampling for imbalanced data classification', *IJKESDP*, vol. 3, no. 1, p. 4, 2011, doi: 10.1504/IJKESDP.2011.039875.

[96] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, 'Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem', in *Advances in Knowledge Discovery and Data Mining*, vol. 5476, T. Theeramunkong, B. Kijsirikul, N. Cercone, and T.-B. Ho, Eds., in Lecture Notes in Computer Science, vol. 5476. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 475–482. doi: 10.1007/978-3-642-01307-2_43.

[97] G. Douzas, F. Bacao, and F. Last, 'Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE', *Information Sciences*, vol. 465, pp. 1–20, Oct. 2018, doi: 10.1016/j.ins.2018.06.056.

[98] N. Bichraoui, B. Guillaume, and A. Halog, 'Agent-based Modelling Simulation for the Development of an Industrial Symbiosis - Preliminary Results', *Procedia Environmental Sciences*, vol. 17, pp. 195–204, Dec. 2013, doi: 10.1016/j.proenv.2013.02.029.

[99] E. Bonabeau, 'Agent-based modeling: Methods and techniques for simulating human systems', *Proc Natl Acad Sci U S A*, vol. 99, no. Suppl 3, pp. 7280–7287, May 2002, doi: 10.1073/pnas.082080899.

[100] M. A. Niazi and A. Hussain, 'Agent-based computing from multi-agent systems to agent-based Models: a visual survey', *Scientometrics*, vol. 89, no. 2, pp. 479–499, Nov. 2011, doi: 10.1007/s11192-011-0468-9.

[101] C. Macal and M. North, 'Tutorial on agent-based modelling and simulation', *J. Simulation*, vol. 4, pp. 151–162, Sep. 2010, doi: 10.1057/jos.2010.3.

[102] I. Goodfellow *et al.*, 'Generative adversarial networks', *Commun. ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020, doi: 10.1145/3422622.

[103] D. P. Kingma and M. Welling, 'An Introduction to Variational Autoencoders', *Found. Trends Mach. Learn.*, vol. 12, no. 4, pp. 307–392, 2019, doi: 10.1561/2200000056.

[104] J. Ho, A. Jain, and P. Abbeel, 'Denoising Diffusion Probabilistic Models', in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan, and H.-T. Lin, Eds., 2020. Accessed: Jan. 24, 2024. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html

[105] I. Goodfellow *et al.*, 'Generative Adversarial Nets', in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds., Curran Associates, Inc., 2014. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

[106] J. Yoon, D. Jarrett, and M. van der Schaar, 'Time-series generative adversarial networks', in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA: Curran Associates Inc., 2019, pp. 5508–5518.

[107] C. Esteban, S. L. Hyland, and G. Rätsch, 'Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs', *CoRR*, vol. abs/1706.02633, 2017, Accessed: Jan. 29, 2024. [Online]. Available: http://arxiv.org/abs/1706.02633

[108] E. Brophy, Z. Wang, Q. She, and T. Ward, 'Generative Adversarial Networks in Time Series: A Systematic Literature Review', *ACM Comput. Surv.*, vol. 55, no. 10, p. 199:1-199:31, Feb. 2023, doi: 10.1145/3559540.

[109] O. Mogren, 'C-RNN-GAN: Continuous recurrent neural networks with adversarial training', *CoRR*, vol. abs/1611.09904, 2016, Accessed: Jan. 29, 2024. [Online]. Available: http://arxiv.org/abs/1611.09904

[110] H. Ni, L. Szpruch, M. Wiese, S. Liao, and B. Xiao, 'Conditional Sig-Wasserstein GANs for Time Series Generation', *CoRR*, vol. abs/2006.05421, 2020, Accessed: Jan. 29, 2024. [Online]. Available: https://arxiv.org/abs/2006.05421

[111] A. Vaswani *et al.*, 'Attention is All you Need', in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 5998–6008. Accessed: Jan. 29, 2024. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[112] 'dblp: Search for "An image is worth 16x16 words: Transformers for image recognition at scale"'. Accessed: Jan. 29, 2024. [Online]. Available: https://dblp.uni-trier.de/search?q=An%20image%20is%20worth%2016x16%20words%3A%20Transformers%20for%20image%20recognition%20at%20scale

[113] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., Association for Computational Linguistics, 2019, pp. 4171–4186. doi: 10.18653/V1/N19-1423.

[114] Y. Jiang, S. Chang, and Z. Wang, 'TransGAN: Two Pure Transformers Can Make One Strong GAN, and That Can Scale Up', in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021, pp. 14745–14758. Accessed: Jan. 29, 2024. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/7c220a2091c26a7f5e9f1cfb099511e3-Abstract.html

[115] X. Li, V. Metsis, H. Wang, and A. H. H. Ngu, 'TTS-GAN: A Transformer-Based Time-Series Generative Adversarial Network', in *Artificial Intelligence in Medicine - 20th International Conference on Artificial Intelligence in Medicine, AIME 2022, Halifax, NS, Canada, June 14-17, 2022, Proceedings*, M. Michalowski, S. S. R. Abidi, and S. Abidi, Eds., in Lecture Notes in Computer Science, vol. 13263. Springer, 2022, pp. 133–143. doi: 10.1007/978-3-031-09342-5_13.

[116] X. Li, A. H. H. Ngu, and V. Metsis, 'TTS-CGAN: A Transformer Time-Series Conditional GAN for Biosignal Data Augmentation'. arXiv, Jun. 27, 2022. doi: 10.48550/arXiv.2206.13676.

[117] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, 'Spectral Normalization for Generative Adversarial Networks', in *6th International Conference on Learning Representations, ICLR 2018,*

Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings, OpenReview.net, 2018. Accessed: Jan. 29, 2024. [Online]. Available: https://openreview.net/forum?id=B1QRgziT-

[118] C. Chu, K. Minami, and K. Fukumizu, 'Smoothness and Stability in GANs', in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, OpenReview.net, 2020. Accessed: Jan. 29, 2024. [Online]. Available: https://openreview.net/forum?id=HJeOekHKwr

[119] Z. Xiao, K. Kreis, and A. Vahdat, 'Tackling the Generative Learning Trilemma with Denoising Diffusion GANs', in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, OpenReview.net, 2022. Accessed: Jan. 29, 2024. [Online]. Available: https://openreview.net/forum?id=JprM0p-q0Co

[120] H. Li, S. Yu, and J. Principe, 'Causal recurrent variational autoencoder for medical time series generation', in *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, in AAAI'23/IAAI'23/EAAI'23, vol. 37. AAAI Press, Feb. 2023, pp. 8562–8570. doi: 10.1609/aaai.v37i7.26031.

[121] B. Cai, S. Yang, L. Gao, and Y. Xiang, 'Hybrid variational autoencoder for time series forecasting', *Knowl. Based Syst.*, vol. 281, p. 111079, 2023, doi: 10.1016/J.KNOSYS.2023.111079.

[122] D. Lee, S. Malacarne, and E. Aune, 'Vector Quantized Time Series Generation with a Bidirectional Prior Model', in *International Conference on Artificial Intelligence and Statistics, 25-27 April 2023, Palau de Congressos, Valencia, Spain*, F. J. R. Ruiz, J. G. Dy, and J.-W. van de Meent, Eds., in Proceedings of Machine Learning Research, vol. 206. PMLR, 2023, pp. 7665–7693. Accessed: Jan. 29, 2024. [Online]. Available: https://proceedings.mlr.press/v206/lee23d.html

[123] Z. Chang, S. Liu, R. Qiu, S. Song, Z. Cai, and G. Tu, 'A VAE-based Model for Incomplete Time Series Modeling using a Time-aware Encoder and Neural ODE'. 2022. doi: 10.21203/rs.3.rs-2344250/v1.

[124] D. Kavran, B. Žalik, and N. Lukač, 'Comparing Beta-VAE to WGAN-GP for Time Series Augmentation to Improve Classification Performance', in *Agents and Artificial Intelligence*, A. P. Rocha, L. Steels, and J. van den Herik, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2022, pp. 51–73. doi: 10.1007/978-3-031-22953-4_3.

[125] I. Naiman, N. B. Erichson, P. Ren, M. W. Mahoney, and O. Azencot, 'Generative Modeling of Regular and Irregular Time Series Data via Koopman VAEs', *CoRR*, vol. abs/2310.02619, 2023, doi: 10.48550/ARXIV.2310.02619.

[126] A. Desai, C. Freeman, Z. Wang, and I. Beaver, 'TimeVAE: A Variational Auto-Encoder for Multivariate Time Series Generation'. arXiv, Dec. 07, 2021. doi: 10.48550/arXiv.2111.08095.

[127] P. Dhariwal and A. Q. Nichol, 'Diffusion Models Beat GANs on Image Synthesis', in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021, pp. 8780–8794. Accessed: Jan. 29, 2024. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html

[128] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis, 'Structure and Content-Guided Video Synthesis with Diffusion Models', in *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, IEEE, 2023, pp. 7312–7322. doi: 10.1109/ICCV51070.2023.00675.

[129] P. Yu *et al.*, 'Latent Diffusion Energy-Based Model for Interpretable Text Modelling', in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, Eds., in Proceedings of Machine Learning Research, vol. 162. PMLR, 2022, pp. 25702–25720. Accessed: Jan. 29, 2024. [Online]. Available: https://proceedings.mlr.press/v162/yu22h.html

[130] Y. Li, X. Lu, Y. Wang, and D. Dou, 'Generative Time Series Forecasting with Diffusion, Denoise, and Disentanglement', in *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., 2022. Accessed: Jan. 29, 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2022/hash/91a85f3fb8f570e6be52b333b5ab017a-Abstract-Conference.html

[131] Y. Tashiro, J. Song, Y. Song, and S. Ermon, 'CSDI: Conditional Score-based Diffusion Models for Probabilistic Time Series Imputation', in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021, pp. 24804–24816. Accessed: Jan. 29, 2024. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/cfe8504bda37b575c70ee1a8276f3486-Abstract.html

[132] H. Lim, M. Kim, S. Park, and N. Park, 'Regular Time-series Generation using SGM', *CoRR*, vol. abs/2301.08518, 2023, doi: 10.48550/ARXIV.2301.08518.

[133] Single European Sky ATM Research 3 Joint Undertaking (EU body or agency), *Digital European sky: strategic research and innovation agenda*. LU: Publications Office of the European Union, 2020. Accessed: Jan. 29, 2024. [Online]. Available: https://data.europa.eu/doi/10.2829/117092

[134] 'ARTIMATION, Validation report', D 6.2.

[135] C. Westin and C. Borst, 'Personalized and transparent AI support for ATC conflict detection and resolution: an empirical study',

[136] M. Cocchioni, S. Bonelli, C. Westin, A. Ferreira, and N. Cavagnetto, 'Guidelines for Artificial Intelligence in Air Traffic Management: a contribution to EASA strategy', presented at the 14th International Conference on Applied Human Factors and Ergonomics (AHFE 2023), 2023. doi: 10.54941/ahfe1003008.

[137] 'Final Project Report PJ.04-TAM: Total Airport Management', SESAR JU, Industrial research, D 1.2.

[138] SESAR JU, 'AI in ATM: transparency, explainability, conformance, situation awareness and trust : A white paper', 2022.

[139] SESAR JU, 'https://www.sesarju.eu/projects/trusty', EXPLORATORY RESEARCH PROJECT TRUSTY- Trustworthy intelligent system for remote digital tower.

[140] 'https://www.sesarju.eu/projects/MultiModX', EXPLORATORY RESEARCH PROJECT - Integrated Passenger-Centric Planning of Multimodal Transport Networks.

[141] 'https://www.sesarju.eu/projects/ASTRA', EXPLORATORY RESEARCH PROJECT ASTRA- AI-enabled tactical FMP hotspot prediction and resolution.

[142] 'https://www.sesarju.eu/projects/JARVIS', EXPLORATORY RESEARCH PROJECT JARVIS- Just a rather very intelligent system.

[143] SESAR JU, 'https://www.sesarju.eu/news/ai-advance-single-pilot-operations', AI to advance single pilot operations.

9 List of acronyms

Term	Definition
ADS-B	Automatic Dependent Surveillance - Broadcast
A-CDM	Airport Collaborative Decision-Making
AIRAC	Aeronautical Information Regulation And Control
ANN	Artificial Neural Network
ANSP	Air Navigation Service Provider
AOCC	Airline Operation Control Centre
ARIMA	Autoregressive Integrated Moving Average
ATFM	Air Traffic Flow Management
EOBT	Estimated Off-Block Time
GAN	Generative Adversarial Network
GARCH	Generalised Autoregressive Conditional Heteroskedasticity
GDPR	General Data Protection Regulation
GRU	Gated Recurrent Units
IFR	Instrument Flight Rules
LSTM	Long Short-Term Memory
ML	Machine Learning
MLP	Multilayer Perceptron
NAA	National Aviation Authority
NM	Network Manager
PSAP	Process Structure Aware Prediction
SARIMA	Seasonal Autoregressive Integrated Moving Average
SDG	Synthetic Data Generation
TOBT	Target Off-block Time
TTS-GAN	Transformer-based Time-Series Generative Adversarial Network
VAE	Variational Autoencoder
SRIA	Strategic Research and Innovation Agenda

Table 9: List of acronyms