**Proceedings of the ASME 2025**
**International Mechanical Engineering Congress and Exposition**
**IMECE2025**
**November 16–20, 2025, Memphis, TN**

# IMECE2025-165140

# GENERATIVE AI APPROACH FOR SYNTHETIC AVIATION FUELS DATA GENERATION

**Mohammed I. Radaideh [1,†], Majdi I. Radaideh [1], Angela Violi [1,*]**

[1]University of Michigan, Ann Arbor, MI

## ABSTRACT

*Data scarcity is a common problem related to aviation fuels. Many studies are still developing predictive models to predict aviation fuel properties using fuel composition to help in Sustainable Aviation Fuels (SAF) certification. However, due to the data scarcity, most of these models were trained using limited experimental data and are prone to overfitting. To this end, we utilize generative AI models as a data augmentation technique to address the data scarcity of aviation fuels. The tabular data of aviation fuels considered includes experimental data for fuel composition and nine thermophysical properties. This data is used to train ForestDiffusion, a recently developed diffusion generative AI model for tabular data generation without the need for many training samples. The performance of ForestDiffusion is compared with another simple model, Synthetic Minority Oversampling TEchnique (SMOTE), that is commonly used as a baseline. Different metrics are used to assess the quality of the synthetic data generated by ForestDiffusion and SMOTE. Results showed that the simple model SMOTE provided better synthetic data fidelity, diversity, and machine learning utility. Still, it has lower real data privacy protection than ForestDiffusion, making the latter a more desirable choice regardless of the advantages that SMOTE provided, and showing the importance of considering multiple assessment metrics to have more balanced performance conclusions.*

**Keywords: Generative AI, Diffusion Models, Forest Diffusion, Sustainable Aviation Fuels**

## NOMENCLATURE

***Roman letters***

| | |
|---|---|
| HOC | Heat of Combustion [MJ/kg] |
| $n$ | Number of |
| $Sp$ | Smoke Point [mm] |
| $T$ | Temperature [$^o$C] |

*Greek letters*

| | |
|---|---|
| $\alpha$ | Synthetic data fidelity metric |
| $\beta$ | Synthetic data diversity metric |
| $\nu$ | Kinematic viscosity [mm$^2$/s] |

***Superscripts and subscripts***

| | |
|---|---|
| 10 | 10% of the distillate is collected |
| 50 | 50% of the distillate is collected |
| 90 | 90% of the distillate is collected |
| $b$ | Boiling point |
| $f$ | Flash point |
| $fr$ | Freezing Point |
| $t$ | noise level |
| $noise$ | duplicates of original |

***Abbreviations***

| | |
|---|---|
| ATJ | Alcohol To Jet |
| ASTM | American Society for Testing and Materials |
| CFM | Conditional Flow Matching |
| DCR | Distance to Closet Records |
| FAA | Federal Aviation Administration |
| FT | Fischer Tropsch |
| GAN | Generative Adversarial Networks |
| GC×GC | 2-dimensional Gas Chromatography |
| HEFA | Hydroprocessed Esters and Fatty Acids |
| Gen-AI | Generative Artificial Intelligence |
| MAPE | Mean Absolute Percentage Error |
| ODE | Ordinary Differential Equations |
| RP | Rocket Propulsion |
| SAF | Sustainable Aviation Fuels |
| Sos | Score-based oversampling |
| SMOTE | Synthetic Minority Oversampling TEchnique |
| VAE | Variational Auto Encoders |
| XGBoost | Extreme Gradient Boosting |

## 1. INTRODUCTION

Sustainable Aviation Fuels (SAF) serve as an immediate alternative to conventional jet fuels to reduce the harmful emissions of conventional jet fuels, including greenhouse gases and soot [1]. SAF can be produced from different sustainable feedstocks

[†]Joint first authors
[*]Corresponding author: avioli@umich.edu
Documentation for `asmeconf.cls`: Version 1.41, September 11, 2025.

like biomass, fatty acids, oils, corn, sugars, etc., using certified technologies like Alcohol to Jet (ATJ) [2]. However, the produced SAF must be certified using the procedures identified by the American Society for Testing and Materials (ASTM). The first phase of the certification process includes reporting many thermophysical properties of the produced SAF, like flash point and distillation temperatures, and ensuring they meet the ASTM's identified constraints [3].

To this end, many predictive models were developed to help predict SAF properties using the fuel composition as an input to reduce the need for experiments to measure SAF properties. These models include linear regression models [4, 5], partial least squares [6, 7], and machine and deep learning models [8–11]. However, all the models were trained using limited data (fewer than 100 samples), and the models are prone to overfitting. Additionally, the relevant aviation fuels data are *tabular data* that include the fuel composition obtained by 2-dimensional Gas Chromatography (GC×GC) and fuel properties. GC×GC provides the weight of the many hydrocarbon classes in the aviation fuel. The number of features that GC×GC can provide can exceed 200, like the ones used in this study [8]. Accordingly, the predictive models are highly prone to overfitting when the number of features is significantly higher than the number of training samples [12].

The data ideally can be augmented by conducting more experiments to measure the fuel composition and properties, which are expensive and time-consuming. Alternatively, Generative Artificial Intelligence (Gen-AI) can create quality synthetic data that mimics the real data. A number of Gen-AI models were developed for tabular data generation. The recent models include Variational Auto Encoders (VAE) [13], Generative Adversarial Networks (GAN) [14], transformers and Large Language Models [15], and diffusion models [16].

Diffusion models have recently been gaining more attention for Gen-AI tasks because of their advantages over other generative models. Diffusion models are more stable than GAN models and capture the full data distribution; more details can be found in this review [17]. A number of diffusion models were developed for tabular data generation. Among these models are Score-based oversampling (Sos) [18], STaSy [19], Codi [20], TabDDPM [21], and recently Forest-Diffusion [22] and TabSyn [23].

In this study, we will utilize diffusion models to create synthetic aviation fuel data using real experimental data to overcome the common aviation fuel data scarcity issue. We will consider many fuel properties that ASTM requires, including kinematic viscosity, distillation temperatures (10%, 50%, 90%), boiling point, freezing point, flash point, heat of combustion, and smoke point. The quality of the synthetic data will be assessed and compared to the real data using high-order metrics that assess synthetic data fidelity, diversity, and privacy. As another performance metric, the synthetic data will be used to train a machine learning model and compare its performance with the same machine learning model trained on real data, where both will be tested on the same real data. This is crucial to develop more accurate, well-trained predictive tools using abundant, high-quality data.

Section 2 describes the dataset used in this study, the Gen-AI models used to generate synthetic data, and the assessment metrics of the synthetic data generated, Section 3 has the results for the four assessment metrics considered (fidelity, diversity, privacy, machine learning utility), and Section 4 contains the conclusions of this study.

## 2. DATASET AND METHODS

### 2.1. Dataset Description

The dataset includes the fuel composition for aviation fuels obtained by (GC×GC) as an input, and their properties as an output, both obtained from the Federal Aviation Administration (FAA) Alternative Jet Fuel Test Database (ASCENT) [24] and data published by [11]. The datasets include various types of aviation fuels, including conventional jet fuels (Jet A1, JP-5, JP-8), Sustainable Aviation Fuels (HEFA, ATJ, FT), Rocket Propulsion (RP) fuels, and blends (SAF/conventional and SAF/SAF).

The (GC×GC) data is used as an input and includes the volume fraction of hydrocarbon families, including 1. alkylbenzenes, 2. diaromatics, 3. cycloaromatics, 4. n-cycloalkanes, 5. di-cycloalkanes, 6. iso-alkanes, and 7. n-alkanes. The volume fraction of each family is distributed among classes classified based on the number of carbon atoms. For example, the volume of the diaromatics family is distributed among four classes: $C_{10}$ to $C_{14}$. Accordingly, the number of classes for the rest of the families is shown in Table 1. The volume fraction of each class is used as an input feature, and the total number of features is 64. Some families, like tri-cycloalkanes and alkenes (olefins), are dropped because their volume fractions for most of the fuels are negligible (less than $< 0.1\%$). We did this to reduce the number of features because the number of samples is limited, as shown in Table 2, and, hence, to minimize overfitting. Accordingly, the sum of the volume fractions will not sum to 100% (most sum to 99%), so we normalized the fraction of the 64 classes to sum to 100%.

**TABLE 1: THE HYDROCARBON FAMILIES AND CLASSES CONSIDERED IN THIS STUDY. EACH CLASS RESEMBLES AN INPUT FEATURE. THE TOTAL NUMBER OF FEATURES IS 64.**

| Family | Classes (# Features) |
|---|---|
| alkylbenzenes | $C_7$-$C_{16}$ (10) |
| diaromatics | $C_{10}$-$C_{14}$ (5) |
| cycloaromatics | $C_9$-$C_{15}$ (7) |
| n-cycloalkanes | $C_7$-$C_{17}$ (11) |
| di-cycloalkanes | $C_9$-$C_{16}$ (8) |
| iso-alkanes | $C_7$-$C_{18}$ (12) |
| n-alkanes | $C_7$-$C_{17}$ (11) |

Table 2 shows the properties considered in this study and the number of samples for each property. The samples of the first four properties ($\nu$, HOC, $T_f$, $T_{fr}$) were collected from both ASCENT [24] and [11], and include RP fuel samples, while the samples of the rest were collected from ASCENT. Each sample contains the volume fractions of the classes shown in Table 1 and the related property in Table 2.

**TABLE 2: PROPERTIES CONSIDERED IN THIS STUDY AND THE NUMBER OF SAMPLES FOR EACH PROPERTY.**

| Property | Number of Samples |
|---|---|
| Kinematic Viscosity ($\nu$) [mm$^2$/s] | 145 |
| Heat of Combustion (HOC) [MJ/kg] | 144 |
| Flash Point ($T_f$) [$^o$C] | 151 |
| Freezing Point ($T_{fr}$) [$^o$C] | 136 |
| Boiling Point ($T_b$) [$^o$C] | 47 |
| Smoke Point ($Sp$)[mm] | 33 |
| Distillation Temperature ($T_{10}$) [$^o$C] | 48 |
| Distillation Temperature ($T_{50}$) [$^o$C] | 48 |
| Distillation Temperature ($T_{90}$) [$^o$C] | 48 |

### 2.2. Diffusion Gen-AI Models

Diffusion Gen-AI Models are commonly used for image generation tasks and outperform VAE and GAN models. The studies [18–23] also demonstrated the superior performance of diffusion models over GAN and VAE models in tabular data generation. The diffusion model training consists of two processes: (1) Forward diffusion model and (2) reverse denoising process. In the forward process, the data is corrupted by gradually adding noise to the original samples. In contrast, in the reverse process, the diffusion model learns how to denoise the samples and generate synthetic but realistic data [17].

The majority of the diffusion models for tabular data generation [18–21, 23] are based on deep networks, which are data-hungry and require thousands of training data to perform well. However, the number of samples shown in Table 2 is very limited, and such models are likely to fail if trained on this limited data. Consequently, the *ForestDiffusion* model [22] that has been developed recently is based on Gradient Boosting algorithms, particularly, Extreme Gradient Boosting (XGBoost). Still, other boosting algorithms like LightGBM and CatBoost can be used. Such models (gradient boosting + diffusion) are less data-hungry compared with the deep neural networks and less likely to overfit [25, 26], making them more suitable for limited data applications compared with neural network models. The *ForestDiffusion* uses a diffusion process with Conditional Flow Matching (CFM). The CFM uses Ordinary Differential Equations (ODEs) to obtain the data distribution [27]. Neural networks ordinarily approximate the solution of ODEs, but in *ForestDiffusion*, XGBoost is used [22]. In this study, we will use *ForestDiffusion* to create synthetic aviation fuel data, and the quality of the synthetic data will be assessed using different metrics that will be defined next. Synthetic Minority Oversampling TEchnique (SMOTE) is a common baseline model used for comparison with other tabular data generative models [21, 23]. SMOTE was introduced to address the class imbalance in classification problems, where it creates synthetic data for the class with fewer samples [28]. Accordingly, it can also be used to generate synthetic tabular data by putting the real data (both input and output) in class 0 and creating **twice** the amount of random data and putting it in label 1, SMOTE will create synthetic data for class 0 to balance the random data in class 1. This simple interpolation model can quickly generate synthetic data without the need for powerful computing resources, and since it is an interpolative model, it is less sensitive to the data size. While SMOTE is not a true generative model, its interpolation-based approach provides a computationally cheap baseline for fidelity-focused tasks, albeit at the cost of diversity and privacy. It is commonly used as a baseline for performance comparison, and it can provide synthetic data metrics comparable to the more advanced Gen-AI models, as illustrated by [21] and [23].

### 2.3. Assessment Metrics

We will use four assessment metrics that are commonly used to assess the quality of the synthetic data; these metrics are [17]:

1. **Fidelity Metrics**: These metrics are used to assess the similarity between the synthetic data generated by the Gen-AI model and the real data used to train the model. $\alpha-$Precision is a high-order fidelity metric that indicates whether the synthetic data resembles the real one. This metric has the range [0,1], where higher values are desirable and indicate more similarities between the synthetic and real data.

2. **Diversity Metrics**: These metrics are used to assess to which extent the synthetic data could capture the distribution of the real data. $\beta$-Recall is a popular diversity metric that assesses the extent to which the synthetic data covers the distribution of the real data. Like $\alpha$-Precision, the values of $\beta$-Recall are in [0,1], where higher values are desirable and indicate higher diversity captured from the real data.

3. **Privacy Metrics**: The metric assesses whether the synthetic data is randomly sampled according to the distribution density and not copied from the training data. Distance to Closet Records (DCR) is a common metric to assess the privacy protection of real data, where *low* values (close to zero) of DCR are undesirable and indicate that the Gen-AI is copying from the real data and violating its privacy.

4. **Utility Metrics**: Also known as machine learning efficiency or utility. In this metric, both the real and synthetic data are used to train a machine learning model and test it on the same *real* data. As a result, the performance of the machine learning trained on the synthetic data should be *comparable* or even better than that trained on the real data.

## 3. RESULTS AND DISCUSSIONS

The two key hyperparameters in *Forestdiffusion* Python model are the number of noise levels ($n_t$), and the number of original data duplicates ($n_{noise}$). $n_t$ controls the level of noise added to the original data, while $n_{noise}$ is used because gradient boosting algorithms do not learn by batches like neural networks; they use the whole dataset. According to the developers, increasing both would give better performance, but this increase is accompanied by a computing time increase [22]. We set $n_t$ = 1,500 and $n_{noise}$ = 2,000 - significantly higher than default values ($n_t$ = 50, $n_{noise}$ = 100) - to compensate for limited training data and high feature dimensionality (64 inputs). Pilot tests

on viscosity confirmed these values improved machine learning utility (a regression model gave better performance when trained on synthetic data) by gradually increasing both parameters over defaults, consistent with [22] observation that larger $n_t/n_{noise}$ enhance sample quality at the cost of linear compute scaling.

It should be noted that testing different values using a grid search will be very time-consuming for generative models, considering that we have nine aviation fuel properties shown in Table 2. 80% of the data were used to train *ForestDiffusion* model, which generates the same number of synthetic samples. The remaining 20% of the data are used for the machine learning efficiency test. The computing time for forestdiffusion varies depending on the number of samples, but the computing time ranges from 1 hour to 3 hours for each property shown in Table 2 using 300/384 cores of two Dual-Thread AMD EPYC 9654 96-Core Processors.

The input features are volume fractions of different hydrocarbon classes, which sum to 100% as mentioned in Section 2.1. While *Forestdiffusion*'s synthetic data did not strictly enforce this constraint; the pre-normalization sums deviated by 6% or less on average, indicating near-preservation of compositional structure. Post-hoc normalization was thus applied to align synthetic data with real-data characteristics, with negligible impact on metrics (see Section 3.1).

### 3.1. Fidelity, Diversity, and Privacy Metrics

The synthetic data fidelity is assessed using the metric $\alpha$-Precision, diversity using $\beta$-Recall, and privacy using DCR. The metrics were calculated using 80% of the number of samples given in Table 2 and the same number of synthetic data. The $\alpha$-Precision and $\beta$-Recall can be obtained using the Python package Synthcity [29], while we followed the method used by [21] for DCR.

Table 3 shows the three metrics to assess the synthetic data generated by ForestDiffusion. For $\alpha$-Precision, the values are comparable for most properties, except for the smoke point ($Sp$), which has the lowest metric. The values are comparable because the input features are the same (GC×GC) and the only difference is the output. The synthetic data of Freezing point $T_{fr}$ has the best fidelity score, while the worst one belongs to the smoke point ($Sp$) synthetic data, as mentioned. The smoke point is a height in mm and a crucial parameter to calculate the soot formation rate [30]. The number of training samples is the lowest for $Sp$, but this cannot explain the low fidelity score, because the diversity ($\beta$-Recall) and privacy (DCR) metrics are the highest for $Sp$ synthetic data. Figure 1 shows the probability density plot of real smoke point data and the synthetic data generated by ForestDiffusion. The mean and standard deviation of the synthetic data (29.67±6.34) are comparable to those of real data (30.32±8.60), which explains the high diversity metric. However, the difference in the shapes of the density plot is notable and this is reflected on the low fidelity metric. While the results follow the known trade-off between fidelity and diversity, where the increase in one is accompanied by the other [31], insufficient training samples can lead to misleading or overly optimistic metrics, where more samples should be included for more robust conclusions. This is related to the definition of precision and recall metrics,

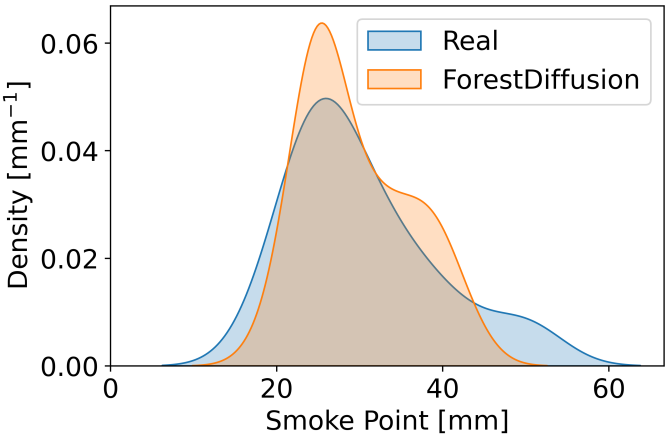


**FIGURE 1: THE PROBABILITY DENSITY PLOT OF REAL SMOKE POINT DATA (REAL) AND SYNTHETIC DATA GENERATED BY FORESTDIFFUSION.**

where the number of samples is a factor as illustrated by [31]. The $Sp$ synthetic data has the best diversity and privacy metrics; the worst diversity metric belongs to the distillation temperature $T_{10}$ (0.667). In contrast, the worst privacy metric belongs to the synthetic data of the freezing point $T_{fr}$ (0.491).

**TABLE 3: THE METRICS TO ASSESS SYNTHETIC DATA FIDELITY ($\alpha$-PRECISION), DIVERSITY ($\beta$-RECALL), AND PRIVACY (DCR) GENERATED BY *FORESTDIFFUSION*.**

| Property | $\alpha$-Precision | $\beta$-Recall | DCR |
|---|---|---|---|
| $\nu$ | 0.664 | 0.699 | 0.599 |
| HOC | 0.666 | 0.732 | 0.538 |
| $T_f$ | 0.704 | 0.705 | 0.596 |
| $T_{fr}$ | 0.735 | 0.722 | 0.491 |
| $T_b$ | 0.694 | 0.689 | 0.682 |
| $Sp$ | 0.555 | 0.826 | 0.718 |
| $T_{10}$ | 0.659 | 0.667 | 0.522 |
| $T_{50}$ | 0.670 | 0.688 | 0.589 |
| $T_{90}$ | 0.646 | 0.712 | 0.553 |

Table 4 shows the three metrics of fidelity, diversity, and privacy for synthetic data generated by SMOTE. Compared with Table 3, SMOTE can provide comparable and even better fidelity and diversity metrics than ForestDiffusion, and SMOTE can provide synthetic data within a few seconds. However, the common drawback of the SMOTE model is the low privacy protection of real data, as can be seen from the significantly lower DCR values for most of the properties compared with ForestDiffusion, as shown in Table 4. As a result, SMOTE is copying from the real data more than sampling from a learned distribution, where the usage of diffusion and other Gen-AI models becomes justifiable. The low privacy protection of SMOTE was also noticed by [21, 23]. The exception is the smoke point, but this high DCR value is over-optimistic and likely due to the small sample size. This shows the importance of considering multiple metrics to assess the quality of synthetic data and justifies the need to develop Gen-AI models for tabular data rather than using simple

models like SMOTE, which preserves local fidelity but leaks real data and violates its privacy. ForestDiffusion, on the other hand, samples globally but may sacrifice fidelity for privacy.

**TABLE 4: THE METRICS TO ASSESS SYNTHETIC DATA FIDELITY ($\alpha$-PRECISION), DIVERSITY ($\beta$-RECALL), AND PRIVACY (DCR) GENERATED BY *SMOTE*.**

| Property | $\alpha$-Precision | $\beta$-Recall | DCR |
|---|---|---|---|
| $\nu$ | 0.830 | 0.796 | 0.357 |
| HOC | 0.751 | 0.806 | 0.326 |
| $T_f$ | 0.704 | 0.785 | 0.300 |
| $T_{fr}$ | 0.729 | 0.777 | 0.360 |
| $T_b$ | 0.650 | 0.715 | 0.433 |
| $Sp$ | 0.642 | 0.766 | 0.692 |
| $T_{10}$ | 0.583 | 0.711 | 0.494 |
| $T_{50}$ | 0.524 | 0.671 | 0.404 |
| $T_{90}$ | 0.552 | 0.744 | 0.365 |

### 3.2. Machine Learning Utility (Efficiency)

Machine learning utility is the most important metric for developing property prediction models. As mentioned, a machine learning model is trained using real and synthetic data, where both are tested using the same real data. Accordingly, we trained a CatBoost regression model [32], a gradient boosting model commonly used for machine learning efficiency tests [26]. Hyperparameter tuning is performed using grid search on two catboost parameters: number of iterations (trees) and the learning rate. The values considered for the number of iterations are {50, 100, 200, 500, 1000, 2000, 5000}, while learning rate values are {0.01, 0.1, 0.15, 0.2}. 80% of the real data is used to train catboost (which were used before to train ForestDiffusion and SMOTE). This is the "Real" case shown in the second column of Table 5. For the "ForestDiffusion" and "SMOTE" columns, catboost is trained using the same amount of synthetic data generated by these two models. The predictions for all three cases were evaluated using 20% of the real data, and the metric used is Mean Percentage Absolute Error (MAPE), where lower error is desirable. Both catboost regression and grid search hyperparameter tuning were performed using the scikit-learn package in Python 3.11.

Table 5 shows the MAPE for the three cases, and the numbers in bold are the cases where the synthetic data is better than the real one. For all cases except $Sp$, catboost trained by SMOTE synthetic data has better metrics than ForestDiffusion and even better metrics than the real case in predicting Heat of Combustion (HOC), freezing point ($T_{fr}$), and distillation temperature $T_{10}$. The comparable performance to the Real case is expected since SMOTE is more copying from the real data as indicated by the low privacy metrics shown in Table 4. It can be seen that considering machine learning efficiency alone can be misleading, as the numbers in Table 5 show that using the diffusion models in general is unjustifiable for this case, and the simple model SMOTE is enough. ForestDiffusion provides better metrics than the Real case for the smoke point ($Sp$) only, but comparable performance with a slight increase in MAPE between 0.05% - 3% for all other

**TABLE 5: MAPE OF THE CATBOOST REGRESSION MODEL TRAINED USING REAL AND SYNTHETIC DATA GENERATED BY FORESTDIFFUSION AND SMOTE. THE NUMBERS IN BOLD INDICATE THE CASES WHERE THE SYNTHETIC DATA IS BETTER THAN THE REAL ONE.**

| Property | Real | ForestDiffusion | SMOTE |
|---|---|---|---|
| $\nu$ | 7.068 | 9.763 | 8.667 |
| HOC | 0.272 | 0.321 | **0.263** |
| $T_f$ | 4.805 | 6.198 | 5.919 |
| $T_{fr}$ | 26.793 | 29.131 | **22.336** |
| $T_b$ | 3.839 | 5.954 | **2.542** |
| $Sp$ | 13.545 | **10.822** | 14.031 |
| $T_{10}$ | 4.990 | 6.796 | **4.773** |
| $T_{50}$ | 6.508 | 7.301 | 6.924 |
| $T_{90}$ | 6.853 | 8.294 | 7.392 |

properties, which is a very good sign of the performance of the Gen-AI model. The better performance of SMOTE than "Real" on multiple properties is possibly due to the linear interpolation done by SMOTE between the real data samples, reducing the impact of the noise inherent in the real data [33], but again cannot be used as a real generative model due to the data redundancy.

For cases like distillation temperature ($T_{10}$), the fidelity score of data generated by SMOTE (0.583) is considerably lower than that of kinematic viscosity ($\nu$, 0.830) as shown in Table 4. However, CatBoost trained by data generated by SMOTE gave better performance than the real data for $T_{10}$ while the synthetic data for $\nu$ did not, as shown in Table 5. This highlights that high synthetic data fidelity or similarity to the real data is not always desirable in the context of tabular data generation, especially in our case, where machine learning utility is the most crucial aspect to develop more reliable predictive models. More distributed synthetic data (higher diversity metrics) compared with real data will help the machine learning model to generalize better. The high fidelity score can be desirable in image generation applications, where synthetic images very similar to the real ones are desirable.

The major limitation of this work is the limited data used, which prevents the training of other diffusion models and Gen-AI models, including GAN and others. Such models require a massive training sample size to perform well and may provide metrics better than those shown in Table 3 for ForestDiffusion.

## 4. CONCLUSION

In this work, two Gen-AI models were used to generate synthetic data to address the aviation fuels data scarcity problem. ForestDiffusion provided a more balanced quality of synthetic data compared with SMOTE, where ForestDiffusion provided better real data privacy protection. SMOTE performed better in fidelity, diversity, and machine learning utility measures, but had low privacy protection for the real data. This highlights the need to consider multiple metrics to assess the quality of the synthetic data. For future work, we will work on a pipeline based on linear blending rules to provide thousands of training data points to train other diffusion models based on neural networks, with the goal of getting better synthetic data quality for aviation fuels. The

blending rules are used to predict aviation fuel properties using the fuel composition obtained by GC×GC as an input. We will calibrate these rules using experimental data and then use them to predict properties for thousands of random fuel compositions. Then, we will use this data to train diffusion models with the hope that they learn to generate realistic data that matches the experimental one and provide better metrics.

## ACKNOWLEDGMENTS

This work was partially supported by (1) the Michigan Institute for Computational Discovery and Engineering (MICDE) Research Scholar Program, (2) the Michigan Institute for Data and AI in Society (MIDAS) Propelling Original Data Science (PODS) grant number: MIDASPODS2448, and (3) Michigan Transport Research And Commercialization (MTRAC) and Advanced Transportation Innovation Hub. This research made use of Idaho National Laboratory's High-Performance Computing systems located at the Collaborative Computing Center and supported by the Office of Nuclear Energy of the U.S. Department of Energy and the Nuclear Science User Facilities under Contract No. DE-AC07-05ID14517.

## REFERENCES

[1] Wandelt, Sebastian, Zhang, Yahua and Sun, Xiaoqian. "Sustainable aviation fuels: A meta-review of surveys and key challenges." *Journal of the Air Transport Research Society* (2025): p. 100056.

[2] Undavalli, Vamsikrishna, Olatunde, Olanrewaju Bilikis Gbadamosi, Boylu, Rahim, Wei, Chuming, Haeker, Josh, Hamilton, Jerry and Khandelwal, Bhupendra. "Recent advancements in sustainable aviation fuels." *Progress in Aerospace Sciences* Vol. 136 (2023): p. 100876.

[3] "Standard Practice for Evaluation of New Aviation Turbine Fuels and Fuel Additives (D4054-22)." *ASTM International* (2022).

[4] Boddapati, Vivek, Ferris, Alison M and Hanson, Ronald K. "Predicting the physical and chemical properties of sustainable aviation fuels using elastic-net-regularized linear models based on extended-wavelength FTIR spectra." *Fuel* Vol. 356 (2024): p. 129557.

[5] Boehm, Randall C, Yang, Zhibin and Heyne, Joshua S. "Threshold sooting index of sustainable aviation fuel candidates from composition input alone: progress toward uncertainty quantification." *Energy & Fuels* Vol. 36 No. 4 (2022): pp. 1916–1928.

[6] Shi, Xiangpeng, Li, Haijing, Song, Zhaoyu, Zhang, Xiangwen and Liu, Guozhu. "Quantitative composition-property relationship of aviation hydrocarbon fuel based on comprehensive two-dimensional gas chromatography with mass spectrometry and flame ionization detector." *Fuel* Vol. 200 (2017): pp. 395–406.

[7] Berrier, Kelsey L, Freye, Chris E, Billingsley, Matthew C and Synovec, Robert E. "Predictive modeling of aerospace fuel properties using comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry and partial least squares analysis." *Energy & Fuels* Vol. 34 No. 4 (2020): pp. 4084–4094.

[8] Hall, Clemens, Rauch, Bastian, Bauder, Uwe and Aigner, Manfred. "Comparison of probabilistic jet fuel property models for the fuel screening and design." *Fuel* Vol. 351 (2023): p. 128965.

[9] Oh, Ji-Hun, Oldani, Anna, Solecki, Alex and Lee, Tonghun. "Learning to predict sustainable aviation fuel properties: A deep uncertainty quantification viewpoint." *Fuel* Vol. 356 (2024): p. 129508.

[10] Liu, Ziyu and Yang, Xiaoyi. "Insight of low flammability limit on sustainable aviation fuel blend and prediction by ANN model." *Energy and AI* Vol. 18 (2024): p. 100423.

[11] Shao, Yitong, Yu, Mengxian, Zhao, Mengchao, Xue, Kang, Zhang, Xiangwen, Zou, Ji-Jun and Pan, Lun. "Comprehensive accurate prediction of critical jet fuel properties with multiple machine learning models." *Chemical Engineering Science* Vol. 304 (2025): p. 121018.

[12] Subramanian, Jyothi and Simon, Richard. "Overfitting in prediction models–is it a problem only in high dimensions?" *Contemporary clinical trials* Vol. 36 No. 2 (2013): pp. 636–641.

[13] Kingma, Diederik P, Welling, Max et al. "An introduction to variational autoencoders." *Foundations and Trends® in Machine Learning* Vol. 12 No. 4 (2019): pp. 307–392.

[14] Goodfellow, Ian J, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron and Bengio, Yoshua. "Generative adversarial nets." *Advances in neural information processing systems* Vol. 27 (2014).

[15] Borisov, Vadim, Seßler, Kathrin, Leemann, Tobias, Pawelczyk, Martin and Kasneci, Gjergji. "Language models are realistic tabular data generators." *arXiv preprint arXiv:2210.06280* (2022).

[16] Sohl-Dickstein, Jascha, Weiss, Eric, Maheswaranathan, Niru and Ganguli, Surya. "Deep unsupervised learning using nonequilibrium thermodynamics." *International conference on machine learning*: pp. 2256–2265. 2015. pmlr.

[17] Li, Zhong, Huang, Qi, Yang, Lincen, Shi, Jiayang, Yang, Zhao, van Stein, Niki, Bäck, Thomas and van Leeuwen, Matthijs. "Diffusion Models for Tabular Data: Challenges, Current Progress, and Future Directions." *arXiv preprint arXiv:2502.17119* (2025).

[18] Kim, Jayoung, Lee, Chaejeong, Shin, Yehjin, Park, Sewon, Kim, Minjung, Park, Noseong and Cho, Jihoon. "Sos: Score-based oversampling for tabular data." *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*: pp. 762–772. 2022.

[19] Kim, Jayoung, Lee, Chaejeong and Park, Noseong. "Stasy: Score-based tabular data synthesis." *arXiv preprint arXiv:2210.04018* (2022).

[20] Lee, Chaejeong, Kim, Jayoung and Park, Noseong. "Codi: Co-evolving contrastive diffusion models for mixed-type tabular synthesis." *International Conference on Machine Learning*: pp. 18940–18956. 2023. PMLR.

[21] Kotelnikov, Akim, Baranchuk, Dmitry, Rubachev, Ivan and Babenko, Artem. "Tabddpm: Modelling tabular data with diffusion models." *International Conference on Machine Learning*: pp. 17564–17579. 2023. PMLR.

 

[22] Jolicoeur-Martineau, Alexia, Fatras, Kilian and Kachman, Tal. "Generating and imputing tabular data via diffusion and flow-based gradient-boosted trees." *International Conference on Artificial Intelligence and Statistics*: pp. 1288–1296. 2024. PMLR.

[23] Zhang, Hengrui, Zhang, Jiani, Srinivasan, Balasubramaniam, Shen, Zhengyuan, Qin, Xiao, Faloutsos, Christos, Rangwala, Huzefa and Karypis, George. "Mixed-type tabular data synthesis with score-based diffusion in latent space." *arXiv preprint arXiv:2310.09656* (2023).

[24] Oh, Jihun, Oldani, Anna, Lee, Tonghun and Shafer, Linda. "Deep learning algorithms for assessing sustainable jet fuels from two-dimensional gas chromatography." *AIAA Science and Technology Forum and Exposition, AIAA SciTech Forum 2022*. 2022. American Institute of Aeronautics and Astronautics Inc, AIAA.

[25] McCluskey, William J, Daud, Dzurllkanian Zulkarnain and Kamarudin, Norhaya. "Boosted regression trees: An application for the mass appraisal of residential property in Malaysia." *Journal of Financial Management of Property and Construction* Vol. 19 No. 2 (2014): pp. 152–167.

[26] Hollmann, Noah, Müller, Samuel, Purucker, Lennart, Krishnakumar, Arjun, Körfer, Max, Hoo, Shi Bin, Schirrmeister, Robin Tibor and Hutter, Frank. "Accurate predictions on small data with a tabular foundation model." *Nature* Vol. 637 No. 8045 (2025): pp. 319–326.

[27] Tong, Alexander, Malkin, Nikolay, Huguet, Guillaume, Zhang, Yanlei, Rector-Brooks, Jarrid, Fatras, Kilian, Wolf, Guy and Bengio, Yoshua. "Conditional flow matching: Simulation-free dynamic optimal transport." *arXiv preprint arXiv:2302.00482* Vol. 2 No. 3 (2023).

[28] Chawla, Nitesh V, Bowyer, Kevin W, Hall, Lawrence O and Kegelmeyer, W Philip. "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research* Vol. 16 (2002): pp. 321–357.

[29] Qian, Zhaozhi, Cebere, Bogdan-Constantin and van der Schaar, Mihaela. "Synthcity: facilitating innovative use cases of synthetic data in different data modalities." *arXiv preprint arXiv:2301.07573* (2023).

[30] Chen, Zhibin, Wen, Jennifer, Xu, Baopeng and Dembele, Siaka. "Extension of the eddy dissipation concept and smoke point soot model to the LES frame for fire simulations." *Fire safety journal* Vol. 64 (2014): pp. 12–26.

[31] Naeem, Muhammad Ferjad, Oh, Seong Joon, Uh, Youngjung, Choi, Yunjey and Yoo, Jaejun. "Reliable fidelity and diversity metrics for generative models." *International conference on machine learning*: pp. 7176–7185. 2020. PMLR.

[32] Prokhorenkova, Liudmila, Gusev, Gleb, Vorobev, Aleksandr, Dorogush, Anna Veronika and Gulin, Andrey. "CatBoost: unbiased boosting with categorical features." *Advances in neural information processing systems* Vol. 31 (2018).

[33] Xu, Pengfei and Jia, Yinjie. "SNR improvement based on piecewise linear interpolation." *J. Electr. Eng* Vol. 72 No. 5 (2021): pp. 348–351.