Scientific Research Publishing

Journal of Data Analysis and Information Processing, 2025, 13(3), 281-297
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

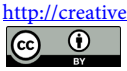
Data Evaluation in Artificial Intelligence

Philip de Melo

Department of Nursing and Allied Health, Norfolk State University, Norfolk, VA, USA
Email: ferndemelo@gmail.com

How to cite this paper: de Melo, P. (2025) Data Evaluation in Artificial Intelligence. *Journal of Data Analysis and Information Processing*, **13**, 281-297.
https://doi.org/10.4236/jdaip.2025.133017

Received: May 16, 2025
Accepted: August 4, 2025
Published: August 7, 2025



Abstract

High-quality data is essential for hospitals, public health agencies, and governments to improve services, train AI models, and boost efficiency. However, real data comes with challenges: strict privacy laws, high storage costs, legal constraints, and issues like bias or incompleteness. These can reduce the reliability of AI systems. As a result, artificial datasets are gaining importance. Synthetic and augmented data offer alternatives, yet their differences and potential are not fully understood. Data quality refers to how well a dataset is suited for its intended purpose in an AI pipeline. Key attributes include Accuracy—How correct and error-free the data is; completeness—Whether all required data fields are present; Consistency—Uniformity across datasets (e.g., same format or scale); Timeliness—Relevance of the data in time (significant in real-time systems), Validity—Whether the data follows defined formats or constraints; Uniqueness—Absence of duplicate records. This paper examines how both types of data are generated and used, showcasing their characteristics through practical examples.

Keywords

Artificial Intelligence, Accuracy, PM GenAI Algorithm

1. Introduction

Data in healthcare refers to the collection, storage, analysis, and use of various information generated within the healthcare system. This data is essential for improving patient outcomes, supporting clinical decision-making, enhancing operational efficiency, and advancing research.

Clinical data includes Electronic Health Records (EHRs), laboratory results, medical imaging, and prescription records. Patient-generated data is collected from wearables, mobile health applications, or patient surveys (e.g., step count, sleep patterns). Genomic data is derived from DNA sequencing, supporting personalized medicine and genetic research. Public health data encompasses disease surveillance, vac-

cination records, and population-level health statistics.

Healthcare data use focuses on clinical decision support: AI-driven tools that assist clinicians with diagnosis and treatment recommendations, operational efficiency, streamlining hospital workflows, optimizing staff scheduling, and managing effective research and innovation, facilitating the development of new treatments and understanding disease mechanisms and progression. The collected artificial data plays a pivotal role in Population Health Management, identifying and managing at-risk populations to prevent chronic disease and improve community health outcomes, and in personalized medicine, customizing treatment plans based on individual genetic, environmental, and lifestyle factors.

The major challenges in healthcare data include privacy and security, ensuring the protection of sensitive patient information in compliance with regulations such as HIPAA (Health Insurance Portability and Accountability), and data interoperability, which facilitates effective communication and data exchange between different health information systems.

One of the major challenges is data quality, which impacts the accuracy, completeness, consistency, and reliability of collected data, while ethical concerns address issues of fairness, bias in AI algorithms, and responsible use of patient data.

Technologies involved in healthcare data management can be listed as follows:

- Electronic Health Records (EHRs);
- Health Information Exchanges (HIEs);
- Artificial Intelligence and Machine Learning;
- Blockchain (for secure and transparent data sharing);
- Big Data Analytics.

Real vs. Synthetic Healthcare Data

The differences between real and artificial data are presented in Table 1.

Table 1. Real and artificial data characteristics.

Aspect	Real Data	Artificial Data
Privacy & Security	Contains identifiable information; higher risk of breaches and regulations	Artificially generated; no real personal data, reducing privacy concerns
Availability	Often limited due to cost, time, and legal/ethical constraints	It can be generated quickly, offering scalability and flexibility
Accessibility	Restricted access to protect patient privacy	Easier to share and use for development, testing, and training

Real datasets may exhibit bias due to the methods used in their collection, which can result in the underrepresentation of certain groups or skewed distributions. In contrast, artificial data can be deliberately engineered to reduce biases and ensure a more equitable representation of diverse populations.

Sectors such as healthcare and finance are bound by stringent data protection regulations like the European GDPR (General Data Protection Regulation) and American HIPAA. Artificial data offers a valuable solution by enabling compliance with these laws while maintaining the utility of the data for research, analysis, and development purposes.

Real-world data often contains inconsistencies, missing values, or errors, all of which can undermine the quality of analysis. Artificial datasets, however, can be systematically designed to uphold high standards of consistency, accuracy, and relevance for specific applications.

Unfortunately, the literature equates synthetic data with augmented data. We will show that these kinds of data are very different. Deep learning techniques, particularly Convolutional Neural Networks (CNNs), have revolutionized numerous computer vision tasks using large-scale, annotated datasets. However, acquiring such datasets in the medical field is particularly challenging. In [1], Frid-Adar *et al.* introduced methods for generating synthetic medical images using Generative Adversarial Networks (GANs). These synthetic images were shown to enhance CNN performance in medical image classification. Traditional data augmentation alone achieved 78.6% sensitivity and 88.4% specificity, while incorporating synthetic augmentation improved these metrics to 85.7% and 92.4%, respectively.

de Melo [2] described a new algorithm that uses augmented data to significantly improve the accuracy of lung cancer detection.

Shorten *et al.* [3] explored augmentation strategies for deep learning using a complete data likelihood function analogous to weighted least squares regression. This approach allows explicit uncertainty modeling at each neural network layer and supports diverse regularization schemes. It was applied across common activation functions like ReLU, leaky ReLU, and logit, offering a comprehensive framework for deep neural network training and inference. Wang *et al.* [4] investigated the use of data augmentation in deep learning.

The most basic and widely used data augmentation based on geometric transformation techniques is affine transformations, which include rotation, sharing, translation, scaling (resizing without zooming or cropping), mirroring, reflection, and flipping. While zooming and cropping are common image scaling techniques, they are not classified as affine transformations. Rotations, reflections, and translations form a subset of affine transformations known as Euclidean transformations [5]. Despite their simplicity, these methods have been shown to be highly effective in a variety of computer vision tasks [6] [7]. Due to their ease of implementation and proven effectiveness, they are often employed as the initial step in data augmentation before applying more advanced techniques [8].

Non-affine transformations enable the simulation of complex geometric distortions, often essential in specialized fields such as medical imaging [9] and document analysis. Unlike affine transformations, they can handle intricate and non-uniform deformations [10].

A common form of non-affine transformation is the projective or perspective

transformation, which maps points from an original image to a new reference frame, simulating different viewing angles or observer perspectives. These transformations are particularly valuable in applications like satellite imagery, UAV surveillance, and omnidirectional Field-of-View (FoV) systems, where wide-angle distortions occur. Such augmentations help models trained on standard image datasets generalize better to geometrically distorted or deformed inputs [11] [12].

Another key non-affine transformation is nonlinear deformation, which introduces variable transformation strength across different image regions. This approach increases the degrees of freedom beyond fundamental affine transformations, making it well-suited for simulating non-rigid deformations such as those caused by body movements or lens distortions. It's beneficial for augmenting data where natural variability or hardware-induced artifacts affect appearance [13]-[15].

In this paper, we explore the generation of synthetic and augmented data and highlight their significant differences. Synthetic data refers to artificially created datasets that can supplement or even replace real-world data in machine learning and other computational applications. Its primary aim is to address issues related to data scarcity and mitigate privacy and security concerns associated with the use of real data. Synthetic data can be generated through various techniques, including simulations, generative models (e.g., GANs), or rule-based data generation algorithms.

In contrast, data augmentation is a technique used to enhance the size and variability of an existing dataset, particularly in the context of deep learning. It involves applying a range of transformations—such as rotation, scaling, flipping, or noise injection—to original data samples, thereby creating new, diverse training examples that help improve model generalization and robustness.

2. Data Augmentation

2.1. Gaussian Augmentation

The greatest advantage of data augmentation is that it only requires the original training data, making it a cost-effective approach to increasing the size and diversity of the training data. Data augmentation is a powerful technique for mitigating overfitting—a prevalent challenge in deep learning where models become overly tailored to the training data and fail to generalize to unseen inputs. By generating additional, varied training samples, data augmentation encourages the model to learn broader data patterns, thereby enhancing its generalization capabilities and overall performance on new data.

Class imbalance, where certain classes have significantly fewer examples than others, can lead to biased model predictions. Data augmentation provides an effective strategy to counter this issue by generating synthetic examples for underrepresented classes, promoting a more balanced training process and improving classification accuracy across all classes.

By introducing a wider range of variations in the training dataset, data augmen-

tation increases the diversity of data the model encounters during training. This expanded exposure helps the model become more robust to variations in input and prevents overfitting to specific patterns or artifacts within the original dataset.

Gaussian augmentation is a probabilistic model used in statistics and machine learning to represent a distribution of data as a combination of multiple Gaussian (normal) distributions.

Mathematically, a Gaussian Mixture Model is defined as:

$$p(x)=\sum_{k=1}^{N}\pi_k \Psi\left(x_k \mu,\Sigma_k\right) \tag{1}$$

In the case of binary problems, the Gaussian distributions can be expressed as:

$$P\left(x\middle|y=k\right)=\sum_{j=1}^{M_k}\pi_{kj} \Psi\left(x\middle|\mu_{ki},\Sigma_{kj}\right) \tag{2}$$

In these formulas:

$$\psi\left(x\middle|\mu_k,\Sigma_k\right)=\frac{1}{(2\pi)^{d/2}\left|\Sigma\right|^{1/2}}\exp\left(-1/2(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\right) \tag{3}$$

which is the multivariate normal distribution with mean μ and covariance matrix Σ. K is the number of Gaussian components and π_k is a coefficient for k-Gaussian distribution:

$$\sum_{k=1}^{N}\pi_k=1 \tag{4}$$

2.2. Gibbs Augmentation

Gibbs data augmentation, like Gaussian augmentation, aims to:

1) Increase training data diversity;

2) Reduce overfitting by enforcing invariance or equivariance in models;

3) Improve generalization by simulating variations the model may encounter.

Let us denote d as an original data set (input) and y as a label:

$$x\in\chi,\quad y\in\Omega$$

The original data can be presented as a distribution of pairs:

$$(x,y)\sim P(x,y),$$

where P is a distribution. The augmented data set can be expressed as:

$$\tilde{x}=T(x)$$

where T is transformation operator. If T is a probability distribution over possible transformations, the augmented data would be:

$$(\tilde{x},y)=P_T(x,y)$$

A new distribution $P_T(x,y)$ is constructed such that:

$$P_T(x,y)=\int\delta\left(x-T(\tilde{x})\right)P(\tilde{x},y)\mathrm{d}T \tag{5}$$

where δ is the Dirac delta function and T is sampled from a distribution over al-

lowable transformation. de Melo [2] showed that Gibbs statistics can be an optimal choice for calculating the augmented data. This is because Gaussian statistics minimizes the information integral, though it can be used in many applications. The mathematics of data augmentation using Gibbs distributions can be formalized through probabilistic modeling, where the original data point d and its label y are part of a joint distribution, and augmented samples are generated in a way that preserves this distribution.

A Gibbs distribution is defined as:

$$p(x,y)=\frac{1}{z}\exp\left(-\beta E(\tilde{x},y)\right) \tag{6}$$

where $E(x, y)$: Energy function (often related to loss or negative log-likelihood);

β: Controls sharpness;

Z: Partition function (normalization constant).

The augmented samples can be derived from the conditional probability (to ensure correspondence of augmented samples and labels y):

$$p(\tilde{x}|y)=\frac{1}{z(y)}\exp\left(-\beta E(\tilde{x},y)\right) \tag{7}$$

In this expression,

$$Z(y)=\int\exp\left(-\beta E(x,y)\right)\mathrm{d}x \tag{8}$$

Data augmentation involves applying transformations to existing real-world data to create new, slightly modified versions. This technique is commonly used in fields like computer vision, natural language processing, and audio processing.

3. Data Augmentation for Labeled and Unlabeled Data

Data augmentations can be used to model the distribution of the training data and generate synthetic samples. The basic process is: Fit a Gaussian or Gibbs model to the real data, Sample new data points x' from the learned distribution $p(x)$, and use synthetic samples x' to augment the dataset. This is especially useful in low-data regimes or imbalanced datasets and appears in techniques like: Probability-based oversampling, Data augmentation for generative modeling and Anomaly detection via probabilistic likelihoods.

Let us consider a few examples: **Figure 1(a)** shows the original data. **Figure 1(b)** shows the original plus augmented data (red dots). **Figure 2** shows the dependence of information score on the Gaussian number of elements. As we increase the number of GMM components, the model can better fit the data. The log-likelihood increases, so BIC (Bayesian Information Score) decreases. However, with a higher number of Gaussian components, the model becomes overfitted. AIC stands for Akaike Information Criterion. It is a metric used in statistical model selection to compare how well different models fit a dataset while penalizing model complexity. The best choice would be the number of Gaussian elements, which is 4.

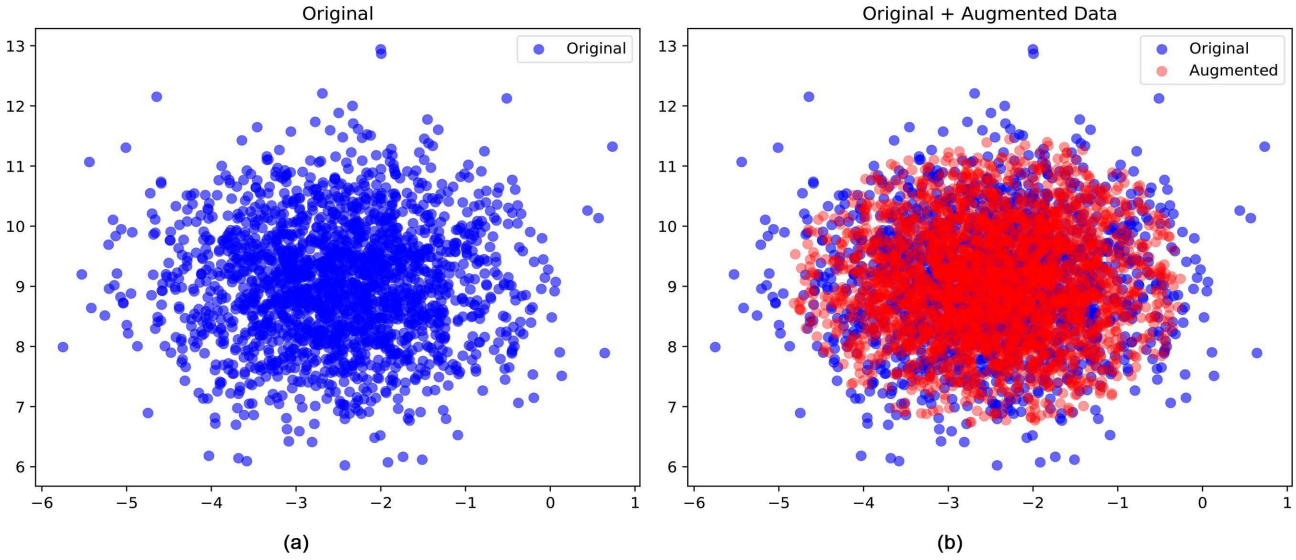


Figure 1. (a) Original data display; (b) Original plus augmented data.

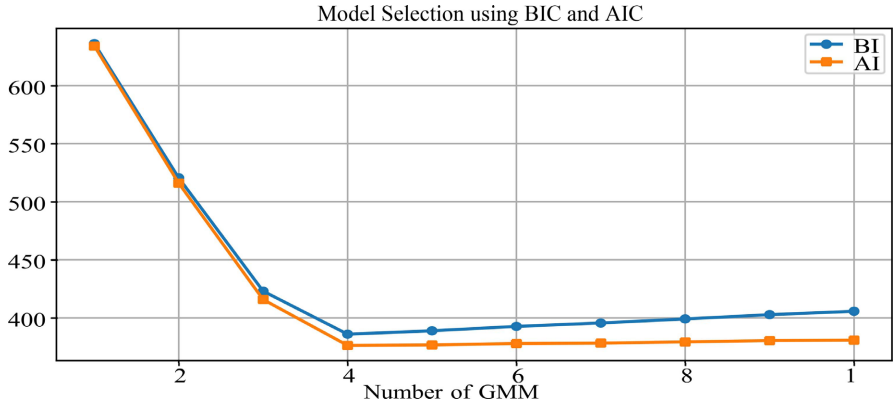


Figure 2. BIC and AIC dependance on the number of GMM components.

Figure 3 shows the data augmentation (original + augmented) in 4 clusters. **Figure 4** is the data augmentation for unlabeled data ((a) low probability, (b) high probability). **Figure 5** is the original and augmented data for a heart PQRT pulse.

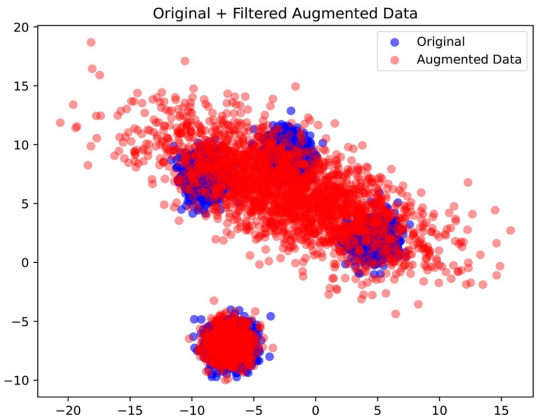


Figure 3. Original + augmented data with 4 clusters.

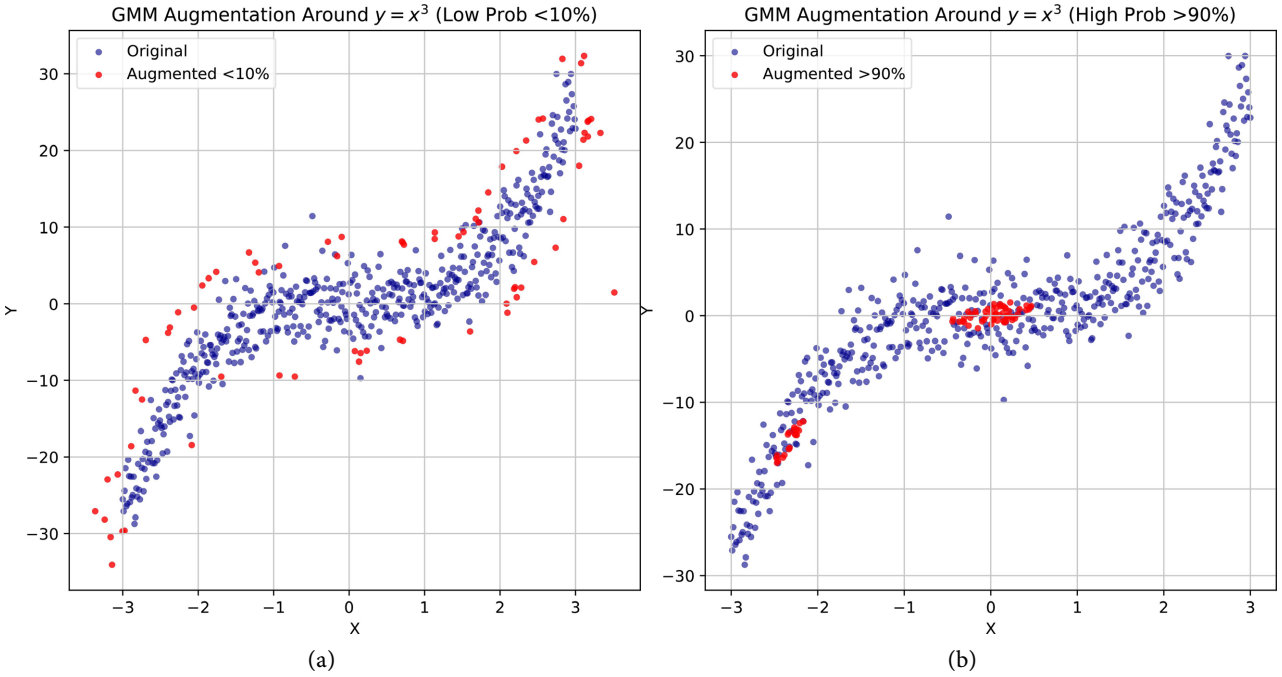


Figure 4. (a) Data augmentation of unlabeled data; (b) Data augmentation of unlabeled data with threshold > 90%.

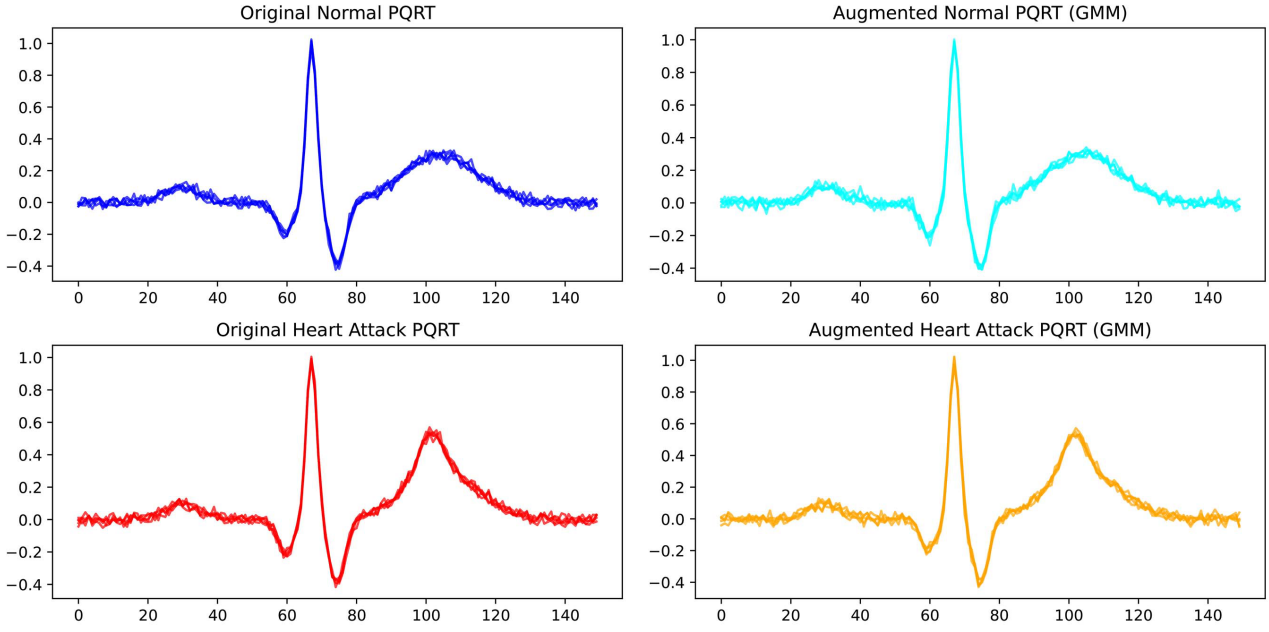


Figure 5. Original and data augmentation of heart pulse.

4. Synthetic Data Generation

4.1. Augmented vs. Synthetic Data Generation

Synthetic data generation involves creating entirely new data samples that don't originate from real data but are generated using models or simulations designed to replicate real-world distributions. The objective of synthetic data generation is to supplement or replace real data when it's scarce, expensive, or sensitive (e.g., in

healthcare, finance, or autonomous vehicle training). Figure 6 shows the histograms of original data (purple) and augmented (red). The augmented data algorithm captures well the pattern of the original data set.

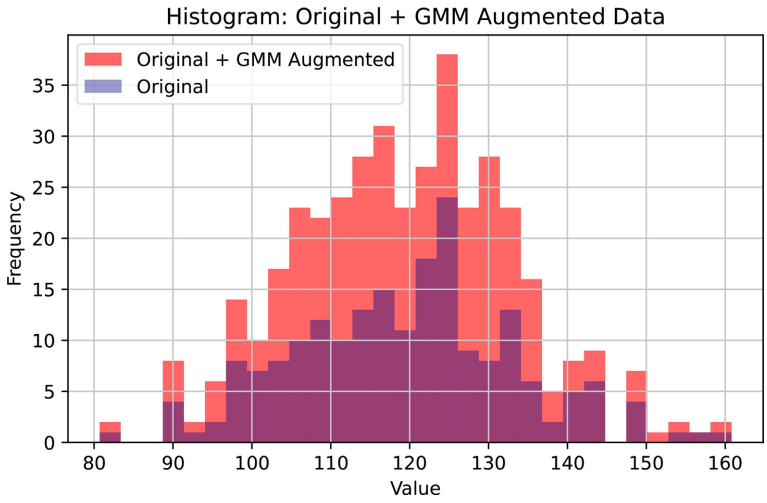


Figure 6. Histograms of original data (purple) and original + augmented (red).

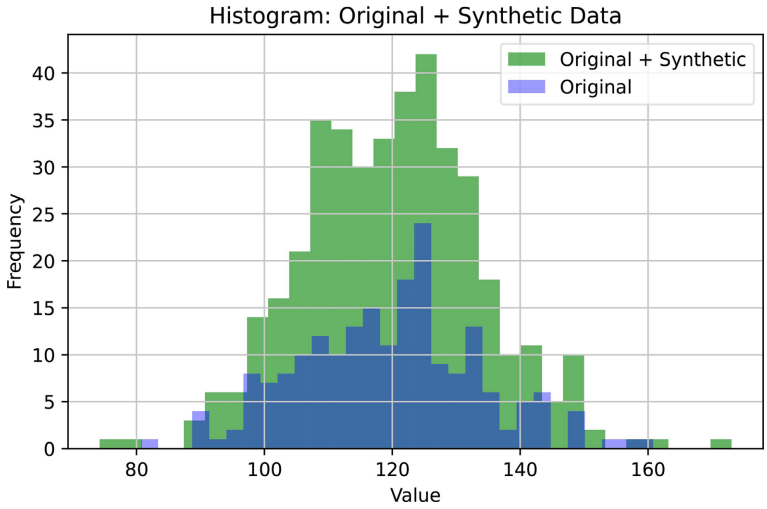


Figure 7. Histograms of original + synthetic data (green) and original (blue).

Augmented data was generated using the PM GenAI algorithm [4] that divides the data set into batches and then uses a combination of multiple mini-batches to mitigate overfitting. It learns the structure of the original data, possibly multiple clusters or distributions. Each sample is drawn from one of the mini-batches, using learned weights. It shows more flexible and powerful. It captures multimodal patterns and structures in data that are more representative of the complexity of the original distribution.

Synthetic data was generated based on a single Gaussian distribution. Uses only the meaning and standard deviation of the original data. It assumes the data follows a unimodal normal distribution while ignoring any clustering or multiple modes

in the original data. Usually (as seen in Figure 7), it fails to capture complex structures or multiple peaks in data. Table 2 shows the differences between augmented and synthetic data.

Table 2. The difference between synthetic and augmented data.

Augmented Data	Synthetic Data
Data augmentation is a technique used to expand and diversify the training dataset for deep learning models by applying various transformations to the original data. These transformations generate new, modified versions of existing samples, helping the model generalize more effectively and reducing the risk of overfitting. Common augmentation methods for image data include flipping, rotation, scaling, and adding noise. In addition to improving generalization, data augmentation can help address class imbalance by creating more examples of underrepresented classes. A major benefit of this approach is that it enhances the dataset without the need for collecting new data, making it a cost-efficient solution.	Synthetic data refers to data that is artificially created rather than collected from real-world events. It is used to address challenges such as data privacy, security, and limited access to real data. By leveraging techniques like simulations, generative models, and algorithmic data generation, synthetic data can serve as a substitute or complement to real datasets in machine learning and other domains. Its usefulness depends heavily on how accurately it reflects the patterns and characteristics of real-world data. Synthetic data is particularly valuable in fields like medical imaging and autonomous driving, where obtaining real data can be difficult or impractical. Additionally, it can be used to enrich existing datasets by introducing diverse examples with varied attributes.

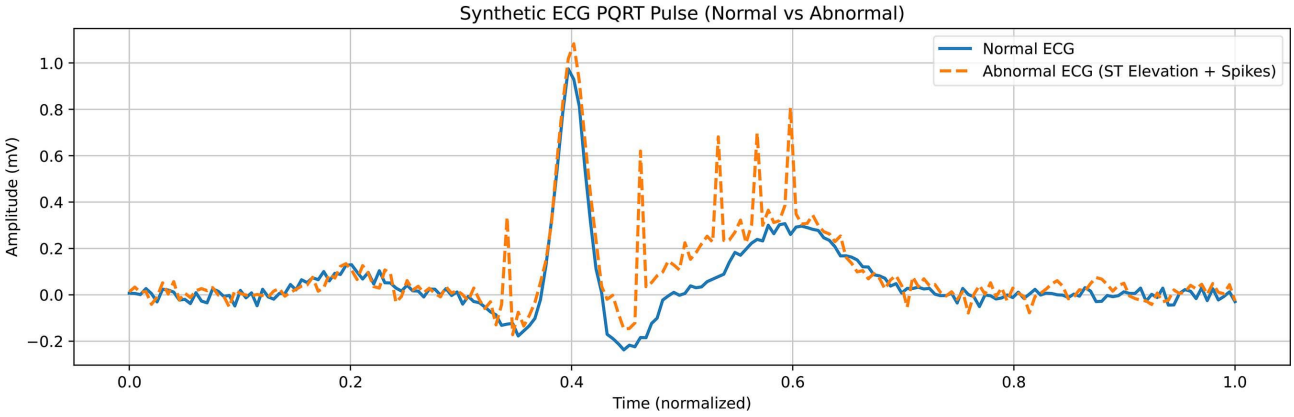


Figure 8. Generation of the ECG synthetic PQRT pulses for normal and abnormal heart.

Figure 8 shows a synthetic data set for ECG time series. Although synthetic data and data augmentation both aim to expand and diversify training datasets, they differ fundamentally in how they achieve this. Synthetic data is created entirely from scratch using simulations, generative models, or algorithms, whereas data augmentation modifies existing data to produce new examples. Figure 9 shows augmented data for normal heart and the ECG corresponding to heart abnormality (heart attack) (orange). Figure 10 depicts normal ECG and its wavelet transform. Figure 11 depicts abnormal ECG (heart attack) and its wavelet transform.

Synthetic data offers added advantages, such as improved privacy, enhanced security, and the ability to address data scarcity. However, if not carefully designed, it can introduce bias or lack realism. In contrast, data augmentation is

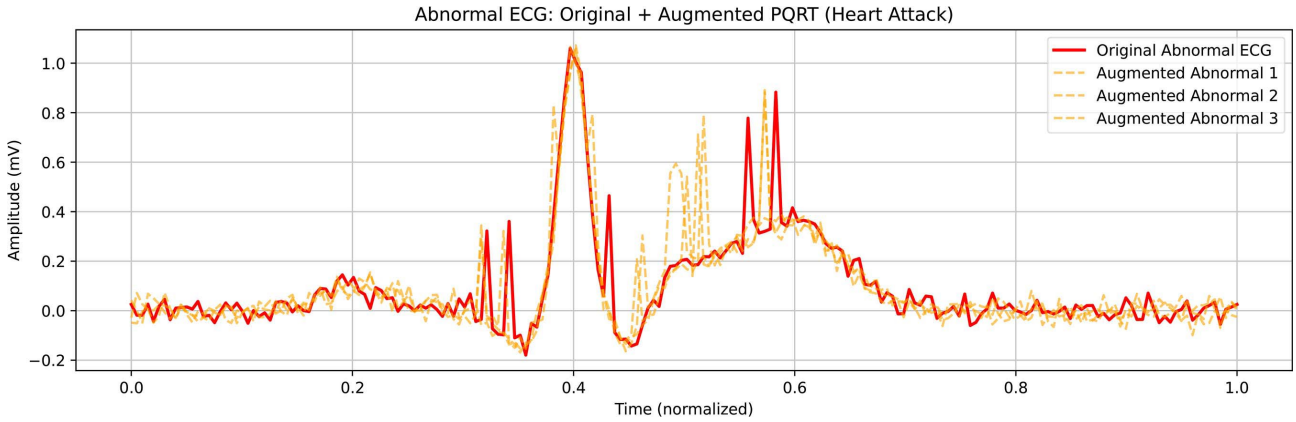


Figure 9. Abnormal (heart attack) ECG with augmented data.

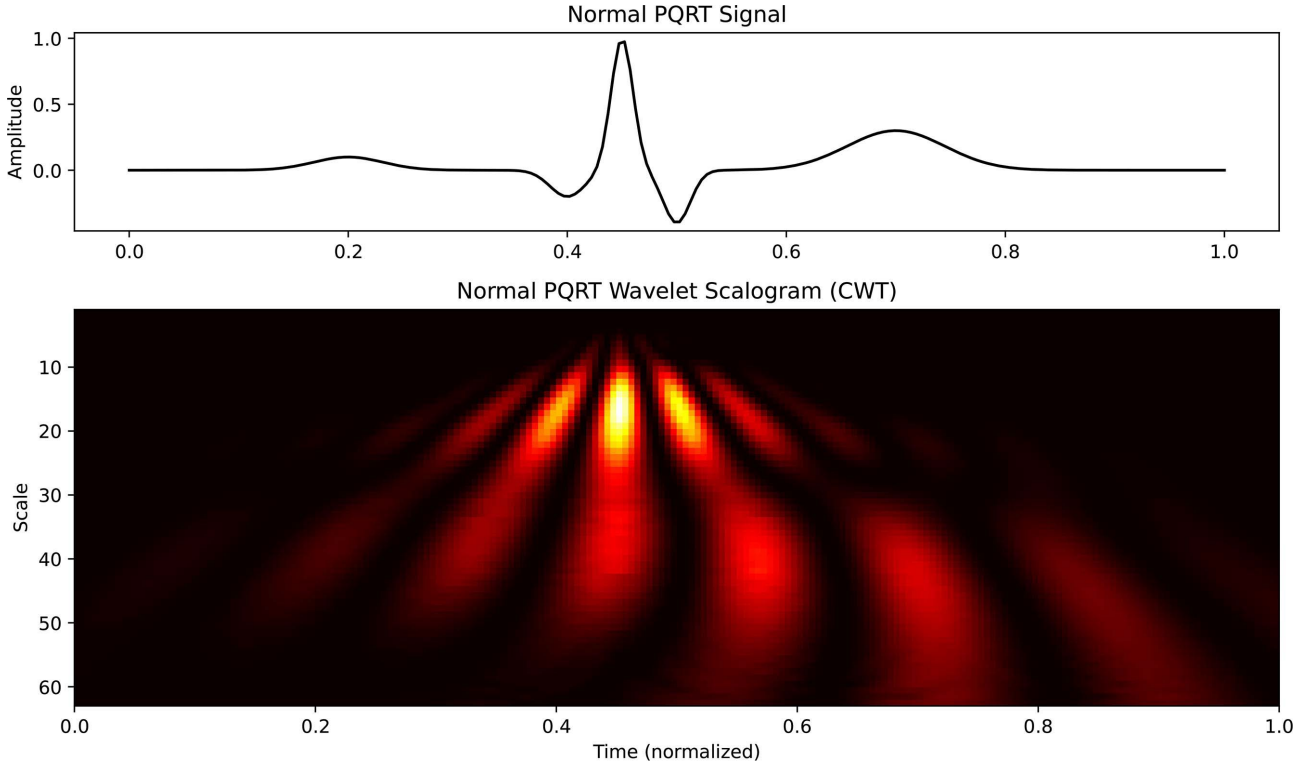


Figure 10. Normal synthetic ECG and its wavelet transform.

constrained by the quality and variety of the original dataset. When used together, these approaches can complement each other and enhance the performance of deep learning models.

Table 3 depicts the major characteristics of normal and abnormal PQRT pulse.

Let us examine how synthetic data impacts the original data set. Figure 12 shows the data set and the probability of distribution of elements.

Let us compute KDE for original, synthetic, and augmented data.

Definition of KDE

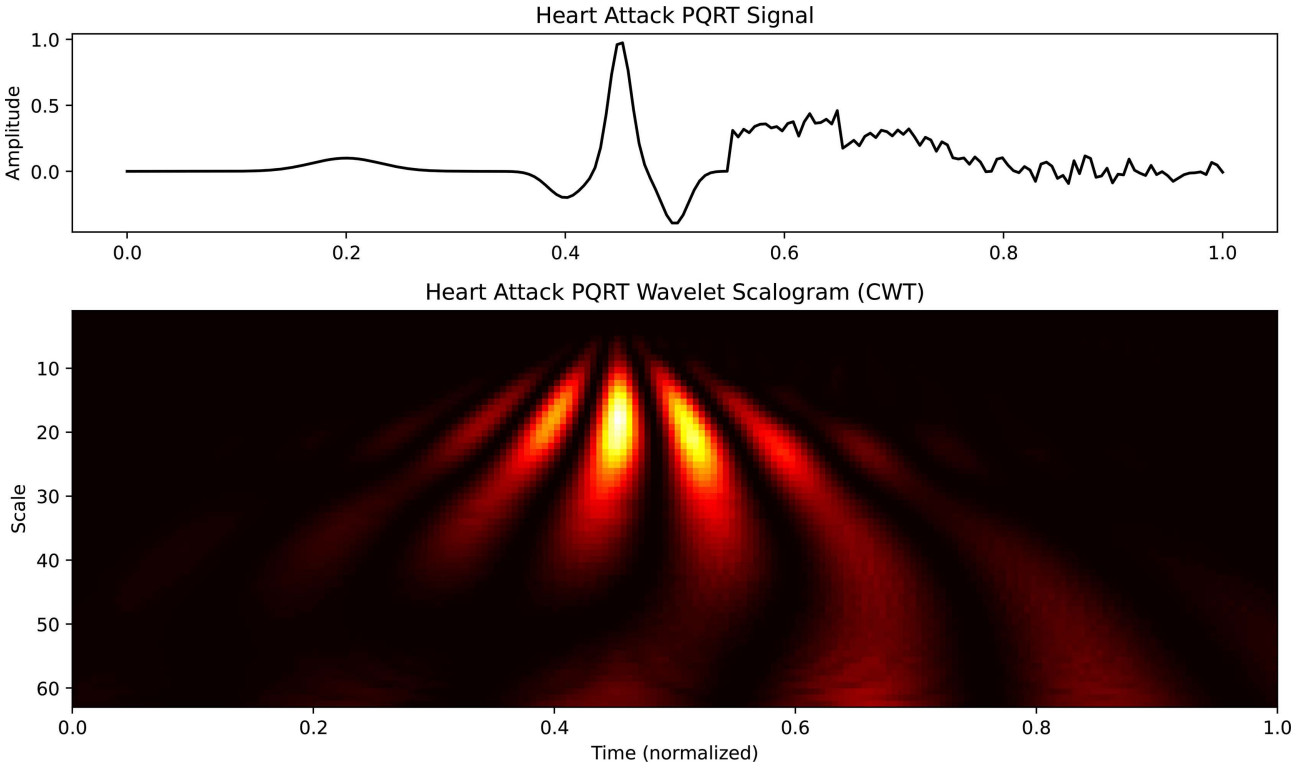


Figure 11. Abnormal synthetic ECG (heart attack) and its wavelet transform.

Table 3. Summary of the differences between normal and abnormal ECG that supports Figure 10 and Figure 11.

Feature	Normal PQRT	Heart Attack PQRT
Signal shape	Smooth Gaussian-like waves	Irregular, asymmetric, spiky
Frequency content	Dominated by low frequencies	Contains more high-frequency energy
Baseline	Stable and flat between waves	May be elevated or undulating (ST elevation, noise)
Energy distribution	Concentrated in R-wave	Spread across many unpredictable components

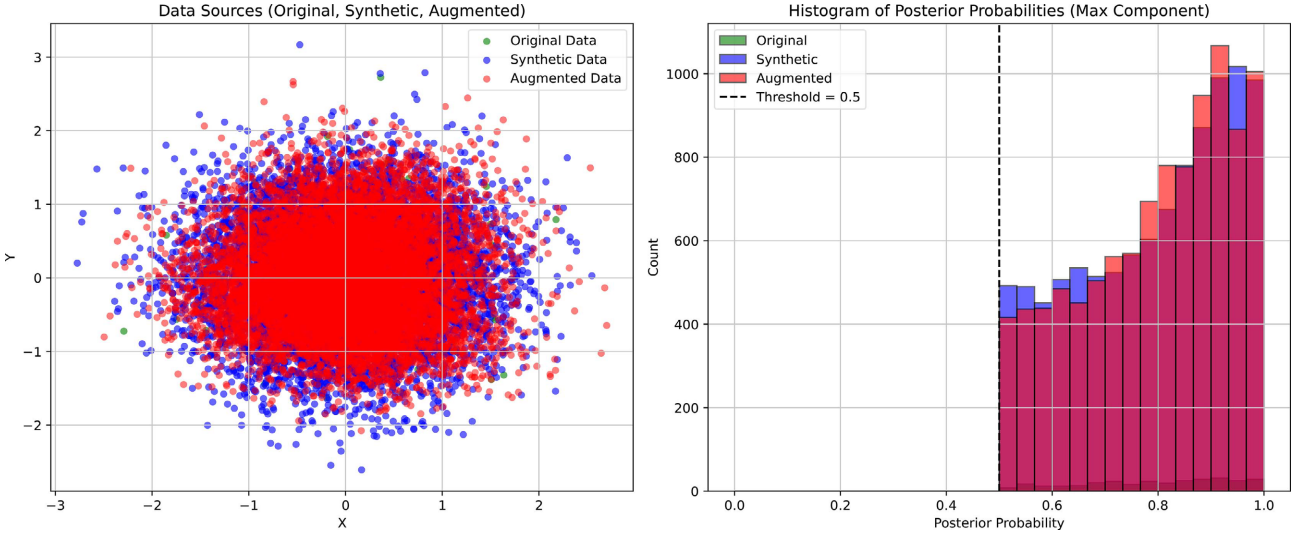


Figure 12. A combination of original, augmented, and synthetic data.

Given a dataset of n observations $\{x_1, x_2, \cdots, x_n\}$ the kernel density estimator at point $x \in R$ is defined as:

$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{9}$$

where:

- $\widehat{f}_h(x)$ is the estimated density at point x;
- K is the kernel function (e.g., Gaussian, Epanechnikov);
- $n > 0$, $h > 0$ is the bandwidth (smoothing parameter);
- x_i are the sample points.

Figure 13 shows the KDE graphs for original, synthetic and augmented data.

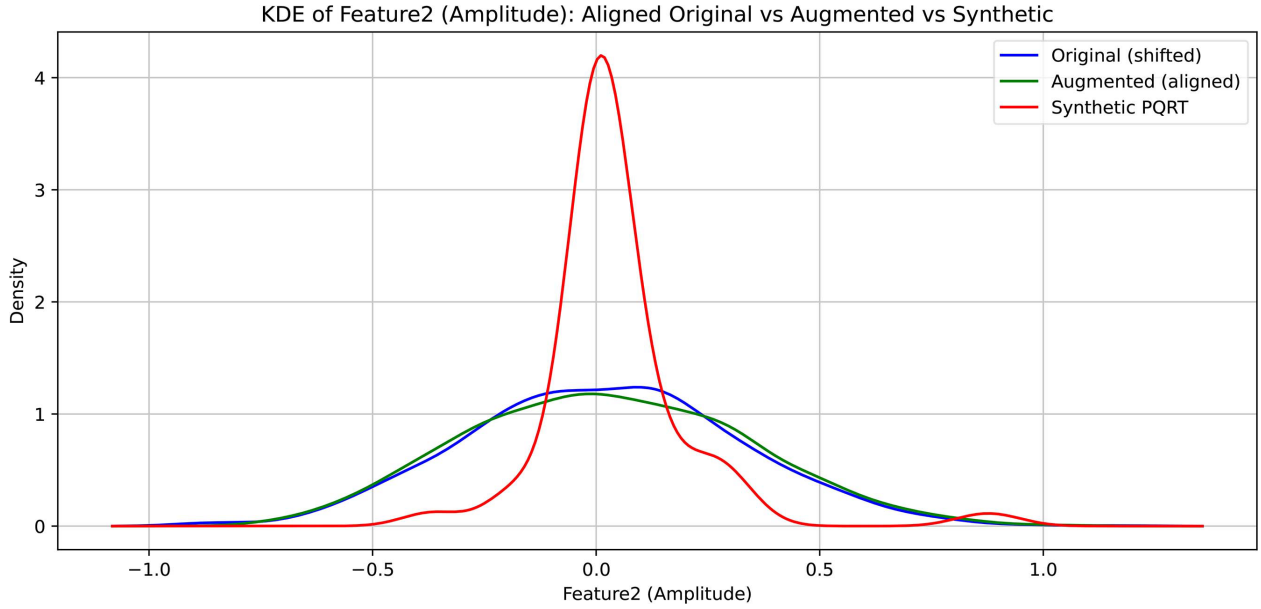


Figure 13. A KDE plot shows excellent approximation of the original data by augmented data but not by synthetic data.

4.2. Comparison of Gaussian and Gibbs Statistics

We will investigate an important feature of Gibbs and Gaussian statistics. [16] and [17] demonstrated the use of Gaussian statistics in health care application.

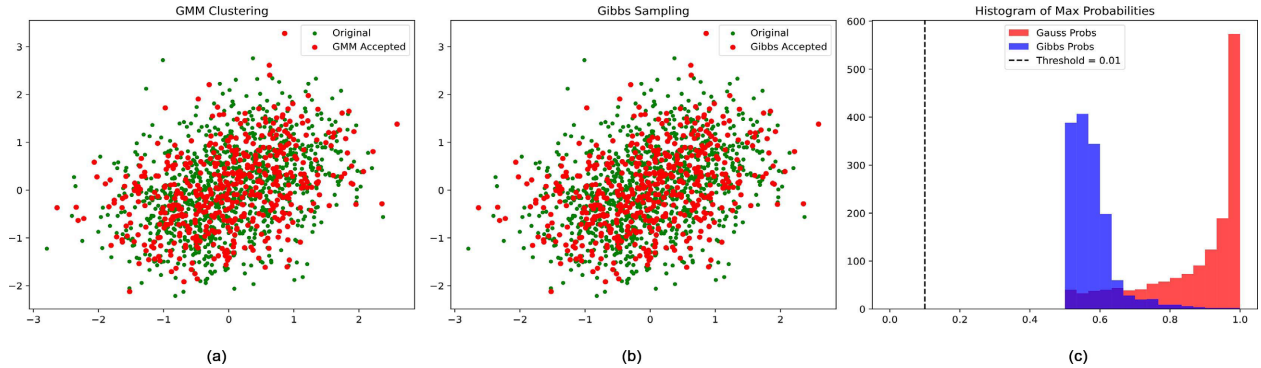


Figure 14. (a) Gaussian statistics used in data augmentation; (b) Gibbs statistics in data augmentation; (c) Probability distribution of augmented data with the threshold 0.5.

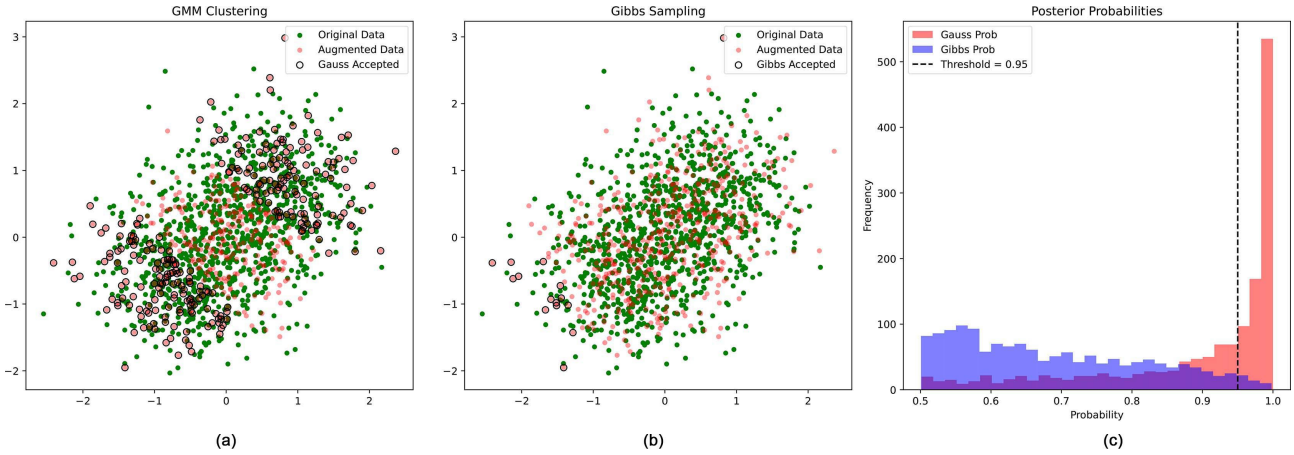


Figure 15. (a) Gaussian statistics used in data augmentation; (b) Gibbs statistics in data augmentation; (c) Probability distribution of augmented data with the threshold 0.9.

Figure 14 and Figure 15 show that the Gaussian augmentation directly maximizes the likelihood of data distribution. The Gaussian augmentation uses the Expectation-Maximization (EM) algorithm to find parameters (means, covariances, and weights) that maximize the likelihood of the data. This means it's actively optimizing the model to make the data points fit the Gaussian components as well as possible. The result is often sharper, more confident probability assignments (*i.e.*, probabilities closer to 1).

Gibbs sampling is a sampling-based technique from Bayesian statistics. Gaussian augmentation samples from a posterior distribution over cluster assignments, not optimized to maximize individual point likelihoods and (Figure 14, Figure 15) demonstrate this feature. The process is noisier and more diffuse, especially in short runs (e.g., 50 iterations), which leads to lower or more uncertain probability estimates.

Gibbs augmentation computes posterior probabilities from sampled parameters without guaranteeing global maximum likelihood, which may lead to underconfident (*i.e.*, lower) values.

While Gaussian augmentation fits covariances carefully for each component, in the Gibbs sampler, covariance updates are simplified (especially when points per cluster are small), which can reduce the accuracy of the likelihood computation.

5. Conclusions

The era of artificial data is rapidly unfolding, with transformative implications for society and industry alike. As we advance deeper into the digital age, artificial data promises not only to address pressing challenges—such as data privacy, regulatory compliance, and the training of artificial intelligence—but also to fundamentally reshape how we perceive, manage, and utilize information.

This paper explores the evolving role of artificial data, highlighting the distinction between synthetic and augmented data and examining the use of various statistical frameworks such as Gaussian and Gibbs distributions. While the concept

of artificial data has been around for some time, its adoption is now accelerating at an unprecedented pace.

In the coming years, artificial data is expected to become an integral part of daily business operations, scientific research, and technology development. From generating training data for AI models to simulating complex scenarios in sectors like healthcare, finance, and autonomous systems, synthetic data will offer scalable, cost-effective solutions to some of the most pressing challenges facing modern industries.

One of the synthetic data's most significant future advantages lies in its capacity to uphold data privacy. With regulations like the General Data Protection Regulation (GDPR) in Europe and similar frameworks emerging globally, there is a growing demand for data solutions that minimize the risk of personal data exposure. Synthetic data provides a compelling answer: it enables organizations to work with realistic, representative datasets without compromising user privacy.

In parallel, the contribution of synthetic data to improving AI model performance is critical. Deep learning models often require vast amounts of labeled data, which is not always readily available or affordable in many domains. Synthetic data can fill this gap by generating large volumes of high-quality training data, helping models learn more effectively and generalize better to real-world conditions. Over time, blending synthetic with real-world data will lead to more robust and accurate AI systems.

The impact of synthetic data on innovation is particularly evident in fields like autonomous driving, where it is used to simulate real-world environments without the safety risks of live testing. These applications can significantly reduce development costs and accelerate the deployment of new technologies.

On a global scale, synthetic data has the potential to democratize access to high-quality information and drive advancements in research, policy, and development. In healthcare, for instance, it can enable privacy-preserving analysis and enhance diagnostic tools, even in regions where real data is scarce or sensitive.

As the field evolves, staying engaged with thought leaders and domain experts is essential. Platforms such as Twitter, LinkedIn, and Medium offer valuable perspectives from researchers and practitioners shaping synthetic data's future. Their insights into emerging trends, use cases, and ethical considerations are vital resources for navigating this fast-moving landscape.

Synthetic data and data augmentation are complementary strategies to address the challenge of limited training data in deep learning. While synthetic data generates entirely new examples that mimic real-world data distributions, data augmentation enhances existing datasets by introducing variability through transformations. Both approaches contribute to building more effective and generalizable models. A hybrid approach—combining synthetic data generation with augmentation—often yields the best performance in practical applications.

Our journey with synthetic data is just beginning. As technological capabilities expand, synthetic data will play an increasingly vital role in addressing complex

challenges and driving innovation across sectors. From safeguarding privacy to enabling breakthroughs in AI and scientific research, its potential is vast and far-reaching [17].

Researchers, developers, and policymakers must work together to harness this potential to ensure that synthetic data is used responsibly and ethically. Those who invest in understanding and advancing this technology today will be better prepared to capitalize on the opportunities it presents tomorrow. The future of artificial data is not only promising but inevitable.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J. and Greenspan, H. (2018) GAN-Based Synthetic Medical Image Augmentation for Increased CNN Performance in Liver Lesion Classification. *Neurocomputing*, **321**, 321-331.

[2] de Melo, P. (2025) Augmented and Synthetic DATA in Artificial Intelligence. *International Journal of Artificial Intelligence & Applications*, **16**, 93-108. https://doi.org/10.5121/ijaia.2025.16307

[3] Shorten, C. and Khoshgoftaar, T.M. (2019) A Survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, **6**, 1-48.

[4] Wang, Y., Polson, N. and Sokolov, V.O. (2023) Data Augmentation for Bayesian Deep Learning. arXiv: 1903.09668.

[5] Ryan, P.J. (1986) Euclidean and Non-Euclidean Geometry. Cambridge University Press. https://doi.org/10.1017/cbo9780511806209

[6] Xu, Y., Jia, R., Mou, L., Li, G., Chen, Y., Lu, Y., *et al.* (2016) Improved Relation Classification by Deep Recurrent Neural Networks with Data Augmentation. arXiv: 1601.03651.

[7] Wong, S.C., Gatt, A., Stamatescu, V. and McDonnell, M.D. (2016) Understanding Data Augmentation for Classification: When to Warp? 2016 *International Conference on Digital Image Computing: Techniques and Applications* (DICTA), Gold Coast, 30 November 2016-2 December 2016, 1-6. https://doi.org/10.1109/dicta.2016.7797091

[8] Dong, X., Potter, M., Kumar, G., Tsai, Y., Saripalli, V.R. and Trafalis, T. (2022) Optimizing Data Augmentation Policy through Random Unidimensional Search. In: Simos, D.E., Rasskazova, V.A., Archetti, F., Kotsireas, I.S., Pardalos, P.M., Eds., *Lecture Notes in Computer Science*, Springer International Publishing, 306-318. https://doi.org/10.1007/978-3-031-24866-5_23

[9] Milletari, F., Navab, N. and Ahmadi, S. (2016) V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. 2016 *Fourth International Conference on* 3*D Vision* (3*DV*), Stanford, 25-28 October 2016, 565-571. https://doi.org/10.1109/3dv.2016.79

[10] Simard, P.Y., Steinkraus, D., Platt, J.C., *et al.* (2003) Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis. 7*th International Conference on Document Analysis and Recognition*, **3**, 958-963.

[11] Wang, K., Fang, B., Qian, J., Yang, S., Zhou, X. and Zhou, J. (2020) Perspective Transformation Data Augmentation for Object Detection. *IEEE Access*, **8**, 4935-4943. https://doi.org/10.1109/access.2019.2962572

[12] Franke, M., Gopinath, V., Reddy, C., Ristic-Durrant, D. and Michels, K. (2021) Bounding Box Dataset Augmentation for Long-Range Object Distance Estimation. 2021 *IEEE/CVF International Conference on Computer Vision Workshops* (*ICCVW*), Montreal, 11-17 October 2021, 1669-1677. https://doi.org/10.1109/iccvw54120.2021.00192

[13] Ronneberger, O., Fischer, P. and Brox, T. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A., Eds., *Lecture Notes in Computer Science*, Springer International Publishing, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[14] Jaderberg, M., Simonyan, K., Zisserman, A., *et al.* (2015) Spatial Transformer Networks. *Advances in Neural Information Processing Systems*, **28**, 2017-2025.

[15] Karargyris, A. (2015) Color Space Transformation Network. arXiv: 1511.01064.

[16] de Melo, P. and Davtyan, M. (2023) High Accuracy Classification of Populations with Breast Cancer: SVM Approach. *Cancer Research Journal*, **11**, 94-104.

[17] de Melo, P. (2024) Public Health Informatics and Technology. Library of Congress.