Synthetic Flight Data Generation Using Generative Models

Aly, Karim; Sharpanskykh, Alexei

DOI
10.1109/ICNS65417.2025.10976960

Publication date
2025

Document Version
Final published version

Published in
ICNS 2025 - Integrated Communications, Navigation and Surveillance Conference

Citation (APA)
Aly, K., & Sharpanskykh, A. (2025). Synthetic Flight Data Generation Using Generative Models. In *ICNS 2025 - Integrated Communications, Navigation and Surveillance Conference: Integrated CNS: Towards Innovative and Efficient CNS Service Provision* (Integrated Communications, Navigation and Surveillance Conference, ICNS). IEEE. https://doi.org/10.1109/ICNS65417.2025.10976960

Synthetic Flight Data Generation Using Generative Models

Karim Aly
Faculty of Aerospace Engineering
Delft University of Technology (TU Delft)
Delft, The Netherlands
k.y.s.b.aly@tudelft.nl

Alexei Sharpanskykh
Faculty of Aerospace Engineering
Delft University of Technology (TU Delft)
Delft, The Netherlands
o.a.sharpanskykh@tudelft.nl

2025 Integrated Communications, Navigation and Surveillance Conference (ICNS) | 979-8-3315-3473-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICNS65417.2025.10976960

Abstract—**The increasing adoption of synthetic data in aviation research offers a promising solution to data scarcity and confidentiality challenges. This study investigates the potential of generative models to produce realistic synthetic flight data and evaluates their quality through a comprehensive four-stage assessment framework. The need for synthetic flight data arises from their potential to serve as an alternative to confidential real-world records and to augment rare events in historical datasets. These enhanced datasets can then be used to train machine learning models that predict critical events, such as flight delays, cancellations, diversions, and turnaround times. Two generative models, Tabular Variational Autoencoder (TVAE) and Gaussian Copula (GC), are adapted to generate synthetic flight information and compared based on their ability to preserve statistical similarity, fidelity, diversity, and predictive utility. Results indicate that while GC achieves higher statistical similarity and fidelity, its computational cost hinders its applicability to large datasets. In contrast, TVAE efficiently handles large datasets and enables scalable synthetic data generation. The findings demonstrate that synthetic data can support flight delay prediction models with accuracy comparable to those trained on real data. These results pave the way for leveraging synthetic flight data to enhance predictive modeling in air transportation.**

Index Terms—**Generative Artificial Intelligence, Variational Autoencoders, Gaussian Copula, Synthetic Flight Information, Synthetic Data Quality Assessment, Flight Delay Prediction, Air Traffic Management, Air Transportation Deep Learning, Statistical Modeling**

I. Introduction

The aviation industry increasingly relies on artificial intelligence (AI) and machine learning (ML) to optimize air transport operations, enhance efficiency, reduce delays, and improve decision-making processes. These technologies enable airlines, airports, and air traffic controllers to make data-driven decisions that improve scheduling, fuel efficiency, and passenger experiences. However, a key challenge remains: the scarcity of comprehensive, high-quality datasets due to limited data collection possibilities, strict data privacy regulations, commercial competition, proprietary restrictions, and regulatory barriers. Many critical flight-related events, such as delays, diversions, and cancellations, occur infrequently, making them rare in historical datasets. This scarcity complicates the study and prediction of such events, a challenge known as class imbalance in machine learning. These constraints hinder the development of accurate predictive models and the generalization of machine learning-based solutions, ultimately limiting their applicability to real-world aviation scenarios. A promising approach to addressing these challenges is synthetic data generation (SDG), which creates artificial datasets that closely replicate real-world scenarios while maintaining essential statistical properties [1]. SDG can not only supplement existing datasets but also help augment machine learning training data, addressing the issues of class imbalance and the underrepresentation of rare events.

Recent advancements in generative AI have produced powerful models capable of generating high-fidelity synthetic tabular data. Notable approaches include probabilistic models like the Gaussian Copula (GC) [2] and deep learning-based methods such as the Tabular Variational Autoencoder (TVAE) [3]. These techniques use different strategies to model real-world data distributions. GC relies on statistical modeling, while TVAE employs neural network architecture. Despite their promise, the application of these approaches in the area of air transportation remains largely underexplored, particularly regarding their ability to preserve data fidelity and enhance predictive modeling. Although synthetic data offers clear benefits, its effectiveness in aviation-related tasks remains uncertain, highlighting the need for rigorous benchmarking against real data through statistical and machine learning assessments to validate its applicability in this domain. Furthermore, aviation data poses unique challenges, such as complex temporal dependencies and operational constraints, which necessitate careful model selection and evaluation.

Since access to key attributes of European flight data, such as flight schedules, statuses, delays, and diversion information, is restricted, this study aims to evaluate the feasibility of using generative models to produce realistic synthetic flight information. Specifically, we examine the potential of TVAE and GC in generating synthetic flight data, with a particular focus on their impact on flight delay prediction accuracy. To achieve this, we conduct five experiments using different sets of features as input to the generative models, aiming to identify the optimal set of features and data types that allow the model to learn the underlying characteristics of real-world flight data.

We propose a four-stage evaluation framework that assesses the statistical similarity, fidelity, diversity, and predictive per-

Funded by SESAR 3 Joint Undertaking, co-funded by the European Union.

formance of the generated data. Our findings reveal a trade-off between the size of the synthetic data that can be produced and the utility of these data for predictive tasks. While GC demonstrates superior statistical similarity, its high computational demand limits scalability, resulting in smaller synthetic datasets that may not be ideal for training predictive models due to the underrepresentation of various flight patterns. In contrast, TVAE offers greater scalability, capable of being trained on large datasets that include all flight patterns and generating large synthetic datasets that retain these patterns. However, it exhibits higher sensitivity to feature selection and data types.

Despite these challenges, our results suggest that synthetic data can effectively support predictive modeling in aviation, providing a viable solution to the data access challenges that hinder research in air transportation. By enabling the development of more robust machine learning models, synthetic data generation can help address key limitations associated with real-world datasets.

The remainder of this paper is structured as follows: Section II reviews key methodologies for synthetic tabular data generation. Section III describes the methodological framework employed to generate synthetic flight data using TVAE and GC, covering all stages from preprocessing the raw historical data to evaluating the synthetically generated datasets. Section IV presents the results of our comparative experiments, assessing both the statistical fidelity and the predictive utility of the generated data across multiple evaluation metrics. Section V highlights key insights, discusses limitations, and explores practical considerations for deploying synthetic data in air transport applications. Finally, Section VI provides concluding remarks and outlines future research directions, focusing on further improving the quality and realism of synthetic flight data.

II. Related work

Synthetic data generation for tabular datasets has advanced considerably, evolving from traditional statistical techniques to sophisticated deep learning-based models. This section reviews key methodologies, highlighting their applications and limitations while identifying gaps in the literature concerning the utilization of synthetic flight data in air transportation and air traffic management (ATM).

Statistical and probabilistic methods like Gaussian Copulas (GC) remain powerful tools for synthetic data generation, effectively decomposing multivariate distributions into marginal distributions and dependency structures to capture complex relationships efficiently [2]. Their flexibility has been further enhanced through extensions such as Archimedean and Vine Copulas, which allow for more adaptable hierarchical dependency modeling in high-dimensional datasets. Additionally, oversampling techniques like the Synthetic Minority Oversampling Technique (SMOTE) [4] and the Adaptive Synthetic (ADASYN) sampling approach [5] were originally designed to address class imbalances by generating synthetic samples for minority classes, but have since been repurposed for general synthetic data generation.

Deep learning has revolutionized synthetic data generation, particularly through Variational Autoencoders (VAEs) [6], which have been adapted for tabular data via architectures like the Tabular Variational Autoencoder (TVAE) [3]. TVAEs introduce modifications that accommodate mixed categorical and continuous variables, improving their ability to preserve complex feature relationships. Similarly, Generative Adversarial Networks (GANs) have gained traction since their introduction by Goodfellow et al. [7], leading to tabular adaptations such as Table-GAN and TGAN [8], [9]. Notably, CTGAN [3] effectively addresses challenges associated with categorical variables through conditional generation and mode-specific normalization, while medGAN [10] pioneered the use of GANs for discrete electronic health records. Other refinements, such as VeeGAN [11], introduce mechanisms to mitigate mode collapse, further improving the robustness of synthetic data generation.

Although synthetic data is widely used in domains such as finance, healthcare, cybersecurity, computer vision, and manufacturing, its adoption in air transportation remains limited. Most existing research focuses on downstream machine learning tasks, such as predicting flight delays, cancellations, diversions, and turnaround times, relying solely on historical data. However, these datasets often suffer from significant class imbalances, as critical events like cancellations and diversions occur far less frequently than on-time flights, making accurate predictions challenging. Rather than augmenting these datasets with synthetic flight records, research efforts have primarily concentrated on refining predictive models while overlooking the fundamental limitations of the data itself. Beyond augmenting rare events in historical datasets, the potential for conditionally generating synthetic flight data to simulate hypothetical scenarios—such as extreme weather conditions or congested airspace—remains largely unexplored.

Prior work that references synthetic flight data has primarily focused on generating synthetic flight trajectories [12]–[14]. In contrast, this study defines flight data more broadly, encompassing key flight attributes such as flight number, airline and aircraft information, origin and destination airports, scheduled and actual departure and arrival times, air time, and operational flight logs indicating delays, cancellations, and diversions. Many of these features, particularly for European flights, are restricted or unavailable in public datasets, limiting researchers' ability to develop robust machine learning models. This study seeks to bridge this gap by demonstrating how generative models can supplement real-world flight data, providing a scalable and practical solution to data scarcity challenges in ATM.

Despite the growing number of generative models and their continuous evolution, no single approach can be universally regarded as the best. The performance of synthetic data generation methods is highly dataset-dependent, and in many cases, well-established techniques still outperform newer variants on specific datasets. Historical flight data encompasses flights between numerous airports at different times of the day and under varying conditions, requiring generative models capable of capturing this complexity. These models must generate synthetic data that reflects the full variability of real-world flight patterns without overfitting to a specific subset. For this reason, we adopt Gaussian Copulas (GC) from statistical modeling and the Tabular Variational Autoencoder (TVAE) from deep learning to generate synthetic flight data. While each method brings different advantages to the analysis, both are recognized for their stability and reduced susceptibility to mode collapse, a common issue in GAN-based approaches, making them well-suited for our application.

III. Methodology

This section describes our analysis framework, covering data collection, preprocessing, and feature engineering. It also details the generative models used in five experiments with three different sets of input features. Finally, we outline the four key evaluation criteria applied to assess the experimental results. Fig. 1 provides an overview of the entire analysis process.

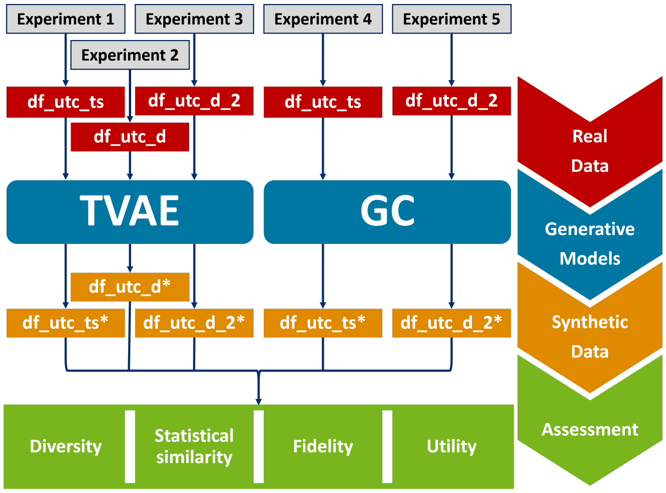


Fig. 1: Overview of the analysis framework.

A. Data and Preprocessing

This study uses the publicly available "TranStats Database for Airline On-Time Performance" [15] from the Bureau of Transportation Statistics (BTS) [16]. The data covers U.S. domestic flights and provides detailed information on flight delays, cancellations, diversions, and their causes. This level of detail makes it a valuable resource for modeling various use cases in air transportation.

To make the task more challenging for the generative models, we used flight data for all arrivals and departures in New York State during January 2023, rather than limiting the data to flights between two specific airports. This resulted in a dataset with 109 features (columns) and approximately 61,000 flights (rows), spanning 113 airports and 508 routes.

The data underwent extensive exploratory analysis, preprocessing, and feature engineering to ensure clean and well-structured input for the generative models. Randomly missing values, such as missing "Tail number", were removed from the dataset. However, other missing values, like arrival times for cancelled flights, were retained, as they carry meaningful information and were incorporated into the training of the generative model.

Departure and arrival time features were initially represented as local times in integer format (HHMM) without the associated date components. During the preprocessing phase, the date component from the "Flight Date" feature was combined with the "Scheduled Departure Time". Using additional duration-based features, such as "Air Time (min)" and "Scheduled Elapsed Time (min)", the appropriate date components were calculated and appended to the remaining time features. To ensure consistency, all time features were converted into timezone-aware datetime objects based on the time zones of the origin and destination airports. Finally, they were standardized to Coordinated Universal Time (UTC) to achieve a unified temporal representation throughout the dataset.

Based on their relevance to this study, the number of features was reduced to 30, encompassing categorical, numerical, and datetime features, along with relational attributes that can be derived from others. This combination of different data types and the complex relationships among them presents challenges for synthetic data generation. Directly including all 30 features as input to the generative models may disrupt inherent dependencies, leading to inconsistencies. For instance, the "Origin Airport ID" could be incorrectly paired with an unrelated "Origin City", or the "Scheduled Elapsed Time (min)" might not align with the difference between "Scheduled Arrival Time UTC" and "Scheduled Departure Time UTC".

To preserve these dependencies and enhance the fidelity of the generated flight information, the dataset was organized into three distinct DataFrames: *"df_utc_ts"*, *"df_utc_d"*, and *"df_utc_d_2"*. Each DataFrame incorporates different sets of features as input to the generative models. These DataFrames were used in five generation tests, as described in Section III-C, to assess the impact of various feature combinations on the quality of the generated data.

Table I provides an overview of the content of each DataFrame, specifying the features included as input to the generative models and the relational features that can be computed or inferred after generation. The first DataFrame primarily consists of time-related features represented as timestamps (datetime format), while the second and third focus mainly on time durations in minutes (numeric format), with a limited number of timestamps included. This structured approach allows for a systematic analysis of how different feature representations influence the ability of the generative models to learn and replicate the underlying patterns in real-world flight data.

By testing multiple feature configurations, we aim to determine the optimal set of attributes that maximize the realism and fidelity of the generated data while preserving essential relationships among features. Ensuring these dependencies remain intact is crucial for maintaining the operational correctness of synthetic flight records, particularly for downstream machine learning tasks such as flight delay prediction. The last column of Table I specifies the input features used in the predictive models discussed in Section III-D, which serve as a benchmark for assessing the practical utility of the synthetic data in real-world aviation scenarios.

TABLE I: Features used in this analysis, categorized as: included (✓), excluded (✗), or calculated post-generation (🖩).

Features	df_utc_ts	df_utc_d	df_utc_d_2	df_prediction
Unique Carrier Code	✓	✓	✓	✓
Tail Number	✓	✓	✓	✓
Origin Airport ID	✓	✓	✓	✗
ICAO Origin Airport	🖩	🖩	🖩	✓
Origin City	🖩	🖩	🖩	✗
Origin State Code	🖩	🖩	🖩	✗
Origin State Name	🖩	🖩	🖩	✗
Destination Airport ID	✓	✓	✓	✗
ICAO Destination Airport	🖩	🖩	🖩	✓
Destination City	🖩	🖩	🖩	✗
Destination State Code	🖩	🖩	🖩	✗
Destination State Name	🖩	🖩	🖩	✗
Quarter	🖩	🖩	🖩	✓
Day of Week	🖩	🖩	🖩	✓
Scheduled Departure Time UTC	✓	✓	✓	✓
Actual Departure Time UTC	✓	🖩	✓	✓
Departure ΔT (min)	🖩	✓	✓	✓
Departure Delay Label	🖩	🖩	🖩	✗
Taxi Out Time (min)	🖩	✓	✓	✓
Wheels Off Time UTC	✓	🖩	🖩	✓
Wheels On Time UTC	✓	🖩	🖩	✗
Taxi In Time (min)	🖩	🖩	✓	✗
Scheduled Arrival Time UTC	✓	🖩	🖩	✓
Actual Arrival Time UTC	✓	🖩	🖩	✗
Arrival ΔT (min)	🖩	🖩	✓	**Target**
Arrival Delay Label	🖩	🖩	🖩	✗
Scheduled Elapsed Time (min)	🖩	✓	✓	✓
Actual Elapsed Time (min)	🖩	✓	✓	✗
Air Time (min)	🖩	✓	✓	✗
Distance (miles)	🖩	🖩	🖩	✓

B. Generative Models

In this study, we employed the Tabular Variational Autoencoder (TVAE) and Gaussian Copula (GC) models to generate synthetic flight data, including trip logs that indicate whether flights departed or arrived on time or were delayed.

One of the primary challenges in generating tabular data is managing the variety of feature types (e.g., numerical, categorical, datetime) and handling missing values that might carry additional information about the dataset. Since synthetic data must replicate the structure of the original data, any missing values in the original dataset must be mirrored in the generated data. Both TVAE and GC models assume that the data columns are fully populated with numerical values. In cases where these assumptions do not hold, a preprocessing step is required. This step modifies the data by transforming columns of one type into one or more columns of another type, as outlined in Table II.

To address columns with missing values, each of such columns is split into two: a column of the same type, where missing values are filled by randomly selecting non-missing values from the same column, and a categorical column indicating whether the original data was present ("Yes") or missing ("No") for each row. This approach ensures that the original column is fully populated, while also accounting for the presence of missing values in the original dataset [2].

TABLE II: Data conversion during preprocessing to handle non-numeric and missing values, adapted from [2].

Original Column Type	Replaced Column(s) Type
Categorical	Number
Datetime	Number
Number w/Missing Values	Number & Categorical
Categorical w/Missing Values	Categorical & Categorical
Datetime w/Missing Values	Datetime & Categorical

The TVAE is a deep learning model designed to extend the functionality of traditional autoencoders by incorporating probabilistic modeling tailored to tabular data [17]. The model includes an encoder neural network that maps the input data, x, into a probabilistic distribution over the latent space, z, represented as $q(z|x)$. A decoder network reconstructs the data as the conditional distribution $p(x|z)$. This process allows the TVAE to learn the underlying patterns and relationships within the data, enabling the generation of synthetic data that closely mimics the original data [18]. The objective function of TVAE is to maximize the Evidence Lower BOund (ELBO) on the log-likelihood of data [19], denoted by:

$$ELBO = \mathbb{E}_{z\sim q(z|x)}[\log p(x|z)] - D_{\mathrm{KL}}(q(z|x)\|p(z)) \quad (1)$$

Maximizing the data reconstruction likelihood $\log p(x|z)$, represented as the expectation $\mathbb{E}_{z\sim q(z|x)}[\log p(x|z)]$, ensures that the decoder accurately reconstructs the input from the latent representation. Simultaneously, minimizing the Kullback-Leibler (KL) divergence $D_{\mathrm{KL}}(q(z|x)\|p(z))$ ensures that the approximate posterior $(q(z|x)$ aligns closely with the prior $p(z)$, promoting a coherent and structured latent space [6], [20].

Through our analysis, we adapted the same TVAE model structure as in [3]. The TVAE was trained for 300 epochs using Adam optimizer with a learning rate 1e-3 and ELBO loss (1).

The Gaussian Copula is a statistical model that allows for the modeling of complex dependencies between variables while preserving their marginal distributions [21]. It operates by transforming the marginal distributions of the data into uniform distributions through their cumulative distribution functions (CDFs), and then combining them using a Gaussian Copula function [22]–[24]. The Gaussian Copula is defined by the correlation structure of the underlying multivariate normal distribution. Specifically, for random variables $X_1, X_2, \ldots, X_d$ with marginals $F_1(x_1), F_2(x_2), \ldots, F_d(x_d)$, the copula C_θ captures the joint distribution as:

$$\begin{aligned} &C_\theta\left(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)\right) = \\ &\quad \Phi_\theta\left(\Phi^{-1}(F_1(x_1)), \Phi^{-1}(F_2(x_2)), \ldots, \Phi^{-1}(F_d(x_d))\right) \quad (2) \end{aligned}$$

where Φ_θ represents the joint CDF of the multivariate normal distribution with correlation matrix θ, and Φ^{-1} is the inverse of the standard normal CDF. This copula model enables the generation of synthetic data that captures realistic dependencies while maintaining the marginal distributions of the variables.

This research combines Gaussian Copulas with Kernel Density Estimation (KDE) from [25], using a Gaussian kernel to estimate the marginal distributions of each variable. This method allowed us to model the interdependencies between variables and generate synthetic data that closely reflects the joint distribution of the original dataset.

C. Experiments

Five experiments were conducted to systematically evaluate the impact of input data types and feature selection on the quality of synthetic flight information generated by Tabular Variational Autoencoder (TVAE) and Gaussian Copula (GC). The experiments were designed as follows:

- ***Experiment 1: TVAE with df_utc_ts***
- ***Experiment 2: TVAE with df_utc_d***
- ***Experiment 3: TVAE with df_utc_d_2***
- ***Experiment 4: GC with df_utc_ts***
- ***Experiment 5: GC with df_utc_d_2***

Table III presents the input data sizes used to train the generative models in the different experiments, along with the size of the synthetic datasets sampled from the learned distributions. TVAE was trained on approximately 61,000 flights, whereas GC was limited to 5,000 flights due to memory constraints. Similarly, sampling with GC is computationally expensive, so only 5,000 flights were generated.

TABLE III: Data sizes

Experiments	Input (real)	Sampled (syn.)	Reconstructed	Cleaned
Experiment 1	(60767, 10)	(60000, 10)	(60000, 30)	(24422, 30)
Experiment 2	(60767, 10)	(60000, 10)	(60000, 30)	(50954, 30)
Experiment 3	(60767, 13)	(60000, 13)	(60000, 30)	(51685, 30)
Experiment 4	(5000, 10)	(5000, 10)	(5000, 30)	(2433, 30)
Experiment 5	(5000, 13)	(5000, 13)	(5000, 30)	(2414, 30)

After generation, we reconstructed relational features inferred from other variables and applied rejection sampling to remove synthetic routes that did not exist in historical data. The cleaned synthetic datasets were used in the evaluation framework described in the next section.

D. Evaluation Framework

Evaluating synthetic data is more critical than generating it, as unreliable synthetic data can lead to incorrect conclusions. Without rigorous validation, synthetic data cannot be trusted to be used for downstream tasks. A standard evaluation step is to assess the validity and structure of the generated data by ensuring that the number of generated features matches the real data, continuous values remain within the observed min/max range, and discrete values correspond to the original categories. Beyond validity, the evaluation framework focuses on four aspects: data diversity, to ensure the synthetic dataset captures the variability of the real data; statistical similarity, by comparing distributions and correlations; fidelity, measuring how well synthetic samples preserve patterns from real data; and utility, determining how well models trained on synthetic data perform in comparison to those trained on real data.

The diversity assessment included applying Principal Component Analysis (PCA) to project both real and synthetic data into a two-dimensional space [26], allowing for a visual evaluation of whether the synthetic data captured the different distributions and clusters present in the real data. Additionally, the class balance was examined by inspecting the arrival and departure delay labels to ensure that the synthetic data maintained a similar or nearly identical ratio of on-time to delayed flights as the real data. Maintaining diversity is crucial, as insufficient variability in synthetic data may lead to biased or unrepresentative models in downstream tasks.

The statistical assessment of the synthetic data was performed both visually and numerically. Visually, we compared the marginal and bivariate distribution plots of the real and synthetic data to identify discrepancies in statistical patterns. To compare individual distributions numerically, we used the Total Variation Distance (TVD) to quantify the divergence between the probability distributions of boolean and categorical columns [27], and for numerical and datetime columns, we calculated the statistical similarity using the Kolmogorov-Smirnov test [28], which measures the difference between the cumulative distribution functions (CDFs) of the real and synthetic datasets. To compare relationships between feature pairs, Correlation Similarity [29] and Contingency Similarity [30] were employed.

To assess the fidelity of the synthetic data, seven classifiers were trained for a binary classification task to distinguish between real and synthetic data. Each model brings unique strengths to the assessment: Random Forest and Gradient Boosting capture complex, non-linear relationships [31], [32]; K-Nearest Neighbors (KNN) and Decision Trees are effective for detecting local patterns [33], [34]; Naive Bayes provides insights into probabilistic dependencies [35]; Logistic Regression models linear relationships [36]; and Stochastic Gradient Descent (SGD) is well-suited for handling large-scale data [37]. Stratified KFold cross-validation was employed to ensure class distribution was preserved across folds [38], with five splits and shuffling to enhance model robustness. Performance was evaluated by averaging accuracy and F1 scores across all classifiers [39], providing a comprehensive measure of their ability to differentiate between the real and synthetic datasets. A lower classification accuracy indicates higher similarity between synthetic and real data, as the models struggle to distinguish between them.

The utility of the synthetic data was assessed by evaluating whether it preserved or enhanced the predictive characteristics compared to real data. Accurate prediction from models trained on synthetic data indicates that it can be reliably used as a substitute for real data in downstream tasks and decision-making. For this reason, sixteen regression models were trained to predict flight arrival delays in minutes, with each offering distinct advantages for the evaluation. Similar to the choices of the classification models, the selection of the regression models was strategically diverse, encompassing algorithms capable of capturing both linear and non-linear relationships, handling high-dimensional feature spaces, and identifying local patterns within the data. Additionally, several ensemble learning techniques were incorporated to leverage their enhanced predictive performance, while other models were specifically chosen to assess the impact of dimensionality reduction. Each experiment from Section III-C, included two testing scenarios: (1) Train-Real-Test-Real (TRTR), which established the baseline performance using historical data, and (2) Train-Synthetic-Test-Real (TSTR), which evaluated the utility of synthetic data for model training. The input features used for these regression models are outlined in the last column of Table I, with a careful exclusion of time-related variables that could directly infer the actual arrival time at destination airports, ensuring that the models were trained on information that did not directly relate to the target variable. Model performance was quantified using three complementary metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R^2) [40]–[42]. These evaluation metrics were averaged across all models to provide a robust assessment of synthetic data quality. The comparative analysis of the TRTR and TSTR results offers valuable insights into the machine learning utility of synthetic data, as well as its potential as a replacement for historical data in downstream predictive tasks.

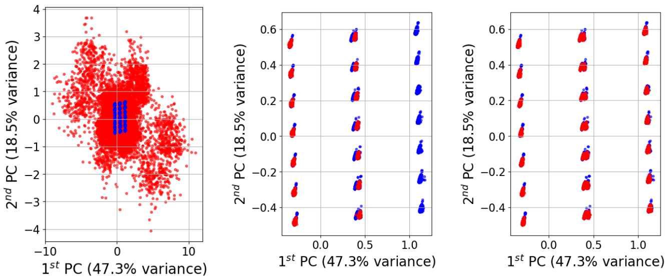
Due to the varying data sizes used for training TVAE (in Experiments 1, 2, and 3) and GC (in Experiments 4 and 5), along with the differences in the sizes of the sampled data (see Table III), we expect the PCA plots in Section IV-A to exhibit distinct clustering patterns. Likewise, we anticipate notable differences in the distribution plots presented in Section IV-B. Moreover, these discrepancies in data size are likely to impact the overall utility of the synthetic datasets for machine learning predictive tasks, underscoring the trade-offs between scalability and utility in generative modeling.

IV. Results

This section evaluates the quality of the synthetic flight information generated by the adapted Tabular Variational Autoencoder (TVAE) and Gaussian Copula (GC) models through the five experiments described in Section III-C. The assessment follows the evaluation framework detailed in Section III-D.

A. Diversity Assessment

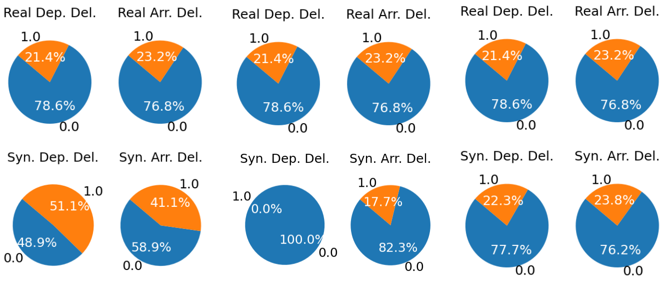
The performance of TVAE exhibited significant dependency on feature selection and data type representation across three experimental configurations: *"df_utc_ts"*, *"df_utc_d"* and *"df_utc_d_2"* in Experiments 1, 2, and 3, respectively. This limitation is likely attributed to poor latent space learning, a known issue in VAEs referred to as posterior collapse [43]–[47]. When temporal features were input in datetime format, the synthetic data failed to accurately replicate the statistical structure of the original data, as illustrated in Fig. 2a.



(a) Experiment 1 (b) Experiment 2 (c) Experiment 3

Fig. 2: *TVAE* - PCA analysis of real (blue) vs. synthetic (red) flight information.

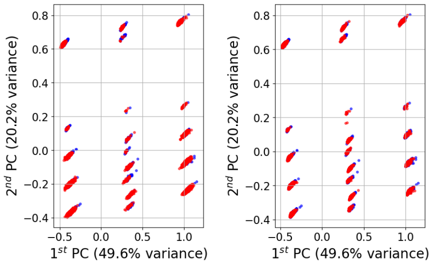
Replacing datetime features with numerical time duration values led to a partial improvement in capturing data variability. However, Fig. 2b indicates that certain clusters present in the real dataset remain absent in the synthetic data. This observation is further supported by Fig. 3b, which shows that the generated dataset contains no instances of delayed departures (0%). These findings suggest that the generative model failed to learn and reproduce departure-related patterns, with the missing clusters in the PCA representation likely corresponding to departure-related information. To address this limitation, we refined the feature set by transitioning from *"df_utc_d"* to *"df_utc_d_2"*, incorporating an additional temporal feature—"Actual Departure Time UTC"—at the departure airport. This modification aims to enhance the model's ability to capture departure-related patterns and improve the representation of delayed departures in the synthetic dataset. The impact of this adjustment is evident in Figs. 2c and 3c.



(a) Experiment 1 (b) Experiment 2 (c) Experiment 3

Fig. 3: *TVAE* - Class balance analysis of real (top) vs. synthetic (bottom) departure and arrival delay labels (1 = delayed, 0 = on time).

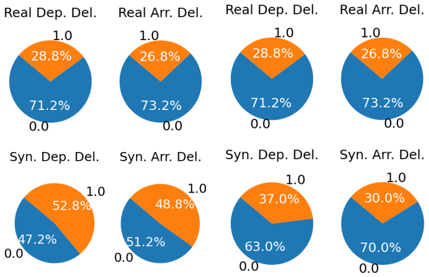
Another significant distinction between *"df_utc_d"* and *"df_utc_d_2"* lies in the inclusion of the "Taxi In Time (min)" feature. This feature can be calculated as the time difference between "Actual Arrival Time UTC" and "Wheels On Time UTC". However, while this feature depends on other time-related variables, it is also strongly influenced by the arrival airport's characteristics. Calculating it post-generation would only account for its temporal dependencies while largely neglecting its relationship with the arrival airport. Therefore, we explicitly incorporated it into *"df_utc_d_2"*, enabling the model to learn both the correlation between "Taxi In Time (min)" and the arrival airport, as well as its relationships with other temporal features.



(a) Experiment 4 (b) Experiment 5

Fig. 4: *GC* - PCA analysis of real (blue) vs. synthetic (red) flight information.

Unlike TVAE, the CG model demonstrated greater robustness to feature selection and data types across both experimental configurations: *"df_utc_ts"* in Experiment 4 and *"df_utc_d_2"* in Experiment 5. As shown in Fig. 4, the synthetic flight data generated by CG effectively preserved the full variability of the real dataset, even when temporal features were represented in datetime format, as in Experiment 4. Furthermore, when using *"df_utc_d_2"* as input, the GC-generated synthetic data exhibited a class distribution more closely matching that of the real data, as illustrated in Fig. 5b. The cluster patterns in Fig. 4 differ from those in Fig. 2, due to the different data sizes used with TVAE and GC.



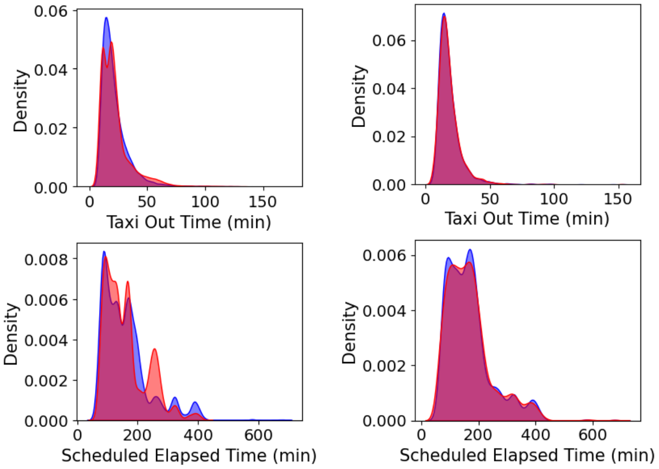
(a) Experiment 4 (b) Experiment 5

Fig. 5: *GC* - Class balance analysis of real (top) vs. synthetic (bottom) departure and arrival delay labels (1 = delayed, 0 = on time).

The diversity analysis demonstrated that using the *"df_utc_d_2"* DataFrame as input for both the TVAE and GC generative models resulted in improved diversity coverage and a class distribution more closely aligned with the real data. Consequently, the subsequent evaluation will focus exclusively on Experiments 3 and 5.

B. Statistical Assessment

Both TVAE and GC generated synthetic flight data that closely matched the real data's individual feature distributions and pairwise feature relationships. Fig. 6 illustrates this similarity by comparing the distributions of two features between real and synthetic data. The distributional differences observed between Fig. 6a and Fig. 6b stem from the varying training dataset sizes used for TVAE and GC, necessitated by computational constraints.



(a) Experiment 3 (b) Experiment 5

Fig. 6: Similarity of marginal distributions for real (blue) vs. synthetic (red) data.

This statistical similarity was quantitatively validated through the Total Variation Distance (TVD) and Kolmogorov-Smirnov test results, which showed minimal divergence between the real and synthetic datasets across both categorical and numerical features. The Correlation Similarity and Contingency Similarity metrics confirmed that both generative models effectively preserved inter-feature statistical dependencies, as detailed in Table IV.

TABLE IV: Quantitative evaluation of the statistical similarity between synthetic and real data.

Statistical similarity	Experiment 3	Experiment 5
Marginal distribution	87.53%	94.41%
Bivariate distributions	75.41%	83.7%
Average	81.47%	89.06%

The Gaussian Copula model demonstrated superior performance compared to TVAE, achieving higher accuracy in replicating the statistical properties of the real data. GC's enhanced capability was particularly evident in its reproduction of both univariate distributions and bivariate relationships, establishing it as the more effective model for replicating the statistical characteristics of the real flight data.

C. Fidelity Assessment

As anticipated from the previous assessment, which demonstrated approximately 90% similarity between the Gaussian Copula-generated synthetic data and real data, the seven selected classifiers showed reduced accuracy in distinguishing between real and GC-generated synthetic flight data in Experiment 5. Table V quantifies this performance, showing lower average accuracy (66.93%) and F1 score (54.72%) compared to Experiment 3, further validating the high quality of the GC-generated synthetic data.

TABLE V: Classification performance metrics for distinguishing real from synthetic flight data.

Discriminative score	**Experiment 3**	**Experiment 5**
Average Accuracy	78.39%	66.93%
Average F1 Score	74.41%	54.72%

D. Utility Assessment

The regression models trained on TVAE-generated synthetic data in Experiment 3 achieved comparable or superior accuracy when tested on real data (TSTR), with mean absolute errors around 11 minutes for arrival delay predictions, as shown in Table VI. However, despite producing statistically superior synthetic data that was harder to distinguish from real data, GC-generated data was less effective for training predictive models. This limitation arose from computational constraints that restricted GC's synthetic dataset to approximately 2,000 flights, compared to TVAE's 52,000 flights (Table III).

To illustrate the impact of this difference, assume the real dataset contains 500 unique flight routes, with flights evenly distributed among them. Under this assumption, TVAE-generated data would provide 104 flights per route, whereas GC-generated data would yield only 4 flights per route. This small sample size limits the model's ability to learn flight delay patterns and capture route-specific variability. While real-world distributions are more uneven, this simplified example highlights why the smaller GC dataset results in weaker predictive performance despite its stronger statistical resemblance to real data.

TABLE VI: Predictive performance of machine learning models trained on real vs. synthetic flight data for arrival delay prediction.

Predictive score	**Experiment 3**		**Experiment 5**	
	TRTR	**TSTR**	**TRTR**	**TSTR**
Average RMSE	15.50	14.72	12.89	20.11
Average MAE	11.50	11.06	9.48	14.66
Average R²	0.76	0.79	0.86	0.66

V. Discussion

An important observation is that assessing only the marginal distributions or the statistical similarity of individual features is insufficient. It is equally essential to visually and numerically examine the joint distributions between pairs of variables. For instance, unlike air time, the distances between airport pairs are not explicitly provided as direct inputs to the generative model; instead, they are inferred post-generation based on airport ID pairs. As a result, the generator can only establish relationships between air time and airport ID pairs, rather than directly modeling the distance. By analyzing the correlation between distance and air time in Fig. 7, it is evident that the synthetic data contains some incorrect air time values (highlighted in green circles) that are not proportional to the corresponding airport distances. To address this, the generation process should be refined to minimize the occurrence of these incorrect values, and any remaining inaccuracies should be filtered out during post-generation cleaning.

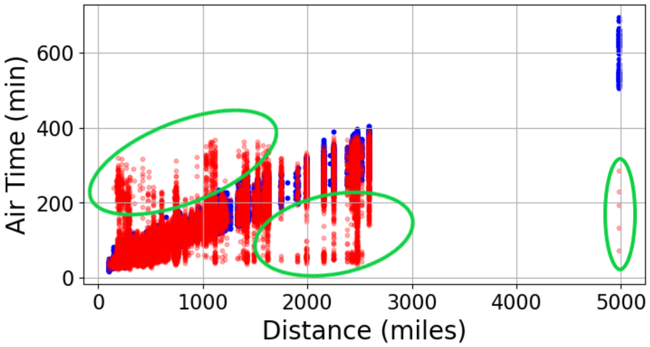


Fig. 7: Correlation between distance and air time for real (blue) vs. synthetic (red) data.

Some variation in the marginal distributions between real and synthetic data is acceptable and does not necessarily indicate an issue with the synthetic data. For instance, synthetic data might reflect a higher number of flights between airports X_1 and Y_1 (1000 miles apart) and fewer flights between airports X_2 and Y_2 (500 miles apart) compared to real data. This discrepancy would cause a shift in the peaks of the marginal distribution for the distance feature, showing fewer instances of 500-mile flights and more instances of 1000-mile flights. However, it is crucial that the bivariate distributions (i.e., the correlations between features) remain consistent between the real and synthetic data in order to maintain operational validity.

Another key takeaway is that although the Gaussian Copula (GC) model demonstrated higher statistical similarity and fidelity than the Tabular Variational Autoencoder (TVAE), it was less effective in terms of synthetic data utility. This underscores the importance of a multi-faceted evaluation framework. Furthermore, it highlights GC's limitations in handling large datasets and the critical role that dataset size plays in capturing and accurately predicting flight delay patterns.

VI. Conclusions & Future Work

In this study, we explored the use of generative models for producing realistic synthetic flight information and established a rigorous four-stage evaluation process to assess the statistical similarity, fidelity, diversity, and predictive utility of the synthetic data. While both TVAE and GC models demonstrated the ability to generate high-quality synthetic data, TVAE was sensitive to data types and feature selection, which affected its performance in certain cases. On the other hand, GC achieved higher statistical similarity and fidelity. However, GC's computational limitations restricted its application to larger datasets, ultimately affecting the utility of the GC-generated data for predictive modeling. In contrast, TVAE was capable of handling larger datasets efficiently and, once trained, enabled fast and scalable sampling of synthetic data, making it more practical for large-scale applications.

Despite these limitations, our findings indicate that synthetic data can be effectively used to train flight delay prediction models, achieving accuracy comparable to models trained on real data. This brings us one step closer to providing the aviation community with an abundant source of reliable synthetic flight data, adaptable to different operational scenarios.

This study lays the foundation for future research on synthetic data generation methods specifically tailored to address the unique challenges of air transport applications. Future work will focus on refining the generative process by leveraging increased computational power to test GC on larger datasets, similar to TVAE. Additionally, strategies to mitigate posterior collapse in TVAE will be explored, along with an assessment of their impact on TVAE's sensitivity to data types and feature selection. Hyperparameter tuning will also be investigated to optimize feature correlations and improve the operational correctness of the generated data. Moreover, in this analysis, rejection sampling was applied to filter out synthetic routes that did not exist in historical data. Future research will analyze these newly generated routes to assess their plausibility and potential insights. Finally, the scope of this work will be expanded by incorporating additional flight attributes, such as diversions and cancellations, to further enhance the applicability of synthetic flight data in air transportation research.

Acknowledgment

This paper is based on the work done in the SynthAIr project. SynthAIr has received funding from the SESAR Joint Undertaking under the European Union's Horizon Europe research and innovation programme under grant agreement No 101114847. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or SESAR 3 Joint Undertaking. Neither the European Union nor SESAR 3 Joint Undertaking can be held responsible for them.

References

[1] A. Figueira and B. Vaz, "Survey on synthetic data generation, evaluation methods and gans," *Mathematics*, vol. 10, no. 15, 2022. [Online]. Available: https://www.mdpi.com/2227-7390/10/15/2733

[2] N. Patki, R. Wedge, and K. Veeramachaneni, "The synthetic data vault," in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2016, pp. 399–410.

[3] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional gan," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/254ed7d2de3b23ab10936522dd547b78-Paper.pdf

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

[5] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. Ieee, 2008, pp. 1322–1328.

[6] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations (ICLR)*, 2013. [Online]. Available: https://arxiv.org/abs/1312.6114

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 27. Curran Associates, Inc., June 2014. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

[8] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim, "Data synthesis based on generative adversarial networks," *arXiv preprint arXiv:1806.03384*, 2018.

[9] L. Xu and K. Veeramachaneni, "Synthesizing tabular data using generative adversarial networks," *arXiv preprint arXiv:1811.11264*, 2018.

[10] E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, and J. Sun, "Generating multi-label discrete patient records using generative adversarial networks," in *Machine learning for healthcare conference*. PMLR, 2017, pp. 286–305.

[11] A. Srivastava, L. Valkov, C. Russell, M. U. Gutmann, and C. Sutton, "Veegan: Reducing mode collapse in gans using implicit variational learning," *Advances in neural information processing systems*, vol. 30, 2017.

[12] D. Liu, J. Zhang, J. Cui, S. X. Ng, R. G. Maunder, and L. H. Hanzo, "Deep-learning-aided packet routing in aeronautical ad hoc networks relying on real flight data: From single-objective to near-pareto multiobjective optimization," *IEEE Internet of Things Journal*, vol. 9, pp. 4598–4614, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:238673995

[13] Q. Zhang and J. H. Mott, "An exploratory assessment of llm's potential toward flight trajectory reconstruction analysis," *ArXiv*, vol. abs/2401.06204, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:266977542

[14] S. Wijnands, A. Sharpanskykh, and K. Aly, "Generation of synthetic aircraft landing trajectories using generative adversarial networks," 2024. [Online]. Available: https://zenodo.org/doi/10.5281/zenodo.14774664

[15] B. of Transportation Statistics, "Transtats database for airline on-time performance," 2023, accessed: 2025-01-23. [Online]. Available: https://transtats.bts.gov/Tables.asp?QO_VQ=EFD&QO_anzr=Nv4yv0r%FDb0-gvzr%FDcr4s14zn0pr%FDQn6n&QO_fu146_anzr=b0-gvzr

[16] U.S. Department of Transportation. Bureau of transportation statistics. [Online]. Available: https://www.bts.gov/

[17] L. Yang and A. Shami, "Towards autonomous cybersecurity: An intelligent automl framework for autonomous intrusion detection," in *AutonomousCyber@CCS*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:272423857

[18] Y. Shen, A. Sudjianto, R. ArunPrakash, A. Bhattacharyya, M. Rao, Y. Wang, J. Vaughan, and N. Zhou, "Towards a framework on tabular synthetic data generation: a minimalist approach: theory, use cases, and limitations," *ArXiv*, vol. abs/2411.10982, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:274131324

[19] M. Al-Shedivat, A. Dubey, and E. P. Xing, "The intriguing properties of model explanations," *arXiv preprint arXiv:1801.09808*, 2018.

[20] H. Akrami, S. Aydöre, R. M. Leahy, and A. A. Joshi, "Robust variational autoencoder for tabular data with beta divergence," *ArXiv*, vol. abs/2006.08204, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:219687586

[21] R. B. Nelsen, *An introduction to copulas*. Springer, 2006.

[22] H. Khosravi, S. Das, A. Al-Mamun, and I. Ahmed, "Binary gaussian copula synthesis: A novel data augmentation technique to advance ml-based clinical decision support systems for early prediction of dialysis

among ckd patients," *ArXiv*, vol. abs/2403.00965, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268230538

[23] H. J. Asghar, M. Ding, T. Rakotoarivelo, S. Mrabet, and M. A. Kâafar, "Differentially private release of high-dimensional datasets using the gaussian copula," *ArXiv*, vol. abs/1902.01499, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:59604403

[24] Y. Jiang, L. Mosquera, B. Jiang, L. Kong, and K. E. Emam, "Measuring re-identification risk using a synthetic estimator to enable data sharing," *PLoS ONE*, vol. 17, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:249748022

[25] S. D. Vault, "Copulas documentation," 2025, accessed: 2025-01-27. [Online]. Available: https://sdv.dev/Copulas/index.html

[26] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[27] J. Knoblauch and L. Vomfell, "Robust bayesian inference for discrete outcomes with the total variation distance," *ArXiv*, vol. abs/2010.13456, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:225066970

[28] T. Viehmann, "Numerically more stable computation of the p-values for the two-sample kolmogorov-smirnov test," *arXiv preprint arXiv:2102.08037*, 2021.

[29] S. Developers, "Correlationsimilarity," 2023, accessed: 2025-02-03. [Online]. Available: https://docs.sdv.dev/sdmetrics/metrics/metrics-glossary/correlationsimilarity

[30] ——, "Contingencysimilarity," 2023, accessed: 2025-02-03. [Online]. Available: https://docs.sdv.dev/sdmetrics/metrics/metrics-glossary/contingencysimilarity

[31] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[32] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[33] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.

[34] L. Breiman, *Classification and regression trees*. Routledge, 2017.

[35] D. D. Lewis, "Naive (bayes) at forty: The independence assumption in information retrieval," in *European conference on machine learning*. Springer, 1998, pp. 4–15.

[36] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 20, no. 2, pp. 215–232, 1958.

[37] S. learn Developers, "Sgdclassifier - scikit-learn 1.3.0 documentation," n.d., accessed: 2025-01-31. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html

[38] A. Ahmadi, S. S. Sharif, and Y. M. Banad, "A comparative study of sampling methods with cross-validation in the fedhome framework," *ArXiv*, vol. abs/2406.01950, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:270226221

[39] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information processing & management*, vol. 45, no. 4, pp. 427–437, 2009.

[40] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance," *Climate research*, vol. 30, no. 1, pp. 79–82, 2005.

[41] L. Mentch and G. Hooker, "Quantifying uncertainty in random forests via confidence intervals and hypothesis tests," *Journal of Machine Learning Research*, vol. 17, no. 26, pp. 1–41, 2016.

[42] A. C. Cameron and F. A. Windmeijer, "An r-squared measure of goodness of fit for some common nonlinear regression models," *Journal of econometrics*, vol. 77, no. 2, pp. 329–342, 1997.

[43] Y. Zhao, P. Yu, S. Mahapatra, Q. Su, and C. Chen, "Discretized bottleneck in vae: Posterior-collapse-free sequence-to-sequence learning," *ArXiv*, vol. abs/2004.10603, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:216056569

[44] G. G. Tong, C. A. S. Long, and D. E. Schiavazzi, "Invaert networks: A data-driven framework for model synthesis and identifiability analysis," *Computer Methods in Applied Mechanics and Engineering*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:261697481

[45] Y. Hao, P. Zhao, J. Fang, J. Qu, G. Liu, F. Zhuang, V. S. Sheng, and X. Zhou, "Meta-optimized joint generative and contrastive learning for sequential recommendation," *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pp. 705–718, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:264426259

[46] J. He, D. M. Spokoyny, G. Neubig, and T. Berg-Kirkpatrick, "Lagging inference networks and posterior collapse in variational autoencoders," *ArXiv*, vol. abs/1901.05534, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:58014132

[47] A. Kuzina and J. M. Tomczak, "Discouraging posterior collapse in hierarchical variational autoencoders using context," *arXiv preprint arXiv:2302.09976*, 2023.