AUGMENTED AND SYNTHETIC DATA IN ARTIFICIAL INTELLIGENCE

Philip de Melo

Department of Nursing and Allied Health , Norfolk State University , 700 Park Avenue Norfolk, VA 23504

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. This paper examines how both types of data are generated and used, showcasing their characteristics through practical examples.

Data generation techniques—such as Gaussian Mixture Models (GMM), Generative Adversarial Networks (GANs), and Gibbs sampling—enable the creation of realistic, privacy-preserving patient records that mimic the statistical properties of real data. Data augmentation, commonly used in image and signal analysis, is increasingly applied to structured electronic health records (EHRs), laboratory values, and time-series data to enhance model robustness and generalizability.

This paper explores mathematical foundations, methodological frameworks, and real-world applications of synthetic and augmented data in healthcare. We highlight how these techniques improve disease prediction, mitigate bias, and enable high-performance machine learning models, particularly in low-resource or imbalanced clinical domains. By expanding the effective size and diversity of training datasets, synthetic and augmented data serve as critical enablers for equitable, scalable, and data-driven healthcare systems.

KEYWORDS

Artificial intelligence, accuracy, PM GenAI algorithm.

1. INTRODUCTION

Data in healthcare refers to the collection, storage, analysis, and use of various types of 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 data, vaccination records, and population-level health statistics.
Uses of healthcare data focus 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 resources effectively and research and

Innovation: Facilitating the development of new treatments and understanding disease mechanisms and progression. The collected artificial data plays the 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 that facilitates effective communication and data exchange between different health information systems.

One of the major challenges is data 1uality that 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
Understanding the differences between real and artificial data is crucial for assessing their applications and benefits:

Table 1. Real and artificial data.

Aspect	Real Data	Artificial Data
Privacy & Security	Contains identifiable information; higher risk of breaches and regulation	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 underrepresentation of certain groups or skewed distributions. In contrast, artificial data can be deliberately engineered to reduce such 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, on the other hand, can be systematically

designed to uphold high standards of consistency, accuracy, and relevance for specific applications.

Unfortunately, in the literature synthetics data is equated to augmented data. We will show that these kinds of data are well 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], Mayan 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 the use of augmented data to significantly boost the accuracy of lung cancer detection.

Shorten et al. [3], explored augmentation strategies for deep learning using a full data likelihood function analogous to weighted least squares regression. This approach allows for 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. Y. Wang et al. [4] investigated the use of data augmentation in deep learning.

The most basic and widely used data augmentation is based on geometric transformation techniques are affine transformations, which include operations such as rotation, shearing, 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, which are 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 basic affine transformations, making it well-suited for simulating non-rigid deformations such as those caused by body movements or lens distortions. It’s especially useful for augmenting data where natural variability or hardware-induced artifacts affect appearance [13],[14],[15].

In this paper, we explore the generation of synthetic and augmented data, highlighting their significant differences. Synthetic data refers to artificially created datasets that can be used to 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 to mitigate privacy and security concerns associated with the use of real data. Synthetic data can be generated through a variety of 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 augmentation 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:

$$\boldsymbol{p}(x) = \sum_{k=1}^{N} \pi_k(x_k \mu, \Sigma_k) \qquad (1)$$

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

$$\begin{aligned} p(\mathbf{x} \mid y=0) &= \sum_{k=1}^{K_0} \pi_k^{(0)} \mathcal{N}(\mathbf{x} \mid \mu_k^{(0)}, \Sigma_k^{(0)}) \\ p(\mathbf{x} \mid y=1) &= \sum_{k=1}^{K_1} \pi_k^{(1)} \mathcal{N}(\mathbf{x} \mid \mu_k^{(1)}, \Sigma_k^{(1)}) \end{aligned} \qquad (2)$$

In these formulas:

$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \quad (3)$$

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:

$$\sum_{k=1}^{N} \pi_k = 1 \quad (4)$$

2.2. Gibbs Augmentation

Gibbs data augmentation like Gaussian augmentation aims to:

- Increase training data diversity.
- Reduce overfitting by enforcing invariance or equivariance in models.
- Improve generalization by simulating variations the model may encounter. Let us denote d as an original data set (input) and y is a label:

$$\boldsymbol{x} \in \boldsymbol{\chi} \ , \ \boldsymbol{y} \in \boldsymbol{\Omega}$$

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

$$(\boldsymbol{x}, \boldsymbol{y}) \sim \boldsymbol{P}(\boldsymbol{x}, \boldsymbol{y}),$$

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

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

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

$$(\tilde{\boldsymbol{x}}, \boldsymbol{y}) = \boldsymbol{P}_{\boldsymbol{T}}(\boldsymbol{x}, \boldsymbol{y})$$

A new distribution $\boldsymbol{P}_{\boldsymbol{T}}(\boldsymbol{x}, \boldsymbol{y})$ is constructed such that:

$$\boldsymbol{P}_{\boldsymbol{T}}(\boldsymbol{x}, \boldsymbol{y}) = \int \boldsymbol{\delta}(\boldsymbol{x} - \boldsymbol{T}(\tilde{\boldsymbol{x}}))\boldsymbol{P}(\tilde{\boldsymbol{x}}, \boldsymbol{y}) \, d\boldsymbol{T} \quad (5)$$

where δ is the Dirac delta function and T is sampled from a distribution over allowable transformation. De Melo [4] 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:

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

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):

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

In this expression,

$$Z(y) = \int exp(-\beta E(x, y)\, dx \qquad (8)$$

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

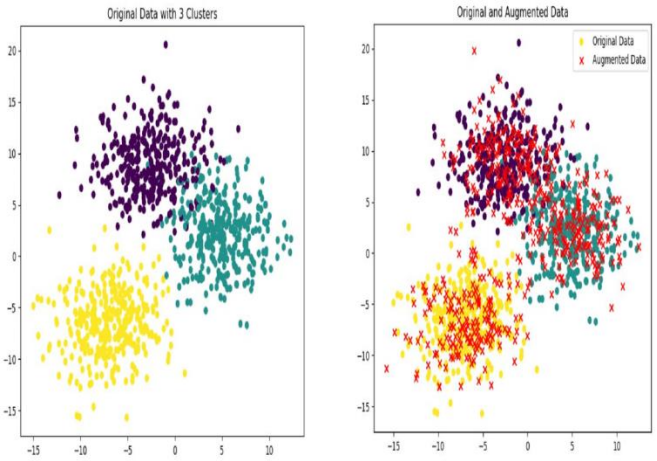
Data augmentations can be used to model the distribution of the training data and generate synthetic samples. The basic process is:

Fit a Gaussian or Gibbs model to the real data.
Sample new data points x′ from the learned distribution p(x).
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.

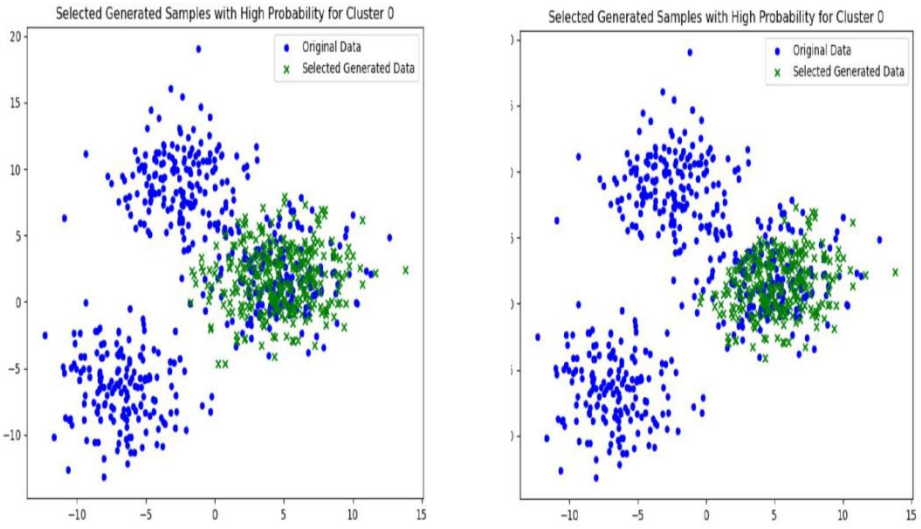
Let us consider a few examples: Figure 1a shows the original data with 3 clusters. Figure 1b shows the original plus augmented data (red dots). Figure 2a shows the second cluster with augmented data (green) with low threshold, while Figure 2b shows the second cluster with augmented data (green) with high threshold.

Figure 3 shows the data augmentation for unlabeled data: a) no threshold and b) threshold > 90%.



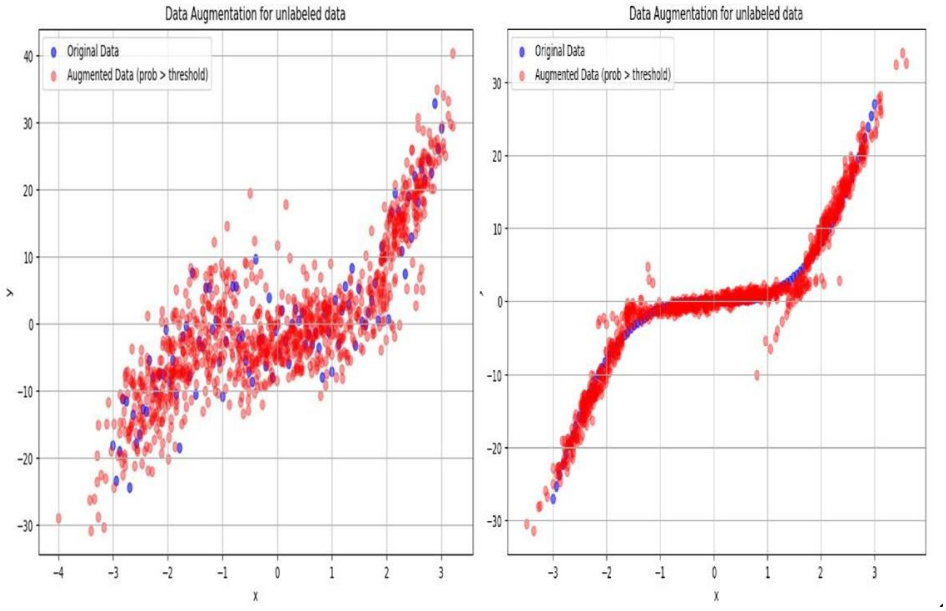
a) b)

Figure 1 (a) original data display of 3 clusters. b) Original plus augmented data



a) b)

Figure 2 (a) Second cluster with low threshold. b) Second cluster with higher threshold.



(a) (b)

Figure 3 (a) Data augmentation of unlabeled data. b) Data augmentation of unlabeled data with threshold >90%.

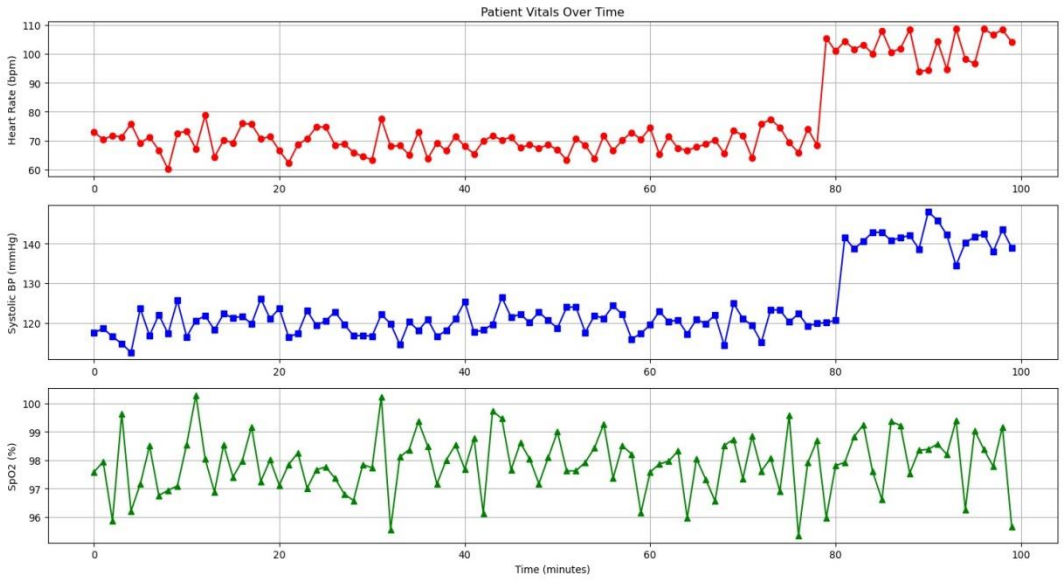


Figure 4 (a) Data augmentation of heart pulse. b) Data augmentation for systolic pressure, c) Data augmentation for oxygen saturation.

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 5 shows the histograms of original data (green) and augmented (red). The augmented data algorithm captures well the pattern of the original data set.

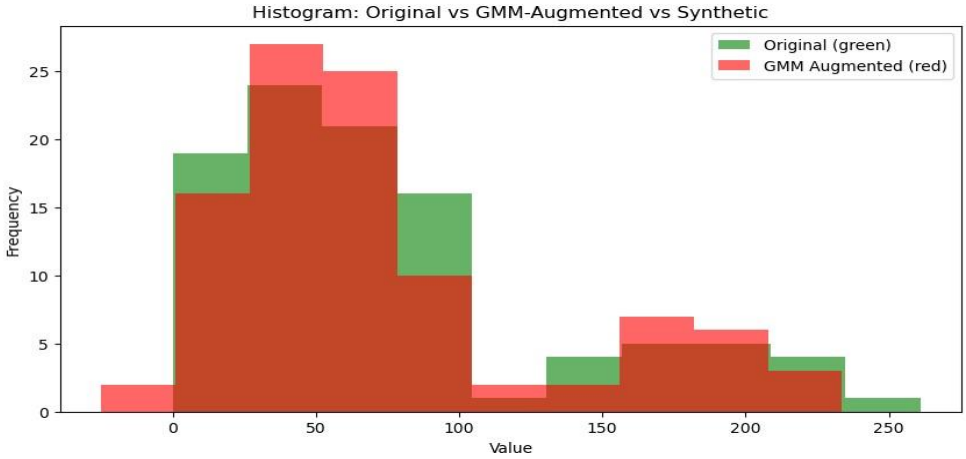


Figure 5: Histograms of original data (green) and augmented (red). The augmented data algorithm captures well the pattern of the original data set.

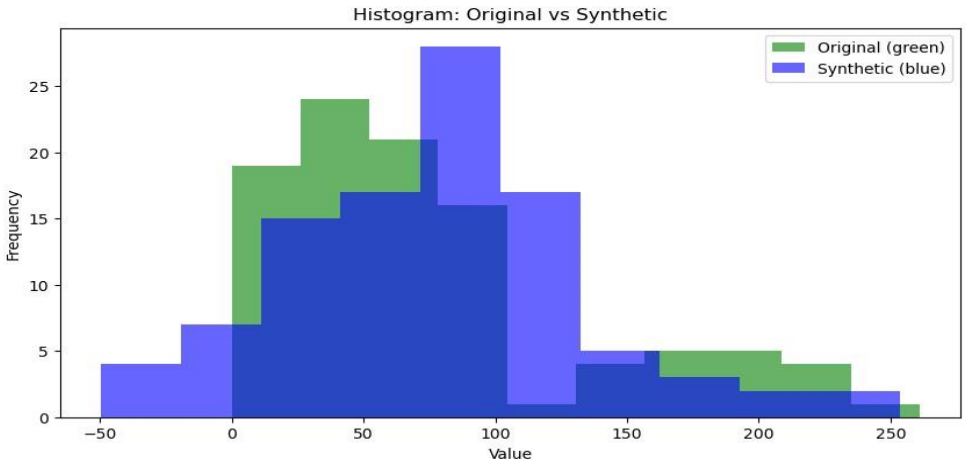


Figure 6: Histograms of original data (green) and synthetic (blue). The augmented data algorithm does not capture well the pattern of the original data set.

Augmented data was generated using the PM GenAI algorithm [4] that divides the data set in 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 being 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 it is seen in Figure 6, fails to capture complex structures or multiple peaks in data.
Table 1 summarizes 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.

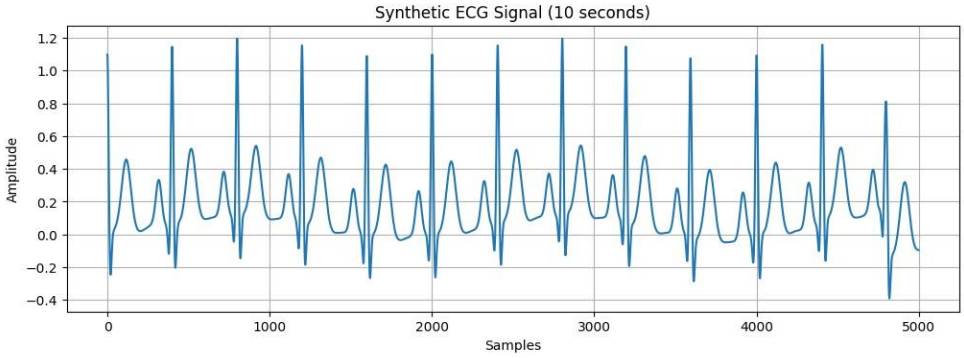


Figure 7: Generation of the ECG synthetic data.

Figure 7 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 8 shows synthetic data for normal heart (a) and the ECG corresponding to heart abnormality (b).

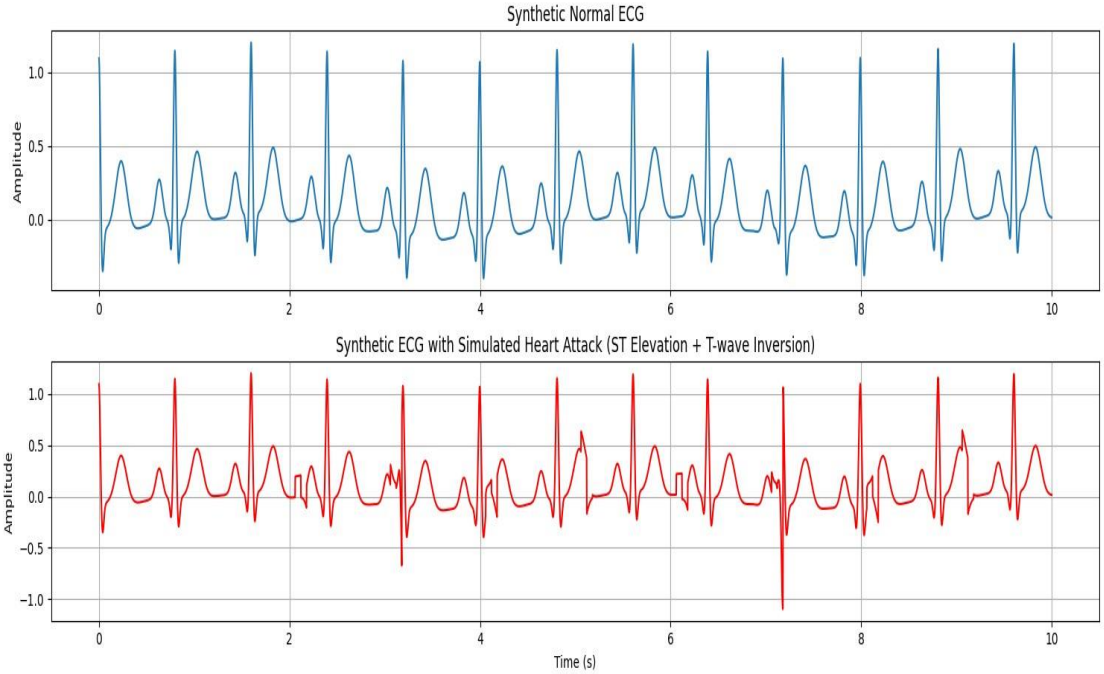
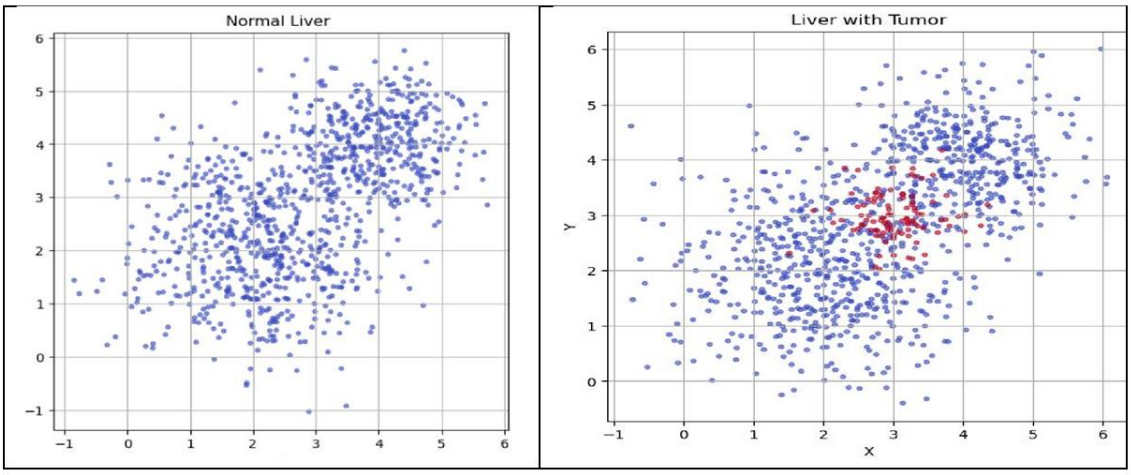


Figure 8: Generation of the normal and abnormal ECG synthetic data.

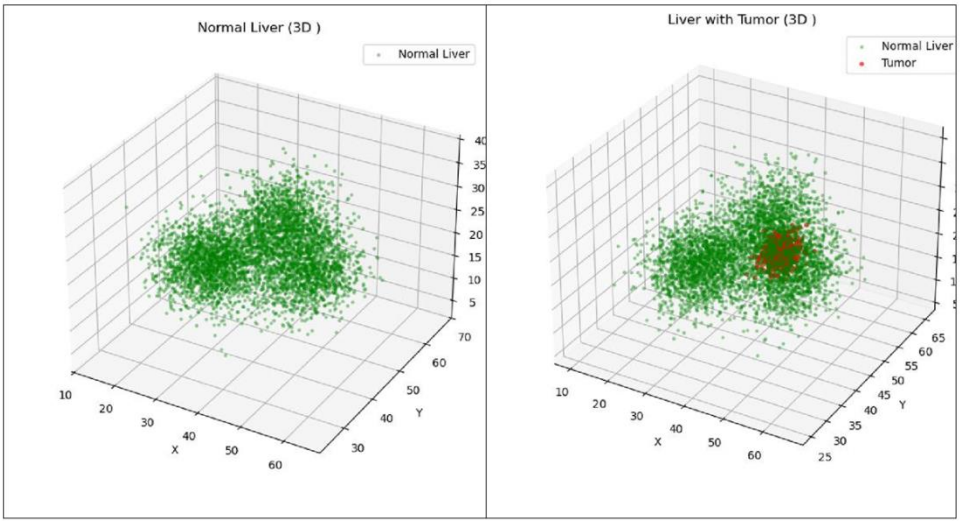
Synthetic data offers added advantages, such as improved privacy, enhanced security, and the ability to address data scarcity. However, if not carefully designed, it can introduce bias or lack realism. In contrast, data augmentation is constrained by the quality and variety of the original dataset. When used together, these approaches can complement each other and enhance the performance of deep learning models.



(a) (b)

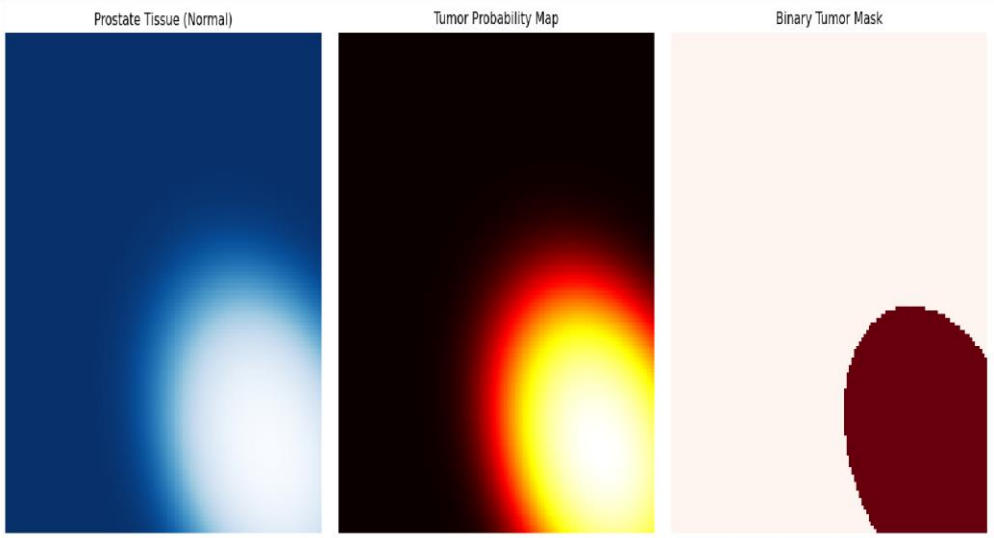
Figure 9: Synthetic model for normal liver (a) and with tumor (b).

Figure 9 (a and b) depicts normal liver and liver with tumor, while Figure 10 (a and b) shows the synthetic model for normal and tumor present in liver. Figure 11 shows a synthetic image of a normal prostate (a), with a lesion (b), and the lesion mask (c).



a) b)

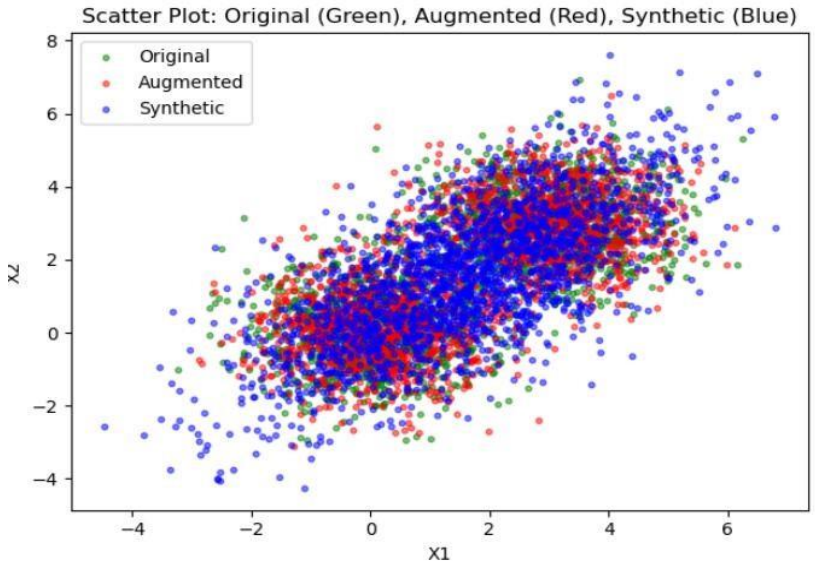
Figure 10: Gaussian synthetic model for normal liver (a) and with tumor (b).



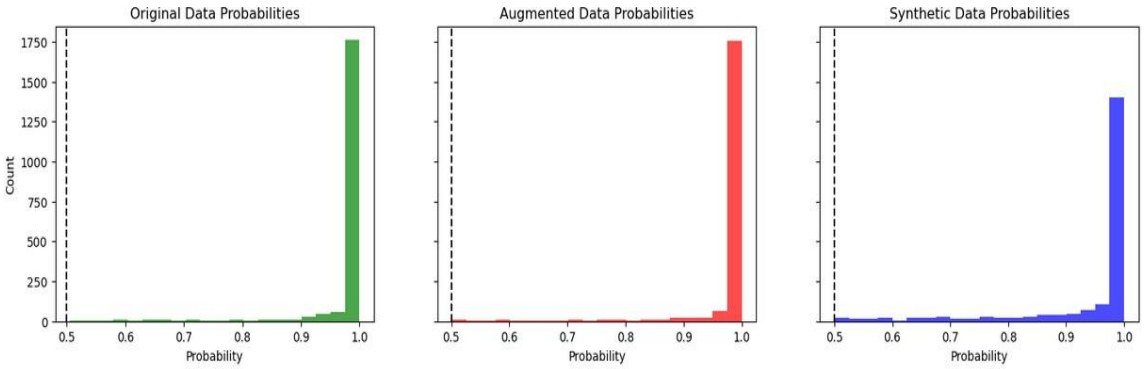
a) b) c)

Figure 11: a) Normal tissue, b) Occurrence of tumor, c) Binary tumor mask.

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



a)

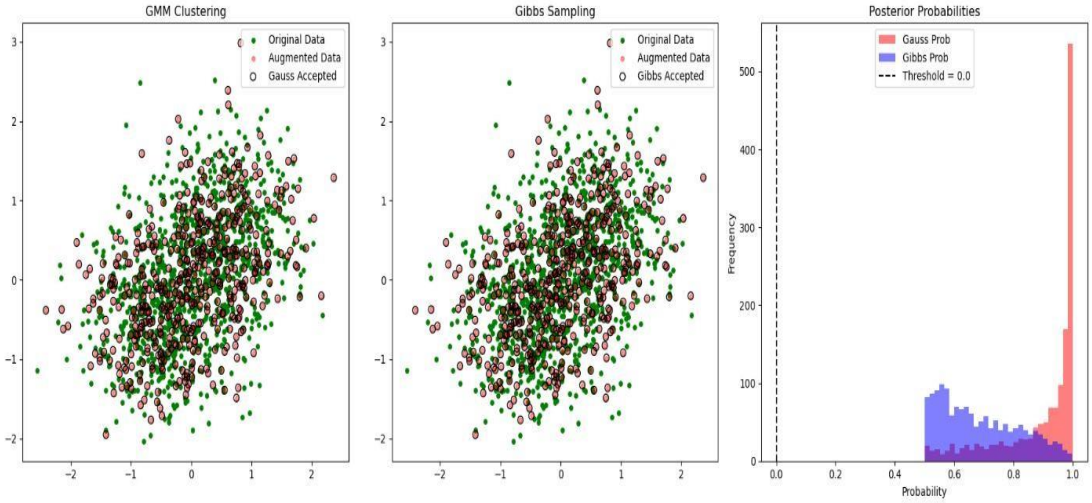


b)

Figure 12: a) A combination of original, augmented, and synthetic data b) Probability distribution of synthetic data and augmented data showing higher accuracy in augmented 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.



a) b) c)

Figure 13: a) Gaussian statistics used in data augmentation, b) Gibbs statistics in data augmentation, c)

Probability distribution of augmented data with the threshold 0.5

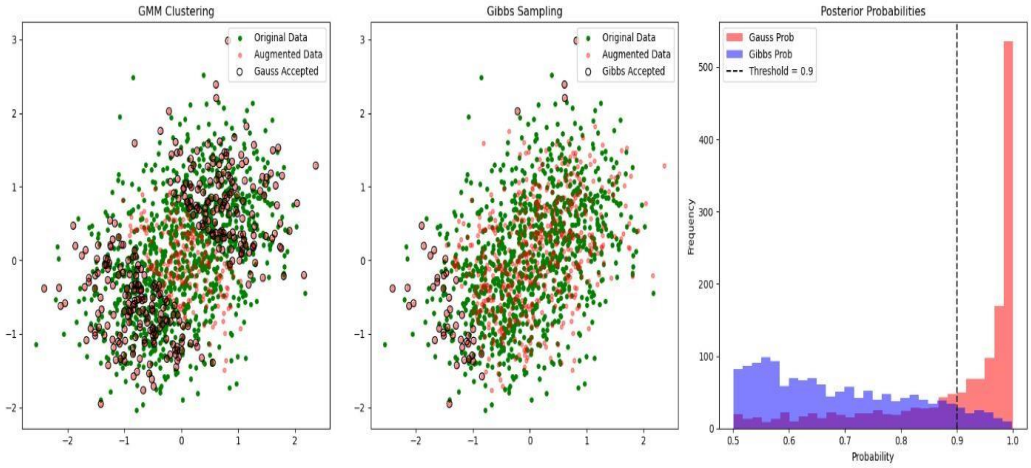


Figure 14: a) Gaussian statistics used in data augmentation, b) Gibbs statistics in data augmentation, c) Probability distribution of augmented data with the threshold 0.9

Figure 13-14 shows 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 13-14 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, Gibbs sampler, covariance updates are simplified (especially when points per cluster are small), which can reduce the accuracy of the likelihood computation.

5. CONCLUSION

The revolution of artificial data is in full swing, and its impact on society and industries will be monumental. As we move further into the digital age, artificial data will not only offer practical solutions to immediate problems such as privacy, regulatory compliance, and enhancing artificial intelligence but also has the potential to transform how we perceive, manage, and use data. In this paper, we reflected on the future of artificial data, the difference between synthetic and augmented data and the choice of various statistics (Gaussian and Gibbs).

Although the concept of artificial data has existed for several years, its widespread adoption and current relevance are experiencing exponential growth. In the future, artificial data will become even more integrated into daily business operations, scientific research, and the development of innovative technologies. From generating data to train artificial intelligence models to simulating

complex scenarios in sectors like healthcare, finance, and automotive, the versatility of synthetic data will enable a revolution in how industries approach their most urgent challenges.

A key aspect of the future of synthetic data will be its ability to ensure privacy. Data protection regulations, such as the General Data Protection Regulation (GDPR) in Europe and other global regulations, will continue to drive the demand for solutions that minimize the risk of exposure to personal data. Synthetic data offers an ideal solution, allowing organizations to work with realistic data without compromising user privacy. Furthermore, the ability of synthetic data to improve the quality of AI models will be crucial.

Today, many AI models require large volumes of labeled data to train effectively. Synthetic data can generate these volumes without the need to collect real-world data, thus enabling access to high-quality information in sectors where real data may be limited or expensive to obtain. Over time, the combination of real and synthetic data will result in more robust and accurate AI models. The impact of synthetic data on technological innovation will also be profound. In areas like autonomous driving, synthetic data creation allows the simulation of real-world scenarios to train vehicles without safety risks. This type of application in the automotive sector is just one example of how synthetic data can reduce development costs and accelerate the progress of new technologies.

The Global Impact of Synthetic Data On a global level, synthetic data has the potential to democratize access to valuable information and enable significant breakthroughs in research, development, and public policy. For example, in the field of healthcare,

Follow Thought Leaders and Experts: As the field of synthetic data continues to evolve, it is crucial to stay informed about the contributions of experts and thought leaders. Following AI researchers and industry professionals on platforms like Twitter, LinkedIn, and Medium will provide deeper insights into emerging trends, best practices, and the potential implications of synthetic data. Some of the most influential names in AI and data science regularly share their discoveries and experiences, which is an excellent learning resource.

Synthetic data and data augmentation are two different approaches to address the challenge of limited data in deep learning. Synthetic data can be used as an alternative to real-world data, while data augmentation can be used to increase the size and diversity of the training data. Both methods can play an important role in the development of effective deep learning models, and the choice of approach will depend on the specific requirements and constraints of the problem at hand. In general, a combination of both synthetic data and data augmentation can provide the best results in deep learning applications. By generating additional, diverse, and representative data, these techniques can help deep learning models to learn more effectively and generalize better to new, unseen data.

The journey towards understanding and implementing synthetic data is just the beginning. As we move forward in technology, synthetic data will continue to play a crucial role in solving complex problems and creating innovative solutions that will improve all aspects of our lives. From protecting privacy to enhancing AI models and fostering scientific research, synthetic data has the power to transform entire industries. It is crucial that we continue to explore, learn, and collaborate to understand how synthetic data can be used ethically and effectively. The future is full of opportunities, and those who invest in studying this technology will be better positioned to take advantage of the breakthroughs to come.

REFERENCES

[1] Maayan Frid-Adar,Idit Diamant,Eyal Klang,Michal Amitai,Jacob Goldberger,Hayit Greenspan,"GAN-based Synthetic Medical Image Augmentation for increased CNN Performance in Liver Lesion Classification", https://doi.org/10.1016/j.neucom.2018.09.013

[2] Philip de Melo, 2025, "Accurate Diagnostics of Lung Cancer Using Prime Model Generative AI", Cancer Research Journal (in press).

[3] Connor Shortenand Taghi M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning", Journal of Big Data, DOIhttps://doi.org/10.1186/s40537-019-0197-0

[4] Yuexi Wang , Nicholas Polson and Vadim O. Sokolov, "Data Augmentation for Bayesian Deep Learning", https://doi.org/10.48550/arXiv.1903.09668, 2022.

[5] Ryan P.J., "Euclidean and non-Euclidean geometry: An analytic approach", Cambridge University Press (1986), Google Scholar

[6] Xu Y., Jia R., Mou L., Li G., Chen Y., Lu Y., et al., "Improved relation classification by deep recurrent neural networks with data augmentation", (2016) arXiv preprintarXiv:1601.03651

[7] Wong S.C., Gatt A., Stamatescu V., McDonnell M.D., "Understanding data augmentation for classification: when to warp?", 2016 international conference on digital image computing: techniques and applications, IEEE (2016), pp. 1-6

[8] Dong X., Potter M., Kumar G., Tsai Y.-C., Saripalli V.R., "Automating augmentation through random unidimensional search", (2021) arXiv preprintarXiv:2106.08756

[9] Milletari F., Navab N., Ahmadi S.-A., "V-net: Fully convolutional neural networks for volumetric medical image segmentation", 2016 fourth international conference on 3D vision (3DV), IEEE (2016), pp. 565571

[10] Simard PY, Steinkraus D, Platt JC, et al. "Best practices for convolutional neural networks applied to visual document analysis", In: Icdar, vol. 3, no. 2003. 2003.

[11] Wang K., Fang B., Qian J., Yang S., Zhou X., Zhou J., "Perspective transformation data augmentation for object detection", IEEE Access, 8 (2019), pp. 4935-4943

[12] Franke M, Gopinath V, Reddy C, Ristić-Durrant D, Michels K. Bounding Box Dataset Augmentation for Long-range Object Distance Estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021, p. 1669–77.

[13] Ronneberger O., Fischer P., Brox T., "U-net: Convolutional networks for biomedical image segmentation", International conference on medical image computing and computer-assisted intervention", Springer (2015), pp. 234-241

[14] Jaderberg M., Simonyan K., Zisserman A., et al." Spatial transformer networks", Adv Neural Inf Process Syst", 28 (2015)

[15] Karargyris A. "Color space transformation network", (arXiv preprintarXiv:1511.01064, 2015

[16] Philip de Melo and Mane Davtyan. "High Accuracy Classification of Populations with Breast Cancer", SVM Approach. Cancer Research Journal, *11*(3), 94-104. https://doi.org/10.11648/j.crj.20231103.1317)Philip de Melo, "Public Health Informatics and Technology", ISBN 13 9798893729535, 2024.