Enhancing Detector Performance through Synthetic Data Generation Using GANs

Abstract: **Particle detectors are essential in high-energy physics and medical imaging, and material sciences, just like their testing and calibration, which is very resource-intensive. The project presented herein, Enriching Detector Performance by Synthetic Data Generation Using GANs, to overcome these challenges using synthesis data generated via the 2D Convolutional Generative Adversarial Networks. Using voxelized particle shower datasets coming from the CERN Open Data Portal, the project transforms energy deposition data into 2D projections for effective synthetic data generation. In this project, the GAN model is trained in order to replicate some key characteristics of the particles' interactions, while statistical metrics on realism will be assessed by comparing synthetic and real data.**

A. INTRODUCTION

Title of the project: Enhancing Detector Performance through Synthetic Data Generation Using GANs

Project Team Members: Sai Varun Nimmagadda [KL37536], Sanjay Varatharajan [UL66332]

In disciplines including high-energy physics, medical imaging, and material sciences, particle detectors are essential instruments that allow researchers to investigate the basic characteristics of matter and energy. However, the expense and scarcity of high-quality experimental data frequently hinder the development, testing, and calibration of these detectors. A viable answer to these problems is synthetic data generation, which provides a scalable and affordable way to supplement experimental datasets while preserving the realism needed for insightful analysis.

Particle detectors are vital instruments that facilitate important discoveries and developments in the fields of material sciences, high-energy physics, and medical imaging. These systems' creation, testing, and calibration, however, are expensive and resource-intensive processes that frequently call for substantial amounts of experimental data. A potential remedy for these issues is synthetic data creation, which simulates real data for detector improvement. Generative Adversarial Networks (GANs) are used in this research, Enhancing Detector Performance via Synthetic Data Generation Using GANs, to produce high-quality synthetic data that accurately simulates particle interactions in the real world. The initiative streamlines processing and model training by converting three-dimensional data into simplified two-dimensional projections using voxelized particle shower datasets from the CERN Open Data Portal. The project intends to lower costs, speed up discovery, increase measurement accuracy, and improve the detection capabilities of commercial and scientific systems by offering a strong framework for creating synthetic data.

B. DATA SOURCES

This project utilizes voxelized particle shower data from the CERN Open Data Portal, specifically curated from the publicly available datasets at CERN Open Data Portal. These datasets include:

Photon Samples:

photon_samples.tgz: Lower-statistics photon events.
photons_samples_highStat.tgz: Higher-statistics photon events.

The datasets capture energy deposits in the ATLAS calorimeter, originally represented as three-dimensional voxel data. To facilitate the generation of synthetic data, this project transforms the 3D voxel data into two-dimensional projections. These 2D representations retain the essential characteristics of particle interactions while simplifying data processing and enhancing the efficiency of model training. A 2D Generative Adversarial Network (GAN) is employed to generate realistic synthetic projections, which replicate key patterns in particle shower events for subsequent analysis and evaluation.

C. RELATED RESEARCH STUDIES

Monte Carlo (MC) simulations are used to model particle interactions in the ATLAS experiment at CERN's Large Hadron Collider (LHC). Approximately 75% of the simulation time is spent simulating the detector's reaction, particularly in the calorimeters. These simulations currently make use of the Geant4 toolset, which yields incredibly precise results but demands a large amount of processing power. The need for speedier simulations has increased as the LHC gets ready for future runs with bigger data rates. Although current technologies such as FastCaloSim and FastCaloSimV2 are designed to expedite this procedure, they still lack precision for some tasks and require further calibration effort.

FastCaloGAN, a machine learning-based solution utilizing Generative Adversarial Networks (GANs), was created to address these issues. FastCaloGAN enhances training stability and performance by utilizing Wasserstein GANs with Gradient Penalty (WGAN-GP), a more sophisticated variant of GANs. Across a range of energies, this tool has effectively replicated detector responses for many particles, including electrons, photons, and pions. It achieves excellent accuracy and faster computation by splitting the detector into slices and training a different GAN for each slice.

My project's goal is to develop a 2D GAN framework for producing artificial projections of particle shower occurrences. I used the above research as my reference. In particular, I used the CERN Open Data Portal's voxelized particle shower. Through the process of transforming three-dimensional voxel data into two-dimensional projections, the method streamlines data organization, improves computing effectiveness, and preserves essential characteristics for detector calibration. The versatility of GAN-based simulation techniques in furthering particle physics research is highlighted by this methodology, which not only adheres to the concepts shown in FastCaloGAN but also adapts them to the project's practical and educational restrictions.

D. MODELING APPROACH

This project's modeling strategy uses a 2D Generative Adversarial Network (GAN) to create realistic-looking synthetic data from datasets of voxelized particle showers. A discriminator and a generator are the two main parts of the GAN framework. These elements compete with one another to create synthetic data that closely mimics the original data. Below is a description of the procedure:

1. Data Preprocessing

Preprocessing is used to convert the 3D voxel data into 2D projections from the raw voxelized particle shower datasets. This simplifies the intricate three-dimensional structure into a more understandable format while maintaining crucial aspects of the particle interactions. The data is better suited for analysis thanks to this transformation, which also lowers the processing overhead. Next, the data is standardized to remove any potential biases, guarantee consistent scaling across all features, and improve the stability and effectiveness of the GAN training procedure. The preparation of high-quality input data that enables precise and trustworthy synthetic data synthesis depends on these procedures.

2. GAN Architecture

Generator : A crucial aspect of the GAN framework, the generator is made to produce 2D artificial visuals that closely mimic actual particle shower projections. It starts with a random noise vector as input and uses a number of convolutional layers to turn it into structured data. The generated images closely resemble the properties of real particle showers because these layers are tuned to capture the fine spatial patterns and nuances found in the original data.

Discriminator: The discriminator functions as a classifier that evaluates the authenticity of the input data. It uses convolutional layers to extract and analyze features from the input, distinguishing between real and synthetic images. By providing feedback on the quality of the generated data, the discriminator helps refine the generator's performance, fostering an adversarial learning process that drives the GAN to produce increasingly realistic synthetic data.

3. Training Procedure

Using a Generative Adversarial Network (GAN), this project tackles the problem of producing synthetic data for particle physics simulations by producing artificial 2D projections of particle shower data. The generator and the discriminator are the two primary parts of the GAN architecture. The generator uses convolutional and fully linked layers to capture complex spatial patterns and converts random noise vectors into realistic 2D visuals that resemble actual particle shower data. The discriminator processes characteristics through fully linked layers with regularization and activation functions to determine if

the input data is synthetic or real, functioning as a binary classifier. These elements work together in an adversarial process to iteratively enhance the synthetic data's quality.

To guarantee efficiency and stability, the training procedure incorporates multiple techniques. In addition to Adam optimizers with adjusted learning rates that provide balanced updates to both networks, Binary Cross-Entropy Loss directs the generator and discriminator. In order to prevent overfitting and promote better convergence, learning rate schedulers—such as a ReduceLROnPlateau scheduler and a step decay—adjust the learning rates according to performance indicators. In order to preprocess data, guarantee consistent batch sizes, resize data as necessary, and enable effective training on big datasets, a dynamic batch loader is utilized.

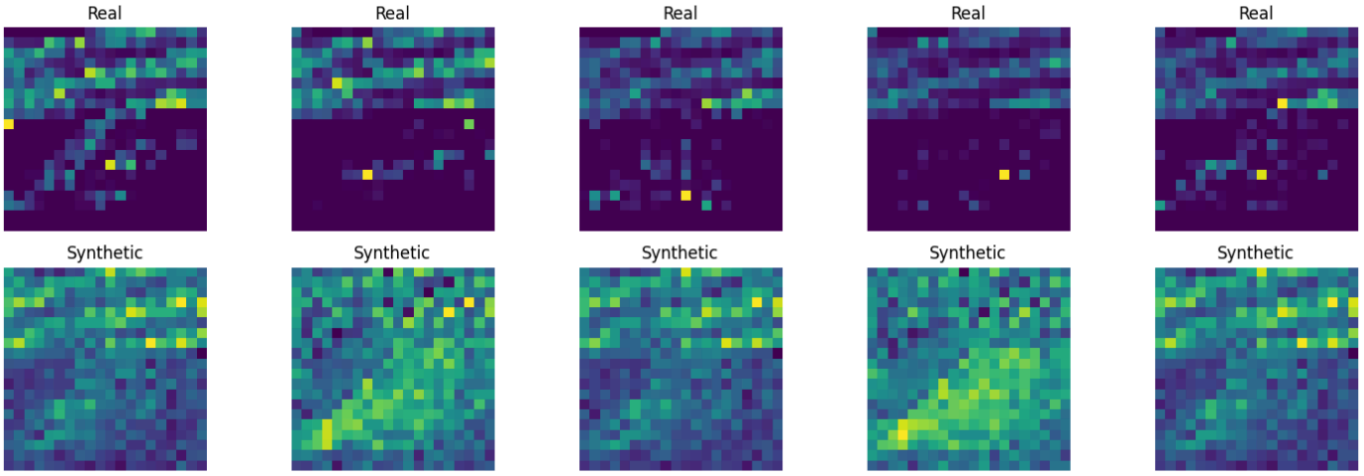
4. *Evaluation*

Three main measures were used to assess the synthetic data produced by the trained Generative Adversarial Network (GAN): Wasserstein Distance, Mean Squared Error (MSE), and Structural Similarity Index (SSIM). Batches of actual data were first loaded from the designated directory, preprocessed, then concatenated to the necessary size and shape for comparison. At the same time, randomly sampled latent points were used to create synthetic data of the same size using the GAN's generator. To avoid inconsistencies due to hardware differences and guarantee consistency in computations, both synthetic and actual datasets were loaded onto the same computing device.

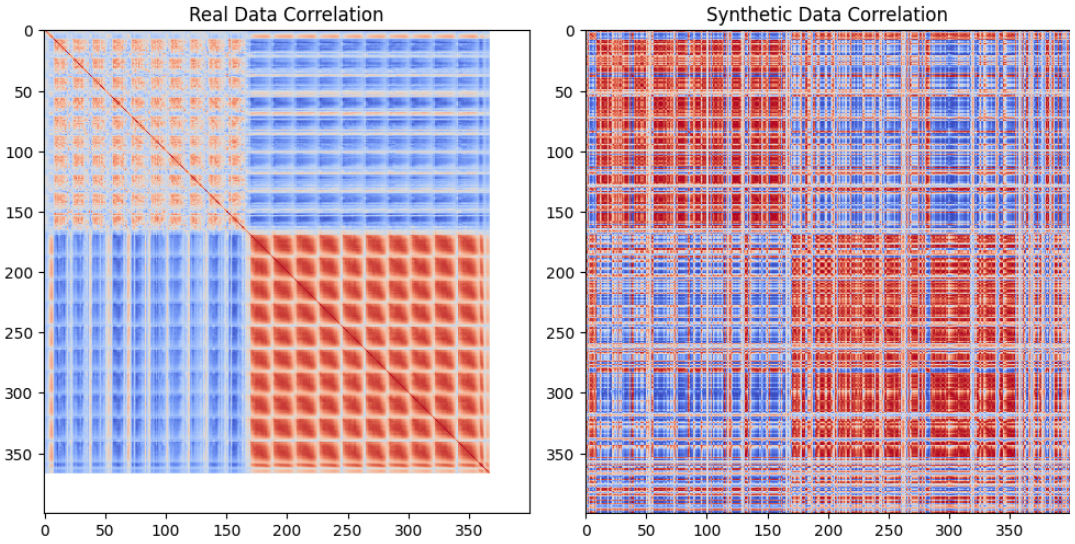
To evaluate the quality of the generated data, three metrics were calculated for each sample in the datasets. MSE calculated the average squared difference between the synthetic and real samples to offer a pixel-wise similarity metric. SSIM was utilized to assess perceptual similarity, concentrating on structural differences through data analysis of the first channel. The distributional alignment between real and synthetic data samples was also measured using the Wasserstein Distance, after both datasets were flattened into a one-dimensional format for compatibility.

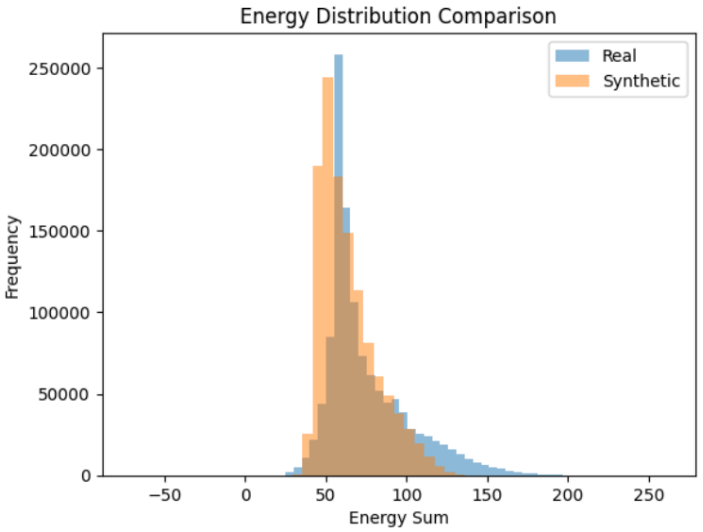
In order to provide a thorough evaluation of the generator's performance, the outcomes for each parameter were then averaged across all samples. Combining these criteria allowed the evaluation method to capture several dimensions of similarity, such as distributional alignment, structural coherence, and pixel-level precision, and it produced a reliable study of how well the synthetic data mirrored real-world traits. The identification of places where the generator could be further optimized is made possible by this comprehensive evaluation framework.

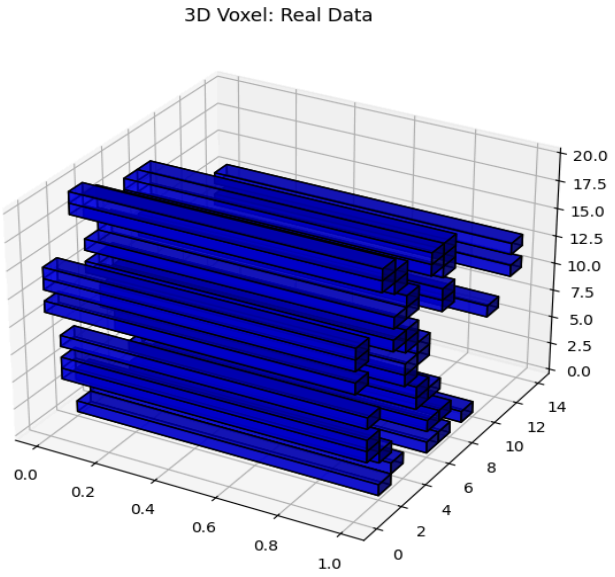
5. *Results and Output*

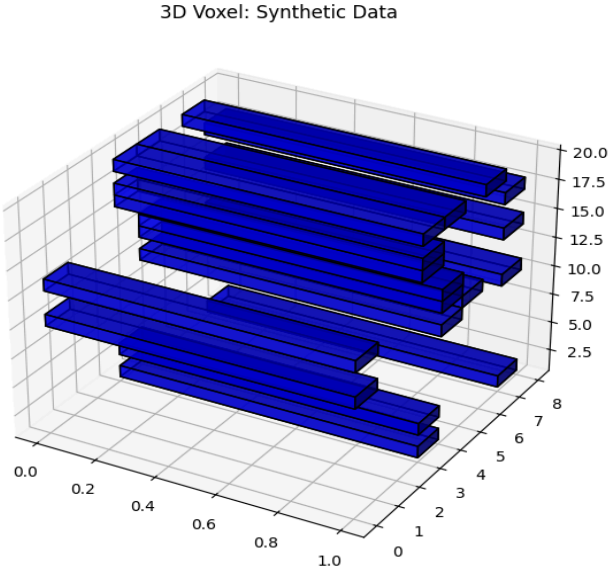


DATA 602 PROJECT









Summary of Results - The evaluation's findings offer important new information on how well the GAN model performs while creating artificial data. Pixel-wise similarity was measured by the Mean Squared Error (MSE), which came out at 0.0685. This suggests that the model did a respectable job of capturing broad patterns in the actual data. Nonetheless, the Structural Similarity Index (SSIM) was comparatively low at 0.2366, indicating difficulties in reproducing spatial coherence and finer structural details. With a distributional alignment measurement of 0.0979, the Wasserstein Distance indicated a reasonable approximation of the overall distribution of the real data. These results were corroborated by visualizations including heatmaps, energy distributions, and 3D voxel plots, which revealed differences in finer structural elements but high alignment in general patterns.

Interpretation and Observations - The comparatively low SSIM suggests that the model had trouble capturing more intricate spatial relationships, even if the synthetic data generally matches the real data quite well. Visual examinations of voxel architectures and correlation matrices show that while the GAN was successful in recreating general trends, it was not as accurate in reproducing specific feature interdependencies. These findings emphasize both the GAN's limits in terms of

structural fidelity and its ability in simulating broad patterns. These findings suggest that in order to increase structural correctness and perceptual similarity, the generator and discriminator need to be further optimized.

E. CONCLUSION and FUTURE SCOPE

The study's findings show that GAN-based models may produce synthetic data that closely resembles real data in terms of general trends and overall distribution. Visualizations demonstrate the model's efficacy in reproducing general trends, while metrics like Mean Squared Error (MSE) and Wasserstein Distance show how well it matches the pixel-level and distributional features of the actual data. Visual differences in finer structural details and the comparatively low Structural Similarity Index (SSIM) point to the model's limits in capturing subtle features and complicated spatial linkages.

Future studies should concentrate on improving GAN designs, adding sophisticated strategies like attention mechanisms, and using bigger and more varied datasets to increase generalization in order to overcome these difficulties. Adopting more thorough evaluation frameworks and domain-specific optimizations will also guarantee that the data produced is appropriate for tasks unique to the application and realistic. By offering a viable substitute for conventional data gathering and simulation techniques, these advancements will greatly increase the usefulness of synthetic data in both scientific and practical fields.

Future Scope - In the Future, we should research sophisticated GAN designs, such as conditional GANs or Wasserstein GANs with gradient penalty, which could boost distribution alignment and structural coherence and increase the fidelity and application of synthetic data generation. The SSIM may be improved by improving focus on important characteristics by the addition of attention mechanisms to the GAN. Increasing or supplementing the dataset may also provide the model access to a wider range of patterns, which would enhance generalization. Furthermore, the generator might be able to capture finer features if regularization techniques like feature matching or perceptual loss are used. A more thorough evaluation of the produced data would be provided by adding more evaluation measures, such as Fréchet Inception Distance (FID). These developments would expand the uses of synthetic data in scientific inquiry while simultaneously enhancing its quality.

F.. *REFERENCES*

[1] The ATLAS Collaboration. (2020, November 27). *Fast simulation of the ATLAS calorimeter system with Generative Adversarial Networks* (ATL-SOFT-PUB-2020-006). CERN. https://cds.cern.ch/record/2746032/files/ATL-SOFT-PUB-2020-006.pdf