Synthetic Data Generation using GANs

Student Name: Nishchay Krishnappa **Reg. No:** 42308956 **Course:** PCS Project 1

Date: October 22, 2025 **Project Repository:** github.com/nishchaykrishnappaiu/synthetic-data-gan

Problem Statement

The scarcity of accessible, high-quality training data due to privacy constraints, collection costs, and limitations of conventional augmentation methods creates an urgent need for scalable synthetic data generation that preserves semantic richness and underlying correlations across diverse modalities.

Abstract

This project proposes a unified Generative Adversarial Network (GAN) framework for generating high-quality synthetic datasets for both structured (tabular) and unstructured (text, image, audio) data, based on user-defined requirements. The system combines multiple GAN architectures—CTGAN for tabular data, conditional GAN and diffusion-based models for images, and transformer–GAN hybrids for text—under a modular and configurable pipeline. A user requirement parser enables conditional data generation, allowing users to specify schema constraints or descriptive prompts. The framework integrates comprehensive evaluation metrics for fidelity, diversity, and privacy, ensuring that the generated data remains statistically similar yet non-identical to the original dataset. This system facilitates safe data sharing, research reproducibility, and bias-free model training across domains.

Proposed Solution

The proposed pipeline consists of four main modules:

1. **Preprocessing Layer:** Handles feature scaling, encoding, and normalization for structured and unstructured inputs.
2. **Generation Layer:** Implements CTGAN for tabular data and StyleGAN/TextGAN variants for unstructured data, with support for conditional inputs.
3. **Evaluation Layer:** Assesses generated data using statistical similarity (KS-test, correlation metrics), perceptual quality (FID, BLEU), and privacy measures (membership inference tests).
4. **User Interface Layer:** A REST API or Streamlit-based dashboard that allows users to define requirements and retrieve generated datasets and performance reports.

Expected Outcome

The project aims to deliver an end-to-end synthetic data generation framework that:

- Generates realistic structured and unstructured data aligned with user constraints.
- Preserves statistical distributions and data relationships while safeguarding privacy.
- Provides automated evaluation reports quantifying data fidelity and diversity.
- Offers a deployable and scalable solution for researchers and developers to obtain customizable synthetic datasets for model development and testing.