Journal of the Midwest Association for Information Systems

Volume 2025 | Issue 2 Article 3

Date: 07-01-2025

Leveraging Synthetic Data from Generative Models for Snow Detection in Data-Scarce Environments

Ricardo de Deijn
Minnesota State University, Mankato, ricardodedeijn@gmail.com

Rajeev Bukralia
Minnesota State University, Mankato, rajeev.bukralia@mnsu.edu

Abstract

Data scarcity poses a significant challenge for training robust machine learning models in safety-critical applications like snow detection, where real-world data collection is often limited and seasonal. This study explores the potential of synthetic data sets generated by prompt-based image synthesis models to enhance machine learning applications in such data-scarce environments. Using OpenAI's DALL·E 3 and xAI's Aurora, synthetic images of snowy and clear sidewalks were compared against a real-world data set for training image-classification models. The findings reveal that an Aurora-based model achieved the highest F2 scores, excelling in snow detection because of its high photorealism and contextual relevance. However, the real-world data set demonstrated greater accuracy in detecting clear sidewalks, resulting in fewer overall classification errors. These results highlight the potential of synthetic data to supplement real-world data sets, particularly in data-scarce domains, while also emphasizing that real-world data remains crucial for balanced classification. This research underscores the necessity for advancements in generative models to more effectively capture complex environmental conditions and improve the generalizability of AI-generated data sets for scalable and practical machine learning applications.

Keywords: prompt-based image synthesis; synthetic images; snow classification; real-world data

DOI: 10.17705/3jmwa.000095

1. Introduction

Since the increased popularity of generative artificial intelligence (Gen AI) started with the introduction of OpenAI's ChatGPT-3.5 in late November 2022 and following models like ChatGPT-4o, xAI's Grok, Meta's Llama, and Google's Bard / Gemini, numerous fields have been largely impacted in their daily processes. These large Gen AI models have been shown to be powerful in creating text similar to how a human would write these (Bandi et al., 2023; Chen et al., 2023). This phenomenon resulted in a large increase in work productivity, as they are now widely used as productivity enhancers (Haan, 2023; Shaji George et al., 2023; Valeriya et al., 2024) helping employees optimize and automate parts of (repetitive) processes or improving decision-making (Dwivedi et al., 2023; Shaji George et al., 2023; Valeriya et al., 2024). The increase in accuracy that led to the rise of these generative artificial intelligence models can be attributed to numerous factors, such as the invention of how to train using parallelization with large numbers of graphics processing units (GPUs) (Atallah et al., 2023; Bandi et al., 2023; Vaswani et al., 2017), the increase in computing power that a GPU can output within a certain timeframe (Bandi et al., 2023). The increased computing efficiency of the models (Bandi et al., 2023), and the increased high-quality internet data sets that could be used for training (Atallah et al., 2023; Bandi et al., 2023). But similar to these text generators, a surge of high-resolution image Gen AI models was observed in the same time frame, with models such as OpenAI's DALL·E, xAI's Aurora, and Google's Imagen (Meng et al., 2023; Zhou et al., 2023). Showcasing an increase in image synthesis capabilities, trying to close the gap between generated images and real-world pictures (Bandi et al., 2023; Meng et al., 2023; Zhou et al., 2023).

Although applications for text-based Gen AIs are largely being developed by businesses and academia, potential applications for image synthesis models are opening as well. This study explores using synthetic data generated by these models to overcome data scarcity, which is a common issue in many domain-specific applications. If successful, this approach could be applied to other scenarios where data is hard to collect. For example, detecting rare events (like certain weather phenomena or accidents) could similarly benefit from generative data (Chang et al., 2020; Elfeki et al., 2017). When an image synthesis model can accurately mimic the data required to train artificial intelligence (AI) models in these data-scarce sectors, the challenges of imbalanced or small data sets will shrink and increase the potential of model convergence, improving its results (D.-C. Li et al., 2017). Current state-of-the-art (SOTA) models like DALL·E and Aurora show a possibility to accurately generate images using textual descriptions that the generator tries to simulate with very minimal costs. It shows that if these models create images indistinguishable from real-world pictures, training data sets can be expanded significantly for costs far cheaper than manual data collection and annotation (Arif & Mahalanobis, 2021; Barbosa et al., 2018).

This study investigates the effectiveness of two prominent prompt-based synthetic image generators, DALL·E 3 and Aurora, in creating training data for a snowy sidewalk image classification task, benchmarking their performance against real-world images. Employing an image classification architecture from de Deijn (2024), the research evaluates three data sets: 500 real-world images of snowy and clear sidewalks, and two synthetic data sets of 500 images each, generated by OpenAI's DALL·E 3 and xAI's Aurora. The choice of DALL·E 3 and Aurora over other popular models like Midjourney or Stable Diffusion was deliberate: DALL·E 3 was selected for its advanced prompt fidelity and ability to generate highly realistic and consistent snowy and clear sidewalk scenes, while Aurora was chosen for its capacity to create diverse, contextually relevant synthetic images, aligning with the study's focus on viable real-world data substitutes. The primary aim is to assess how well models trained on synthetic data can differentiate snow and clear sidewalks in real-world conditions compared to a model trained on real-world images, with all models evaluated on unseen data. We hypothesize that synthetic image generators, such as DALL·E 3 and Aurora, do not yet produce data sets sufficient to match the performance of real-world data in this task, while exploring the alternative possibility that these synthetic data sets are on par with real-world data in terms of model training objectives. This investigation examines the potential of synthetic data sets as practical alternatives to real-world data for training machine learning models, particularly in cases like snow detection, where gathering extensive real-world data sets is often impractical. Performance is assessed using multiple metrics, including precision, recall, accuracy, Area Under the Receiver Operating Characteristic Curve (ROC AUC), and the F2-score, which prioritizes recall over precision to emphasize the importance of detecting snow for safety-critical applications. These comprehensive metrics provide a robust basis for comparing the models' performance in this context and testing the proposed hypothesis.

The remainder of this paper is organized as follows. Section 2 reviews related work, tracing the historical progress of generative AI from generative adversarial networks (GANs) to diffusion models and discussing their applications in synthetic image generation. Section 3 details the methods, including the data set creation process, model training, and evaluation metrics used to compare synthetic and real-world data. Section 4 presents the experimental results, analyzing the performance of models trained on DALL·E 3, Aurora, and real-world data sets, with statistical comparisons to test the hypotheses. Section 5 concludes with key findings, highlighting the potential and limitations of synthetic data for snow detection. Section 6 discusses challenges encountered, such as prompt dependency and data set diversity, and suggests directions for future research. Appendices provide additional details, including confusion matrices for model performance.

2. Related Work

Generative artificial intelligence (Gen AI) is a concept referring to deep-learning models that can generate new-like, meaningful content. This content can include multiple modalities, such as text, images, audio, and video (Bandi et al., 2023; Feuerriegel et al., 2024). These models are designed to mimic human creativity by learning patterns and structures from large data sets, enabling them to create outputs that resemble real-world data or entirely imaginative constructs (Creswell et al., 2018; Donahue et al., 2016; Dumoulin et al., 2016; Feuerriegel et al., 2024; Goodfellow et al., 2014; Mescheder et al., 2017; Mirza & Osindero, 2014; Nichol & Dhariwal, 2021; Ramesh et al., 2021; Sohl-Dickstein et al., 2015; Zhou et al., 2023). Gen AI has seen widespread adoption in various domains, such as automated content creation, image synthesis, natural language processing, music composition, and video generation (Bandi et al., 2023; Zhang et al., 2023).

Understanding the historical progress of Gen AI is essential to appreciating its current capabilities and challenges. The development of key frameworks, from foundational generative adversarial networks (GANs) to more recent diffusion models, reveals the iterative process through which researchers have enhanced the quality, stability, and applicability of generative models. This historical perspective not only highlights the technical milestones achieved but also underscores the limitations that continue to drive innovation in this space and where we are right now.

2.1 Historical Progress of Generative AI

Generative AI has been in development for multiple decades, but its significance surged with the introduction of generative adversarial networks (GANs) (Bandi et al., 2023; de Deijn et al., 2024; K. Wang et al., 2017), as originally proposed by Goodfellow et al. (2014). GANs create new synthetic data using an example training data set built of real-world or authentic data and minimax two-player game concepts using fully connected neural networks. This concept assumes that there are two components: component A, a generator block, and component B, a discriminator block. The generator block's mission is to generate new data that mimics the distribution of the original real-world data, while the discriminator block tries to distinguish whether the data shown is from the original data set or from the generator. As each component wants its score to be as good as possible, it tries to outperform the other component during its training. This adversarial process should *theoretically* lead to a point in which neither component can improve its score, as the data achieves a high level of realism that the discriminator can't identify synthetic data from authentic data (Creswell et al., 2018; Goodfellow et al., 2014; K. Wang et al., 2017).

Early implementations of GANs performed well on simple data sets such as MNIST and CIFAR-10 but struggled with more complex scenes due to limited understanding of image semantics and spatial relationships (Creswell et al., 2018; de Deijn et al., 2024; S. Wang et al., 2023). To address this, researchers incorporated latent space inference, enabling the generator to capture abstract representations of data. Notable examples include Adversarially Learned Inference (ALI) and Bidirectional GANs (BiGANs), which added encoders to transform input images into latent feature vectors, improving the generator's understanding of image structure and context (Donahue et al., 2016; Dumoulin et al., 2016). This helped guide generation and discrimination more meaningfully than relying on noise vectors alone.

Despite these improvements, GANs continued to exhibit issues such as training instability and mode collapse, especially with high-dimensional images. To counter these limitations, researchers explored probabilistic modeling approaches like variational auto-encoders (VAEs), which aim to encode input data into a latent distribution rather than a fixed vector. VAEs optimize a loss function combining a reconstruction term (which measures how well the model can reconstruct the input) and a Kullback-Leibler (KL) divergence term (which ensures that the learned latent space remains close to a prior distribution, typically Gaussian) (Kingma & Welling, 2013). While VAEs improve training stability, they often yield blurry outputs due to their emphasis on reconstruction fidelity rather than adversarial realism.

To bridge the strengths of both models, Adversarial Variational Bayes (AVB) was introduced as a hybrid framework that integrates the probabilistic inference of VAEs with the generative sharpness of GANs. AVB replaces the traditional VAE encoder with one trained adversarially to match a posterior distribution, thereby preserving sample quality while improving stability Mescheder et al. (2017). This hybrid approach reduces the tension between quality and training convergence and marks a key moment in Gen AI's trajectory. This showed that integrating probabilistic reasoning into adversarial frameworks can lead to more robust generative models.

However, even these advanced models faced limitations in capturing fine-grained visual complexity and maintaining training stability at scale. This led to the rise of diffusion models, which offered a fundamentally different approach. Rather than pitting two networks against each other, diffusion models learn to gradually corrupt input data with noise and then reconstruct it step-by-step in reverse. The process is optimized using reconstruction loss and KL divergence terms similar to VAEs, but avoids adversarial training entirely. This method forms the basis for many recent high-fidelity image generators and sets the stage for the diffusion-based models explored in the next section (Ho et al., 2020; Nichol & Dhariwal, 2021).

2.2 Diffusion Models

A diffusion model often uses UNet UBlocks (Ho et al., 2020; Saharia et al., 2022), ResNet blocks (Nichol & Dhariwal, 2021), as well as down sample and attention blocks such as self-attention (Ho et al., 2020; Saharia et al., 2022), multi-head attention (Nichol & Dhariwal, 2021), or cross attention (Saharia et al., 2022) to forward diffuse images with noise. To reverse diffuse, a model uses a combination of up sample blocks, skip connections to retain information across layers, and more attention blocks. By applying attention to multiple depths, such as 64x64, 32x32, 16x16, and 8x8, the model can better understand spatial relationships in early layers while recognizing global structures on deeper layers (Ho et al., 2020; Nichol & Dhariwal, 2021; Saharia et al., 2022).

Shortly after the introduction of ChatGPT by OpenAI in 2022, large language model (LLM) applications became popular and found their way into diffusion models (Saharia et al., 2022; Zhang et al., 2023). LLMs showed that models can mimic real-life conversations, with near-accurate responses to questions and comments from users. As a result of the transformative research on LLMs such as GPT (Radford et al., 2018), BERT (Devlin et al., 2018), and T5 (Raffel et al., 2020), large text-to-image synthesis models were developed that utilized LLMs to create images (Ramesh et al., 2021; Saharia et al., 2022; Zhang et al., 2023). These diffusion models take a textual prompt and transform it into vector embeddings through pre-trained embedders such as Contrastive Language-Image Pre-Training (CLIP). Adding these text embedders allows the diffusion model to create images for many classes, as the text informs the model during both the forward and reverse diffusion sections on what it tries to create (Radford et al., 2021; Ramesh et al., 2022; Saharia et al., 2022). Using large data sets of web-scraped data with both images and image descriptions, large diffusion models, such as Imagen (Saharia et al., 2022), Stable Diffusion (Rombach et al., 2021), and DALL·E (Ramesh et al., 2021) were created with the ability to handle both high-end input prompts and to create high-quality photorealistic images within seconds (Ramesh et al., 2021, 2022; Saharia et al., 2022).

However, current research shows that a lot of current generations are more likely to look like artwork instead of real-world images (Hao et al., 2023; Hossain et al., 2024; Zhang et al., 2023). This is related to the limited availability of realistic annotated data sets and the constraints of model sizes. Additionally, image quality is further influenced by prompt engineering, a process of crafting input text to optimize the output of generative AI models, as well as by model architecture, fine-tuning processes, and prompt formulation (Zhang et al., 2023). Because of the way textual embeddings are integrated into the models' architecture, text-to-image models are directly dependent on the quality of the input prompt to generate better results. With well-engineered prompts, a model can emulate real-world situations, possibly even functioning as an emulator for real-world training data in scarce domains. But the opposite is true as well; if a prompt is not well-defined and of low quality, results can be nothing like the desired training set (Hao et al., 2023; Zhou et al., 2023). Next to that, there is also a risk that the model would like the prompt to be received in a certain way and performs worse when it sees a word that it has not been trained on. This creates different challenges, as it is often hard to know which words a model prefers to see and which words receive preferred results for the user (Hao et al., 2023; Zhang et al., 2023). For this reason, approaches such as Diffusion Inversion (Zhou et al., 2023) or prompt adaptation frameworks (Hao et al., 2023) are proposed to overcome these challenges. These considerations emphasize the importance of addressing prompt dependency and data set limitations to advance the generation of more realistic and contextually accurate AI outputs, setting the stage for further exploration in this study.

Although diffusion models have become the dominant architecture for high-resolution and prompt-driven image generation due to their stability and effective noise-based reconstruction, they are not the only approach shaping the field of generative modeling. Another important category, autoregressive models, is more upcoming and have shown strong performance by generating images in a sequential manner, predicting one pixel or image patch at a time based on prior context. This approach differs fundamentally from the denoising process used in diffusion models, emphasizing local coherence and detail through ordered prediction. While often more computationally intensive, autoregressive models excel in structured generation tasks, particularly when combined with language modeling techniques. The next section introduces these models and examines their role in advancing the realism and utility of synthetic image generation.

2.3 Autoregressive Models

Autoregressive (AR) models represent distinctive approaches to generative modeling, diverging fundamentally from adversarial and diffusion-based methods. Unlike diffusion models, which iteratively refine noisy data, or generative adversarial networks (GANs) that pit a generator against a discriminator to learn image distributions, AR models generate images sequentially. They model the conditional probability of each pixel or image patch based on all previously generated content. This sequential process ensures that each step is informed by prior steps, fostering strong local coherence and enabling highly structured outputs (Gu et al., 2025). The autoregressive framework excels at capturing intricate dependencies within an image, making it particularly effective for generating detailed and contextually consistent visuals. However, this approach has historically faced challenges, including slow generation speeds and difficulties in modeling both local textures and global structures simultaneously, which have limited its scalability for high-resolution image synthesis (Gu et al., 2025).

Recent advancements have significantly mitigated these limitations, pushing AR models toward greater efficiency and quality. One notable innovation is the Multi-scale Vector-Autoregressive (M-VAR) framework, which introduces scale-separated modeling to decompose image generation into intra-scale and inter-scale processes (Ren et al., 2024). Intra-scale modeling focuses on capturing fine-grained details within a given resolution, while inter-scale modeling addresses long-range dependencies across different resolutions. This decomposition allows M-VAR to efficiently model complex global structures without sacrificing high-fidelity textures, resulting in substantial performance improvements on large-scale benchmarks like ImageNet-256. By enabling faster and more robust generation, M-VAR demonstrates the potential of AR models to compete with diffusion-based methods in high-resolution image synthesis, particularly for applications requiring both local precision and global coherence (Ren et al., 2024).

In addition to improvements in generation speed and flexibility, AR models have evolved to support continuous data representations, moving beyond the limitations of vector-quantized image tokens. Traditional token-based approaches often introduce quantization artifacts, which can degrade image quality, especially for high-resolution outputs. The Multimodal Autoregressive (MAR) model addresses this by operating in continuous latent spaces and applying autoregressive loss functions informed by diffusion-based likelihoods (T. Li et al., 2024). This integration reduces tokenization overhead while preserving the sequential modeling benefits of AR frameworks, resulting in sharper and more natural images. By leveraging continuous representations, MAR achieves a balance between computational efficiency and visual fidelity, making it a compelling choice for applications where artifact-free generation is critical.

Moreover, advances in multimodal alignment have enhanced the reasoning capabilities of AR models, enabling them to generate images that align closely with complex input prompts. Techniques such as chain-of-thought prompting and preference alignment allow models to perform stepwise reasoning during generation, improving both the interpretability and accuracy of outputs (Guo et al., 2025). For instance, when conditioned on detailed textual descriptions, these models can produce images that reflect nuanced semantic relationships, such as specific object placements or environmental conditions. This capability is particularly advantageous for tasks requiring precise control over generated content, as it allows AR models to translate abstract instructions into visually coherent results. Such advancements underscore the potential of AR models to bridge the gap between generative modeling and structured reasoning (Guo et al., 2025).

In the context of synthetic image generation for data-scarce domains like snow detection, AR models offer unique advantages. Their ability to produce highly structured, prompt-conditioned images makes them well-suited for creating synthetic data sets that capture realistic spatial relationships and environmental features, such as snow accumulation patterns or terrain variations. Unlike diffusion models, which often prioritize raw image fidelity, AR models perform well in controllability and semantic precision, enabling them to generate images tailored to specific requirements. While diffusion models remain dominant in applications demanding photorealistic outputs, the structured visual understanding offered by AR models provides a powerful alternative.

3. Methods

Annually, tens of thousands of people end up at the emergency department as a result of winter-related fall and slip injuries, mostly among seniors and the visually impaired. These fall and slip injuries are weather-related, such as snow and ice on sidewalks, which are challenging to detect and are very slippery (Kakara et al., 2021; Mills et al., 2020). Solutions using artificial intelligence, such as image classification networks, might help these demographics to promptly recognize and avoid dangerous spots with timely alerts. These networks can be trained on sidewalk images with snow and ice, and be deployed to mobile phones and/or other wearables (de Deijn, 2024; de Deijn & Bukralia, 2024) but require large quantities of data to achieve high accuracy (de Deijn, 2024; Lecun et al., 2015). This might be one of the bigger challenges for sidewalk snow and ice detection, as these images are not as commonly available on the internet and are season-dependent to collect (de Deijn, 2024; de Deijn & Bukralia, 2024). Next to that, snow and ice have very unique optical characteristics, meaning that light reflection depends on a large number of factors such as lighting type, snow depth, environmental conditions (de Deijn, 2024; Warren, 2019), snow grain size (Dang et al., 2016; de Deijn, 2024; Zhuravleva & Kokhanovsky, 2011), light brightness, and solar zenith angle (Dang et al., 2016; de Deijn, 2024) to name a few. Synthetic data could offer a solution, as it enables researchers to generate new training data cheaply throughout the year. It has the potential to simulate all types of different environments under different reflection conditions. For synthetic data to be used as training data for a snow detection network in the real-world, photorealistic images are necessary. For this reason, a model is required to simulate the real-world, rather than creating cartoon-like images (Hossain et al., 2024).

This paper compares two prompt-based synthetic image generators: OpenAI's DALL·E 3, which is a diffusion-based image generator, and xAI's Aurora, which is an autoregression-based image generator. For the comparison, each model was used to generate 250 images of snowy sidewalks using the following prompt:

> *Generate a realistic winter scene from a first-person pedestrian perspective looking down at a cracked concrete sidewalk. The sidewalk is partially covered with icy, slushy snow, concentrated along the edge where snow transitions into ice. Footprints are visible in the icy patches. The left side of the image shows accumulated snow with small patches of brown grass peeking through. The lighting is natural, with soft daylight, capturing the texture of the snow, ice, and sidewalk in a cold, wintry outdoor setting.*

Similarly, another 250 images of clear sidewalks were generated by each model using this prompt:

> *Generate a realistic outdoor scene from a pedestrian's perspective looking down at a clean, cracked concrete sidewalk on a clear day. The sidewalk is bordered by a patch of dry grass and a landscaped area with mulch, dormant shrubs, and a large rock. Pink stakes are visible marking parts of the landscaped area. The lighting is natural and sunny, casting soft shadows and highlighting the textures of the sidewalk, grass, and surroundings. The scene captures a calm, real-world outdoor setting without including any visible objects like a smartphone.*

These prompts were designed by OpenAI's ChatGPT-4o after analyzing some example real-world sidewalk images, as shown in Figure 1, from a data set created by de Deijn & Bukralia (2024).

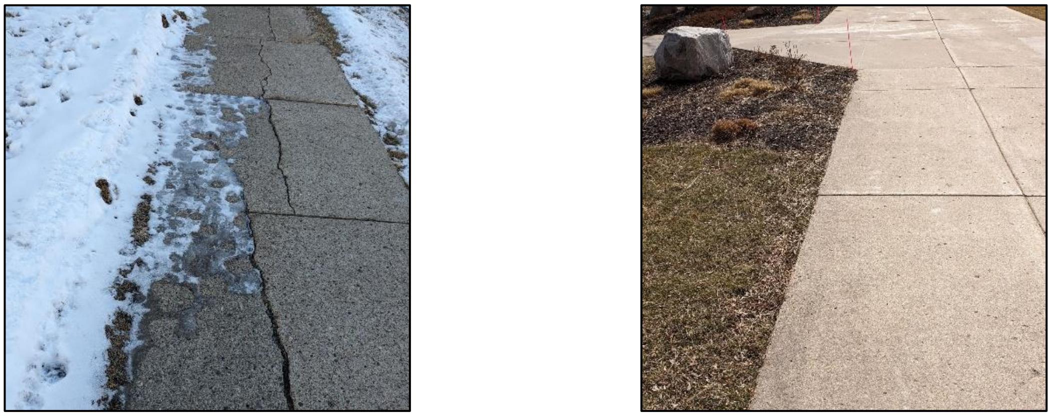
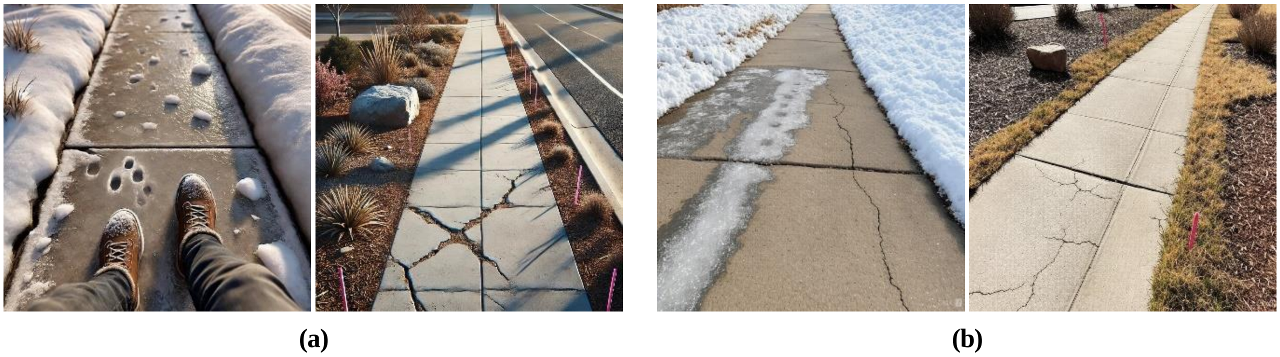


Figure 1. Examples of Real-World Snowy Sidewalk Data Set By de Deijn & Bukralia (2024), Used By ChatGPT-4o to Generate Image Generation Prompts

To evaluate the generated images, as shown in Figure 2, we trained models using the snowy sidewalk image classification architecture proposed by de Deijn (2024). This convolutional neural network (CNN) incorporates a spatial attention mechanism with combined average and max pooling. We configured the model with a learning rate of 0.001, a batch size of 4, and trained it for 25 epochs due to the limited data set size. The Adam optimizer and cross-entropy loss function were employed. Images were resized to 224 × 224 pixels, with data augmentation including random rotations of up to 90 degrees and random horizontal flips. By training a model on each synthetic data set, we assessed their performance against real-world data. The evaluation was conducted using a separate, unseen real-world test data set to ensure unbiased comparison.



(a)

(b)

Figure 2. Examples of the Synthetic Images Created Using (a) DALL·E 3 and (b) Aurora

Additionally, another model was trained on different real-world data of snowy and clear sidewalks, comprising 250 images per category as well. All models were evaluated using a test data set that contains 59 images per category. Despite its small size, this data set provides an effective means of evaluating the performance of the synthetic data. Model performance is compared using the F2 score, which prioritizes recall over precision, highlighting accurate snow detection over clear sidewalk detection, as snow has a higher hazard level for pedestrian safety when not correctly identified as compared to clear sidewalks.

4. Experiment Results

Three similar networks, built on de Deijn's (2024) architecture, were trained using three distinct data sets: synthetic data sets from DALL·E 3 and Aurora, and a real-world training data set created for this study. Each model underwent twelve prediction rounds, with the real-world test data set randomly split into batches to mitigate variability from its limited size. Performance was assessed on an unseen real-world test data set using key metrics: recall (proportion of true snow instances detected), precision (accuracy of snow predictions), F2 score (emphasizing recall over precision to prioritize snow detection while considering clear sidewalk accuracy), accuracy (overall correct predictions), and ROC AUC (ability to distinguish snow from clear sidewalks). To measure statistical significance and the magnitude of performance differences, p-values (with $p < 0.05$ indicating significance) and Cohen's d (standardized effect size, where positive values favor the synthetic model and negative values favor the real-world model) were calculated relative to the real-world model's test results. The results of these tests are presented in Table 1.

Metric	Model	Mean (± Std)	P-value (vs. Real-World)	Cohen's *d* (vs. Real-World)	Interpretation
F2 score	DALL·E 3	82.49% (±0.80)	5.8871×10^{-11}	-7.4000	Significantly Worse
	Aurora	**94.40% (±0.23)**	6.6502×10^{-10}	5.9085	**Significantly Better**
	Real-World	89.53% (±0.91)	-	-	-
Precision	DALL·E 3	62.73% (±1.21)	1.8152×10^{-15}	-19.1604	Significantly Worse
	Aurora	77.13% (±0.77)	1.3991×10^{-14}	-15.9049	Significantly Worse
	Real-World	**94.28% (±0.59)**	-	-	**-**
Recall	DALL·E 3	89.55% (±0.94)	0.0128	0.8944	Significantly Better
	Aurora	**100.00% (±0.00)**	2.3771×10^{-12}	9.9440	**Significantly Better**
	Real-World	88.42% (±1.16)	-	-	-
Accuracy	DALL·E 3	68.15% (±1.40)	1.5834×10^{-14}	-15.7263	Significantly Worse
	Aurora	85.17% (±0.65)	2.7259×10^{-12}	-9.8198	Significantly Worse
	Real-World	**91.53% (±0.51)**	-	-	-
ROC AUC	DALL·E 3	67.52% (±0.68)	9.7188×10^{-21}	-57.8231	Significantly Worse
	Aurora	95.54% (±0.59)	2.0715×10^{-9}	-5.3125	Significantly Worse
	Real-World	**98.09% (±0.22)**	-	-	-

Table 1. Statistical Comparison of Models Trained on DALL·E 3, Aurora, and Real-World Data Across Twelve Test Runs. Means and standard deviations (± Std) are reported, with p-values and Cohen's *d* calculated relative to the Real-World model. A p-value < 0.05 indicates statistical significance, and Cohen's *d* reflects effect size (positive = better than Real-World, negative = worse).

Table 1 shows that the Aurora-trained model achieved the highest F2 score (94.40%) and perfect recall (100%), with t-test p-values of $p = 6.65 \times 10^{-10}$ and $p = 2.38 \times 10^{-12}$, respectively, confirming its mean scores significantly surpass the real-world model (F2: 89.53%, recall: 88.42%). To complement statistical significance, we also report Cohen's *d*, a standardized effect size that quantifies the magnitude of differences across twelve independent runs, showing very large positive values (F2: $d = 5.91$; recall: $d = 9.94$), which underscores Aurora's pronounced advantage in snow detection. This performance likely stems from Aurora's ability to generate highly photorealistic images that closely mimic real-world variability. Conversely, the real-world model led in precision (94.28%), accuracy (91.53%), and ROC AUC (98.09%), with synthetic-data models showing significantly lower means and large negative effect sizes (e.g., Aurora precision: $p = 1.40 \times 10^{-14}$, $d = –15.90$; DALL·E 3 precision: $p = 1.82 \times 10^{-15}$, $d = –19.16$). DALL·E 3 trailed across most metrics (e.g., F2: 82.49%; $p = 5.89 \times 10^{-11}$, $d = -7.40$), likely due to less realistic image detail as verified by visual data set analysis.

Surprisingly, the real-world model did not outperform synthetic models in all areas, an observation reflected not only in p-values but also in moderate to large Cohen's *d* values that highlight substantial performance gaps. Confusion matrices (see Appendix) reveal Aurora's model perfectly identifies snow (recall: $p = 2.38 \times 10^{-12}$, $d = 9.94$) but occasionally misclassifies clear sidewalks, while the real-world model balances both classes with fewer errors. The F2 score's recall bias enhances Aurora's edge (F2: $p = 6.65 \times 10^{-10}$, $d = 5.91$), with effect sizes confirming its superior snow detection. DALL·E 3's consistent underperformance (for example, ROC AUC: $p = 9.72 \times 10^{-21}$, $d = -57.82$) suggests it struggles with photorealistic detail critical for this task.

The test data set's small size (59 images per category) and limited variability (e.g., lighting, snow texture) may have impacted results. A larger, more diverse test set could improve generalizability. Aurora's stable performance, with significant mean differences ($p < 0.05$), highlights the value of high-quality synthetic data for applications like pedestrian safety, while DALL·E 3's significant gaps indicate areas for refinement.

5. Conclusion

This study highlights the transformative potential and current limitations of generative AI in creating synthetic data sets for snow detection. By comparing models trained on real-world data with those trained on synthetic data sets from DALL·E 3 and Aurora, the results reveal significant performance differences, as confirmed by t-test p-values ($p < 0.05$), showing that mean metric scores for synthetic models diverge markedly from the real-world model. Aurora's data set achieved the highest F2 score (94.40%, $p = 6.65 \times 10^{-10}$) and perfect recall (100%, $p = 2.38 \times 10^{-12}$) versus the real-world model (F2: 89.53%; recall: 88.42%), driven by its ability to catch every snow patch, as F2 weights recall more heavily. However, Aurora's lower precision (77.13%, $p = 1.3991 \times 10^{-14}$) due to higher false positives indicates it is less effective at distinguishing snow from clear sidewalks compared to the real-world model. In contrast, DALL·E 3's lower F2 score (82.49%, $p = 5.89 \times 10^{-11}$), ROC AUC (67.52%, $p = 9.7188 \times 10^{-21}$), and accuracy (68.15%, $p = 1.5834 \times 10^{-14}$) highlight its struggles with capturing real-world variability in snow appearance, making it unreliable as a standalone data source for this task.

The real-world model, despite a lower F2 score and recall, outperforms or matches both synthetic-data models across all metrics (F2, precision, recall, accuracy ROC AUC) when considering both false positives and false negatives, achieving higher precision (94.28%) and accuracy (91.53%). This balanced performance suggests that evaluation metrics and data set diversity are critical for assessing generalization. The test data set's small size (59 images per category) may exaggerate synthetic data's apparent superiority, emphasizing the need for larger, more varied data sets to validate these findings.

While generative AI has made remarkable strides, it is not yet a full substitute for real-world data. Aurora's success suggests that synthetic data sets can effectively complement traditional data, particularly for tasks like snow detection, where real-world collection is challenging. However, DALL·E 3's significant performance gap indicates that further refinements are needed. Future research should prioritize enhancing generative models to capture environmental complexities, potentially through physics-based lighting models, increased scenario variability, or advanced prompt engineering, to produce synthetic data sets that meet the demands of robust machine learning applications.

Ultimately, this study underscores that generative AI is narrowing the gap with real-world data. The significant differences in performance, backed by t-test results, point to a future where synthetic data, as exemplified by Aurora, can augment real-world data sets, enabling scalable, effective solutions for machine learning tasks like pedestrian safety.

6. Discussion

This study faced several challenges that influenced its results. One of the primary limitations lies in the dependency of synthetic image generation models on the quality and precision of the prompts provided by the researcher or user. Poorly constructed prompts can negatively impact the quality and photorealism of the generated images. This issue was highly present in this study, as pictures produced by OpenAI's DALL·E were not as photorealistic as those generated by xAI's Aurora, despite both data sets being created using identical prompts. The two prompts were generated by ChatGPT based on some example images, with the thought that this would return the most optimally engineered prompt for the model, but this did not optimally deliver for DALL·E. This underscores the importance of the model architecture, the underlying data used for training, and the quality of prompts in determining the quality and relevance of generated images.

Additionally, while synthetic data sets addressed data scarcity, they still could face challenges in generalizing real-world scenarios. The constrained nature of prompts and the uniformity of generated data sets can lead to overfitting, where models perform well on test sets with similar characteristics but struggle with diverse or unforeseen real-world conditions. This highlights a key limitation of synthetic data: its ability to replicate specific scenarios but not the variability and complexity of real-world environments. For instance, while Aurora generated more realistic images, the synthetic data sets lacked diversity in surface textures, lighting conditions, and environmental contexts, such as the appearance of other white objects that you might see outside, such as chalk, paint, or dandelion puffs, which pose critical factors for robust snow detection models. A solution would be to expand the data set with more variable data or to apply domain adaptation by training a model on synthetic images and fine-tuning it on real images. This would reduce the need for large data sets of real-world data.

Google's Imagen3 model was also evaluated using the prompts. While it successfully generated images for the snow-covered sidewalk prompt, it failed to produce images for the clear pavement prompt, citing policy restrictions. Consequently, this limitation prevented Imagen3 from being included in the comparative analysis. Visually, the snow images generated by Imagen3 appeared more realistic than those from DALL·E but less so than Aurora. However, because of its inability to generate a complete data set under identical conditions, a formal evaluation of Imagen3's performance could not be conducted in this study.

Future research should address these limitations by focusing on improving both the diversity and quality of synthetic data sets. Refining prompt engineering techniques and developing frameworks that automate prompt optimization could reduce the reliance on user expertise, making synthetic data generation more accessible and reliable. Furthermore, as AI-generated images become increasingly indistinguishable from real-world data, it is important to think about mechanisms such as permanent watermarks on AI-produced images. These identifiers would ensure transparency of synthetic data usage, mitigating risks associated with misuse or misrepresentation. Such watermarks must be robust and impervious to tampering, striking a balance between maintaining the utility of the data for machine learning applications and addressing ethical concerns around authenticity.

Acknowledgments

This experiment was conducted without any external funding or grants.

7. References

Arif, M., & Mahalanobis, A. (2021). Infrared Target Recognition Using Realistic Training Images Generated by Modifying Latent Features of an Encoder–Decoder Network. *IEEE Transactions on Aerospace and Electronic Systems*, *57*(6), 4448–4456. https://doi.org/10.1109/TAES.2021.3090921

Atallah, S. B., Banda, N. R., Banda, A., & Roeck, N. A. (2023). How large language models including generative pre-trained transformer (GPT) 3 and 4 will impact medicine and surgery. *Techniques in Coloproctology*, *27*(8), 609–614. https://doi.org/10.1007/s10151-023-02837-8

Bandi, A., Adapa, P. V. S. R., & Kuchi, Y. E. V. P. K. (2023). The Power of Generative AI: A Review of Requirements, Models, Input–Output Formats, Evaluation Metrics, and Challenges. *Future Internet*, *15*(8), 260. https://doi.org/10.3390/fi15080260

Barbosa, I. B., Cristani, M., Caputo, B., Rognhaugen, A., & Theoharis, T. (2018). Looking beyond appearances: Synthetic training data for deep CNNs in re-identification. *Computer Vision and Image Understanding*, *167*, 50–62. https://doi.org/10.1016/j.cviu.2017.12.002

Chang, D.-L., Yang, S.-H., Hsieh, S.-L., Wang, H.-J., & Yeh, K.-C. (2020). Artificial Intelligence Methodologies Applied to Prompt Pluvial Flood Estimation and Prediction. *Water*, *12*(12), 3552. https://doi.org/10.3390/w12123552

Chen, B., Wu, Z., & Zhao, R. (2023). From fiction to fact: the growing role of generative AI in business and finance. *Journal of Chinese Economic and Business Studies*, *21*(4), 471–496. https://doi.org/10.1080/14765284.2023.2245279

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative Adversarial Networks: An Overview. *IEEE Signal Processing Magazine*, *35*(1), 53–65. https://doi.org/10.1109/MSP.2017.2765202

Dang, C., Fu, Q., & Warren, S. G. (2016). Effect of Snow Grain Shape on Snow Albedo. *Journal of the Atmospheric Sciences*, *73*(9), 3573–3583. https://doi.org/10.1175/JAS-D-15-0276.1

de Deijn, R. (2024). *Developing a Snow Detection Algorithm Using Spatial Attention for Pedestrian Safety* [Master's thesis, Minnesota State University, Mankato]. https://www.proquest.com/openview/e98fa5068183de2880a438d6fd5ad9c9/1

de Deijn, R., Batra, A., Koch, B., Mansoor, N., & Makkena, H. (2024). Reviewing FID and SID Metrics on Generative Adversarial Networks. *AI, Machine Learning and Applications*, 111–124. https://doi.org/10.5121/csit.2024.140208

de Deijn, R., & Bukralia, R. (2024, June 30). Image Classification for Snow Detection to Improve Pedestrian Safety. *Proceedings of MWAIS 2024*. http://arxiv.org/abs/2407.00818

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv Preprint ArXiv:1810.04805*. https://doi.org/10.48550/arXiv.1810.04805

Donahue, J., Krähenbühl, P., & Darrell, T. (2016). Adversarial feature learning. *ArXiv Preprint ArXiv:1605.09782*. http://arxiv.org/abs/1605.09782

Dumoulin, V., Belghazi, I., Poole, B., Mastropietro, O., Lamb, A., Arjovsky, M., & Courville, A. (2016). Adversarially Learned Inference. *ArXiv Preprint ArXiv:1606.00704*. https://doi.org/10.48550/arXiv.1606.00704

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Elfeki, A., Masoud, M., & Niyazi, B. (2017). Integrated rainfall–runoff and flood inundation modeling for flash flood risk assessment under data scarcity in arid regions: Wadi Fatimah basin case study, Saudi Arabia. *Natural Hazards*, *85*(1), 87–109. https://doi.org/10.1007/s11069-016-2559-7

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *Business & Information Systems Engineering*, *66*(1), 111–126. https://doi.org/10.1007/s12599-023-00834-7

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014, June 10). Generative Adversarial Networks. *Proceedings of Advances in Neural Information Processing Systems*. http://arxiv.org/abs/1406.2661

Guo, Z., Zhang, R., Tong, C., Zhao, Z., Gao, P., Li, H., & Heng, P.-A. (2025). Can We Generate Images with CoT? Let's Verify and Reinforce Image Generation Step by Step. *ArXiv Preprint ArXiv:2501.13926*. http://arxiv.org/abs/2501.13926

Haan, K. (2023, April 25). *24 Top AI Statistics and Trends In 2024*. Forbes Advisor. https://www.forbes.com/advisor/business/ai-statistics/

Hao, Y., Chi, Z., Dong, L., & Wei, F. (2023). Optimizing Prompts for Text-to-Image Generation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Proceedings of Advances in Neural Information Processing Systems* (pp. 66923–66939). Curran Associates, Inc. http://arxiv.org/abs/2212.09611

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. *Proceedings of Advances in Neural Information Processing Systems*, 6480–6851. https://doi.org/10.48550/arXiv.2006.11239

Hossain, S., Ibrahim Protik, T., & Sazid, G. (2024). Potential Risk of Faking Microscopic Images in the Characterization of Nanomaterials Through Artificial Intelligence. In *Proceedings of International Conference on Physics*. https://doi.org/10.13140/RG.2.2.25498.45766

Kakara, R. S., Moreland, B. L., Haddad, Y. K., Shakya, I., & Bergen, G. (2021). Seasonal variation in fall-related emergency department visits by location of fall – United States, 2015. *Journal of Safety Research*, *79*, 38–44. https://doi.org/10.1016/j.jsr.2021.08.002

Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes. *ArXiv Preprint ArXiv:1312.6114*. https://doi.org/10.48550/arXiv.1312.6114

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. In *Nature* (Vol. 521, Issue 7553, pp. 436–444). Nature Publishing Group. https://doi.org/10.1038/nature14539

Li, D.-C., Hu, S. C., Lin, L.-S., & Yeh, C.-W. (2017). Detecting representative data and generating synthetic samples to improve learning accuracy with imbalanced data sets. *PLOS ONE*, *12*(8), e0181853. https://doi.org/10.1371/journal.pone.0181853

Li, T., Tian, Y., Li, H., Deng, M., & He, K. (2024). Autoregressive Image Generation without Vector Quantization. *ArXiv Preprint ArXiv:2406.11838*. http://arxiv.org/abs/2406.11838

Meng, C., Rombach, R., Gao, R., Kingma, D., Ermon, S., Ho, J., & Salimans, T. (2023). On Distillation of Guided Diffusion Models. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14297–14306. https://doi.org/10.1109/CVPR52729.2023.01374

Mescheder, L., Nowozin, S., & Geiger, A. (2017). Adversarial Variational Bayes: Unifying Variational Autoencoders and Generative Adversarial Networks. *Proceedings of International Conference on Machine Learning*, 2391–2400. https://doi.org/10.48550/arXiv.1701.04722

Mills, B., Andrey, J., Doherty, S., Doberstein, B., & Yessis, J. (2020). Winter Storms and Fall-Related Injuries: Is It Safer to Walk than to Drive? *Weather, Climate, and Society*, *12*(3), 421–434. https://doi.org/10.1175/WCAS-D-19-

0099.1

Mirza, M., & Osindero, S. (2014). Conditional Generative Adversarial Nets. *ArXiv Preprint ArXiv:1411.1784*. http://arxiv.org/abs/1411.1784

Nichol, A., & Dhariwal, P. (2021). Improved Denoising Diffusion Probabilistic Models. *Proceedings of International Conference on Machine Learning*, 8162–8171. https://doi.org/10.48550/arXiv.2102.09672

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *Proceedings of International Conference on Machine Learning*, 8748–8763. https://doi.org/10.48550/arXiv.2103.00020

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. In *ArXiv*. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. In *Journal of Machine Learning Research* (Vol. 21). http://jmlr.org/papers/v21/20-074.html.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents. *ArXiv Preprint ArXiv:2204.06125*. https://doi.org/10.48550/arXiv.2204.06125

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-Shot Text-to-Image Generation. *Proceedings of International Conference on Machine Learning*, 8821–8831. https://doi.org/10.48550/arXiv.2102.12092

Ren, S., Yu, Y., Ruiz, N., Wang, F., Yuille, A., & Xie, C. (2024). M-VAR: Decoupled Scale-wise Autoregressive Modeling for High-Quality Image Generation. *ArXiv Preprint ArXiv:2411.10433*. http://arxiv.org/abs/2411.10433

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2021). High-Resolution Image Synthesis with Latent Diffusion Models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695. https://doi.org/10.48550/arXiv.2112.10752

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., & Norouzi, M. (2022). Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *Proceedings of Advances in Neural Information Processing Systems*, 36479–36494. https://doi.org/10.48550/arXiv.2205.11487

Shaji George, A., Hovan George, A. S., & Gabrio Martin, A. S. (2023). A Review of ChatGPT AI's Impact on Several Business Sectors. *Partners Univers Int Innov J*, *1*(1), 9–23. https://doi.org/10.5281/zenodo.7644359

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., & Ganguli, S. (2015). Deep Unsupervised Learning using Nonequilibrium Thermodynamics. *Proceedings of International Conference on Machine Learning*, 2256–2265. https://doi.org/10.48550/arXiv.1503.03585

Valeriya, G., John, V., Singla, A., Yamini Devi, J., & Kumar, K. (2024). AI-Powered Super-Workers: An Experiment in Workforce Productivity and Satisfaction. *BIO Web of Conferences*, *86*, 01065. https://doi.org/10.1051/bioconf/20248601065

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. In U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, & R. Fergus (Eds.), *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 6000–6010). Curran Associates Inc. http://arxiv.org/abs/1706.03762

Wang, K., Gou, C., Duan, Y., Lin, Y., Zheng, X., & Wang, F.-Y. (2017). Generative adversarial networks: introduction and outlook. *IEEE/CAA Journal of Automatica Sinica*, *4*(4), 588–598. https://doi.org/10.1109/JAS.2017.7510583

Warren, S. G. (2019). Optical properties of ice and snow. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *377*(2146), 20180161. https://doi.org/10.1098/rsta.2018.0161

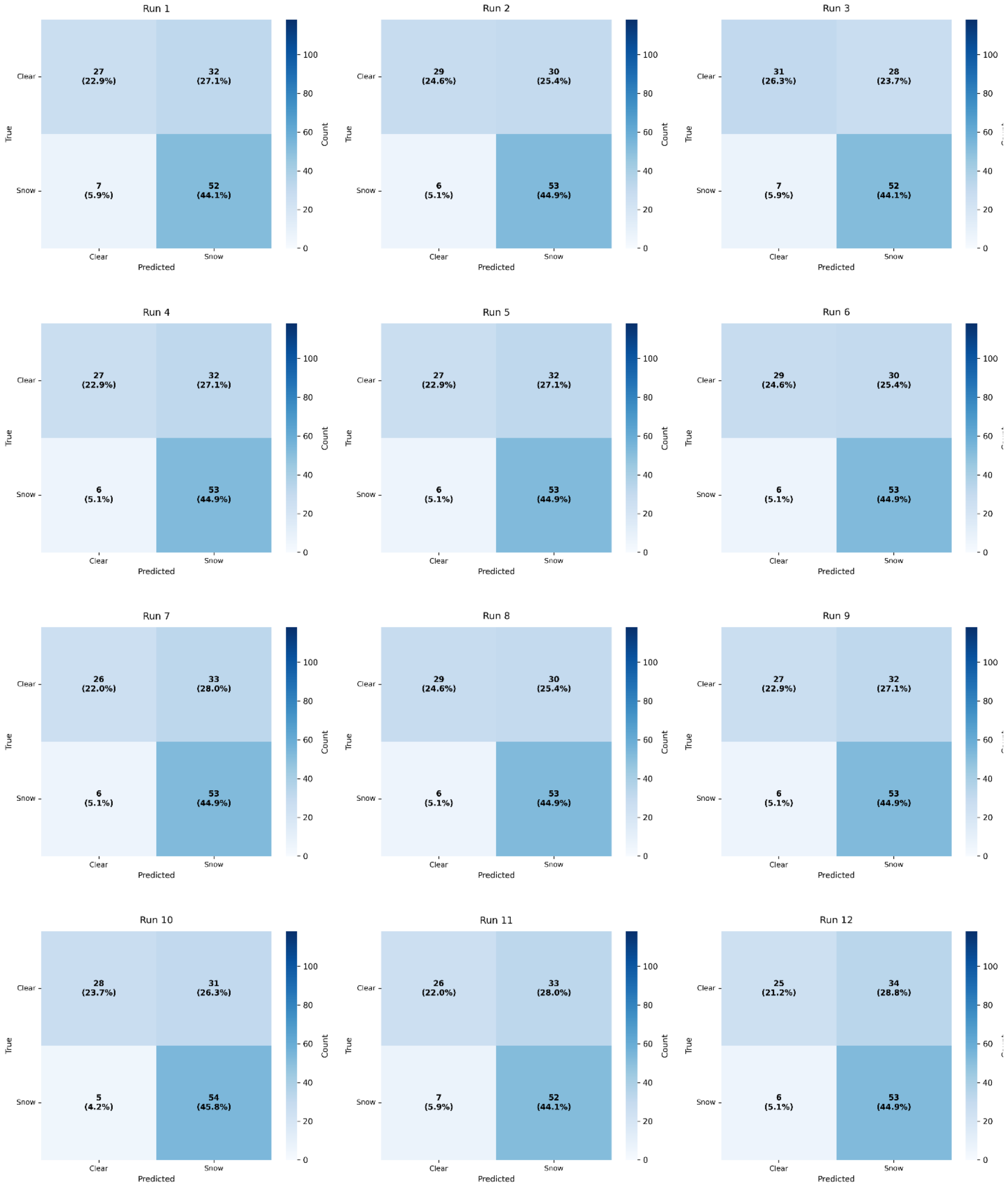
Zhang, T., Wang, Z., Huang, J., Tasnim, M. M., & Shi, W. (2023). A Survey of Diffusion Based Image Generation Models: Issues and Their Solutions. *ArXiv Preprint ArXiv:2308.13142*. https://doi.org/10.48550/arXiv.2308.13142

Zhou, Y., Sahak, H., & Ba, J. (2023). Training on Thin Air: Improve Image Classification with Generated Data. *ArXiv Preprint ArXiv:2305.15316*. https://doi.org/10.48550/arXiv.2305.15316

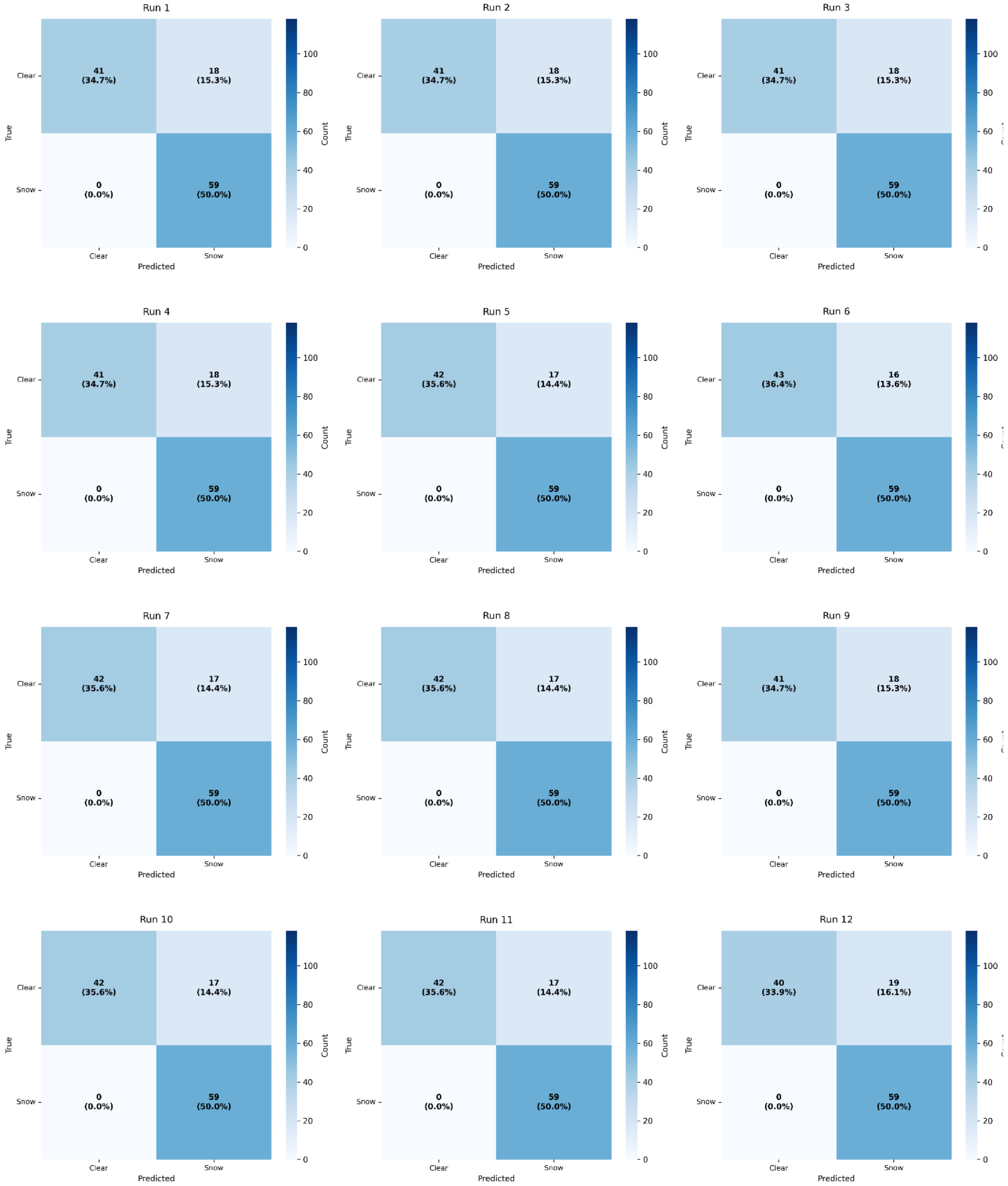
Zhuravleva, T. B., & Kokhanovsky, A. A. (2011). Influence of surface roughness on the reflective properties of snow. *Journal of Quantitative Spectroscopy and Radiative Transfer*, *112*(8), 1353–1368. https://doi.org/10.1016/j.jqsrt.2011.01.004

Appendix

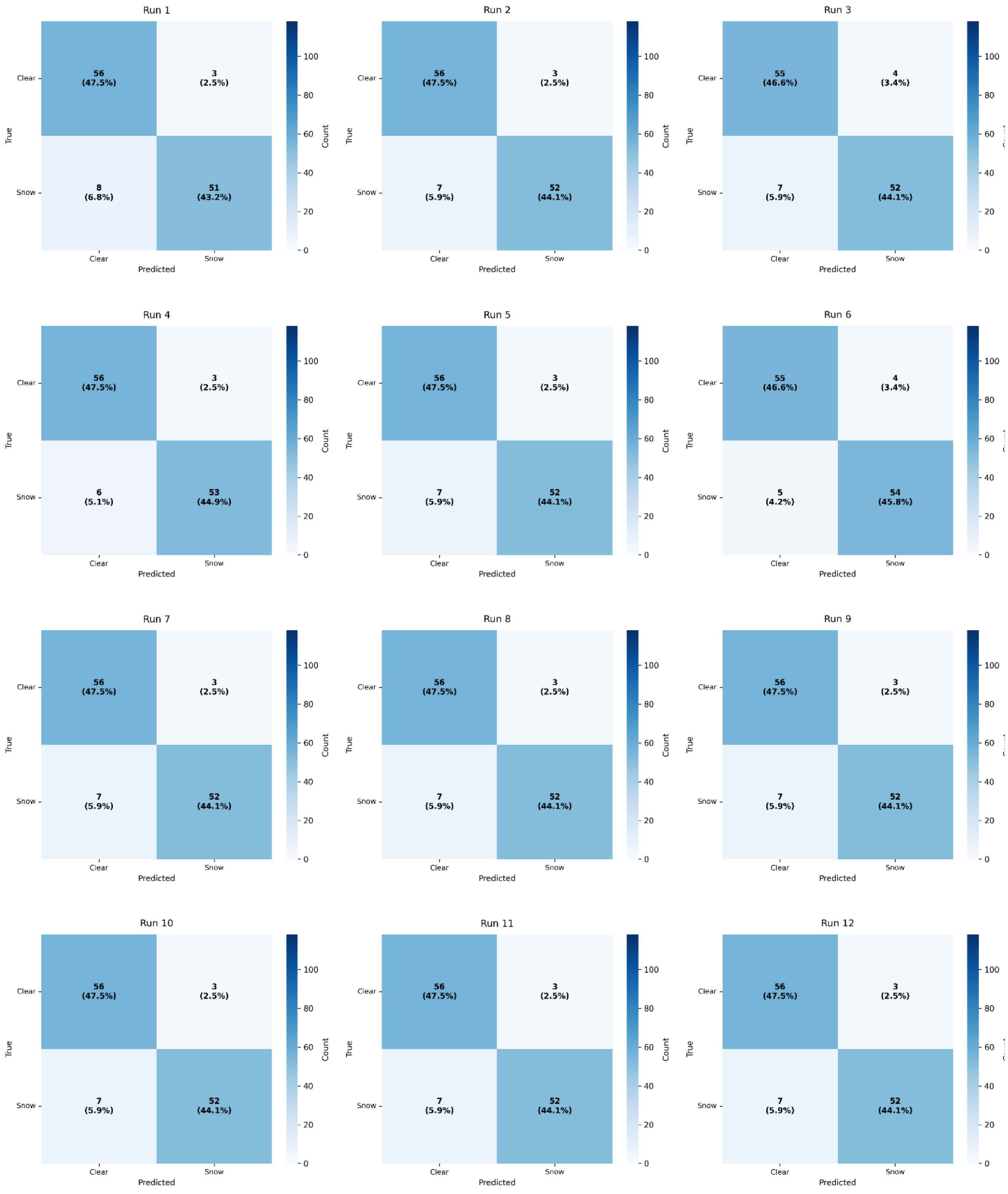
Appendix I. Model Confusion Matrices



Appendix I Fig. 1. DALL·E 3 Test Data Confusion Matrices



Appendix I Fig. 2. Aurora Test Data Confusion Matrices



Appendix I Fig. 3. Real-World Test Data Confusion Matrices

Author Biographies

Ricardo de Deijn is a data scientist and researcher based in Minnesota, with a Master of Science in Data Science from Minnesota State University, Mankato. Originally from the Netherlands, his fascination with snow and winter safety began after experiencing a Minnesota winter where snowfall reached up to his head. His research focuses on computer vision, real-time edge AI, and synthetic data generation for public safety. His work has been featured at academic conferences such as MWAIS, AIMLA, and CADSCOM, where he received a best paper award. He is also interested in exploring emerging and underutilized techniques, including mobile AI deployment and federated learning, for on-device inference in safety-critical environments.

Rajeev Bukralia is a tenured associate professor and the founding director of the Data Science and Artificial Intelligence programs in the Department of Computer Information Science at Minnesota State University, Mankato. He is a recipient of Minnesota’s Tekne Tech Educator of the Year Award and the Minnesota State Board of Trustees' Outstanding Educator Award.