A black rectangle with a black background

Description automatically generated with low confidenceLogo, company name

Description automatically generatedA person in a suit and bow tie

Description automatically generated

DAEN 690

Project Report

Naga Pranavi Kandarpa

Neerajkumar Neela

Pravallika Avula

Sai Yeshwanth Reddy Chinakota

Junaid Mohammed

Lalithanjali Yarrapatruni

Spring 2024

**Generating Synthetic Flight Tracks Using Generative Adversarial Networks**

**About the Cover**

Professor Berlin is an instructor at the George Mason University College of Engineering and Computing, Volgenau School of Engineering, MS Data Analytics Engineering (DAEN) program. He began working with the DAEN program as an adjunct faculty member in 2012 and became a fulltime faculty member in 2016. He is a passionate contributor to the program and a devoted mentor to his students.

His passion for new value creation is built on over 50 years of professional experience – innovating and advocating for innovators applying leading-edge digital solutions to mission challenges. He has served with outstanding teams in various roles, including senior strategy executive, consultant, and mentor; applied information and systems technologist; collaborative leader; computer scientist, and public policy entrepreneur.

He serves as a strategy advisor and mentor to public and private sector innovators and entrepreneurs and as a public speaker (emerging challenges, innovation opportunities, and ethics). His core interests include public policy, high-performance computing, cyber, emerging big data, health informatics, and digital economy and governance challenges.

In addition to teaching and mentoring, Professor Berlin seeks new engagements with high-quality, core-value-centered innovation teams – collaborating to address societal and market challenges with cyber-physical and policy innovation. Specifically, sustainable solutions can be delivered at the intersection of innovative value creation, human aspiration, and strategic vision.

Professor Berlin is a graduate of the US Air Force Academy with a Bachelor of Science in Computer Science and Mathematics as well as a graduate of the University of Texas at Austin with a Master of Arts in Computer Science. He is also attended the USAF Air Command and Staff College for leadership and strategy training.

Contents

Table of Contents

[Section 1: Problem Definition 3](#_Toc165035383)

[1.1 Background 3](#_Toc165035384)

[1.2 Problem Space 5](#_Toc165035385)

[1.3 Research 5](#_Toc165035386)

[1.4 Solution Space 7](#_Toc165035387)

[1.5 Project Objectives 7](#_Toc165035388)

[1.6 Primary User Stories 8](#_Toc165035389)

[1.7 Product Vision 8](#_Toc165035390)

[**1.7.1** **Scenario #1** 8](#_Toc165035391)

[**1.7.2** **Scenario #2** 9](#_Toc165035392)

[Section 2: Datasets 9](#_Toc165035393)

[2.1 Overview 9](#_Toc165035394)

[2.2 Field Descriptions 10](#_Toc165035395)

[2.3 Data Context 10](#_Toc165035396)

[2.4 Data Conditioning 11](#_Toc165035397)

[2.5 Data Quality Assessment 12](#_Toc165035398)

[2.6 Other Data Sources 13](#_Toc165035399)

[2.7 Storage Medium 13](#_Toc165035400)

[2.8 Storage Security 13](#_Toc165035401)

[2.9 Storage Costs 14](#_Toc165035402)

[Section 3: Algorithms & Analysis / ML Model Exploration & Selection 15](#_Toc165035403)

[3.1 Solution Approach 15](#_Toc165035404)

[**3.1.1** **Systems Architecture** 15](#_Toc165035405)

[**3.1.2** **Systems Security** 16](#_Toc165035406)

[**3.1.3** **Systems Data Flows** 17](#_Toc165035407)

[**3.1.4** **Algorithms & Analysis** 17](#_Toc165035408)

[3.2 Machine Learning 18](#_Toc165035409)

[**3.2.1** **Model Exploration** 18](#_Toc165035410)

[**3.2.2** **Model Selection: Rationale for Employing GANs for Flight Track Generation** 20](#_Toc165035411)

[Section 4: Visualizations / ML Model Training, Evaluation, & Validation 20](#_Toc165035412)

[4.1 Overview 21](#_Toc165035413)

[4.2 Visualizations 21](#_Toc165035414)

[4.3 Machine Learning 24](#_Toc165035415)

[**4.3.1** **Model Training** 24](#_Toc165035416)

[**4.3.2** **Model Evaluation** 24](#_Toc165035417)

[Section 5: Summary 29](#_Toc165035418)9

[Section 6: Future Work 29](#_Toc165035419)9

[Appendix A: Glossary](#_Toc165035420) 31

[Overview: 32](#_Toc165035421)2

[GitHub Repository Link: 32](#_Toc165035422)2

[GitHub Repository Contents: 32](#_Toc165035423)2

[Appendix C: Risks 36](#_Toc165035424)6

[Sprint 1 Risks 36](#_Toc165035425)6

[Sprint 2 Risks 37](#_Toc165035426)7

[Sprint 3 Risks 38](#_Toc165035431)8

[Sprint 4 Risks 39](#_Toc165035435)9

[Sprint 5 Risks](#_Toc165035436) 40

[Appendix D: Agile Development 41](#_Toc165035437)

[Scrum Methodology](#_Toc165035438) 41

[Sprint Analysis](#_Toc165035439) 42

[Sprint 1 Analysis:](#_Toc165035440) 42

[Sprint 2 Analyis:](#_Toc165035441) 43

[Sprint 3 Analysis:](#_Toc165035442) 43

[Sprint 5 Analysis:](#_Toc165035443) 44

Table of Figures

**Figure 1: Accidents by calendar year……………………………………………………...……………3  
Figure 2: Accident rate by Calendar Year…………………..…………………………………………4  
Figure 3: Example of the original tracks (left) and the synthetic tracks (right)……………….…….6  
Figure 4: Zurich airport Runway 34…………………………………………………………………...9  
Figure 5: Runway 34 Zurich Airport (LSZH)………………………………………………….…….11  
Figure 6: System architecture…………………………………………………….…………………....16  
Figure 7: Altitude distribution………………………………………………..………………………..21  
Figure 8: Groundspeed distribution…………………………………………………………………..22  
Figure 9: Correlation between variables……………………………………………………………...22  
Figure 10: Altitude over time for flight AAL92\_016………………………………………..………..23  
Figure 11: Flight Path……………………………………………….………………………………….23  
Figure 12: Synthetic flight routes – MINMAX SCALING…………………………………………..25  
Figure 13: Altitude distribution – MINMAX SCALING………………….………………...……….25  
Figure 14: Comparison – MINMAX SCALING…………………………………………….…..……26  
Figure 15: Synthetic flight routes – STANDARD SCALING……………………………….……….26  
Figure 16: Altitude distribution – STANDARD SCALING…………………….…………………...27  
Figure 17: Comparison – STANDARD SCALING…………………….…………………………….27**

**Figure 18: Synthetic Flight Routes – Wasserstein GAN’s …………….…………………………….28**

**Figure 19: Altitude distribution – Wasserstein GAN’s …...…………………….…………………...28**

**Figure 20: Comparison – Wasserstein GAN’s …...…………………….…………………………….29**

[**Table 1: Glossary Table**………………………………………………………………………………..31](#_Toc165237911)

[**Table 2: Sprint 1 Risks** 36](#_Toc165237912)

[**Table 3: Sprint 2 Risks** 37](#_Toc165237913)

[**Table 4: Sprint 3 Risks** 38](#_Toc165237914)

[**Table 5: Sprint 4 Risks** 39](#_Toc165237915)

[**Table 6: Sprint 5 Risks** 40](#_Toc165237916)

[**Each Sprint Timelines** 42](#_Toc165237917)

**This page intentionally left blank**

Abstract

This project focuses on the generation of synthetic flight tracks using Generative Adversarial Networks (GANs) based on historical flight data from Zurich Airport. The dataset is based on runway 34 at Zurich Airport for the dates 3 October 2019 to 30 November 2019. The dataset includes information such as timestamp, altitude, callsign, and other relevant fields. The project aims to overcome the limitations of traditional simulation methods, which often rely on simplified models that may not accurately represent the complexities of real-world flight trajectories.

The proposed system architecture comprises several key components, including a data source (OpenSky Network), data storage (Amazon S3), a computational cluster, a development environment (Jupyter Notebook), and the GAN model itself. The GAN is trained on historical flight data to learn the underlying patterns and dynamics of flight trajectories, allowing it to generate diverse synthetic trajectories that closely mimic real-world behavior. The generated flight tracks can be used for various applications, including air traffic simulation, flight path optimization, and anomaly detection. The proposed system architecture leverages cloud computing and AWS services for scalability and flexibility. The workflow includes data collection, preprocessing, model training, and generation of realistic flight trajectories. Security measures such as encryption and access control ensure the integrity and confidentiality of the data. The project aims to address the scarcity of data on new technologies and procedures in air traffic management, enabling scenario forecasting and innovation in airspace design. The significance of this research lies in its potential to enhance safety and efficiency in air traffic management through the generation of realistic synthetic flight tracks.

Overall, this project aims to contribute to the field of air traffic management by providing a novel approach to generating synthetic flight tracks that can be used to improve the safety and efficiency of aviation operations.

**This page intentionally left blank**

Report

# Problem Definition

## Background

Globalization has significantly transformed the landscape of international air travel, redefining the dynamics of connectivity and movement across borders. What was once considered a luxury experience, international flights now epitomize a convergence of commerce, tourism, and cultural exchange, intricately weaving a seamless tapestry in our interconnected world. In the year 2023, a staggering 4.5 billion passengers embarked on international journeys, a volume equivalent to filling the Sydney Opera House approximately 183,000 times.

Amidst the exhilarating experience of traversing the skies, a silent guardian diligently ensures the safety of our airways: collision risk modeling. This sophisticated technological system processes an extensive array of data, encompassing aircraft speed, meteorological patterns, avian migration routes, and even the presence of blinking lights from nearby drones. It meticulously generates a real-time representation of potential hazards in every corner of the airspace. Undoubtedly, collision risk modeling stands as the unsung hero of aviation safety, discreetly averting countless near misses and ensuring the secure navigation of our aerial pathways.[1]

The threat of mid-air collisions is one of the most serious problems facing the air traffic control system and has been studied by many researchers. A mid-air collision is an accident that occurs when two or more aircraft come into unexpected contact while in flight.[2] Due to the very high speeds involved and the potential of subsequent impact with the earth or sea, at least one of the aircraft frequently sustains catastrophic damage or is destroyed. Miscommunication, mistrust, navigation errors, departures from flight plans, a lack of situational awareness, and the absence of collision-avoidance technologies all raise the risk of a mid-air collision. Almost all current large aircraft are equipped with a traffic collision avoidance system (TCAS), which is intended to help prevent mid-air collisions. Based on signals from aircraft transponders, the system informs pilots when a collision with another aircraft is imminent. Despite its shortcomings, it is thought to have significantly reduced midair collisions.[3]

On average, there are approximately 28,000 flights that take off and land in the United States each day. This includes both domestic and international flights. The number of incidents has more than doubled over the past 10 years, with “about 300 accounts of near collisions involving commercial airlines” in the most recently available 12 months of data.[4] From 2016 to 2021, there were 43 reports of midair collisions involving GA operations in the United States, resulting in 79 fatalities. [5]

A graph of blue bars

Description automatically generated

**Figure 1: Accidents by Calendar Year [8]**

The graph depicts a trend in aircraft accidents from 2012 to 2021. Notably, fatal accidents have steadily declined over time, showing a better safety record. In contrast, non-fatal accidents have fluctuated but have largely remained constant. This implies an overall favorable trend in aviation safety, with a considerable decrease in fatal occurrences throughout the indicated time period.

.

A graph of a accident rate

Description automatically generated

**Figure 2: Accident rate by Calendar Year [8]**

The accident rate is presented as accidents per 100,000 flying hours. The black line depicts the total accident rate, while the blue line represents the fatal accident rate. The data show that the total accident rate has progressively decreased from 2012, from 1.10 accidents per 100,000 flight hours in 2012 to 0.24 accidents per 100,000 flight hours in 2021. The fatal accident rate has also decreased but to a lesser amount.

On November 12, 2022, a mid-air collision between a Boeing B-17G and a Bell P-63F at Dallas Executive Airport during an airshow resulted in fatal injuries for all occupants. The aircraft, part of the historical formation, collided as directed by the air boss, leading to a post-impact fire. Both planes broke up and exploded upon impact.[6]

Flight monitoring systems and websites have brought exciting and beneficial improvements to the aviation industry. These offer real-time and historical access to flight information, such as aircraft details, routes, position, speed, altitude, and heading. The broad deployment of ADS-B satellite tracking has made data available to aviation corporations, the media, enthusiasts, and plane spotters, in addition to air traffic control.[7]

ADS-B transponders are installed on aircraft and broadcast radio signals containing information such as the aircraft ID, GPS position, and altitude. These radio emissions are captured by civilian ADS-B receivers stationed near the aircraft. [2]

In aviation, a collision risk model (CRM) is used to track the risk of air-to-air collisions inside airspace and to estimate the likelihood of aircraft collisions. Using probability theory, CRMs evaluate the likelihood of an aircraft collision resulting from random detours.

CRMs compute the probability of an air miss based on two factors:

* The regularity of risk exposure
* The risk of a collision resulting from this exposure

CRMs predict the amount of aircraft collisions and "total" risk for a specific Air Traffic concept-of-operation. The FAA and ICAO established the original CRM, which is used to quantify risk throughout the airborne phase of landing.[9]

## Problem Space

The crucial task of air-to-air collision risk modelling underpins the design and approval of safe and efficient airspace, yet current practices face limitations. Our reliance on historic flight data overlooks the impact of future technologies and procedures, potentially underestimating collision risks and hindering proactive safety measures. Additionally, traditional methods for generating simulated flight tracks struggle to produce the vast sample size needed for comprehensive risk analysis, due to computational constraints.[9]

One solution for this issue is leveraging Generative Adversarial Networks (GANs) to create realistic synthetic flight tracks. GANs possess immense potential to overcome current limitations and revolutionize collision risk modelling. They can learn from existing data and incorporate information about new technologies and regulations, effectively bridging the gap between past and the future. This allows for the generation of synthetic tracks that reflect not only real-world patterns but also the potential impact of upcoming advancements, providing a more complete picture of potential collision scenarios.

Beyond their ability to bridge the data gap, GANs offer remarkable scalability and efficiency. Compared to traditional methods, they can generate a much larger and more diverse library of synthetic tracks, significantly broadening the scope of risk analysis. This comprehensive data allows for the evaluation of numerous scenarios, including extreme weather events, novel flight paths, and the introduction of new aircraft or technologies. By providing a more nuanced understanding of potential risks, GAN-generated data empowers us to proactively implement mitigating measures and safeguard the future of safe airspace. [10]

The impact of this project extends beyond the crucial realm of safety and efficiency. It offers a powerful tool to embrace technological advancements safely and seamlessly, promoting a sustainable and prosperous aviation landscape for the future. By overcoming the limitations of current methods and embracing the potential of GAN-generated data, we can build a future where safety, efficiency, and technological progress go hand in hand, shaping a brighter sky for generations to come.[11]

## Research

The purpose of this study is to conduct GAN experiments to produce realistic airplane trajectories using approach and landing data from airports. Efficiently simulating air traffic flows is crucial for many air traffic management applications. The primary issue with using historical data in a CRM model is that there are frequently not enough tracks available to observe the noteworthy, unusual events. For instance, track data with numerous missed approaches observed must be acquired to determine the collision risk associated with simultaneous missed approaches. Since missed attempts are rare in and of themselves, building a model from a few of these instances could be difficult. Naturally, there are numerous strategies to increase the quantity of tracks that have been detected. The issue with this kind of data set expansion is that trajectories at other airports might not be typical of trajectories at the airport under consideration, and old tracks might not be representative of the current flying environment. The goal of synthetic track data is to address issues with historical and simulated track data. Creating a lot of "realistic" tracks from a small number of observed tracks is the aim of creating synthetic trajectories. According to flight protocols and the principles of physics, the generated tracks ought to be realistic. They ought to have distributions that resemble those of actual paths. Additionally, a wide variety of tracks—including some that have never been heard before—should be included.[12]

The paper, titled 'Generating Synthetic Aircraft Trajectories using Multivariate Density Models,' explores realistic replication of aircraft trajectory patterns during terminal maneuvers, emphasizing multivariate statistical distributions. It aims to generate 2D paths for go-around trajectories at Zurich Airport by amalgamating concepts and extensions of these distributions. The author highlights a tailored discrete representation for effective dimensionality reduction while preserving essential information. While the paper currently focuses on 2D lateral shapes, future work will extend to four-dimensional trajectories, presenting challenges in dimension reduction, potential model shortcomings, and aircraft type influences. Addressing these complexities may involve hybrid methods combining data and model-driven generation to overcome limitations.[13]

In his study "Deep Generative Modelling of Aircraft Trajectories in Terminal Maneuvering Areas," Krauth uses a variational autoencoder to create synthetic tracks based on real tracks close to an airport.[14] This paper explores the use of Variational Autoencoder (VAE) structures and Temporal Convolutional Networks to model 4-dimensional aircraft trajectories. The approach includes a prior distribution called a Variational Mixture of Posteriors (VampPrior) and involves generating artificial data using deep learning techniques like VAEs. However, the effectiveness of this method in safety investigations, particularly in handling unusual events like go-arounds, remains a research question. The paper focuses on whether synthetic flight tracks, generated by VAEs, can provide reliable risk estimates when integrated into a collision risk model, similar to real flight-track data.[12]

A close-up of a person's body

Description automatically generated

**Figure 3: Example of the original tracks (left) and the synthetic tracks (right)**

The ability to randomly generate trajectories may serve to challenge the maximum capacity of airspaces. In general, trajectory-generating methods can be classified into model-driven and data-driven methods. Model-driven approaches use flight dynamic equations to ensure a trajectory's physical qualities while maintaining flight performance. Aircraft performance models such as BADA or OpenAP are especially significant in this context. Although sophisticated solutions exist, model-driven procedures are always based on assumptions that simplify reality. Alternatively, data-driven approaches take advantage of the statistical features of observed paths to excel at duplicating them.[13] Trajectories created using physical and dynamic models in a model-driven generation are unable to account for real-time restrictions like air traffic control or even pilot behavior. Data-driven generation aims to produce more realistic results by considering all the characteristics of actual data. Real airplane models can be directly used in model-driven generation.

Autoencoders excel at learning a given distribution by comparing input to output, making them effective for uncovering hidden data representations. However, they struggle with generating new data as they produce averaged representations, resulting in blurry outputs. In contrast, Generative Adversarial Networks (GANs) take a distinct approach. They employ a Discriminator network to assess the dissimilarity between generated and real data, distinguishing between the two. The Discriminator assigns a number between 0 and 1, indicating fake (0) or real (1) data. The generator's objective is to convincingly generate data that fools the Discriminator into perceiving it as real. These differences make autoencoders more suitable for tasks like data compression or generating semantic vectors, while GANs are better suited for generating new data.[15]

Generative Adversarial Networks (GANs) have become a fascinating tool in the AI toolbox, pushing the boundaries of what machines can create. Their unique "cat-and-mouse" dynamic, where two neural networks compete to generate realistic data and discern it from the real thing, unlock incredible possibilities. Let's explore some existing applications and delve into how we can utilize GANs similarly in other domains.

One well-known application of GANs lies in image generation. Papers like "Progressive Growing of GANs" [16] showcase their ability to produce photorealistic faces, landscapes, and even objects, blurring the lines between the artificial and the authentic. Beyond aesthetics, GANs contribute to advancements in medicine. "Drug Discovery with Deep Generative Models"[17] highlights their potential in designing new molecules with desired properties, accelerating the search for novel drugs and materials.

The creative reach of GANs extends beyond the visual and audio realms. Imagine composing personalized soundtracks with "Text-to-speech with Conditional Generative Adversarial Networks" [18] or even writing original stories guided by text-based GANs. The possibilities are endless! In "A Survey on Text Generation using Generative Adversarial Networks" [19], researchers explore various approaches for text generation with GANs, paving the way for innovative writing tools and content creation.

Looking forward, the potential of GANs remains boundless. From personalized education tailored to individual learning styles to generating synthetic data for scientific simulations, their applications are sure to revolutionize diverse fields. By delving deeper into their inner workings and adapting their principles to new domains, we can unlock a future where creativity and realism seamlessly blend, powered by the ingenuity of GANs.

## Solution Space

Our approach to this project uses advanced machine learning techniques to identify the risks of generating large data of synthetic flight tracks. Our project goal is to develop a model that generates more than one million feasible synthetic flight paths using Generative Adversarial Networks (GANs) and to create a neural network and train it with the historical flight track data to understand and replicate the complexities of flight patterns. This will involve preprocessing the data, exacting design of the GAN architecture, and rigorous training process to maintain the quality of the generated synthetic flight tracks using GANs.

To validate our model's effectiveness, we’ll compare the generated synthetic data against the existing known flight track characteristics and maintain compatibility with CRM requirements. Our project's success is calculated by its ability to generate various sets of flight tracks to improve the safety and efficiency of air traffic management systems.

The intersection of advanced AI techniques and the necessity for improved flight safety measures defines our project solution space. Our model will provide value to airspace management by filling the gaps in current flight track data with synthetic data generated which is similar to realistic trajectories and helps in more comprehensive risk assessments. This will be particularly helpful when new technologies and procedures are introduced especially when not included in historical data. By generating various significant numbers of flight tracks, our model facilitates better and more informed decision-making in airspace management which leads to safer skies.

Our end goal is to develop a robust, scalable model that can be tailored to various aviation scenarios, ensuring the solution remains stable as the air patterns and technology change. From airspace designers to safety analysts, all the users will find it helpful as our method increases the ability to predict and reduce potential collision risks, which leads to a secure aviation environment.

All of this project work from the first design to the last certification is documented and presented at every level. This includes a project report, presentation, and documented GitHub repository which contains all the data and code.

## Project Objectives

To gain a deep knowledge of Generative Adversarial Networks (GANs) and their application in synthesizing flight track data, to learn advanced techniques in data preprocessing, model training, and validation with the flight data, and to achieve insights into the challenges and solutions involved in generating large-scale, realistic datasets for use in risk modeling and airspace design.

To develop a GAN-based model that generates more than one million synthetic flight tracks that are similar to realistic trajectories to produce the data representing new technologies and procedures that are not available in historical flight data and also to deliver a validated model and set of tools and processes which can be used to improve the safety and efficiency of air traffic management.

By the conclusion of the project, it aims to provide a comprehensive understanding of the difficulties involved in flight track data generation and the requirements of CRM, limitations of current simulation techniques and how the AI overcomes these difficulties, and a deep understanding of how the generated data can be utilized to predict and mitigate air traffic risks.

To the world of air aviation safety, this project aims to provide an innovative method to generate various flight track data without extensive computational resources. Also, for airspace designers to safety analysts, this project aims to provide a simulation method to analyze air traffic management scenarios with higher accuracy and detail. These outputs add value to the ongoing research and development in the aviation industry, which leads to safer skies through improved assessment and management.

## Primary User Stories

* As an airspace analyst, I want to give the historical flight track data as input to the synthetic track generation system and get a diverse set of synthetic flight tracks, with that I can perform comprehensive air traffic management simulations and risk assessments for airspace design and safety evaluations.
* As a flight safety engineer, I will utilize a Synthetic Flight Track Generator to create scenarios that include emerging aviation technologies, so that I can assess the impact of these technologies on current air traffic patterns and collision risks.
* As a data scientist in aviation, I will access a diverse dataset on flight tracks generated by GANs to train predictive models for air traffic flow management with more accuracy.
* As an airspace designer, I will incorporate synthetic flight tracks into the design simulations to confirm that new routes and procedures are important for a wide range of operational variables.
* As an air traffic controller, I will utilize the synthetic data generated from GAN model to understand the potential outcomes of unusual flight patterns, which helps to better manage live traffic in difficult scenarios.
* As an aviation regulator, I will input parameters for new flight procedures to the synthetic flight track generator to get synthetic tracks which helps to assess the safety and feasibility of these procedures within the existing airspace structure.
* As an airport operations manager, I will use this synthetic flight data to model potential future traffic scenarios and optimize airport layout and scheduling for better efficiency and safety.
* As a researcher in machine learning, I will compare the synthetic flight tracks with the real-world data to measure the performance and accuracy of the GAN model, which helps me to understand deeper into AI’s potential in data simulation.

## Product Vision

### **Scenario #1**

**For:** Airspace Safety Analysts

**Who**: Wants to assess and improve collision risk models in aviation.

**The:** Synthetic flight track generation system.

**Is a:** tool that uses advanced AI to develop synthetic flight paths that are realistic for the simulation.

**That:** provides a cost-effective and scalable way to get a huge set of diverse flight data.

**Unlike:** traditional simulation methods that depend on historical data and are often computationally intensive and less flexible.

**Our method:** helps analysts test and refine airspace designs and safety measures without the risks of existing traffic patterns or the need for field tests.

**Caveats:** with the notice of the synthetic data validated against real-world scenarios to ensure accuracy and applicability.

### **Scenario #2**

**For:** Aviation researchers concentrating on the integration of new technologies.

**Who:** seeks to learn the impact of drones, urban air mobility, and other emergent technologies on current air traffic.

**The:** GAN based model which generates synthetic flight tracks.

**Is a:** modeling tool that can extrapolate the existing data to forecast future traffic scenarios.

**That:** helps the ability to simulate the incorporation of new flight vehicles and patterns in the existing airspace.

**Unlike:** conventional data extrapolation methods, which do not account for the unique behaviors and routes of new air vehicles.

**Our model:** This model is an analysis tool that can be used in the future for the safety and efficient integration of airspace with new aviation technologies.

**Caveats:** It requires continuous updating and training with the latest data to remain efficient as the technology increases.

# Datasets

## Overview

The dataset for this project contains the airport runaway data. The Zurich airport dataset has air traffic surveillance data that was taken from the OpenSky Network, Which is a non profit community based receiver network collecting the air traffic surveillance data since 2013. The dataset contains information about the runaway 34 at Zurich Airport (LSZH) on the dates 3 October 2019 and 30 November 2019.

A map of an airport

Description automatically generated

**Figure 4: Zurich airport Runway 34**

The dataset consists of 71,400 records. It includes latitude, longitude, icao24, timestamp, altitudeand geo altitude, etc. This dataset helps to analyze and investigate the aviation operations in the area easily by providing the useful insights about the patterns and behaviors of air traffic surrounding runaway 34 at LSZH over the given period.

The developed GAN model which is trained on historical flight data creates new and realistic flight trajectories that resembles the behaviors seen in the existing dataset. Without relying on real time or additional data gathering efforts, this generation of synthetic data helps with the air traffic management research, aviation simulation and to develop innovative approaches for airspace optimization and safety.

## Field Descriptions

**1. timestamp** - Type: datetime - This field indicates the date and time of recording. It is in ISO 8601 format (YYYY-MM-DD Thh:mm:ss+offset).

**2. altitude** – Type: numeric - This field indicates the altitude of the aircraft at the time of the data point recording which is measured in feet.

**3. callsign** – Type: string - This field is a unique identifier assigned to the aircraft which is typically used for communication and identification purposes.

**4. geoaltitude** – Type: numeric - This field indicates the geodetic altitude of the aircraft at the time of the data point recording, measured in feet.

**5. groundspeed** – Type: numeric - This field shows the speed of the aircraft relative to the ground at the time of the data point recording, measured in knots.

**6. icao24** – Type: string - This field is a unique 24-bit address assigned to the aircraft which is generally used for identification in aviation communication protocols.

**7. lastseen** - Type: datetime - This field indicates the date and time when the flight was last detected by the recording system. It is in ISO 8601 format (YYYY-MM-DD Thh:mm:ss+offset).

**8. latitude** – Type: numeric - This field indicates the geographic latitude of the aircraft at the time of the data point recording which is measured in decimal degrees.

**9. longitude** – Type: numeric - This field indicates the geographic longitude of the aircraft at the time of the data point recording which is measured in decimal degrees.

**10. onground** - Type: boolean – This field indicates whether the aircraft is on the ground or in the air at the time of the data point recording. True represents the aircraft being on the ground, while False represents it being in the air.

**11. distance** - Type: numeric – This field tells the distance traveled by the aircraft from a reference point at the time of the data point recording, measured in nautical miles.

**12. flight\_id** - Type: string – This field is a unique identifier assigned to the flight of the aircraft, typically used for tracking purposes.

**13. runway** - Type: string – This field is an identifier of the runway being used by the aircraft at the time of the data point recording, if applicable.

**14. initial\_bearing** - Type: numeric – This field indicates the initial bearing or direction of the aircraft at the time of the data point recording which is measured in degrees clockwise from true north.

**15. initial\_flow** - Type: string – This field indicates the initial flow or direction of air traffic at the time of the data point recording, if applicable.

**16. simple** - Type: string – This field is a simplified representation or categorization of the aircraft's status or activity at the time of the data point recording.

**17. track\_unwrapped** - Type: numeric – This field tells the unwrapped track angle of the aircraft at the time of the data point recording, measured in degrees clockwise from true north.

**18. x** - Type: numeric – This field tells the x-coordinate of the aircraft's position at the time of the data point recording, in a specified coordinate system.

**19. y** - Type: numeric – This field tells the y-coordinate of the aircraft's position at the time of the data point recording, in a specified coordinate system.

**20. timedelta** - Type: numeric – This field tells the time difference between consecutive data points, measured in seconds.

## Data Context

The datasets we are using are a set of data points which represent position of the point at that instance of time and a number of these points can be considered as a flight track. The data was collected by OpenSky Network, the world's largest crowdsourced dataset of real-time and historical flight positions, democratizing air traffic information.

A diagram of a machine

Description automatically generated

**Figure 5: Runway 34 Zurich Airport (LSZH) [20]**

Our dataset offers a unique glimpse into the operations of runway 34 at Zurich Airport (LSZH) for a specific period spanning from October 3rd, 2019, to November 30th, 2019. With 71,400 meticulously cleaned data points, each representing a snapshot of a flight's journey, this resource unlocks valuable insights into various aspects of runway usage. The time frame itself holds significance. Covering late fall, it captures the transition from summer to winter, potentially revealing how changing weather patterns like wind direction and strength impacted runway selection. Examining how wind conditions influenced the distribution of arrivals and departures on runway 34 could expose valuable trends for safety and efficiency optimizations.[21]

## Data Conditioning

The Zurich dataset requires a few conditioning steps to prepare it for analysis and modeling. These procedures include things like dealing with missing values, changing data types, and making sure the data is consistent. The given Zurich dataset focuses on runway 34 only. The missing numbers in the dataset appear to be the primary concern. There are seven aircraft tracks with no geoaltitude in the geoaltitude field, which has 1400 nan values. There are two ways to deal with these missing values: either remove the rows that have missing geoaltitude values or impute the missing values using the mean and median.  Eliminating the nan rows would be appropriate rather than adding inconsequential values that lead to bias since there are 71400 rows total. This also results in the GANs model receiving artificial data input to produce synthetic data.

We then concentrated on each field's datatype and assigned it to its corresponding field. For example, callsign, flight\_id, Icao24, and altitude must be strings; altitude and geoaltitude must be numbers; and the on-ground field, which holds values like True or false, must be Boolean. To ensure that no data values were duplicated or compromised, we additionally verified the data uniqueness of variables such as Icao24, callsign, and flight\_id. We ensured that fields such as geo altitude and altitude were aligned with the ground. That is, anytime the height is zero, the on-ground condition must hold ‘True’. The geo altitude and altitude columns do not include any zero values, and the on-ground field contains only true values. The dataset's integrity needs to be preserved by resolving these discrepancies. After these data conditioning procedures, the Zurich dataset is unique, accurate, and consistent enough to be used in a machine learning model to generate synthetic flight tracks using generative adversarial networks (GANs).

## Data Quality Assessment

**Completeness:** The dataset is highly complete, with all columns containing values except for the altitude field, which has 1400 missing values marked as "nan". However, despite this absence, all other fields are complete with appropriate values. Each attribute, from aircraft identifiers like callsign and icao24 to positional data such as latitude and longitude, contains all the necessary information. Furthermore, all essential fields required for thorough analysis are included, making it a robust dataset for further exploration.

Additionally, there are no missing columns, indicating that the dataset covers all important aspects of flight tracking data as specified in the field descriptions. This level of completeness ensures that the dataset is well-equipped for a variety of analytical tasks, serving as a dependable foundation for gaining insights into aircraft behavior, navigation patterns, and other related analyses.

**Uniqueness:** The dataset shows a high level of uniqueness, with no duplicate columns and each flight consistently represented across 200 rows of data. This guarantees that for every unique flight, identified by attributes like flight\_id and icao24, there are exactly 200 corresponding data points capturing various aspects of its journey. Key flight parameters such as timestamp, altitude, groundspeed, and geographic coordinates (latitude and longitude) show distinct values for each flight, contributing to the dataset's overall uniqueness.

Additionally, attributes like distance, track\_unwrapped, and positional coordinates (x and y) provide further details, ensuring that each data point contributes uniquely to the understanding of individual flight trajectories. Following these uniqueness criteria improves the dataset's dependability and makes it easier to confidently conduct reliable analysis of flying behaviors, navigation patterns, and other aviation-related insights.

**Accuracy:** The dataset shows accuracy in flight characteristics from geographical coordinates to timestamps. Numerical numbers are realistic and free of errors, such as altitude and groundspeed. With accurate insights into aircraft behavior, navigation dynamics, and other crucial aviation parameters, this accuracy creates a solid basis for analytical initiatives.

**Atomicity:** It ensures that each record within the dataset represents a singular, indivisible unit of information pertaining to a specific flight at a precise moment in time. The dataset becomes evident that each row encapsulates a snapshot of a flight's journey, encompassing crucial attributes such as timestamp, altitude, groundspeed, and geographic coordinates. These attributes collectively capture the dynamic nature of aircraft movement and position at distinct points during their travels. By adhering to this atomicity principle, the dataset enables analysts to dissect and analyze flights with granularity, facilitating a detailed understanding of their trajectories, behaviors, and operational characteristics. This granularity allows for precise insights into the temporal and spatial dynamics of aviation activities, empowering stakeholders to make informed decisions and optimizations within the realm of air traffic management, aviation safety, and operational efficiency. Therefore, the dataset's adherence to atomicity principles enhances its utility and reliability for diverse analytical endeavors within the aviation domain.

**Conformity:** This dataset effectively demonstrates conformity by incorporating attributes commonly found in Zurich data, including aircraft identifiers such as callsign and icao24, as well as essential positional data like latitude and longitude, alongside vital flight parameters such as altitude and groundspeed. The dataset's alignment with these standard aviation data elements ensures consistency and compatibility with industry-wide practices, facilitating seamless integration with existing analytical frameworks and tools. Moreover, the meticulous handling of missing values and the maintenance of consistent data types and formats further underscore its conformity to established norms. By upholding these standards, the dataset enables analysts and stakeholders to leverage familiar structures and methodologies, streamlining data processing, interpretation, and decision-making processes within the aviation domain. Thus, the dataset's commitment to conformity not only enhances its interoperability but also bolsters its reliability and utility for a wide array of aviation-related applications and analyses.

**Overall Quality:** The dataset under scrutiny showcases a commendable level of completeness, with the vast majority of attributes populated with values and no glaringly absent columns. This completeness ensures that essential information pertaining to flight tracking, including aircraft identifiers, positional data, and crucial flight parameters, is readily available for analysis. Furthermore, the dataset upholds accuracy standards by providing reliable flight characteristics such as precise geographical coordinates, accurate timestamps, and realistic altitude and groundspeed readings. The meticulous attention to maintaining consistency in data types and formats further enhances the dataset's credibility and usability. Collectively, these attributes contribute to the dataset's high overall quality, positioning it as a robust resource for a diverse range of analytical endeavors in the realm of flight tracking and aviation research. Its completeness, accuracy, and adherence to standards not only instill confidence in the data but also underscore its potential to yield meaningful insights and drive informed decision-making within the aviation domain and beyond.

## Other Data Sources

Currently, we are developing a generative adversarial network (GAN) model to create realistic flight tracks using historical data. For accurate results, our model requires comprehensive flight information, including position (latitude, longitude, altitude), velocity, aircraft identification, and other flight parameters. We began by working with the Zurich dataset from OpenSky Network [21]. However, training solely on the Zurich dataset will likely limit the model's ability to generalize and produce accurate flight tracks for other locations.

To address this challenge, we plan to enhance our machine learning model using hyper tuning and other optimization techniques. Our goal is to generalize the model and test its performance on the historical DCA dataset, also available through OpenSky.[21] This will determine if the model has become dataset specific. While many websites offer real-time flight tracking, historical data sources are limited, with even fewer granting access for research use. Some notable sources include FlightRadar24, Plane Finder, and OAG. We selected OpenSky [21] Network due to its accessibility and suitability for our research purposes.

## Storage Medium

Amazon S3 (Simple Storage Service) is a good storage medium for storing massive datasets, providing scalability, durability, and cost-effectiveness. S3 is widely used across industries and supports a wide range of data types, including multimedia files, backups, logs, and application data [22]. Scalability allows for smooth adjustments to storage capacity without the need for upfront provisioning, making it perfect for changing data requirements. S3's design enables data redundancy across various devices and facilities within an AWS region, improving dependability and reducing data loss due to hardware failures.

Amazon S3 offers a variety of storage classes to accommodate diverse use patterns and cost concerns, improving storage solutions for specific use cases. The Standard class promotes high availability for frequently accessible data, whereas Infrequent Access (IA) and One Zone-IA reduce costs for less often accessed data [23]. Additionally, the Glacier and Glacier Deep Archive classes offer cost-effective long-term archiving solutions. By leveraging these classes, enterprises may efficiently handle massive datasets while limiting costs depending on data access and retention requirements.

To summarize, Amazon S3's advantages, such as scalability, durability, dependability, and cost optimization, make it a popular choice for storing huge datasets. Its flexibility to different data types and access patterns, along with strong data security methods, guarantees a robust and safe storage solution for a wide range of applications and industries.

## Storage Security

AWS S3 (Simple Storage Service) offers strong security measures to safeguard data stored in the cloud. For our project, ensuring the security of data stored on AWS S3 is important. AWS is responsible for safeguarding the infrastructure that underpins AWS services on the AWS Cloud. As part of the AWS compliance initiatives, third-party auditors evaluate and verify security efficacy regularly. Data protection is a shared responsibility and Amazon S3 has several security aspects to consider when creating and enforcing security rules. The following best practices are broad guidelines and do not constitute a complete security solution. To provide optimal security, it is critical to adjust these principles to your specific environment:

Amazon S3 Security Best Practices:

1. Disable access control lists (ACLs): Unless required for specific use cases, it's recommended to disable ACLs and manage access using policies like IAM user policies, S3 bucket policies, VPC endpoint policies, and AWS Organizations service control policies (SCPs). This simplifies permission management and auditing.
2. Ensure correct bucket policies and avoid public access: Use S3 Block Public Access to limit public access to your S3 resources. Review bucket policies to avoid allowing access to "Everyone" or "Any authenticated AWS user."
3. Implement least privilege access: Grant only the permissions necessary for specific tasks. Use tools like IAM roles, bucket policies, user policies, and Service Control Policies to enforce least privilege access.
4. Use IAM roles for applications and AWS services: Manage temporary credentials for applications or services accessing S3 without storing long-term credentials. IAM roles provide temporary permissions for accessing AWS resources.
5. Encrypt data at rest: Use server-side encryption options like SSE-S3, SSE-KMS, and SSE-C to protect data stored in S3. Additionally, consider client-side encryption for added security.
6. Enforce encryption of data in transit: Use HTTPS (TLS) to prevent potential attackers from eavesdropping or manipulating network traffic. Use the aws: Secure Transport condition in S3 bucket policies to allow only encrypted connections.
7. Consider using S3 Object Lock: Implement a "Write Once Read Many" (WORM) models to prevent accidental or inappropriate deletion of data, such as AWS CloudTrail logs.
8. Enable S3 Versioning: Keep multiple variants of an object in the same bucket to easily recover from unintended user actions or application failures.
9. Consider using S3 Cross-Region Replication (CRR): Replicate data between distant AWS Regions to satisfy compliance requirements and enhance data durability.
10. Consider using VPC endpoints for Amazon S3 access: Use VPC endpoints to connect to S3 resources within a VPC, helping prevent traffic from traversing the open internet.
11. Identify and audit all Amazon S3 buckets: Use Tag Editor, S3 Inventory, and resource groups to identify, tag, and audit security-sensitive or audit-sensitive resources.
12. Implement monitoring using AWS tools: Use Amazon CloudWatch metrics for Amazon S3, such as PutRequests, GetRequests, 4xxErrors, and DeleteRequests metrics, to monitor and maintain the reliability, security, availability, and performance of Amazon S3 and other AWS services [24].

In summary, the storage security measures provide a comprehensive approach to protect the integrity and confidentiality of the flight track data used in the project. The data remains secure and reliable for the duration of the research by employing strong encryption, access control and data governance practices.

## Storage Costs

This project requires cloud storage with moderate capacity and flexible access. After careful analysis, AWS S3 Standard has been selected as the most suitable solution. S3's pay-as-you-go model aligns well with the project's requirements. The current small dataset data size makes S3 Standard ideal, as it offers a balance of cost-effectiveness and accessibility. With pricing at $0.023 per GB (US East Coast region), the estimated monthly cost is approximately $0.46, making it a competitive option.

S3 Standard also offers advantages like data backup options for security and seamless integration with EC2 (AWS cloud computing). This compatibility will ensure smooth processes within the project's AWS ecosystem. Overall, S3 Standard emerges as the most appropriate and cost-efficient storage solution for this project [25].

# Algorithms & Analysis / ML Model Exploration & Selection

## Solution Approach

### **Systems Architecture**

The proposed system for generating flight tracks using GANs leverages the power of cloud computing and publicly available datasets to deliver a scalable and flexible solution. The architecture comprises the following core components:

**Data Source (OpenSky Network)**: The OpenSky Network [21] serves as the primary source of historical flight track data. Its extensive repository of real-world flight trajectories provides a rich training dataset for the GAN model. We are using historical data of Zurich dataset from OpenSky.

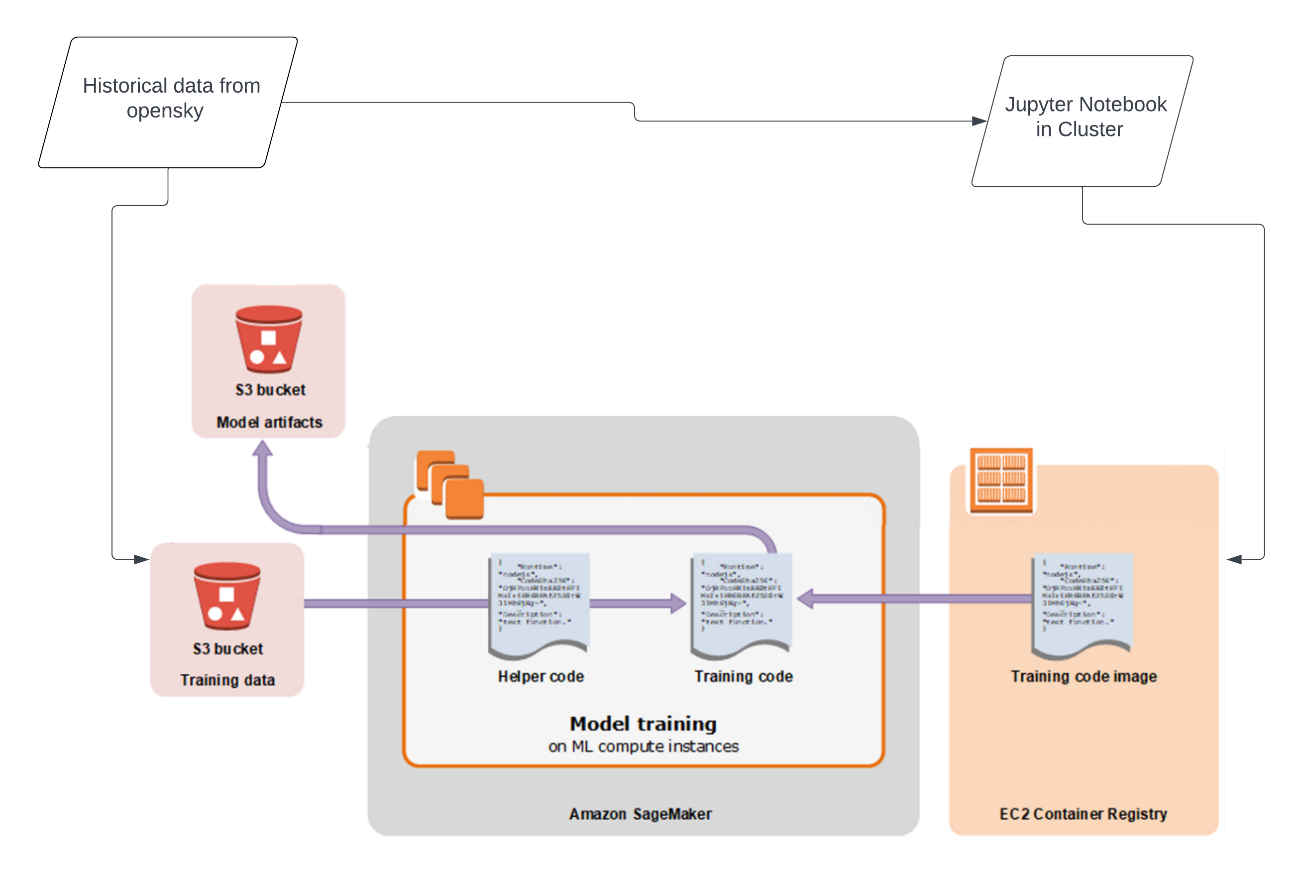
**Data Storage (Amazon S3):** Amazon S3 provides a highly scalable, reliable, and secure object storage service to house the collected flight track data. Data is readily accessible from other AWS components for subsequent processing.

**Computational Cluster:** A cloud-based computational cluster is provisioned to harness the necessary processing power for GAN training. The cluster consists of multiple virtual machines or instances, offering flexibility in selecting instance types (Standard, CPU-optimized, or GPU-optimized) to match the computational demands of the model.

**Development Environment (Jupyter Notebook):** Jupyter Notebooks, hosted within the computational cluster, provide a powerful and interactive development environment. Using preferred deep learning frameworks (TensorFlow, PyTorch, etc.), developers can construct the GAN architecture, comprising both generator and discriminator networks, within this notebook environment.

**GAN Model:** The central component of the system is the Generative Adversarial Network (GAN). The GAN, developed within the Jupyter Notebook, undergoes rigorous training on the historical flight data from OpenSky. The adversarial process between the generator and discriminator networks drives the optimization of flight track generation.

**Generated Flight Tracks:** The trained GAN model produces realistic and plausible new flight tracks. These generated trajectories can be used for various downstream applications, including air traffic simulation, flight path optimization, and anomaly detection.



**Figure 6: System Architecture**

**Workflow**

Data from the OpenSky Network is collected, cleaned, and pre-processed as needed.

* The pre-processed data is uploaded and stored securely in Amazon S3.
* A computational cluster is provisioned, with a suitable configuration based on workload demands.
* A Jupyter Notebook instance is launched within the cluster, offering a development environment.
* The GAN model is designed and implemented within the notebook, leveraging a suitable deep learning framework.
* The GAN training process is initiated, with the model accessing flight track data directly from S3.
* Generated flight tracks are evaluated and visualized for assessment of quality and realism.
* Cloud Provider: The flexibility of this architecture allows the use of various cloud providers (AWS, Azure, GCP). The specific choice may depend on factors such as cost, preferred tools and frameworks, or existing infrastructure. For your project we are going with AWS.

### **Systems Security**

The project to generate synthetic flight tracks using Generative Adversarial Networks (GANs) requires the adoption of strong system security measures. This is critical for maintaining data integrity, ensuring the legitimacy of synthetic flight tracks generated, and protecting any sensitive information contained within the datasets used. We have taken many steps to protect our data and infrastructure. To begin, we store our data on Amazon S3 and encrypt it at rest using server-side encryption (SSE). In addition, we use access control policies such as IAM and bucket policies to restrict access to our S3 buckets. This ensures that only authorized persons have access to the stored data. Furthermore, we employ S3 Access Points to regulate access at the bucket level, allowing us to impose different access regulations for different applications or users.

In our computational cluster, which provides the processing power for training the GAN model, we ensure that the virtual machines or instances are securely configured. We employ security groups or network ACLs to manage inbound and outbound traffic to these instances, reducing the possibility of illegal access.

Our development environment, which is hosted on Jupyter Notebooks, is also secure. We use access controls to restrict who can access and alter the notes. Additionally, we adhere to secure coding practices to prevent vulnerabilities in the GAN model code, ensuring that our development environment remains secure.

Overall, our architecture promotes security at all levels. We regularly upgrade and patch all system components to defend against known vulnerabilities. We also use monitoring and logging to detect and respond to security problems, as well as perform regular security audits and assessments to verify compliance with security best practices and standards.

### **Systems Data Flows**

The data flow within the project encompasses several key components. Firstly, the process begins with the acquisition and preparation of the Zurich airport runway dataset obtained from the OpenSky Network. This dataset, containing crucial air traffic surveillance data, undergoes rigorous cleaning and conditioning procedures to ensure its integrity and consistency. Missing values are addressed, data types are standardized, and any inconsistencies are rectified to prepare the dataset for subsequent analysis and modeling stages.

Following data preparation, the model training and generation phase commence. A Generative Adversarial Network (GAN) model is trained utilizing historical flight data extracted from the cleaned dataset. This model is pivotal in generating synthetic flight trajectories that mimic the behaviors observed in the original dataset. Techniques such as hyperparameter tuning are employed to optimize the model's performance, ensuring accurate and realistic trajectory generation.

Visualizing insights and analysis results derived from the dataset and synthetic flight trajectories constitute the next stage. The project may utilize custom visualization tools or other visualization software tailored to its requirements. These tools connect seamlessly to the dataset stored in Amazon S3, enabling the visualization of valuable insights related to aviation operations and runway usage patterns.

Data storage and security form another critical aspect of the system's data flow. The dataset and synthetic flight trajectories are securely stored in Amazon S3 Standard, leveraging its cost-effectiveness, security features, and seamless integration with other AWS services. Robust security measures, including encryption and access control, are implemented to safeguard the integrity and confidentiality of the stored data, ensuring compliance with AWS security best practices throughout the project's duration.

Finally, cost analysis and optimization are essential considerations in managing the project's resources efficiently. An analysis of the data storage cost on Amazon S3 Standard is conducted, followed by optimization efforts tailored to the project's requirements and budget constraints. This ensures that the project maintains cost-effectiveness while meeting its storage needs, allowing for scalability and flexibility as the project progresses.

Overall, the system's data flow encompasses the acquisition, preparation, modeling, visualization, storage, security, and cost analysis of the Zurich airport runway dataset and synthetic flight trajectories. Each component plays a vital role in ensuring the success and efficiency of the project, from data acquisition to actionable insights and optimization efforts.

### **Algorithms & Analysis**

We initially preprocess the flight data, which comprises timestamps, latitude, longitude, and altitude. The data is grouped by flight ID to guarantee that each route contains 200 consecutive data points. We then use a scaling algorithm to return the data to its original feature range. The GAN has two basic components: the generator and the discriminator. The generator uses a random noise vector as input to construct synthetic routes, while the discriminator determines whether a particular route is real or fake. The GAN is trained in a loop, with iterative updates to the generator and discriminator to enhance the quality of the generated routes. The GAN is trained to construct synthetic flight paths based on patterns seen in real flight data. The discriminator assists the generator in learning to generate more realistic routes by providing feedback on generated routes. By training the GAN on a large dataset of real flight routes, the generator can learn to design routes that mirror actual flight patterns.

The GAN algorithm is implemented in Python, using the Keras module. We store the flight data in AWS S3 and compute it on the Hopper cluster, allowing for efficient GAN training. AWS S3 offers a highly scalable and secure storage solution for massive quantities of flight data needed to train the GAN.   
Utilizing S3 makes it simple to retrieve data from several instances, which supports distributed computing and parallel processing. Because of the cluster's high-performance computing capabilities, the massive amount of data needed to train the GAN can be processed more quickly, resulting in speedier iterations and more effective model building. To minimize the binary cross-entropy loss function, the generator and discriminator networks are optimized throughout the training phase.

We can produce realistic synthetic flight routes once the GAN has been trained. The produced routes show how successful the suggested method is by closely resembling actual flight trajectories. We find that the GAN can capture the underlying patterns and dynamics of real-world flight paths by comparing the generated routes with real flight data. To enhance performance, many topologies for both the generator and discriminator networks are investigated, such as fully connected layers, convolutional neural networks (CNNs), and recurrent neural networks (RNNs). Learning rate, batch size, and network design are all important hyperparameters that influence GAN performance.

Finally, Generative Adversarial Networks provide a promising technique to trajectory prediction in air traffic management, enabling the creation of synthetic aircraft trajectories that closely mirror real-world data patterns. By incorporating new technology and methods, GANs can improve the safety and efficiency of air traffic management, providing significant insights for future airspace design and scenario prediction. Future work could include further developing the GAN architecture and investigating new features to improve the quality of generated routes. Overall, our approach shows potential in creating synthetic flight paths that closely resemble real-world data, which could be useful in a variety of aviation-related applications.

## Machine Learning

### **Model Exploration**

**Time Series Generation Models with GANs**

Time Series Generation Models using Generative Adversarial Networks (GANs) are an innovative approach to generating synthetic time series data that closely mimic real-world data patterns. GANs, a type of deep learning model, consist of two neural networks, a generator and a discriminator, trained adversarially to produce high-quality synthetic samples.

**GAN Architecture for Time Series Generation**

The GAN architecture for time series generation typically involves a generator network that learns to produce synthetic time series data and a discriminator network that learns to distinguish between real and synthetic time series data. The generator network takes random noise as input and generates synthetic time series data, while the discriminator network evaluates the authenticity of the generated data.

**Types of Time Series Models with GANs**

Several types of time series models can be implemented using GANs for various applications:

* Unconditional Time Series Generation: In this approach, the GAN generates time series data without conditioning on any specific input. The generator learns to capture the underlying data distribution and generate diverse samples resembling the training data.
* Conditional Time Series Generation: In conditional time series generation, the GAN takes additional input variables, such as categorical labels or previous time steps, to generate conditioned synthetic time series data. This enables the generation of time series data with specific characteristics or dependencies.
* Multi-modal Time Series Generation: GANs can be extended to generate multi-modal time series data, where the generator learns to produce diverse samples corresponding to different modes or patterns present in the training data. This is useful for capturing the variability and uncertainty inherent in real-world time series data.

**Applications of Time Series Generation Models**

Time series generation models using GANs find applications in various domains, including:

* **Anomaly Detection**: Synthetic time series data generated by GANs can be used to augment training data for anomaly detection models, improving their robustness and generalization to unseen anomalies.
* **Data Augmentation**: GAN-generated time series data can be used to augment limited or imbalanced training datasets, enhancing the performance of machine learning models trained on such data.
* **Scenario Forecasting**: Synthetic time series data generated by GANs can be employed for scenario forecasting and sensitivity analysis in predictive modeling tasks, helping to explore different future trajectories and outcomes.

**Generative Adversarial Networks (GANs)**

Generative Adversarial Networks (GANs) consist of two neural networks – a generator and a discriminator – trained simultaneously through an adversarial process. Here are the key components of GANs that we'll explore:

**Generator Network:**

The generator network takes random noise as input and generates synthetic flight trajectories. We'll explore different architectures for the generator, such as fully connected layers, convolutional neural networks (CNNs), or recurrent neural networks (RNNs), to understand their impact on the quality and diversity of generated tracks.

* **Discriminator Network:**

The discriminator network evaluates the authenticity of generated flight trajectories by distinguishing between real and synthetic tracks. We'll explore variations in discriminator architectures, including CNNs or feedforward networks, to optimize discrimination performance and prevent mode collapse.

* **Training Procedure:**

GANs are trained through an adversarial process where the generator aims to produce realistic flight trajectories to fool the discriminator, while the discriminator aims to distinguish between real and synthetic tracks accurately. We'll explore different training strategies, loss functions (e.g., binary cross-entropy), and optimization techniques (e.g., Adam optimizer) to stabilize training and improve convergence.

* **Hyperparameter Tuning:**

GANs are sensitive to hyperparameters such as learning rate, batch size, and network architecture. We'll explore the effects of varying these hyperparameters on the quality, diversity, and convergence speed of generated flight trajectories. Techniques such as grid search or random search can be employed to find optimal hyperparameter configurations.

* **Evaluation Metrics:**

To assess the performance of GANs in generating flight tracks, we'll explore various evaluation metrics such as Frechet Inception Distance (FID), Inception Score (IS), or Wasserstein distance. These metrics provide quantitative measures of the realism, diversity, and quality of generated trajectories compared to real-world data.

**Trajectory Prediction Models with GANs**

Generative Adversarial Networks (GANs) present a novel approach to trajectory prediction in air traffic management, offering the capability to generate synthetic flight trajectories that closely resemble real-world data patterns. These models hold significant promise for our project objectives, specifically in synthesizing flight track data representing new technologies and procedures not present in historical flight data.

**Architecture:** The GAN architecture for trajectory prediction typically comprises:

* Generator Network: This network learns to generate synthetic flight trajectories based on input noise vectors or conditioning variables, capturing the intricate spatiotemporal dynamics of aircraft movements.
* Discriminator Network: Responsible for evaluating the authenticity of the generated trajectories, distinguishing between real and synthetic trajectories. It provides feedback to the generator to enhance the quality of generated samples.

**Applications:**

The utilization of GANs for trajectory prediction models offers several benefits for our project, including:

* **Data Synthesis:** GANs enable the synthesis of large-scale, realistic datasets of flight trajectories representing new technologies and procedures not present in historical data. This addresses the project's objective of delivering synthetic flight tracks to improve the safety and efficiency of air traffic management.
* **Scenario Forecasting**: Synthetic trajectories generated by GANs can be used for scenario forecasting, allowing us to explore different airspace configurations and traffic scenarios. This provides valuable insights into potential future trajectories and helps in predicting and mitigating air traffic risks.
* **Innovation in Airspace Design:** GAN-generated trajectories facilitate innovation in airspace design by simulating new technologies and procedures. This aids airspace designers and safety analysts in assessing and optimizing airspace management scenarios with higher accuracy and detail.

### **Model Selection: Rationale for Employing Wasserstein GANsfor Flight Track Generation**

Wasserstein Generative Adversarial Networks (WGANs) are selected for this project due to their compelling advantages in synthesizing realistic and diverse flight track data, specifically addressing the project's objectives of representing new technologies and procedures absent from historical data. Key reasons for this choice include:

* **Data Realism and Diversity:** WGANs excel at capturing complex, high-dimensional distributions inherent in real-world flight trajectories. The adversarial training process between generator and discriminator networks drives the generator to produce synthetic flight tracks that are highly indistinguishable from genuine ones. This ensures that the generated data possesses the intricate spatiotemporal patterns characteristic of air traffic movements.
* **Addressing Scarcity of New Technology Data:** WGANs provide a powerful solution for overcoming the lack of historical data concerning new technologies and procedures. Their ability to learn underlying patterns from existing data enables them to simulate plausible flight tracks that align with projected trajectories of novel aircraft or air traffic management protocols.
* **Scenario Forecasting Power:** Generated synthetic flight tracks enable comprehensive scenario forecasting. By simulating diverse and representative air traffic patterns, WGANs facilitate the exploration of various airspace configurations and traffic conditions. This empowers safety and efficiency analyses with insights into potential risks and optimization opportunities.
* **Airspace Design Innovation:** WGAN-produced flight tracks serve as valuable tools for airspace design innovation. Through simulating new technologies and procedures, they provide a sandbox environment for airspace designers and safety analysts to assess, fine-tune, and optimize future airspace management solutions.

**WGANs are the Superior Choice**

WGANs, with their strength in learning complex distributions and generating realistic and diverse data, offer a clear advantage for our project. Their capabilities directly align with the objectives of:

* Representing trajectories shaped by new technologies and procedures
* Enabling scenario forecasting to manage risk and optimize efficiency
* Fostering innovation in airspace design.

By specifically mentioning WGANs, we acknowledge the improvements they offer over traditional GANs, particularly in terms of training stability and convergence.

# Visualizations / ML Model Training, Evaluation, & Validation

## Overview

With the Zurich Dataset from Open Sky Network, we analyzed data and created visualizations to understand the patterns of the data. We compared real data with synthetic data using these visualizations to understand the similarities and differences between the datasets and improve the quality of the synthetic dataset.

## Visualizations

## A comparison of altitude distribution Description automatically generated Figure 7: Altitude distribution

## This histogram illustrates the frequency distribution of altitude measurements across all flights in the dataset. It provides a detailed view of the vertical profile of flights, highlighting common cruising altitudes and the variability in altitude changes during different flight phases such as takeoff and landing. The altitude is at peak when the frequency is above 6000. Within the 12 nautical mile range, the altitude is at peak when the frequency is nearly 4000. A graph of different levels of frequency Description automatically generated with medium confidence

## Figure 8: Groundspeed distribution

**A diagram of a number of blue circles

Description automatically generated**Like the altitude distribution, this histogram plots the groundspeed of aircrafts throughout their flights. Analyzing the distribution of groundspeed helps in understanding typical flight speeds and can aid in identifying outliers or abnormal speed patterns that might indicate specific flight conditions or operational behaviors. The groundspeed peaks between 250-300 when the frequency is between 5500-6000 and the groundspeed – 12 nautical miles peaks when the frequency is nearly 3000.

**Figure 9: Correlation between variables**

This comprehensive scatter plot matrix explores the relationships between multiple flight parameters such as altitude, speed, and time. By examining these correlations, one can infer how changes in one variable might affect others, which is vital for modeling flight dynamics and for predictive analytics in air traffic management.

**A graph showing a line

Description automatically generated  
Figure 10: Altitude over time for flight AAL92\_016**

The above line graph provides a temporal view of altitude changes for a specific flight, showing how altitude varies with time from takeoff to landing. This visualization is particularly useful for studying the operational profile of a single flight, including ascent rates, cruising altitude stability, and descent patterns.

**A graph showing a map of the altitude

Description automatically generated with medium confidence  
Figure 11: Flight Path**

This map visualization traces the actual path flown by an aircraft, illustrating its route from origin to destination. It provides geographic context by showing the trajectory over terrain and can be used to analyze route efficiency, airspace usage, and compliance with flight plans.

## Machine Learning

### **Model Training**

1. **Data Preparation**

**Reshaping Data:** Flight track data is restructured into individual route segments. This prepares the data for model input.

**Feature Scaling:** Features including longitude, latitude, and altitude are scaled (e.g., using MinMaxScaler or StandardScaler). Feature scaling ensures numerical stability and improves model convergence during training.

1. **Model Definition**

**Generator Architecture**: The generator model is constructed using multiple layers with 'relu' activation functions. The final output layer does not use an activation function, providing flexibility and a wider range of values for generated flight routes.

**Discriminator Architecture**: The discriminator is designed to take a flight route as input. It outputs a probability between 0 and 1, indicating its assessment of whether a route is real (1) or generated (0).

1. **GAN Compilation**

**Model Integration**: The generator and discriminator are combined into a single GAN model. During generator training, the discriminator's weights are frozen to focus on improving the generator's ability to deceive the discriminator.

**Optimization and Loss:** The GAN is compiled with an optimizer (commonly 'adam') and a loss function suitable for binary classification ('binary\_crossentropy').

1. **Iterative Training Loop**

**Discriminator Training Phase:** Discriminator training involves feeding it batches of both real and generated flight routes. It learns to differentiate between the characteristics of genuine and synthetic flight paths.

**Generator Training Phase**: The generator learns to produce outputs that are indistinguishable from real flight routes. It is trained in the context of the entire GAN, with the objective of causing the discriminator to misclassify its generated routes as real.

**Monitoring Progress**: Discriminator and generator losses are tracked during each training epoch. Ideally, losses should stabilize over time, implying a healthy training process where both models improve in tandem.

### **Model Evaluation**

**Model Evaluation and Validation for GAN-generated Flight Tracks**

**Introduction**

We have evaluated the performance of a Generative Adversarial Network (GAN) in synthesizing realistic flight tracks. Two scaling techniques, min-max scaling and standard scaling were explored during model development.

**Evaluation**

**Min-Max Scaling**

**Track Smoothness**: Tracks lack ideal smoothness, exhibiting zigzag patterns. This indicates a need for model refinement to better represent continuous flight paths.

**Runway Accuracy**: Runways are accurately depicted by the model , demonstrating successful learning of spatial features.

A graph showing a number of lines and dots

Description automatically generated **Figure 12: Synthetic flight routes – MINMAX SCALING**

**X axis: Scaled Longitude**

**Y axis: Scaled Latitude**

**A red and blue graph

Description automatically generated  
Figure 13: Altitude distribution – MINMAX SCALING**

**X axis: Altitude**

**Y axis: Frequency**

**Altitude Distribution**: Min-max scaling produces a realistic altitude distribution that aligns well with characteristics of real flight data.

**Visual Similarity**: Generated tracks generally resemble real flights, but the lack of smoothness diminishes their realism.

**A screenshot of a computer screen

Description automatically generatedFigure 14: Comparison – MINMAX SCALING**

**X axis: Longitude**

**Y axis: Latitude**

**Standard Scaling**

**Track Smoothness:** Similar to min-max scaling, tracks lack smoothness and appear as zigzags.

**Runway Accuracy**: The model accurately represents runways.

**A graph of a plane

Description automatically generated  
Figure 15: Synthetic flight routes – STANDARD SCALING**

**X axis: Scaled Longitude**

**Y axis: Scaled Latitude**

**Altitude Distribution**: Standard scaling introduces unrealistic altitude values, including outliers and negatives.

**Diversity**: Generated flight paths show more diversity than min-max scaling Addressing the altitude distribution issue could make this a significant advantage.

**A graph of different colored lines

Description automatically generated with medium confidence  
Figure 16: Altitude distribution – STANDARD SCALING**

**X axis: Altitude**

**Y axis: Frequency**

**A screenshot of a graph

Description automatically generated**

**Figure 17: Comparison – STANDARD SCALING**

**X axis: Longitude**

**Y axis: Latitude**

**Wasserstein GANs**

**Track Smoothness:** Compared to mix scaling and standard scaling the flight tracks are more smoother and realistic.

**Runway Accuracy**: The model accurately represents runways and even rest of the path those these path may appear similar but are very different from each other.

A map of a bird

Description automatically generated

**Figure 18: Synthetic flight routes – Wasserstein GANS**

**X axis: Scaled Longitude**

**Y axis: Scaled Latitude**

**Altitude Distribution**: Wasserstein GANs produces a realistic altitude distribution that aligns well with characteristics of real flight data.

**Diversity**: Generated flight paths show more diversity than scaling technique with addressing the altitude distribution issue making a significant advantage. But the generated flight track are similar to each other which show overfitting of model.

A red and blue graph

Description automatically generated

**Figure 19: Altitude distribution – WGANS**

**X axis: Altitude**

**Y axis: Frequency**

A group of graphs with different colored lines

Description automatically generated

**Figure 20: Comparison – WGANS**

**X axis: Longitude**

**Y axis: Latitude**

# Summary

In this project, we've embarked on a fascinating journey to generate synthetic flight tracks using a Generative Adversarial Network (GAN)-based approach. The project involved several key steps, including data preprocessing, model architecture design, training, and route generation. Initially, we preprocessed the flight track data, including scaling using MinMaxScaler, and organized it into suitable formats for training the GAN model. We encountered challenges such as ensuring generated values fall within specific ranges and addressing issues related to model overfitting and lack of diversity in generated tracks.

To tackle these challenges, we made crucial modifications to the GAN architecture, especially focusing on the generator's output layer to ensure generated values align with the desired ranges. we also experimented with different scaling techniques, shifting from StandardScaler to MinMaxScaler for improved normalization. Additionally, we explored ways to enhance diversity in generated tracks, such as incorporating data augmentation techniques and fine-tuning GAN parameters.

# Future Work

Moving forward, several avenues for future work present exciting opportunities to further enhance our project's outcomes. Implementing and evaluating quality metrics such as Frechet Inception Distance (FID) and Inception Score (IS) can provide quantitative insights into the generated flight tracks' realism and diversity. Integrating these metrics into the model evaluation process will enable more robust assessments and help in refining the GAN model's performance.

Exploring advanced techniques like progressive GANs or Wasserstein GANs tailored specifically for flight track generation could lead to more sophisticated and realistic synthetic routes. Additionally, incorporating domain-specific knowledge or constraints, such as airspace regulations or weather conditions, into the GAN training process can enhance the generated tracks' authenticity and relevance to real-world scenarios.

Moreover, considering real-time generation capabilities for synthetic flight tracks opens doors to various applications, including flight simulation, airspace management systems, and training data augmentation for aviation-related machine learning models. Implementing streaming GAN architectures or integrating the GAN model with real-time flight data sources can enable on-the-fly generation of synthetic routes, enhancing our project's practical utility and impact.

Overall, leveraging advanced GAN techniques, integrating quality evaluation metrics, exploring domain-specific constraints, and enabling real-time generation capabilities constitute promising directions for future work, aiming to further improve the realism, diversity, and applicability of synthetic flight track generation using GANs.

Appendix

Appendix A: Glossary

**Table 1: Glossary Table**

|  |  |
| --- | --- |
| Term | Definition |
| TCAS | Traffic Collision Avoidance System |
| ADS-B | Automatic Dependent Surveillance-Broadcast |
| ID | Identification |
| GPS | Global Positioning System |
| CRM | Collision Risk Model |
| GAN's | Generative Adversarial Networks |
| GA | General Aviation |
| VAE | Variational Autoencoder |
| BADA | Base of Aircraft Data |
| LSZH | Zurich Airport (IATA Code) |
| YYYY-MM-DD Thh:mm:ss+offset | Date and Time format with offset |
| ICAO24 | International Civil Aviation Organization 24-bit Aircraft address |
| ISO | International Organization for Standardization |
| DCA | Washington D. C Airport |
| OAG | Official Aviation Guide |
| S3 | Simple Storage Service |
| SSD’s | Solid State Drive’s |
| HDD’s | Hard Disk Drive’s |
| NAS | Network Attached Storage |
| AWS | Amazon Web Services |
| ACL’s | Access Control Lists |
| SCP’s | Service Control Policies |
| IAM | Identity and Access Management |
| VPC | Virtual Private Cloud |
| SSE-S3 | Server-Side Encryption with Amazon S3 |
| SSE-KMS | Server-Side Encryption with AWS Key Management Service |
| SSE-C | Server-Side Encryption with Customer-Provided Keys |
| HTTPS(TLS) | Hypertext Transfer Protocol Secure (Transport Layer Security) |
| WORM | Write Once Read Many |
| CRR | Cross Region Replication |
| GB | Giga byte |
| EC2 | Elastic Compute Cloud |
| GCP | Google Cloud Platform |
| CPU | Central Processing Unit |
| GPU | Graphics Processing Unit |
| SSE | Server-Side Encryption |
| CNN’s | Convolutional Neural Network |
| RNN’s | Recurrent Neural Network |
| FID | Frechet Inception Distance |
| IS | Inception Score |
|  |  |
|  |  |

**Appendix B: GitHub Repository**

**Overview:**

**GitHub Repository Link:**

<https://github.com/SkySquad6/SyntheticFlightTrack>

**GitHub Repository Contents:**

**Generating Synthetic Flight Tracks Using Generative Adversarial Networks (GANS)**

**Capstone Project  
DAEN 690-006  
Team Sky Squad: Neeraj Kumar Neela, Naga Pranavi Kandarpa, Sai Yeshwanth Reddy Chinakota, Pravallika Avula, Junaid Mohammed, Lalithanjali Yarrapatruni**  
**Abstract**

This project focuses on the generation of synthetic flight tracks using Generative Adversarial Networks (GANs) based on historical flight data from Zurich Airport. The dataset is based on runway 34 at Zurich Airport for the dates 3 October 2019 to 30 November 2019. The dataset includes information such as timestamp, altitude, callsign, and other relevant fields. The project aims to overcome the limitations of traditional simulation methods, which often rely on simplified models that may not accurately represent the complexities of real-world flight trajectories. The proposed system architecture comprises several key components, including a data source (OpenSky Network), data storage (Amazon S3), a computational cluster, a development environment (Jupyter Notebook), and the GAN model itself. The GAN is trained on historical flight data to learn the underlying patterns and dynamics of flight trajectories, allowing it to generate diverse synthetic trajectories that closely mimic real-world behavior. The generated flight tracks can be used for various applications, including air traffic simulation, flight path optimization, and anomaly detection. The proposed system architecture leverages cloud computing and AWS services for scalability and flexibility. The workflow includes data collection, preprocessing, model training, and generation of realistic flight trajectories. Security measures such as encryption and access control ensure the integrity and confidentiality of the data. The project aims to address the scarcity of data on new technologies and procedures in air traffic management, enabling scenario forecasting and innovation in airspace design. The significance of this research lies in its potential to enhance safety and efficiency in air traffic management through the generation of realistic synthetic flight tracks. Overall, this project aims to contribute to the field of air traffic management by providing a novel approach to generating synthetic flight tracks that can be used to improve the safety and efficiency of aviation operations.

**Problem Description**

Air-to-air collision Risk Modelling for Airspace design and approval is dependent on accurate flight tracks. CRM typically uses historic flight tracks supplemented with simulated flight tracks. Historic flight tracks cannot represent new technologies and procedures. Simulated flight tracks can represent new technologies and procedures, however, even using super-computers a sufficient sample size of flight tracks cannot be generated. Conventional methods heavily rely on historical flight data, often neglecting emerging technologies and procedures, leading to an incomplete understanding of collision risks. The challenge lies in the scarcity of diverse flight track data for comprehensive risk analysis, constrained by computational limitations. To overcome this, the project proposes leveraging Generative Adversarial Networks (GANs) to synthesize realistic flight tracks. By training the GAN model on historical data, the aim is to produce a comprehensive dataset mirroring real-world trajectories, accommodating both conventional and emerging aviation scenarios. This approach facilitates proactive safety measures and informed decision-making in air traffic management.  
A close-up of a person's body

Description automatically generated

**Project Goals**

The project goals are as follows.

Generate one million realistic synthetic flight tracks using GANs to fill data gaps and assess emerging aviation risks.

Develop validated tools for air traffic management, ensuring safety compliance and informed decision-making.

Gain insights into flight data generation challenges for enhanced airspace design and risk modeling.

**About the Files Attached to the GitHub**

Zurich airport Runway 34 Data set which has the flight details downloaded from the OpenSky Network.

GANS.ipynb - Wasserstein Gans visual using Python Code.

**Tool/algorithm that We Used for the project**

Scaling algorithm to normalize features.

GAN uses noise vectors to create synthetic routes.

**Cloud Platform (AWS)**

Storage: EC2 instance built-in storage  
Computing: EC2 (Windows server)  
AWS S3 and Hopper cluster facilitate GAN training.

**Dataset**

A map of a runway and a map of a runway

Description automatically generated  
The datasets we are using are a set of data points which represent position of the point at that instance of time and a number of these points can be considered as a flight track. The data was collected by OpenSky Network, the world's largest crowdsourced dataset of real-time and historical flight positions, democratizing air traffic information.

Our dataset offers a unique glimpse into the operations of runway 34 at Zurich Airport (LSZH) for a specific period spanning from October 3rd, 2019, to November 30th, 2019. With 71,400 meticulously cleaned data points, each representing a snapshot of a flight's journey, this resource unlocks valuable insights into various aspects of runway usage. The time frame itself holds significance. Covering late fall, it captures the transition from summer to winter, potentially revealing how changing weather patterns like wind direction and strength impacted runway selection. Examining how wind conditions influenced the distribution of arrivals and departures on runway 34 could expose valuable trends for safety and efficiency optimizations.

**Solution Approach**

A diagram of a model training

Description automatically generated  
The proposed system architecture for generating flight tracks using Generative Adversarial Networks (GANs) leverages cloud computing and publicly available datasets for scalability and flexibility. It consists of several core components. Firstly, the OpenSky Network serves as the primary data source, providing historical flight track data, particularly focusing on the Zurich dataset. This data is then securely stored in Amazon S3, a highly scalable and reliable object storage service, facilitating easy access for subsequent processing. A cloud-based computational cluster is provisioned to harness the necessary processing power for GAN training, offering flexibility in selecting instance types based on workload demands.

Within this computational environment, Jupyter Notebooks provide an interactive development environment for constructing the GAN architecture using preferred deep learning frameworks such as TensorFlow or PyTorch. The GAN model, developed within the Jupyter Notebook, undergoes rigorous training on historical flight data from OpenSky, with the adversarial process between the generator and discriminator networks driving optimization of flight track generation. The trained GAN model produces realistic flight tracks, which can be utilized for various downstream applications including air traffic simulation, flight path optimization, and anomaly detection.

The workflow involves data collection, cleaning, and preprocessing from OpenSky, followed by secure storage of preprocessed data in Amazon S3. Subsequently, a computational cluster is provisioned, and a Jupyter Notebook instance is launched for GAN model development. The GAN training process accesses data directly from S3, and the generated flight tracks are evaluated and visualized for quality assessment. Additionally, the architecture offers flexibility in cloud provider selection, with AWS chosen for this project due to factors such as cost and tool compatibility.

**Model Selection**Generative Adversarial Networks (GANs) are chosen for their significant advantages in synthesizing realistic and diverse flight track data, specifically to address the project's objectives of representing new technologies and procedures absent from historical data. GANs excel at capturing complex, high-dimensional distributions inherent in real-world flight trajectories, ensuring that generated data possesses intricate spatiotemporal patterns characteristic of air traffic movements. Moreover, GANs provide a robust solution for overcoming the scarcity of historical data concerning new technologies and procedures by simulating plausible flight tracks aligned with projected trajectories. These synthetic flight tracks enable comprehensive scenario forecasting, empowering safety and efficiency analyses with insights into potential risks and optimization opportunities. Additionally, GAN-produced flight tracks serve as valuable tools for airspace design innovation, providing a sandbox environment for assessing, fine-tuning, and optimizing future airspace management solutions. Overall, GANs stand out as the superior choice due to their capability to represent trajectories shaped by new technologies, enable scenario forecasting, and foster innovation in airspace design.

**Wasserstein GAN  
A map of a plane

Description automatically generated**

Appendix C: Risks

**Sprint 1 Risks**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Risk | Description | Probability | Impact | Mitigation |
| Insufficient knowledge and comprehension of GANs within the group | As GAN technology has only recently been adopted, it is difficult to determine how it works and how it will benefit the current project. | Medium | High | Studying GAN architecture and applications in various sectors will help you learn how to work with GANs and their capabilities. |
| Limited Real-Life information on mid-air collisions | Because there are few sources with accurate information on mid-air crashes that occurred, the project's purpose is also ambiguous. | Medium | High | Thorough research from various sources is required to understand the problem in detail. |
| Dataset | There is an error in creating account on Opensky Network and downloading the dataset. | High | High | Contact the website regarding the problem. Talk to the client about other sources |
| Availability schedules | Members of the team have varying availability times. | Medium | Medium | Discussed with the team and decided to have a virtual meet at a time when everyone is available. |

**Table 2: Sprint 1 Risks**

**Description of Risks:**

Determining the project objectives and problem space was the primary aim of this sprint. Without a clear conversation with the client, it was impossible to grasp the project. Following the client meeting, the client's needs were evident, and the project's objectives were clearly stated.

Among the several problems we faced, acquiring information about mid-air collisions proved difficult due to the abundance of aircraft mishaps and the sheer volume of other instances, making data retrieval for such occurrences more complicated. We looked for different collisions, their causes, and their impacts as a team. To determine their frequency, we located data on mid-air collisions that took place over a ten-year period.

Lack of understanding of GANs was another possible concern. Although GAN technology has recently started to be applied in numerous fields, its full potential and application are still unknown. We conducted research to learn how GANs are applied for various applications and what the outcomes were. We discovered remarkable outcomes and recognized the capabilities of GANs. We also tried to comprehend the architecture of GANs and the various expectations involved. The dataset is the primary issue we encountered. We were unable to obtain the dataset, even though Opensky Network has more precise airport information. The website has a flaw that prevents users from creating new accounts and, data cannot be accessed without an account. We so got in touch with OpenSky and our client to discuss our problem. Although we don't have access to the dataset, we started working on another dataset provided to us by the client.

Even though each team member has varying schedules, we have all talked about this problem and determined that the best way to make time for the project is to provide each team member a daily time slot that works for them. Since the project is still in its early phases, the team has not yet run into any unexpected dangers. Even though the team has been proactive in identifying risks and developing plans to mitigate them, it could have been advantageous to conduct an even more thorough investigation of possible data sources and to consult with domain experts earlier. We are all working well together as a team to make this project successful, not just to get the job done but also merely because we enjoy it.

**Sprint 2 Risks**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Risk | Description | Probability | Impact | Mitigation |
| Generative Modelling | We are new to Generative Modelling | High | High | Start working on GAN’s architecture early. |
| Learn to use AWS and Cluster | To work on the project, we need to learn AWS as backup if the Cluster does not work. | Low | Medium | To watch training tutorials on Cluster and AWS for Self-Training. |
| Missing values in the Zurich dataset | While going through the dataset for Data Quality Assessment and Data Conditioning we identified there are 1400 values missing in the dataset. | High | Low | Removed those missing value rows from the dataset as they do not show much impact on the future work. |
| Visualizations for area under 6 miles | Client advised us to do the visualizations of the flights path under 6 miles for better understanding | Low | Low | We did study the dataset and filtered the data for visualizations under 6 miles |

**Table 3: Sprint 2 Risks**

**Description of Risks:**

One significant risk identified is our novelty in Generative Modelling, posing a high probability and impact on our project's success. To address this, it's imperative to proactively engage with Generative Adversarial Networks (GANs) architecture early in the project timeline. By familiarizing ourselves with GANs and their architecture, we can mitigate potential setbacks and ensure a smoother integration of generative modelling techniques into our workflow. This proactive approach allows us to build competence and confidence in utilizing Generative Modelling effectively, thereby reducing the risk associated with our inexperience in this domain.

Another risk involves the presence of missing values in the Zurich dataset, which presents a high probability of occurrence with a relatively low impact. To mitigate this risk, we've adopted a pragmatic approach of removing rows containing missing values from the dataset during Data Quality Assessment and Conditioning.

Given that the missing values have minimal impact on the future analysis and outcomes, this mitigation strategy ensures the integrity and reliability of our data while streamlining our analysis process. By implementing this solution, we can maintain data consistency and accuracy, mitigating the potential disruptions that missing values could introduce to our project's progress.

**Sprint 3 Risks**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Risk | Description | Probability | Impact | Mitigation |
| Scaling of generator | We need to find scaling of every point at each instance | Medium | High | Trying to index data points |
| ML model selection for data discriminator | We need to select a ML model which best suits our project to use in discriminator | High | Medium | Generally, for a discriminator, CNN is used. But we are exploring other types of CNN like Resnet, VGGNet. |
| Visualizations for different areas | For better understanding we had to consider visualizations for different areas. | High | Low | We had to do multiple visualizations for the areas that cover under 6 miles, 9 miles and 12 miles for better understanding |
| Model Exploration and Analysis | Understanding of model and algorithm | Medium | Medium | As we are in the process of developing the system we require some time to figure out the model in depth. |

**Table 4: Sprint 3 Risks**

**Description of Risks:**

One notable risk in our project involves the scaling of the generator, presenting a medium probability but with potentially high impact. To mitigate this risk, we're implementing a strategy focused on indexing data points, which enables us to efficiently manage and process the scaling of each point at every instance. By employing indexing techniques, we aim to streamline the scaling process, ensuring the smooth operation of our generator component and minimizing the impact of scaling challenges on our project's progress.

Another critical risk pertains to the selection of the machine learning (ML) model for the data discriminator, which carries a high probability and medium impact. To address this risk, we're exploring various ML models beyond the conventional CNN approach, such as Resnet and VGGNet. By conducting thorough experimentation and evaluation, we aim to identify the model that best aligns with our project requirements, optimizing the performance and effectiveness of our discriminator. This proactive approach enables us to mitigate the risk associated with model selection uncertainty, ensuring that our discriminator operates optimally within the project framework.

**Sprint 4 Risks**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Risk | Description | Probability | Impact | Mitigation |
| Loss of details while applying smoothing filter. | Smoothing filters can potentially oversimplify the flight tracks, removing important variations or details. | High | Medium | Experiment with different smoothing filters and their parameters to find a balance between noise reduction and detail preservation. |
| Mode Collapse | If the discriminator becomes too powerful early in training, the generator might struggle to learn diverse patterns and get stuck producing limited types of routes. | Low | High | Start with a simpler discriminator and gradually increase its complexity as the generator improves. |
| Complexity of WGANs | WGANs can be more complex to implement, and tune compared to standard GANs. | High | High | WGANs can be sensitive to hyperparameter choices. Start with well-established configurations and gradually experiment. |
| Increase layers in models | Adding too many layers can lead to the model overfitting. | Low | Medium | Use techniques like dropout or L1/L2 regularization to prevent overfitting. |

**Table 5: Sprint 4 Risks**

**Description of Risks:**

Several risks were found throughout this sprint as the models were developed and implemented. One important risk was the potential loss of detail when applying smoothing filters to flight tracks. Smoothing filters are used to eliminate noise and make recordings more continuous, but they can also oversimplify records by missing essential variations or features. To avoid this risk, we experimented with several smoothing filters and parameters to find an appropriate mix of noise reduction and feature retention.  
  
Another problem we encountered was mode collapse, in which the generator struggles to learn varied patterns and becomes stuck providing only certain types of routes. This might occur if the discriminator gets very powerful early in training. To reduce this risk, we began with a simpler discriminator and gradually raised its complexity as the generator improved, resulting in more diversified pathways.  
  
Furthermore, the intricacy of Wasserstein GANs (WGANs) compared to normal GANs constituted a risk to our research. WGANs can be more difficult to build and modify because of their distinct loss function and training dynamics. To avoid this risk, we began with well-known WGAN configurations and subsequently experimented with various hyperparameters to get the ideal settings for our specific use case.  
  
Finally, adding too many layers to the models increased the risk of overfitting, in which the model learns noise in the training data rather of the underlying patterns. To reduce this risk, we utilized approaches such as dropout or L1/L2 regularization to avoid overfitting and ensure that the models generalize well to new data.  
  
Identifying these risks and implementing appropriate mitigation techniques allowed us to handle possible issues while also improving the performance and robustness of our synthetic flight track generation models.

**Sprint 5 Risks**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Risk | Description | Probability | Impact | Mitigation |
| Completing Report | Completing the report with precise content and proper formatting before the due date is important | Low | High | We as a team worked together every week on parts of report every week. |
| Preparing final presentation | Preparing the final ppt where we explain our problem statement and results clearly within a given time is crucial for everyone to comprehend the project | Low | High | We distributed project parts among ourselves and worked individually on each part to complete and prepare ourselves for the presentation. |
| Exploring WGANs | There are many other parameters in WGANs where we only focused on Wasserstein distance. | Medium | High | Other parameters can be further discussed and employed for more factors and hyper parameter tuning. |

**Table 6: Sprint 5 Risks**

**Description of Risks:**As a team, completing the report with precise content and proper formatting before the due date was crucial. To mitigate this risk, we divided the report into sections and assigned each team member specific parts to work on every week. This allowed us to make steady progress and ensure that the report was completed on time with thorough content and adherence to formatting guidelines.

Preparing the final presentation was another significant risk we identified. It was important for us to create a PowerPoint presentation that clearly explained our problem statement and results within the allotted time. To address this risk, we divided the presentation into sections and assigned each team member specific parts to focus on. This approach allowed us to thoroughly prepare for the presentation and ensure that all aspects of our project were covered in a clear and concise manner.

Lastly, while exploring Wasserstein Generative Adversarial Networks (WGANs), we realized that there were many other parameters and aspects of WGANs that we did not fully explore. Although we focused on the Wasserstein distance, we acknowledged that further exploration of other parameters could enhance our understanding and improve the performance of our model. To mitigate this risk, we planned to discuss and experiment with other parameters in WGANs, such as the architecture of the generator and discriminator, batch size, and learning rate, to gain a more comprehensive understanding and potentially improve our model's performance.

**Appendix D: Agile Development**

**Scrum Methodology**

Our team adopted the Scrum methodology for this generating synthetic flight tracks using Generative Adversarial Networks (GANs) project. Scrum is an agile framework that promotes adaptive planning, early delivery, and continuous improvement through an iterative approach. The project was divided into five sprints, each spanning a fixed duration, to incrementally develop and refine the solution.

**Sprint Timelines**

The project timeline was organized into the following sprints, with each sprint focusing on specific goals and deliverables:

* **Sprint 1:** 01/17/2024 – 02/06/2024

**Goal:** Define project objectives, research mid-air collisions and CRM models, investigate GANs

* **Sprint 2:** 02/07/2024 – 02/27/2024

**Goal:** Study the dataset, provide key insights, perform data conditioning and quality assessment

* **Sprint 3:** 02/28/2024 – 03/19/2024

**Goal:** Explore ML model selection and algorithms, develop system architecture, work on model development

* **Sprint 4:** 03/20/2024 – 04/09/2024

**Goal:** Refine ML models, create visualizations, train and evaluate models, enhance documentation

* **Sprint 5:** 04/10/2024 - 04/30/2024

**Goal:** Finalize model evaluation and validation, prepare final deliverables

**Scrum Ceremonies and Artifacts**

To facilitate effective collaboration and progress tracking, we conducted daily team meetings known as the Daily Scrum or Stand-up meeting. During these meetings, each team member shared their progress, plans for the day, and any impediments they were facing. This helped in synchronizing efforts and promptly addressing any issues.

We maintained a Scrum log to record important discussions, decisions, and action items from the Scrum meetings. This log served as a reference point for the team and stakeholders to track the project's progress and ensure transparency.

The YouTrack tool was utilized to manage the project backlog and assign tasks to team members for each sprint. User stories and tasks were created in YouTrack, and team members updated the status of their assigned tasks as they progressed. The Scrum Master and Product Owner regularly reviewed and verified the task statuses, closing the tasks upon completion.

By employing Scrum methodology and its various ceremonies and artifacts, such as daily team meetings, the Scrum log, and the YouTrack tool, we fostered effective teamwork, maintained transparency, and ensured a structured approach to delivering incremental value throughout the project.

A green arrow with white text

Description automatically generated

**Each Sprint Timelines**

**Sprint Analysis**

**Sprint 1 Analysis:**

The team's main goal during this sprint was to provide the foundation for the project. We initially concentrated on defining each team member's work and contribution to the project. It took some time to learn about everyone's talents and skills, but by the conclusion of the sprint, we had established good collaboration and were able to synchronize our efforts to achieve the project requirements.

Our initial duty was to investigate the causes and consequences of mid-air collisions, which provided us with important insights into the project's aims. Because our project intends to provide data acceptable for CRM models, we undertook considerable study on CRM models to better understand their operation and data requirements. Following that, we looked at several technologies utilized in synthetic flight tracking, which helped us better comprehend the project topic.

Next, we researched Generative Adversarial Networks (GANs) and their potential uses. We investigated GAN architecture, goals, and applications in a variety of disciplines. We identified and recorded user stories based on this knowledge and in alignment with our project objectives.

We divided duties among team members and used tools like YouTrack to track progress and guarantee assignments were completed on time. Daily team meetings were set to review progress, address any issues, and ensure deadlines were met.

Reflecting on our sprint, we identify opportunities for improvement. We could have done a more thorough investigation of prospective data sources and consulted domain experts earlier in the process. Furthermore, while we set aside time for research and planning, a more systematic approach to work assignment and progress tracking could have increased efficiency.

Moving forward, we hope to improve our processes, increase collaboration, and take a more systematic approach to task management. By learning from our sprint experiences, we are better prepared to handle the project's hurdles and achieve successful solutions in succeeding stages.

**Sprint 2 Analyis:**

The team's main goal during this sprint was to study the dataset and provide key insights from the same for future work. Everyone from the team studied the dataset and provided their understandings on the same. We guided each other if anyone one of us is stuck somewhere.

Our interim 1 goal was to provide an overview on the dataset, describe the fields of the dataset and define the context of the dataset. As we have divided 3 tasks for each interim, we assigned each task to 2 people and coordinated on the same in PPT preparation, Report updating and presentations.

Our interim 2 goal was to do Data Conditioning, Data Quality Assessment and provide other data sources if we have referred to any. Similar to previous week we have divided each task for 2 people. While performing this week's tasks we have noticed that there are 1400 missing values for few flights and "nan". We had decided to remove those values and go ahead as there will be not much impact on the project in future.

Our interim 3 goal was to define Storage Medium, Storage Security and Storage Costs. We have refereed to many sources for understanding which storage medium would best fit for our dataset and understood Amazon S3 would be the good one. We have done our research work for this tasks in this area and updated the same in the report.

Overall, in this sprint we have learned about the dataset and its features which laid a great foundation for us on the dataset in further work.

**Sprint 3 Analysis:**

The team’s main goal for this sprint was to explore the ML model selection and algorithm. Everyone in the team provided their understandings and insights on the same. We were able to help each other if anyone is stuck at some point.

Our interim 1 goal was to develop the system architecture and start working on the model development. Besides that we also did the visualizations for the areas under 6 miles, 9 miles, 12 miles. Presented the same to the client and got few valuable insights.

Our interim 2 goal was to develop the model and work on system security and system data flow. We have divided the tasks as 2 persons per task and worked on it. While performing these tasks, we worked on Model development as well.

Our interim 3 goal was to work on Model Exploration, algorithms and analysis and simultaneously working on the model development. We have divided each task for 3 persons and everyone works on model development. The team work resulted in good progress towards the project and updated the same in the project report.

**Sprint 4 Analysis:**

The focus of Sprint 4, Visualizations / ML Model Training and Evaluation, was to refine our approach to machine learning models while enhancing our presentation and documentation skills.

During our interim 1, the team concentrated on selecting the most suitable models for our project, accompanied by the development of visualizations that would effectively communicate our findings. This period also marked the submission of our initial abstract draft, setting a foundational narrative for our project.

In the interim 2, our efforts pivoted towards the training of the chosen models. To enrich our understanding and methodology, we engaged in a thorough examination of relevant IEEE papers, drawing inspiration and guidance for our approach. This phase was also significant for our documentation process, as we prepared the first draft of our final abstract and initiated the layout for our final report, ensuring that our documentation was not only comprehensive but also adhered to professional standards.

The interim 3 was dedicated to the critical tasks of model evaluation and validation, ensuring that our models were not only accurate but also reliable. In addition to revising our abstract to produce a second draft, we concentrated on finalizing the sprint overview, a summary that encapsulates the achievements and learnings of this sprint. The culmination of our efforts was reflected in the meticulous formatting of our report for final submission, symbolizing the readiness to present our project in its entirety.

**Sprint 5 Analysis:**

In Sprint 5, our team continued to embrace Agile development principles, specifically leveraging the Scrum methodology to enhance project management and collaboration. The main goal for this sprint was to finalize model evaluation and validation and prepare the final deliverables, which included contributions to our GitHub repository. We focused on refining our approach to synthetic flight track generation using Generative Adversarial Networks (GANs), ensuring the final model was robust and reliable.

Regular Scrum ceremonies, such as daily stand-ups and sprint reviews, facilitated effective team communication and immediate problem-solving, reflecting a structured and incremental approach to project delivery. This sprint not only marked the culmination of our project's active development phase but also set the stage for future enhancements and potential research.

**References:**

[1] Button, K. (n.d.). The impacts of globalisation - OECD. <https://www.oecd.org/greengrowth/greening-transport/41373470.pdf>

[2] Wikimedia Foundation. (2023, December 17). *Flight tracking*. Wikipedia. <https://en.wikipedia.org/wiki/Flight_tracking#:~:text=Aircraft%20carry%20ADS%2DB%20transponders,the%20vicinity%20of%20the%20aircraft.>

[3] Wikimedia Foundation. (2024, January 21). *Mid-air collision*. Wikipedia. <https://en.wikipedia.org/wiki/Mid-air_collision#:~:text=The%20potential%20for%20a%20mid,lack%20of%20collision%2Davoidance%20systems.>

[4]  Air traffic by the numbers. Air Traffic By The Numbers | Federal Aviation Administration. (n.d.). <https://www.faa.gov/air_traffic/by_the_numbers>

[5] *Midair collision report*. General Aviation Joint Safety Committee. (2022, December 16). <https://www.gajsc.org/mid-air-collision-report/#:~:text=Midair%20collisions%20are%20a%20persistent,States%2C%20resulting%20in%2079%20fatalities.>

[6] Aviation investigation preliminary report - WFAA. (n.d.). <https://interactive.wfaa.com/pdfs/Report_CEN23MA034_106276_11_30_2022-2_31_45-PM.pdf>

[7] Hayward, J. (2022, March 15). *How do flight tracking websites work?*. Simple Flying. <https://simpleflying.com/how-do-flight-tracking-websites-work/>

[8] General aviation accident dashboard: 2012-2021. (n.d.). <https://www.ntsb.gov/safety/data/Pages/GeneralAviationDashboard.aspx>

[9] Air-to-air collision risk models (CRM) in the terminal airspace | IEEE ... (n.d.). <https://ieeexplore.ieee.org/document/10124323/>

[10] Brownlee, J. (2019, July 19). A gentle introduction to generative adversarial networks (Gans). MachineLearningMastery.com. <https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/>

[11] A method for generating synthetic air tracks - IEEE ... - IEEE xplore. (n.d.-a). <https://ieeexplore.ieee.org/document/4700126/>

[12]  Sherry, L., Shortle, J., Payan, A., Harrison, E., Thapa, A. K., Melgar, A. C., & Auguste, Y. (2023, May). Review of Current State of Artificial Intelligence/Machine Learning and Other Advanced Techniques Related to Air-to-Air Collision Risk Models (CRM) in the Terminal Airspace.

[13] Krauth, T., Morio, J., Olive, X., Figuet, B., & Monstein, R. (2021, December 30). *Synthetic aircraft trajectories generated with multivariate density models*. Engineering Proceedings. https://digitalcollection.zhaw.ch/handle/11475/23934

[14] Krauth, T., Lafage, A., Morio, J., Olive, X., & Waltert, M. (2022). Deep Generative Modelling of Aircraft Trajectories in Terminal Maneuvering Areas. doi:org/10.2139/ssrn.4254106

[15] *What are the pros and cons of generative adversarial networks vs variational autoencoders?*. Quora. (n.d.). https://www.quora.com/What-are-the-pros-and-cons-of-Generative-Adversarial-Networks-vs-Variational-Autoencoders

[16] Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018, February 26). *Progressive growing of gans for improved quality, stability, and Variation*. arXiv.org.<https://arxiv.org/abs/1710.10196>

[17] Bian, Y., & Xie, X.-Q. (2020, August 20). *Generative Chemistry: Drug Discovery With Deep Learning Generative Models*. arXiv.org. <https://arxiv.org/abs/2008.09000>

[18] Michelsanti, D., & Tan, Z.-H. (2017, September 7). *Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification*. arXiv.org. <https://arxiv.org/abs/1709.01703>

[19] de Rosa, G. H., & Papa, J. P. (2022, December 20). *A survey on text generation using generative Adversarial Networks*. arXiv.org. <https://arxiv.org/abs/2212.11119>

[20] LSZH - Zurich. My Website. (n.d.). [https://acukwik.com/Airport-Info/icao/lszh](https://acukwik.com/Airport-Info/icao/lszh )

[21] Meides, M. (n.d.). Network opensky explorer emergency alerts coverage & facts receiver ranking VHF/voice feeding. The OpenSky Network. [https://opensky-network.org/](https://opensky-network.org/ ).

[22] [24] Damue, B. (2023, May 17). Exploring the Basics of Amazon Simple Storage Service (S3). Medium. <https://medium.com/@dbrandonbawe/exploring-the-basics-of-amazon-simple-storage-service-s3-f8ad2af0a6f9>

[23] [25] S3 Storage: How It Works, Use Cases and Tutorial. (2024, January 22). Cloudian. https://cloudian.com/blog/s3-storage-behind-the-scenes/

[24] Security best practices for Amazon S3 - Amazon Simple Storage Service. (n.d.-b). <https://docs.aws.amazon.com/AmazonS3/latest/userguide/security-best-practices.html>

[25] AWS.. Amazon S3 Simple Storage Service Pricing - Amazon Web Services. Amazon Web Services, [Inc. https://aws.amazon.com/s3/pricing/?p=pm&c=s3&z=4](https://aws.amazon.com/s3/pricing/?p=pm&c=s3&z=4)

[26] *Cloud Object Storage - Amazon S3 - AWS*. (n.d.). Amazon Web Services, Inc. <https://aws.amazon.com/s3/>   
[27] *Security best practices for Amazon S3 - Amazon Simple Storage Service*. (n.d.). <https://docs.aws.amazon.com/AmazonS3/latest/userguide/security-best-practices.html>

**This page intentionally left blank**