cortAIx

FACULTY

THALES
Building a future we can all trust

Evaluating Synthetic Data's Effectiveness:

Bridging the gap between synthetic and real data and the feasibility of achieving high assurance

WHITEPAPER

Contents

THIS DOCUMENT IS THE PROPERTY OF THALES UK LIMITED

© Thales UK Limited 2025

The copyright herein is the property of Thales UK Limited. It is not to be reproduced, modified, adapted, published, translated in any material form (including storage in any medium by electronic means whether or not transiently or incidentally), in whole or in part nor disclosed to any third party without the prior written permission of Thales UK Limited. Neither shall it be used otherwise than for the purpose for which it is supplied.

The information contained herein is confidential and, subject to any rights of third parties, is proprietary to Thales UK Limited. It is intended only for the authorised recipient for the intended purpose, and access to it by any other person is unauthorised. The information contained herein may not be disclosed to any third party or used for any other purpose without the express written permission of Thales UK Limited.

If you are not the intended recipient, you must not disclose, copy, circulate or in any other way use or rely on the information contained in this document. Such unauthorised use may be unlawful. If you have received this document in error, please inform Thales UK Limited immediately on +44 (0)118 9434500 and post it, together with your name and address, to The Intellectual Property Department, Thales UK Limited, 350 Longwater Avenue, Green Park, Reading RG2 6GF. Postage will be refunded.

01 - EXECUTIVE SUMMARY

In the era of AI development and deployment to enhance, augment and replace conventional software solutions and algorithms, the importance of data has never been greater.

In defence there is a significant challenge in exposing, analysing, and curating suitable data sets due to:

- The sensitivity of the data in the operational context
- The platform's (where the data is generated) ability to persist and return the data to the OEMs
- Challenges around data ownership and storage.

With this backdrop the rapid advancement of synthetic data generation holds significant promise for defence applications, enabling secure, scalable, and privacy-preserving data solutions that can accelerate the deployment of artificial intelligence (AI) across critical missions. However, operationalising synthetic data at scale demands robust assurance methodologies that deliver defensible evidence. These methodologies must enable statistical quantification demonstrating that synthetic data accurately represents the characteristics of real data and matches its quality. Ultimately, this involves presenting performance evaluations that compare synthetic and real data, validating their interchangeability for use in equivalent AI models.

This white paper highlights the pressing need for rigorous data assurance methodologies, arguing that the assurance of AI model generated using synthetic data utilised within defence environments cannot be achieved without first establishing comprehensive synthetic data assurance. Our analysis reveals that while synthetic data is already reshaping the landscape, current approaches to its validation are inconsistent, often lacking statistical rigour and failing to account for the unique requirements of diverse defence datasets. Furthermore, the paper underscores that data assurance protocols effective for one dataset or purpose may not transfer seamlessly to another, reflecting the complexity inherent in military data environments.

In response to these challenges, we present a structured methodology for the evaluation of synthetically generated data. This approach is designed to provide robust, fit-for-purpose assessments that are tailored both to the specific synthetic data generation pathway and the intended operational use case. Through this, we aim to bridge the assurance gap and provide a foundation for both responsible experimentation and future at-scale deployment.

We conclude that, until a universally accepted and rigorous method for synthetic data assurance is established, its widespread operational use in defence will remain constrained. This paper sets out both our latest findings and a proposed pathway for ongoing research and collaborative development, with the goal of supporting the defence chain's safe and effective adoption of synthetic data.

02 - INTRODUCTION

The effectiveness of defence AI systems is constrained by the data on which they are trained.[1] In Defence, data is more than just a technical necessity, it is a strategic asset, that encodes tactics, doctrine, and force structure. Consequently, capturing and sharing real-world data carries inherent security risks. Furthermore, many operational environments are denied or inaccessible, making it technically or ethically challenging to record the data at scale. While commercial open-source datasets may offer some utility, they are susceptible to adversarial poisoning in ways that are notoriously difficult to detect.

Synthetic data provides a method to address these limitations. It improves data access, by broadening the range of data that exists. By generating high-fidelity, representative data within controlled environments, these systems can be modelled without risking operational security. Beyond simply filling gaps, synthetic data enables data to be collected from environments that typically are difficult to access and translate existing datasets to an unfamiliar environment, e.g. data collected in the UK can be simulated as being in snow or a desert.

Synthetic data fundamentally shifts the statistical quality of the training pipeline. Real world data is naturally skewed towards business-as-usual scenarios. While these are common, a model trained exclusively on routine data will lack generalisability. Synthetic data allows us to intentionally broaden the training distribution by injecting rare high consequence edge cases, such as a critical system failures or complex adversary deception techniques. By oversampling these low-probability, high-impact events, we produce robust models that are resilient to the volatility of live operations. This rapid iteration capability ensures that when a threat is identified we can immediately simulate it, rather than waiting months for field data. This allows the new tactics required to counter that threat to be integrated more rapidly than would otherwise be possible.

However, there are significant limitations to the use of synthetic data. Generation of accurate synthetic data is challenging, often requiring as much effort as model training. Then validating and assuring the generated data is, if anything, even more difficult. What, to a human, appears to be superior quality data, is not necessarily data that provides any benefit to algorithm or AI model training. As such, acquiring synthetic data can be expensive, and once acquired there is no guarantee that the data is suitable.

[1] *Possibilities and Challenges for Artificial Intelligence in Military Applications* - Dr Peter Svenmarck, Dr Linus Luotsinen, Dr Mattias Nilsson, Dr Johan Schubert

To begin to address this problem, Thales teams have attempted to map out an assurance approach that focuses specifically on validating synthetic data outputs to determine if they provide sufficient statistical variances to train a machine learning model capable of making predictions in high stakes situations. These models typically have an elevated level of assurance required, which extends to the underlying data also as well as the model outputs.

2.1 - Background

This paper's primary objective is to provide a proof-of-concept (POC) assurance approach for validating synthetic datasets against real-world benchmarks. Our approach lays the foundation for vital assurance measures, including distribution bias, distribution similarity, and correlation matrices.

The project utilises acoustic data, derived from a binary classification activity focused on detecting whether a specific item is present. Because the prevalence of positive examples in this dataset is low and obtaining further real-world samples is difficult, the use of synthetic data is being investigated to augment algorithmic capabilities.

The challenge of generating good synthetic data is amplified by the inherently stochastic nature of acoustic. Echoes, overlapping sound paths, clutter, growth on equipment and system movements introduce complex, non-stationary noise patterns that are difficult to model synthetically. Unlike visually "clean" renders, real objects must be detected and classified against a backdrop of shifting sediment types, variable sound-speed profiles, changing grazing angles, and transient objects, all of which distort target signatures in ways that are neither purely random nor fully deterministic. If synthetic data fails to capture this full variability, models risk learning brittle and rigid decision boundaries that perform well on idealised scenes but break down in real world operations. Ultimately, small fluctuations in the field can be the difference between confidently detecting an object and missing it entirely.

2.1.1 The Operational Objective

While synthetic data is widely used in optical computer vision (such as self-driving cars), replicating acoustic data presents a significantly higher technical barrier. Unlike light, which travels in straight lines with predictable reflections, sound is governed by a highly dynamic and "messy" physics model. Datasets often contain severe class imbalance, with a critical shortage of "Positive" samples (targets) compared to "Negative" samples (i.e. clutter and debris).

In the context of the PoC work, the "cost of failure" is asymmetric. The downstream classification model is optimised for recall - the ability to identify every potential target. While a high false-alarm rate creates an operational burden, a single missed target is catastrophic. Therefore, a key element of the PoC was to determine if the provided synthetic data could harden a model's detection capabilities without introducing "blind spots" caused by the synthetic-to-real domain gap.

03 - DEFINING ASSURANCE

3.1 - Good Data

Defining "good" data is the fundamental challenge of data science. If a researcher could routinely identify which specific data would improve a model, most problems would already be solved. Fundamentally, good data must encompass the entire distribution of real-world inputs. For imagery, this means capturing the full spectrum of lighting, weather, and sensor types the system will encounter. Because there are no automated metrics for "what a system might encounter," assessing data quality often requires measuring second-order effects rather than the raw data itself.

Throughout this report the quality of data is specifically being assessed with respect to model training and validation. This means that a useful second order effect to measure is whether model performance varies based on which dataset is being used. This measures a slightly different parameter for the data however, which is whether the entire distribution of the real dataset is captured inside the synthetic one. This assessment therefore cannot be used instead of the usual data quality checks, but should complement it, for situations where the is insufficient real data across all conditions, and synthetic data is being used to enrich the dataset.

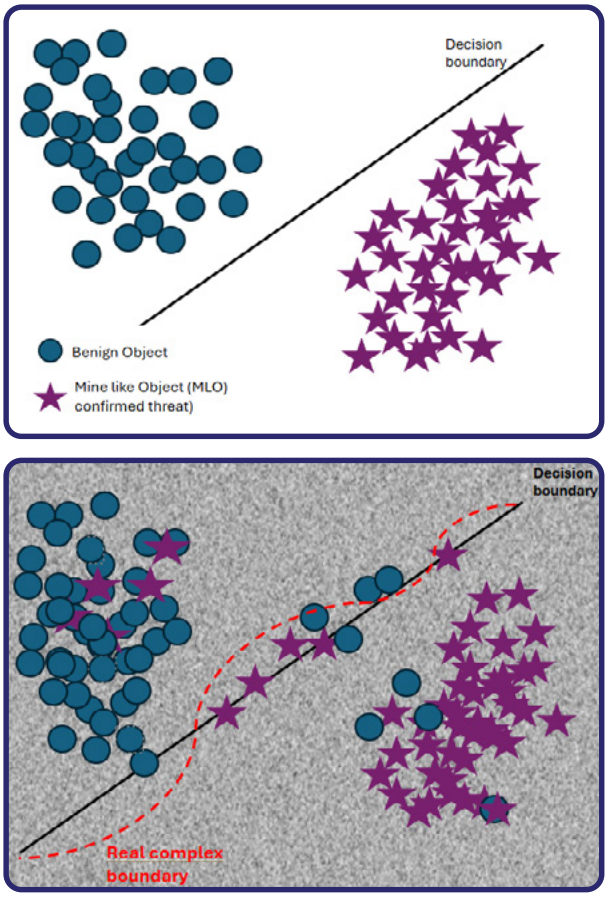

FIGURE 1 - Illustration of Decision Boundary

3.1.1 The Decision Boundary Problem

In life-critical settings, "good-looking" data is insufficient because models do not perceive images as humans do. Instead, they learn statistical patterns to establish a decision boundary, the mathematical "line in the sand" where a model switches its prediction from one class to another (e.g., from "seabed clutter" to "target").

Data that is visually convincing but lacks real-world variability tends to cluster cleanly around ideal examples. This encourages the model to learn an over-simplified separation between classes. While this boundary works perfectly on "clean" data, it collapses when the model encounters the "messy" statistical overlap of the real world, such as degraded sensors, low signal-to-noise targets, or unusual viewing geometries.

Figure 1 illustrates the danger of relying on synthetic data that under-represents these "hard cases." On the top, the simulated data is too clean, allowing the AI to draw a simple, rigid line between threats and benign objects. However, as shown on the bottom, real-world acoustic returns are rarely so distinct.

When a model is trained on a simplified reality, its decision boundary is drawn in the wrong place. In real world operations, this lack of robustness creates a critical dilemma: the model becomes fragile, where a subtle change in background noise or viewpoint can push a true positive across the boundary into the "safe" class, leading to missed contacts or several false alarms.

04 - EXPERIMENTATIONS

4.1 - Data Investigation & Augmentation

Before a detailed analysis could be carried out, it was important to understand the data available. As discussed, there are significant challenges associated with acquiring positive examples of the data under examination. The provided dataset reflected that as well, although the synthetic dataset was more balanced. The data distribution is summarised in Table 1.

TABLE 1 - Data split by category

	+ POSITIVE	- NEGATIVE
REAL	28	1248
SYNTHETIC	1268	1248

The synthetic data was generated using Blender to make an approximation to the real-world objects, including some environmental conditions. A Wasserstein style transfer GAN (Generative Adversarial Network) was then used to modify the blender outputs. This is a common method for generating synthetic data as it allows for a lot of configuration control. This generation method has a few common failure modes, however, such as texture tiling, over-sharpening, and low-frequency bias.

- Texture tiling occurs when the GAN discriminator is overly focused on local patch statistics. The generator then learns to reproduce texture primitives at "safe" frequencies, which fools the critic, but does not actually reproduce an accurate style.
- Over-sharpening is a particular issue in Wasserstein GANs [Banach Wasserstein GAN], as the Wasserstein critic penalises blurriness heavily. This can lead to overly sharp edges, or excessive high frequency noise being introduced during style transfer.
- On the other end of the spectrum, it is common for low frequency patterns to be learned early, as these are coarse patterns over the whole image. This can lead to outputs that look consistent with the desired style, but on examination the fine detail is not correct, as high frequency features have not yet been learned.

Due to the significant imbalances in the data, and to allow for better comparison between the real and synthetic positive sets, a set of augmented data was created. Augmentation was achieved using a random combination of rotation, reflection, and jitter (adding a random value to a subset of individual pixel intensities). As models are typically sensitive to directions and patterns, this kind of augmentation allows for small datasets to be treated similarly to one's many times their size. Figure 2 shows an example of each of these augmentations, and the final combined augmented image when they are chained together.

It is sometimes possible for augmentation to add data in that would be impossible in reality and thereby degrade performance. In this instance, there are examples present in the data of objects at a variety of angles, and with a range of "light source" directions, implying that rotation and reflection are valid augmentations. Jitter is considered valid, as it models noise on the sensor. As a result, a new "augmented real positive" class is obtained, bringing up our number of true positive examples to parity with the other classes available.

FIGURE 2 - Example of each augmentation, and the final combined result

4.2 - Statistical Analysis

To detect if the normal issues the use of GANs can cause, are present, an early direction of investigation was based around statistical analysis of image properties. A wide array of features were considered, but only a subset will be discussed here for brevity. For a more complete list of features potentially of interest, see Considerations. The features were manually selected to give a good coverage of the image properties, with specific features intended to cover the known potential weaknesses.

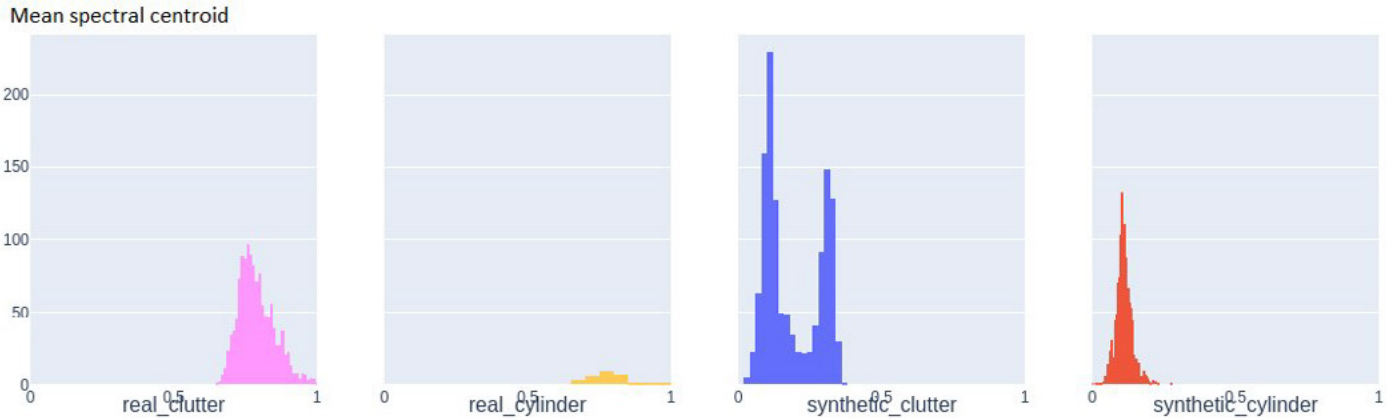

FIGURE 3 - Mean Spectral Centroid across real/synthetic, and positive/negative (cylinder/clutter) classes

To evaluate the level of tiling and low frequency bias, the first feature inspected was the spectral centroid of the data. This is a weighted average of the frequencies present in the image. It represents the centre of mass of the frequency components of the image. A low spectral centroid means that low frequency elements are dominating. This could indicate tiling, as the distinct repeats would each add a low frequency element, and could indicate a low frequency bias. A high spectral centroid may indicate a lot of noise or artefacts within the image, as these are generally high frequency effects. Real data is inherently noisy, and acoustic data can be especially so, due to the many different causes of scatter present within the environment. So, the expectation became that real data would likely have a high spectral centroid.

As shown in Figure 3, the synthetic data typically has a much lower spectral centroid than the real data, and this is true across both positive and negative examples. This is a particularly significant finding for the use of the data in model training, as convolutional neural networks (CNNs) are very frequency sensitive – mathematically they approximate to frequency filters. As such, it is likely that the real and synthetic data would appear very distinct to a machine learning model.

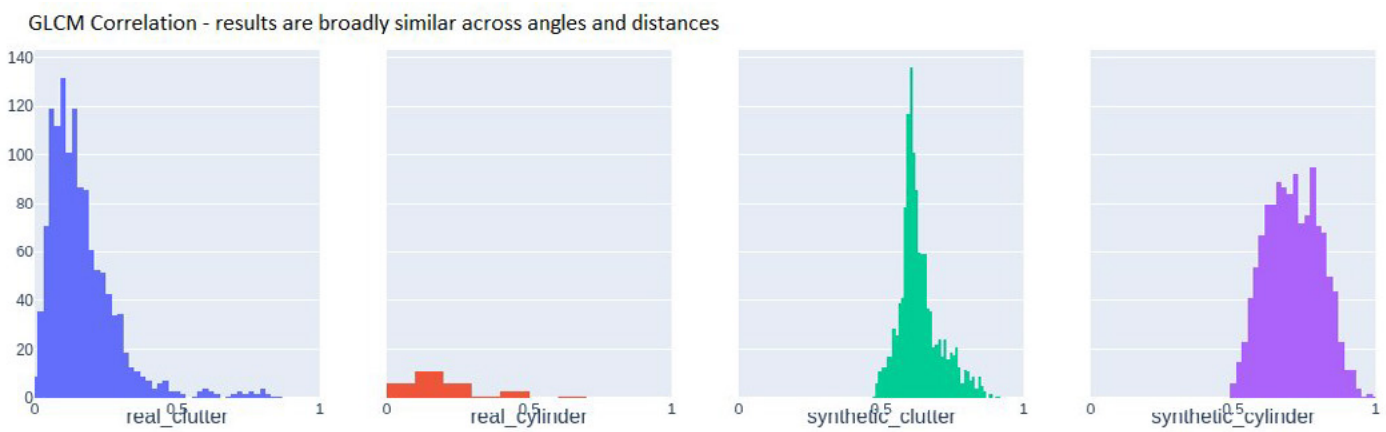

FIGURE 4 - GLCM across real/synthetic and positive/negative (cylinder/clutter) classes

To further assess the likelihood of tiling, the grey level co-occurrence matrix (GLCM) was calculated. The GLCM looks at a pixel within the image, and then for a fixed distance/angle, calculates the similarity between that pixel and the pixel at that distance/angle. When aggregated over a range of angles and distances, this metric can give a measure of how similar areas of the image are – if it is very lopsided towards particular angles and distances, that means that there are features likely to be repeating in their entirety, which could be tiling.

From Figure 4 the GLCM for the real data is lower than that for the synthetic. This indicates that there is more consistency across the image in the synthetic case than there is in the real.

As discussed above, this may be due to reduced noise in the synthetic images, but it may also be due to less background clutter in the scenes.

It bears repeating that differences in these statistical measures do not mean the synthetic data is bad. They instead just indicate areas in which the synthetic data differs from the real data. There are many examples where statistically different data can work equally well for model training – the key element is that the distribution of the real data is captured. However, the differences in these statistical measures do indicate areas in which the data can be adjusted for increased realism – e.g. by adding high frequency features or deliberately adding jitter to reduce the GLCM.

4.3 - Model Training

Statistical analysis of features extracted from the images can give us indications of quantifiable differences in the real and synthetic images. However, it is not always the case that any significant differences will lead to poor downstream model performance. In an ideal situation, any model (regardless of complexity) trained on the real data would achieve similar performance scores as a model trained on the real data, when tested on real data.

To test this, we defined several experiments based on simple image detection models. These experiments varied the proportions of real and synthetic images in the training data, and tested performance on purely real and purely synthetic test sets.

4.3.1 Chosen Model

Our data is labelled as either "positive" or "negative," so our model objective is detection (predicting whether a sample is "positive" or not). These samples are images; hence we need to design model architectures capable of handling this type of data.

The best class of machine learning models to use for this task are Convolutional Neural Networks (CNNs). Neural networks are essential for image detection because they can automatically learn complex, hierarchical patterns, such as edges, textures, shaped and full objects, directly from raw pixels. All modern state-of-the-art image detection models are neural networks. These models work by successively looping over training data and evaluating the outputs for a particular objective. After each loop (epoch), the parameters of the model are tweaked slightly in such a way that the model outputs should perform better on the objective next time round. This process is performed many times, and performance should continue to improve each time until convergence.

Neural networks vary drastically in complexity and size, and large models trained on vast amounts of data require a huge amount of compute resources. However, as mentioned above, the goal here is a comparative study of performance across different data domains, hence we do not need to use a particularly large or complex architecture here. We use small versions of the ResNet[2] models here, adapted to be a binary detection model, but we do compare with another architecture also.

[2] Wider or Deeper: Revisiting the ResNet Model for Visual Recognition - ScienceDirect

4.3.2 Performance Metrics

As the downstream model here is a binary detection one, we will concentrate our performance metrics on classic precision and recall. These metrics are a comparison of the predicted class ("positive" or "negative") against the group truth of that sample.

	+ POSITIVE	- NEGATIVE
+ POSITIVE	True Positives	False Positives
- NEGATIVE	False Negatives	True Negatives

Precision measures the accuracy of the positive predictions made by the model, while recall measures the ability of the model to predict all positive samples:

$$\text{PRECISION} = \frac{TP}{TP + FP} \qquad \text{RECALL} = \frac{TP}{TP + FN}$$

Both precision and recall metrics are between 0 and 1. In different scenarios, it may be better to favour one over the other. For example, if predicting cancer cases from ultrasound images, it is far more desirable to maximise recall over precision, as this minimises false negatives (someone having cancer and being told they do not). On the other hand, if the outcome of a model prediction is the use of lethal force, then precision would be favoured. Often a balance needs to be struck between the two.

A detection model will predict a probability of being "positive" from a sample. Precision and recall metrics rely on "hard" predictions, which mean allocating every sample predicted with a probability above a certain threshold as "positive" and everything else as "negative." Non-parametric scores can be used to measure how well a model predicts "positive" samples with a higher probability than "negative" samples. In this work we use Precision-Recall Area-Under-the-Curve (PR AUC) score, which measures how well a model balances precision and recall across all probability thresholds. It is calculated by sorting predictions by probability, computing precision at increasing levels of recall and then calculating the area under the precision-recall curve.

Once hard predictions are made, we also use a single metric to summarise precision and recall. For our dataset and use case, users' values recall more than precision. Therefore, we summarise precision and recall score using the F_β score:

$$F_\beta = \frac{(1+\beta^2)TP}{(1+\beta^2)TP + FP + \beta^2 FN}$$

With $\beta = 2$, which values recall twice as much as precision.

The aim of this work is to compare model performance when testing on real data and synthetic data. Using the above metrics PR AUC and F2 score, we can evaluate synthetic data quality using the difference in scores between real and synthetic test sets, for various training sets. These scores should be low (near zero), regardless of model, for high quality synthetic data. A low score demonstrates that synthetic and real data are interpreted similarly by the underlying models.

4.3.3 Experiment Design

Six model training experiments were designed. Firstly, a small number of real and synthetic samples are set aside to be used as test data. After this, the remaining data samples are split into training and validation. The training, validation and testing process goes as follows:

01. Train the model using the experiments training set. The training process loops through the training data many times and successively trains a new model each time.

02. The best model is chosen as the one which performs best on the combined (both real and synthetic) validation set, using PR AUC.

03. Individually for both the real and synthetic samples in the validation set, a probability threshold is set to maximise detection performance on those samples, using the F2-score.

04. These real and synthetic probability thresholds are then used to calculate detection metrics on the real and synthetic test sets, respectively.

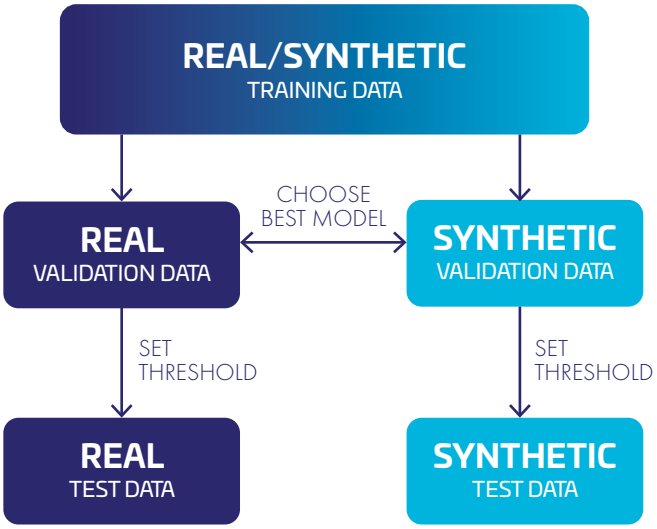

FIGURE 5 -
Schematic of the model training experiment design

Number	Experiment	Training Data				Training Weights	Validation Data
		Real Positives	Real Negatives	Synthetic Positives	Synthetic Negatives		
01	All synthetic train	0	0	1206	1206	-	All synthetic
02	Mixed train - unbalanced real	11	1251	1231	1231	(114, 1, 1, 1)	Equal mixed
03	Mixed train - balanced real	11	12	1231	1231	(112, 103, 1, 1)	Equal mixed
04	Mixed train - augmented real positives	1251*	1251	1231	1231	(2, 2, 1, 1)	Equal mixed
05	Mixed train - augmented real positives, reduced synthetic	1251*	1251	615	615	-	Equal mixed
06	All real train	1251*	1251	0	0	-	All real

TABLE 2 - Training data domain splits for each experiment

*Data is augmented to increase the sample size from the original data.

Table 2 shows how the training data is split for each of the experiments. The aim is to investigate the effect of different distributions of real and synthetic data on the model performance. In all cases, even when there are only an exceedingly small number of real samples in the training data, we attempt to boost the importance of the real data in the training procedure using training weights (via a weighted loss function). Experiments were also done without the weighting but did not perform as well in our test metrics, so are left out here.

05 - FINDINGS

5.1 - Main Results

Table 3 shows the main results from our experiments. These results summarise the key findings and are the main metrics of interest when assessing synthetic data quality. When referring to "mixed" training sets, we are referring to the training experiment which has augmented real positives (experiment 4 in table 2). "Real" training sets refer to the same augmented set without synthetic data (experiment 6). Full results of all 6 experiments can be found in the Appendix.

Training Set	PR AUC			F2-Score		
	Test Real	Test Synthetic	Difference	Test Real	Test Synthetic	Difference
Train Synthetic	0.72	1	**0.28**	0.17	0.98	**0.81**
Train Mix	0.69	1	**0.31**	0.58	1	**0.42**
Train Real	0.52	0.55	**0.03**	0.74	0.83	**0.09**

TABLE 3 - Main results from the model training experiments

All the score difference results should be near zero. This is clearly not the case, except when training on real data only, pointing towards significant differences in model interpretation of the real and synthetic data. We are particularly interested in performance when synthetic data is included in the training data, as not only do we need to assure this data, but we also need a measure of its usefulness in model training contexts. The score differences for training sets with synthetic data included are sub optimal.

These scores on their own are up for interpretation. However, the main takeaways from this are the process of evaluating the synthetic data. Now that we have one set of results, we can baseline quality of synthetic data against this. When new and improved synthetic data becomes available, one can use these metrics to measure against. Continual improvements on the synthetic data quality should push these scores more towards zero.

5.2 - Analysis

To understand the results presented in Table 3 a bit more, we performed some analysis on the model embeddings. The results here will consider the mixed training set model. The model embedding represent how the model has learned to represent the data in a numerical way. These numerical representations are what the model used to make its predictions i.e. they are the output of the penultimate layer of the neural network before the classification head.

To understand the spatial distribution and similarities of embeddings, we transform them into 2-dimensional representations using UMAP. Then plotting these representations gives us a visualisation of how similar or how far apart these embeddings are. Figure 6 shows the UMAP representations of the embeddings of training and validation/test data samples. Considering the training data, the model has clearly learnt to separate the "positive" and "negative" samples from each other. However, it has also learnt another clear separation in the real and synthetic samples, resulting in the 4 clear groups of embeddings. The real decisive test of the model is how this translates to the unseen data. Considering the validation/test data on the right of Figure 6, this shows that the model can clearly differentiate between synthetic "positive" and "negative" classes, but not the real data.

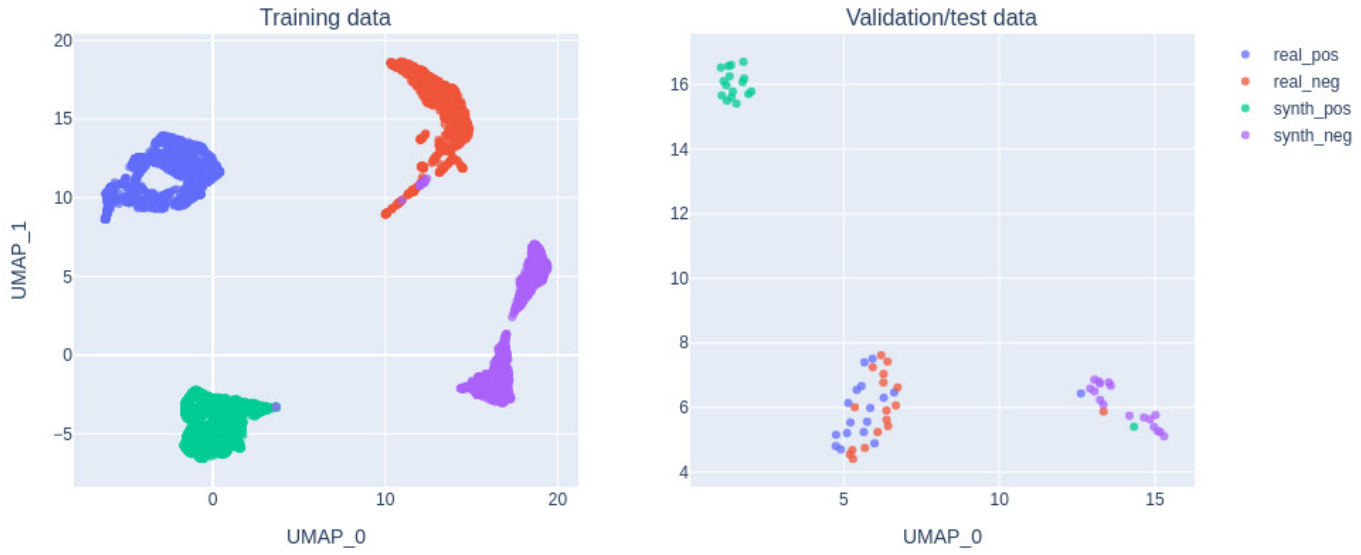

FIGURE 6 -
2-Dimensional UMAP representations of the embeddings of seen (training) and unseen (validation and test) samples, coloured by data class. The distance measure is cosine distance.

There are many reasons why this might be the case. The main one is that the separation between the synthetic "positive" and "negative" samples is far too obvious and well defined. Any model trained on data including those samples will learn this decision boundary very easily. This is not a true representation of real-world data, where the boundary between "positive" and "negative" is fuzzy and far less obvious. Whilst we are not using a particularly sophisticated model here, any more sophisticated model will suffer from the same problems and could even be more profound. Evidence of the underlying statistical differences are highlighted in section 6.2, where we have shown there to be significant differences in the real and synthetic samples with respect to texture and noise levels.

The issues raised here are also common ones seen when using GANs to generate synthetic data. GANs are trained on a discriminator between real and synthetic samples, aiming to decrease the degree of similarity between real and synthetic samples. However, due to the complexity of our data source here, this is incredibly challenging and can often end up with the adverse effect of clear delineation between real and synthetic samples.

06 - CONCLUSIONS

6.1 - Investigation

We have shown there to be significant differences in some of the underlying statistical features of the real and synthetic data samples. Depending on the downstream use case of the synthetic data, this can be a highly informative measure of synthetic data quality. However, in the context of model training this is not always the case. Deep learning model architectures such as CNNs can sometimes greatly benefit from including obscure looking synthetic samples in the training data, as they can help the model generalise better.

To investigate the effectiveness of the synthetic data in a model training context, we trained several image detection models using different mixes of real and synthetic data in the training data. We showed there to be clear differences in performance when testing on a purely real test set and on a purely synthetic test set. Performance was evaluated using PR AUC and F2 scores. These significant differences show that the synthetic data is not being processed and understood by the model in the same way that the real data is. Even if performance were not great on both real and synthetic test sets, that would be a strong indication that they are seen similarly to the model. However, in this case they are not.

Unfortunately, the decision boundary between the synthetic "positive" and "real" examples is far too wide, i.e. it is too easy to predict and not a true representation of the real data. As a result, models trained using this data learns this decision boundary. We have shown by analysis of the model embedding space, that this decision boundary then also clearly divides the data samples into real and synthetic. As such, the model struggles to generalise to real samples.

Due to the scarcity of "positive" real samples, the generation of synthetic data is especially important, and will become more so as state-of-the-art models become more data hungry. However, equally important, especially in defence contexts, is the assurance of this synthetic data. This assessment describes a process for assuring synthetic generated image data in the context of detection models. The ideas and process can easily be adapted for other downstream model use cases, and other data forms.

6.2 - Considerations for automation

To test the synthetic data in a model training perspective, we trained small CNNs. The results reported in this work were trained using a resnet18 (11.7 million parameters) model, but we did also try larger (resnet34) and smaller (mobilenet-v2) models which achieved comparable results. These models were never expected to do particularly well on real data, as developing high performing models on acoustic data is hard and would have required far more compute power than we had available. Larger, more sophisticated models would give stronger evidence on the synthetic data quality. However, the benefits of this needs to be weighed against the increased time to train, as our quality metrics are based on a few training runs based on different training data.

The models we have trained and the metrics we have developed were based on a downstream task of image detection. This implementation would need to be adapted should that downstream task be different. For example, if our downstream task was object detection (i.e. predicting a bounding box around an object in an image), then we would need to adapt the models training in this study for that purpose. An object detection task could also require a different performance metric that that used here, depending on the use case of the model.

This work presents a process and baseline for which to test synthetic data against. It is a (relatively) quick and straightforward process to execute against data to be a strong first pass on assurance. If any generated or acquired synthetic data does not pass this stage of assurance, it is unlikely to be useful for any downstream AI model training tasks. This process could be integrated alongside any generation process to ensure development is pushing the synthetic data quality in the right direction.

There are a few considerations required to best use this work for assurance of synthetic data. These metrics will almost never output a binary "yes this data is suitable" signal and instead will give results that require interpretation. As such, it may be valuable to include this assurance (or a version of) as part of the synthetic data generation pipeline. The benefits of this would be twofold: firstly, it would provide a baseline for continuous development of the generation model, and secondly it provides an evaluation of the synthetic data. To use this framework inside the synthetic data generation pipeline, a real test set would need to be taken out from the real data first. Post generation, take out a similar sized synthetic test set, then proceed to perform the model training tasks, using one purely synthetic training set, a mixed training set and a real training set. Calculate the difference in performance metrics between the real and synthetic test sets (using whichever metric is most appropriate) and log the results.

6.3 - Recommendations

To maximise benefits from this work, it is important to work on both sides of the problem – generating data and building the assurance processes to validate its quality. On the generation side, a key element which is not always captured is what elements of the dataset it is vital to replicate well. This is a fundamental part of requirements capture for synthetic data and is as significant in designing generation systems as the end use case. Many of the failure modes seen in this report appeared because of model selection. The use of a GAN to create the synthetic data was a signal that allowed the detection of issues. Other models have their own failure modes, and it is important that those known failures form part of the assessment criteria for the data. In this specific case, it is likely that Diffusion models would perform better, as they are more stable to train, which can avoid the mode collapse that affects GANs [2307.08702].

It is also often sensible to generate related synthetic data which is not directly relevant to the end use case. In the case of the PoC, there was almost no data covering real, positive examples, but plenty to cover real negative examples. If a synthetic data generation pipeline could be created that could create negative examples indistinguishable from the real ones, then the adjustment to create positive examples is much smaller. Mechanisms for tracking and benchmarking are important in this case, as it allows assurance and limitations of the dataset to be tracked over time. This can help avoid model collapse as the system is fine-tuned for positive examples.

The research into assurance should also continue, with a focus on generating automated pipelines for the assessments in this paper, to allow proper integration into generation pipelines. This would give a methodology to automatically benchmark data as well as models, so that data history can be tracked. Ultimately, the goal would be a data assurance pipeline that could be used across the group, but an initial stage could be an integrated generation and assurance pipeline. This pipeline could be used to demonstrate the speedups from integrated assurance, and the power of the rapid iteration capability that synthetic data introduces. Alongside this work, assurance metrics should be captured and validated, to allow guidance for what level of assurance different uses of data might require. This would allow a standardised approach to data assurance, which would therefore reduce the burden on individual projects to determine how to assess their training data.

07 - APPENDIX

In section 4.3.3 we described 6 different model training experiments. Table 2 shows the setup of each experiment in terms of training data splits between real and synthetic data, whether augmentations were used, and any training weights used. The full results of these experiments are given below.

Number	Experiment	PR AUC			F2-score		
		Test Real	Test Synthetic	Difference	Test Real	Test Synthetic	Difference
01	All synthetic	0.72	1	**0.28**	0.17	0.98	**0.81**
02	Mixed - unbalanced real	0.55	0.86	**0.31**	0.75	0.71	**0.04**
03	Mixed - balanced real	0.57	0.76	**0.19**	0.85	0.6	**0.25**
04	Mixed - augmented real positives	0.69	1	**0.31**	0.58	1	**0.42**
05	Mixed - augmented real positives, reduced synthetic	0.62	1	**0.38**	0.83	0.98	**0.15**
06	All real	0.52	0.55	**0.03**	0.74	0.83	**0.09**

THALES

Building a future we can all trust

Manor Royal
Crawley
West Sussex
United Kingdom
RH10 9HA

+44 (0) 129 358 0000

thalesgroup.com

© Thales 2026 • PCSWF-2347 • Designed by Thales UK Proposal and Creative Solutions