Project Report

Effectiveness of Data Synthesized by Generative AI in Supervised Image Classification

Prepared By:

Group 9: Steffi Dorothy, Diego Pena-Stein, Ehu Shubham Shaw, Bashir Gulistani

Prepared For:

Yanhua Li

CS 539 Machine Learning

Motivation

Collecting large image datasets is expensive and time-consuming, but essential for machine learning model performance. Prototyping datasets from limited or no data is useful for training and testing models without a big investment. Expanding existing datasets, like adding European street signs to a U.S. dataset, can also be beneficial.

The motivation for our project, "Effectiveness of Data Synthesized by Generative AI in Supervised Image Classification," includes:

Enhancing Model Performance: Evaluating if synthetic data can improve supervised image classification, especially when large datasets are hard to obtain.

Cost and Resource Efficiency: Exploring the cost-effectiveness of synthetic data generation compared to real image collection and annotation.

Application in Various Domains: Investigating the broader use of synthetic data techniques in fields like medical imaging and rare species identification.

Advancement in Generative AI: Contributing to the development of generative AI by exploring its real-world applications and limitations.

Problem Definition

This project explores the feasibility and effectiveness of using generative models to enhance supervised image classification. The goal is to improve existing datasets with synthetic data to boost model performance on unseen data, particularly in image classification tasks where generative models can produce realistic images with specific class labels.

In supervised image classification, large, high-quality datasets are essential but often expensive and time-consuming to obtain. Generative AI techniques like Stable Diffusion can help address this challenge by synthesizing diverse, realistic image data.

The project, "Effectiveness of Data Synthesized by Generative AI in Supervised Image Classification," aims to:

- **Evaluate Performance**: Assess how synthetic images impact model performance.
- **Identify Advantages and Limitations**: Explore the benefits and drawbacks of using synthetic data.

The project seeks to determine if synthetic data can reduce dependency on large real datasets, providing a practical solution for image classification tasks.

Example Scenario: Synthetic-Only Training

In cases where real images are impractical or impossible to obtain, this project will explore training a model entirely on synthetic data. For instance, a CNN trained only on synthetic fruit images will be tested to see if it can classify real fruit images, showcasing the viability of synthetic data.

Dataset Description

The dataset consists of fruit images for supervised classification, divided into two groups: real and synthetic images.

Real Images

1. **Source**: The real images are from the publicly available "Fruit Images for Object Detection" dataset on Kaggle, featuring high-resolution fruit images with varied lighting and backgrounds.
2. **Content**: The dataset includes a variety of fruits like apples, bananas, and oranges, with multiple images per category to ensure diversity.
3. **Quantity**: The real dataset contains around 300 images, with roughly 100 images per fruit category.
4. **Annotations**: Each image is labeled with the corresponding fruit type for supervised learning.

Synthetic Images

1. **Generation Method**: Synthetic images are generated using a generative AI model, like Stable Diffusion, trained on real fruit images to replicate their patterns and characteristics.
2. **Content**: The synthetic dataset includes a variety of fruit images, closely resembling real ones in shape, color, texture, and other features.
3. **Quantity**: The synthetic dataset contains about 4,400 images.
4. **Annotations**: Each synthetic image is labeled with the corresponding fruit type based on the generation parameters, matching the labels of the real images.

System Description

Data Splits

To evaluate the effectiveness of synthetic data, the dataset is split into training and test sets with varying ratios of real and synthetic images:

Training Set: Includes combinations of real and synthetic images, such as:

- ★ 100% real
- ★ 50% real, 50% synthetic
- ★ 30% real, 70% synthetic
- ★ 100% synthetic

Validation Set: Contains a balanced mix of real and synthetic images.

Examples

- **Example 1 (Real Image)**: An image of a ripe banana taken in natural light, with a clear background and labeled as "banana."
- **Example 2 (Synthetic Image)**: An AI-generated image of an apple, resembling a real apple in terms of color and texture, labeled as "apple."

Data Preprocessing

1. **Resizing**: All images are resized to a uniform dimension (e.g., 256X256 pixels) to ensure consistency in model training.
2. **Normalization**: Pixel values are normalized to a range of [0, 1] to facilitate better convergence during training.

Synthetic Data Generation Module

1. Generative AI Model:

- The Stable Diffusion model is used to generate synthetic fruit images.
- The model is trained on real fruit images to learn their characteristics and patterns.

2. Image Generation:

- Synthetic images are created by specifying the desired fruit type.
- The generated images are stored with appropriate labels for use in training and evaluation.

Model Training Module

1. CNN Architecture:

- A convolutional neural network (CNN) architecture is designed and implemented using a deep learning framework such as TensorFlow.
- The architecture consists of several convolutional layers, pooling layers, and fully connected layers.

2. Training Configuration:

- The training set is prepared with various combinations of real and synthetic images (e.g., 100% real, 50% real and 50% synthetic).
- Hyperparameters, including learning rate, batch size, and the number of epochs, are set, and configured.

3. Training Process:

- The model is trained on prepared datasets using the configured hyperparameters.
- During training, the model learns to classify the images by minimizing the loss function.

Performance metrics such as accuracy, precision, recall, ROC Score and F1-score are computed to assess the model's performance.

Literature review:

https://ar5iv.labs.arxiv.org/html/2407.00116

This paper focuses on generating synthetic data for medical applications, although inside the medical field it attempts to be multi-modal. This includes images, text, time series, and tabular data formats. It attempts to replicate these with Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), Diffusions Models (DMs), and Large Language Models (LLMs). It explains some of the reasoning behind synthesizing data in the medical field is to augment existing datasets, as well as helping protect patient privacy and anonymity.

The researchers use a variety of data formats in the health field. Electronic health records, clinical text notes, electrocardiogram (EKG) to measure the electrical activity of the heart, dermatoscopic images, mammographic images, ultrasound scans, computed tomography (CT) scans, magnetic resonance imaging (MRI), optical coherence tomography (OCT), and x-rays are all used as mediums in the study. Some outlined reasons that datasets in the medical field can be smaller are duplicate records, wrong language, and having the report request denied. Of the initial 1674 records across 12 data types requested, only 249 were used in the final training of ML models.

The generative models pertinent to our project are split into two categories: Unconditional and conditional. Unconditional generation takes random noise as an input and synthesizes data, which is not useful in our case. Conditional generation, on the other hand, uses noise along with external information or context to generate data.

Two models were used to generate image data: Generative adversarial networks and diffusion models. A Generative Adversarial Network (GAN) is composed of two components: a generator and a discriminator. Both are trained simultaneously to outperform each other, with the generator creating synthetic data that mimics real data as best it can, and the discriminator guessing if data is presented is real or synthetic. While normal GANs are unconditional, a variant of conditional GAN is used to incorporate semantic input into the generation process.

Diffusion models create image data by incrementally adding noise to data and then training a deep neural network to recover the data. Eventually the model is given complete random noise

and iteratively creates data from the random noise. Like a GAN, standard diffusion models are unconditional. However, a variant called stable diffusion uses text input (or embedded vector) to help condition the output. These models are especially useful because they are pre-built for varied use but can also be fine-tuned to a specific dataset.

https://ar5iv.labs.arxiv.org/html/2302.04062v6

Machine Learning for Synthetic Data Generation: A Review, the work I evaluated, explores the use of machine learning-generated synthetic data to address a number of data-related issues. Because real-world applications frequently struggle with data scarcity, privacy concerns, and quality difficulties, synthetic data is a desirable substitute. The study offers a thorough review of synthetic data generation techniques and technologies, including Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), which are extensively employed in domains such as finance, computer vision, healthcare, and natural language processing (NLP).

The use of synthetic data in delicate domains, particularly in the healthcare industry where patient privacy is crucial, is one of the main topics the article examines. GANs may provide realistic medical records and imaging data while maintaining anonymity, as demonstrated by techniques like MedGAN and MMCGAN. Additionally, synthetic data has been helpful in improving model accuracy without running the risk of data privacy violations when training AI models for situations where actual data is scarce.

Fairness and privacy in synthetic data continue to be significant obstacles. The study emphasizes the need for strategies like federated learning and differentiated privacy to lower the possibility of sensitive data leaks. It also highlights how crucial it is to produce fair data that avoids reproducing biases from the source datasets.

In summary, this study recognizes the continuous difficulties in data quality, privacy protection, and fairness while highlighting the promise of synthetic data to further AI research and applications, particularly in regulated industries. These findings imply that future studies should concentrate on creating more reliable assessment techniques for artificial data and investigating sophisticated privacy-preserving models. Machine Learning for Synthetic Data Generation: A Review, the work I evaluated, explores the use of machine learning-generated synthetic data to address data-related issues. Because real-world applications frequently struggle with data scarcity, privacy concerns, and quality difficulties, synthetic data is a desirable substitute. A thorough review of synthetic data generation techniques and technologies, including Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), is given in the study.

https://link.springer.com/article/10.1007/s11042-023-15747-6

The paper titled "Generative Adversarial Network based Synthetic Data Training Model for Lightweight Convolutional Neural Networks" by Ishfaq Hussain Rather and Sushil Kumar reviews the challenges and solutions related to inadequate training data for deep learning models. It emphasizes the limitations of traditional methods like data augmentation and transfer learning, which can only partially address issues of data scarcity.

The authors propose the GAN-ST model, a novel approach that leverages Deep Convolutional Generative Adversarial Networks (DCGAN) and Conditional Generative Adversarial Networks (CGAN) to generate synthetic training data. By training these GANs independently, the GAN-ST model produces diverse and realistic data, which enhances the training of lightweight Convolutional Neural Networks (CNNs).

The paper evaluates this model on the MNIST and CIFAR-10 datasets, achieving high classification accuracy—99.38% and 90.23%, respectively—comparable to models trained on original datasets. The literature review within the paper outlines various GAN-based data augmentation techniques, highlighting their effectiveness in improving classification accuracy in fields such as medical imaging and object detection.

It also discusses the use of GANs for generating synthetic data to address privacy and security concerns, suggesting ensemble methods to avoid overfitting. The results demonstrate that CNNs trained with GAN-ST synthetic data perform better and generalize more effectively than those trained with traditional methods. Overall, the GAN-ST model represents a significant advancement in synthetic data generation, offering a robust solution to the challenges of limited and sensitive data in deep learning applications.

The paper presents experimental results demonstrating the effectiveness of the GAN-ST model. The synthetic data generated by the model not only improves the classification accuracy of lightweight CNNs but also ensures better generalization to new data. The paper compares the performance of the GAN-ST model with several recent data synthesis techniques. The GAN-ST model outperforms many existing methods, achieving higher accuracy and demonstrating the potential of GAN-based synthetic data generation in improving deep learning model performance.

The GAN-ST model effectively addresses the challenge of inadequate training data by generating diverse and realistic synthetic data. The model's ability to enhance data diversity and coverage makes it a valuable tool for training deep learning models, particularly in scenarios with limited or sensitive data.

Key Technologies Used

This project utilizes several advanced technologies:

Generative AI Models

Stable Diffusion: A generative model used to synthesize realistic fruit images by learning from real fruit data.

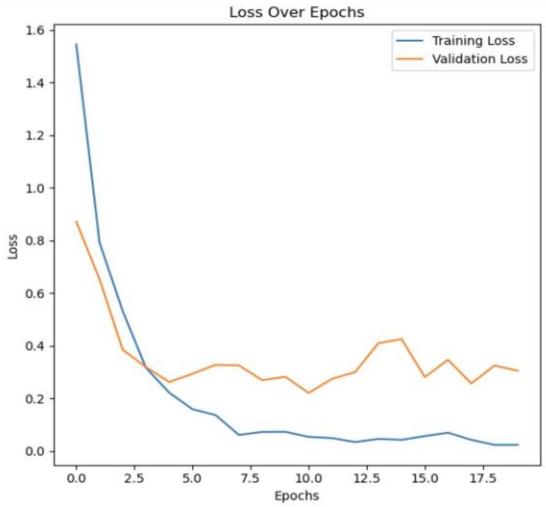
Deep Learning Frameworks

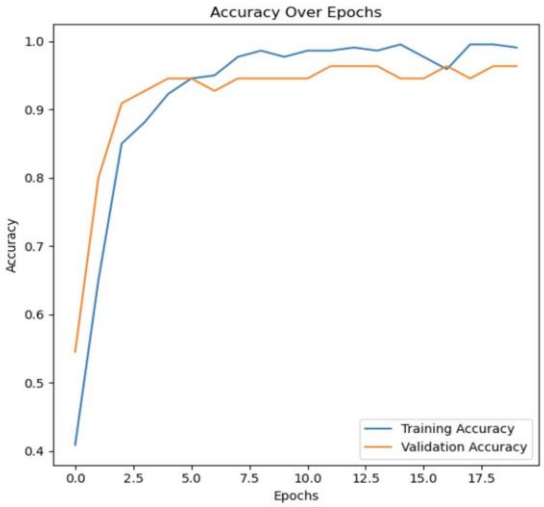
TensorFlow: An open-source framework for building and training convolutional neural networks (CNNs).

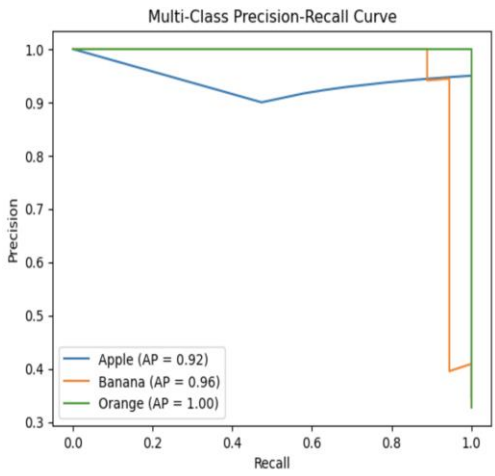
PyTorch: A deep learning framework known for its dynamic graph, ideal for research and development in CNNs.

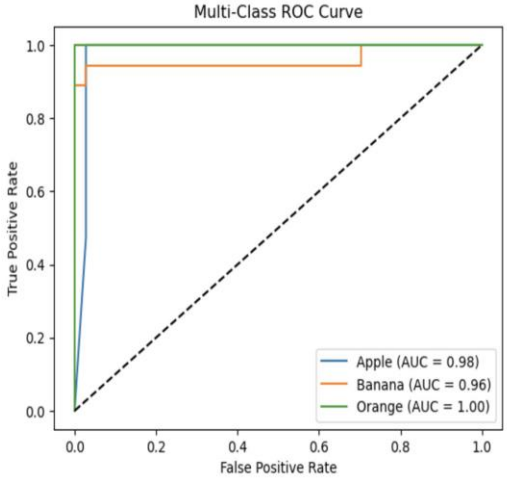
Evaluation

Model1: Real Images 100%



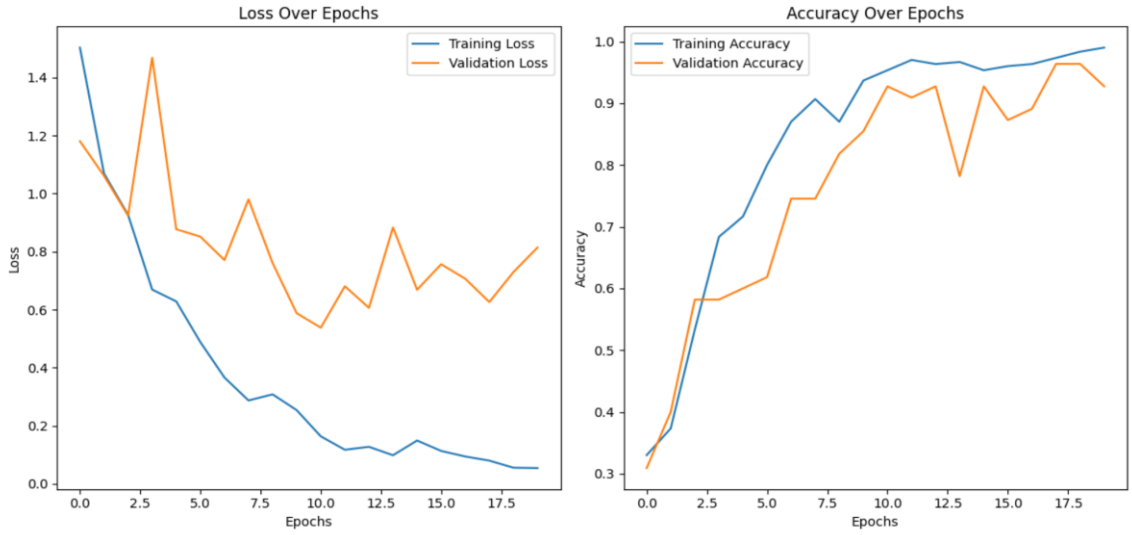


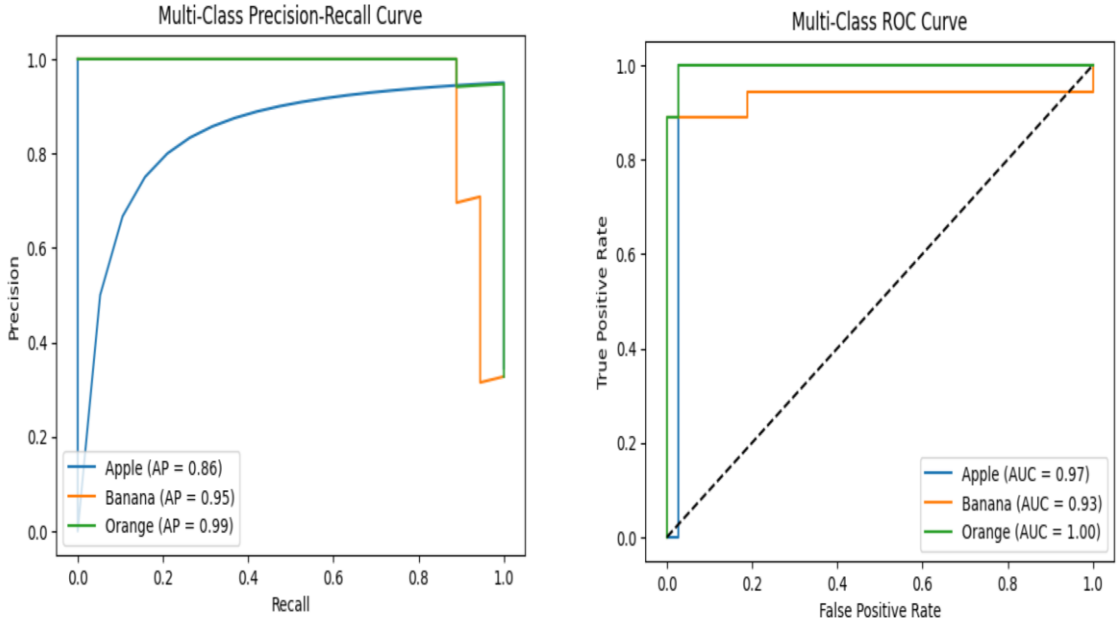




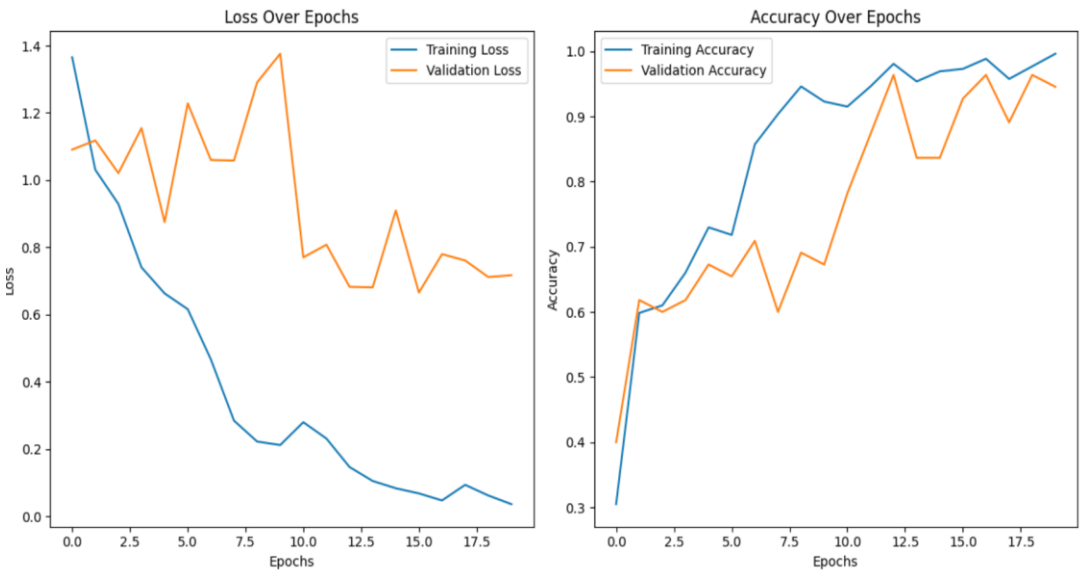

```
Evaluation Metrics:
Accuracy: 0.9636
Precision: 0.9655
Recall: 0.9636
F1-score: 0.9630
ROC-AUC: 0.9865
```

Model2: Synthetic Images 100%



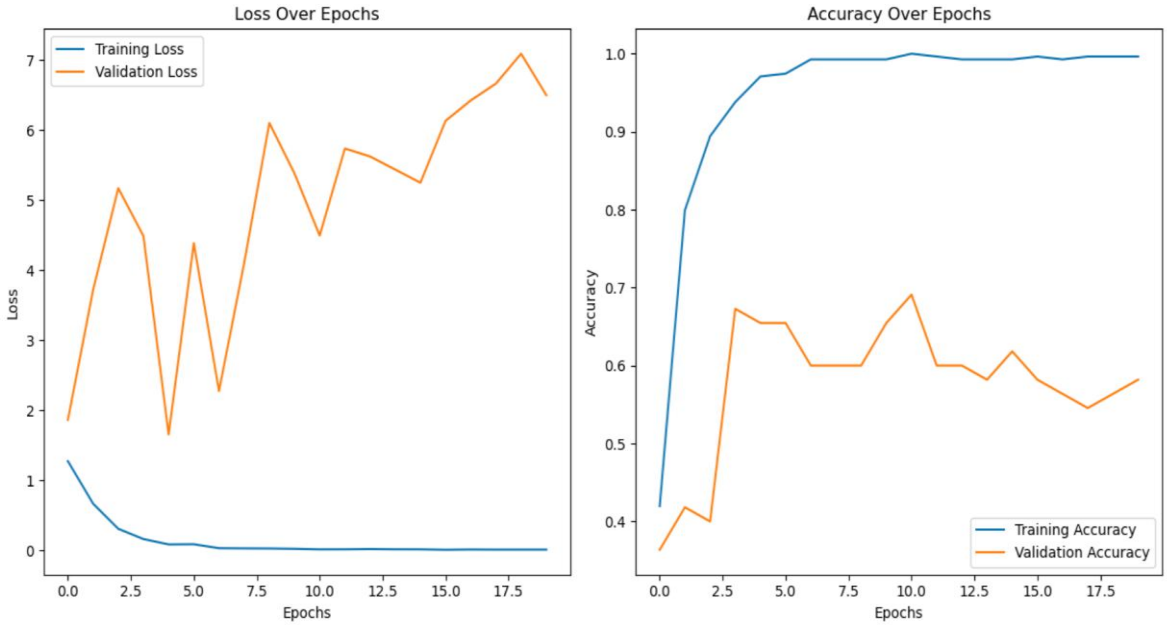


Model3: Real Images 50% Synthetic Images 50%



-

Model4: Real Images 30% Synthetic Images 70%




```
Evaluation Metrics:
Accuracy: 0.7455
Precision: 0.8243
Recall: 0.7455
F1-score: 0.7044
ROC-AUC: 0.9161
```

Resources:
https://github.com/CompVis/stable-diffusion

Papers:
https://ar5iv.labs.arxiv.org/html/2407.00116

- https://ar5iv.labs.arxiv.org/html/2302.04062v6
- https://link.springer.com/article/10.1007/s11042-023-15747-6
- https://arxiv.org/abs/2308.12453 (inpainting and outpainting?)
- https://www.sciencedirect.com/science/article/pii/S0933365723000702?via%3Dihub