GENERATING TISSUE IMAGES FROM RNA-SEQ PROFILES USING DEEP GENERATIVE MODELS

Submitted by
Utsab Sarkar
ID: 221-35-976
Department of Software Engineering
Daffodil International University

Supervised by
Musabbir Hasan Sammak
Senior Lecturer
Department of Software Engineering
Daffodil International University

Thesis submitted in fulfillment of the requirements
for the award of the degree of Bachelor of Science
Department of Software Engineering (Major in Data Science)

Fall 2025

APPROVAL

This thesis titled on "**Generating Tissue Images from RNA-seq Profiles using Deep Generative Models**", submitted by **Utsab Sarkar (ID: 221-35-976)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Chairman

Dr. S M Hasan Mahmud
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 1

A.B.M Shahariar Parvez
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Tapushe Rabaya Toma
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 3

Khalid Been Md. Badruzzaman Biplob
Lecturer (Senior Scale)
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

External Examiner

Dr. Md Sazzadur Rahman
Professor
Institute of Information Technology
Jahangirnagar University, Bangladesh

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is my original work except for quotations and citations which have been duly acknowledged. I also declare it has not been previously or concurrently submitted for any other degree at DAFFODIL INTERNATIONAL UNIVERSITY or any other institution.

Student's Signature
Name: Utsab Sarkar
ID: 221-35-976
Department of Software Engineering (Major in Data Science)
DAFFODIL INTERNATIONAL UNIVERSITY

SUPERVISOR'S DECLARATION

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Science.

Supervisor's Signature
Name: Musabbir Hasan Sammak
Designation: Senior Lecturer
Department of Software Engineering
DAFFODIL INTERNATIONAL UNIVERSITY

ACKNOWLEDGEMENTS

First of all, I want to thank Almighty God for His divine favor, enabling me to complete my undergraduate thesis. It is driven by the interest in how to leverage the potentials of deep learning and generative AI to address pressing real-world problems, particularly in the domain of generating synthetic tissue images from rna data.

I would express my deepest sense of thanks to my thesis supervisor, Musabbir Hasan Sammak, Senior Lecturer of the Department of Software Engineering, for his guidance and all-out support given to me in the entire research work. His intellectual advice and insights have largely shaped this work, and his commitment without wavering has been the inspiration for me to explore the limits of my knowledge and skills.

I would like to express my thanks to Dr. Imran Mahmud, Head of the Department of Software Engineering, Faculty of Science and Information Technology, and my other professors, faculties, and personnel for their kind cooperation and support in the successful completion of my work.

Great thanks to my parents and friends for continuous support, patience, and understanding during these years. Their support has been my stronghold, letting me have the opportunity to weather the turmoil that I went through.

Finally, I would like to acknowledge my batchmates and fellow members of DIU for their kind cooperation and consolation, which helped me reach this goal, as well as the organizations providing data and resources necessary for the research work. Without them, this work was not possible.

ABSTRACT

It is still very hard to combine molecular profiling with histopathological imaging in computational pathology. Reason bulk RNA sequencing only shows the average amount of gene expression and does not give you the spatial information you need to put tissue back together. This study investigates the feasibility of producing histopathological images directly from RNA-seq data using deep generative models, potentially enabling cost-effective imaging of molecular anomalies and the integration of genomic and morphological characterizations. We evaluated four generative architectures—Conditional GAN, Conditional VAE, Conditional Diffusion Model, and a novel Hybrid GAN-VAE—employing a dataset comprising 50 patients with paired RNA-seq profiles (16,383 genes) and H&E-stained histopathology images (128×128 pixels). The dataset was divided at the patient level (70% for training, 15% for validation, and 15% for testing) to make sure that the generalization test was strong. We used RNA conditioning to train the models and then tested them with three different measures: the Structural Similarity Index (SSIM), the Learned Perceptual Image Patch Similarity (LPIPS), and the Fréchet Inception Distance (FID). The Hybrid GAN-VAE had a better perceptual quality (LPIPS: 0.347), which was 16% better than the baseline GAN. The Conditional VAE, on the other hand, had the best statistical fidelity (FID: 275.64). Deep generative models can create realistic histopathological images from RNA-seq data; however, bulk sequencing compromises spatial information, resulting in decreased reconstruction accuracy. The Hybrid architecture's superior perceptual performance demonstrates the efficacy of multi-objective training methodologies that incorporate both adversarial and reconstruction losses. Future research incorporating spatially-resolved transcriptomics could significantly enhance the quality of synthesis, facilitating its application in diagnostic procedures, medical education, and precision medicine research.

Keywords: RNA sequencing, histopathology synthesis, generative adversarial networks, variational autoencoders, diffusion models, computational pathology, multi-modal learning, deep learning

Contents

List of Tables

List of Figures

LIST OF ABBREVIATIONS

Abbreviation	Meaning
GAN	Generative Adversarial Network
VAE	Variational Autoencoder
DDPM	Denoising Diffusion Probabilistic Model
SSIM	Structural Similarity Index Measure
LPIPS	Learned Perceptual Image Patch Similarity
FID	Fréchet Inception Distance
RNA-seq	A technique to measure gene expression by sequencing RNA molecules
Histopathology	The microscopic examination of tissue to study signs of disease
Synthesis	The process of combining elements to form something new

1 INTRODUCTION

1.1 Overview

In this thesis, Generative Adversarial Networks (GANs) are used to turn RNA-Seq data into pictures of tissues. RNA-Seq data and tissue pictures are both very important for studying cancer. They help scientists figure out what cancers look like and how they work at the molecular level. There are times, though, when not all patients can get either type of information. This makes it harder to study cancer. This work tries to fill this gap with GAN models. GANs can learn trends and make new pictures from different types of data. The data for this work comes from RNA-Seq, and the goal is to make it into a useful picture of the tissue. The test is to see if a model can show how patterns in gene expression change the way tissues look.

There are several parts to the study. First, get the RNA-Seq and microscope data ready. The next step is to teach various GAN models how to link RNA-Seq features to picture space. The images are checked to make sure they are real and match what is known about the tissue. You can also compare and check out how well different GAN models work. The main goal of this thesis is to show that deep learning can connect omics data with image data. If this works well, it can help create virtual tissue images for samples where no histopathology image exists. This can support cancer diagnosis, research, data integration, and future work in digital pathology.

1.2 Background

Digital pathology and transcriptomics provide two complementary views of cancer biology: hematoxylin–eosin (H&E) whole-slide images (WSIs) capture spatial morphology at cellular scale, while RNA-Seq summarizes the active molecular programs of the tissue[1]. Over the last decade, deep learning has shown that morphology encodes rich molecular signals—models can predict bulk RNA-Seq profiles from H&E, recover prognostic embeddings, and localize biology across slides[1] [2] [3]. Likewise, spatial-transcriptomics studies confirm that morphology correlates with gene-expression patterns across tissue regions, validating the premise that histology contains latent molecular information[2] [4] [5].

At the same time, pathology models can infer clinically relevant labels (e.g., mutations, subtypes) directly from WSIs, linking image features to genomic status and outcomes[6]

[7]. Together, these results motivate tighter integration between images and omics to improve cancer phenotyping and decision support.

Cancer is a major cause of death worldwide, and the RNA-CDM paper cites GLOBOCAN for this general cancer burden but does not give lung-cancer-specific numbers[8]. Both RNA-CDM and RNA-GAN use this global cancer burden only as background and do not report new lung cancer statistics[8][9]. RNA-CDM gives TCGA sample sizes and shows that TCGA-LUAD has 520 samples, which is the lung dataset used for their tests[8]. RNA-GAN uses healthy lung tiles as one example but does not give lung cancer rates, focusing instead on the lack of paired WSI and omics data[9].

A deep generative model is used to create synthetic whole-slide image tiles from RNA-seq gene expression data. First, a β -VAE is trained on RNA-seq data to learn a low-dimensional latent vector that keeps the differences between tissues. This latent RNA vector is then used as the input of a deep convolutional GAN instead of pure random noise[9]. The GAN is trained with a generator and a discriminator so that the RNA latent space is mapped to 256×256 H&E image tiles of lung and brain cortex tissue. This combined model is called RNA-GAN, since the GAN is guided by RNA features[9]. After training, new RNA latent vectors can be sampled, passed through RNA-GAN, and used to generate realistic synthetic tiles that are consistent with the gene expression profiles and support data augmentation and data imputation[9].

One promising way to boost performance with biology-aware augmentation: generate synthetic H&E tiles conditioned on a compact RNA-Seq latent (∼200-D) using a cGAN baseline and a lightweight diffusion variant. This expands data, oversamples rare molecular phenotypes, and enables pretraining of WSI encoders on transcriptomically aligned morphology. We prevent leakage with patient-wise splits, assess realism (SSIM/-FID/KID/LPIPS/IS), and verify biological faithfulness (cell-mix, immune signatures). Net effect: targeted, RNA-aligned synthetic tiles that strengthen pathology models when paired RNA–image data are scarce[10] [11]. We use conditional GANs to generate synthetic H&E tiles conditioned on a compact RNA-Seq latent (e.g., ∼200-D). This creates morphology that is both realistic and aligned with the sample's transcriptomic profile, letting us augment scarce paired data, oversample rare molecular phenotypes (immune-hot/cold), and pretrain WSI models. Conditioning is injected into the generator (concat/FiLM) and enforced in the discriminator (projection). We keep patient-wise splits and validate with FID/KID/LPIPS plus biological checks (cell-mix, immune signatures) to ensure the synth data actually helps downstream pathology tasks[10].

Beside, diffusion models as the main generative method to create synthetic histology

image tiles. A diffusion model takes a real image and adds random Gaussian noise to it step by step until the image becomes almost pure noise. We then train a neural network to learn the reverse process. It learns how to remove the noise step by step and recover the original image[8]. After training, we can start from pure noise and iteratively denoise it to create new images that follow the same distribution as the training data. In our work, we use a cascaded diffusion framework[8]. First, one denoising diffusion probabilistic model (DDPM) generates a low-resolution image tile. Then a second DDPM takes this output and produces a higher-resolution tile. This second model is conditioned on compact embeddings of the RNA-seq data. In this way, we generate realistic tumour image tiles that are guided by gene expression and can be used for data augmentation and pretraining[8].

However, integrating WSIs with RNA-Seq in practice faces three recurring obstacles. First, paired datasets (slides with matched RNA) are limited and heterogeneous across institutions, staining protocols, and sequencing pipelines, which impedes training data-hungry models and robust evaluation[1] [2] [3] [7]. Second, most available RNA-Seq is bulk (not spatial), so tile-level supervision for image models is missing, making weak-label training standard and potentially noisy[1] [3]. Third, domain shift across hospitals and scanners can degrade generalization, a challenge repeatedly observed in cross-cohort studies in pathology and spatial-omics[2] [3] [7]. These data constraints motivate generative strategies that can synthesize the missing modality to (i) increase sample diversity, (ii) enable cross-modal learning, and (iii) test biological plausibility under controlled conditions.

At the same time, multi-modal discriminative models strengthen the case for image–RNA integration. HE2RNA learned slide-level mappings from WSIs to bulk RNA-Seq across cancers, providing virtual spatialization of immune and proliferation markers and demonstrating that morphology can predict thousands of genes with significant correlations[1]. Transformer-based tRNAsformer jointly optimizes gene prediction and slide representation learning, improving renal-cell carcinoma subtype classification and image retrieval, including on an external cohort despite domain shift[3]. In spatial settings, ST-Net accurately predicts spot-level expression from H&E in breast cancer and generalizes to a different platform (10x Visium), linking morphology to spatial gene programs[2]. Clinical fusion systems that combine WSI with RNA (and other omics) via late-fusion strategies consistently outperform single-modality baselines for non-small-cell lung cancer diagnosis, underscoring that histology and transcriptomics provide complementary evidence[12] [13]. Collectively, these threads support the central hypothesis of this thesis: if we can generate histology from RNA that is visually realistic and biologically faithful, we can unlock new training regimes and evaluation protocols for multi-modal pathology under limited paired data.

1.3 Problem Statement

Given a patient's bulk RNA-Seq and a target cancer cohort, generate realistic H&E tiles whose visual statistics resemble real tiles and whose biological characteristics (e.g., immune cell proportions, pathway activity proxies) align with the conditioning RNA signal. Evaluate both image realism and biological faithfulness, and test whether synthetic data improves downstream pathology tasks (e.g., classification) under low-data regimes.

1.4 Research Gap

Recent work has begun to explore exactly this idea: conditioning image generators on gene-expression representations to produce realistic histology tiles that are aligned with underlying transcriptomic signals. A diffusion framework (RNA-CDM) conditions cascaded denoising models on a β-VAE RNA latent and yields synthetic H&E tiles whose cell-type distributions match real tissues; pretraining on these synthetic tiles improves downstream classification and microsatellite instability (MSI) prediction under data scarcity[8].

A complementary GAN-based approach (RNA-GAN) conditions a DCGAN on an expression latent and reports better visual quality and faster convergence than vanilla GANs, with benefits for downstream tile classification after self-supervised pretraining on synthetic images[9]. Beyond H&E, the single-cell community has shown that GANs can generate realistic scRNA-Seq profiles for augmentation, and GAN-based imputers can recover dropout-suppressed gene signals—evidence that generative modeling can usefully fill in missing data modes or values in omics[14] [15]. These advances suggest that RNA-conditioned histology synthesis can both (a) serve as a research probe for testing morphological consequences of gene programs and (b) act as a practical data-augmentation tool for pathology models.

Despite encouraging progress, key gaps remain. Most RNA-conditioned generators rely on bulk RNA labels, so tile-level ground truth is absent; ensuring that local morphology reflects the intended molecular state is non-trivial and requires explicit biological evaluations (e.g., cell-type composition, immune signatures) beyond standard image metrics like FID[1] [8] [9]. Paired datasets remain small; many prior studies validate internally and need stronger cross-site tests[2] [3] [6] [7]. Normalization choices for RNA-Seq (e.g., log1p-TPM vs. regularized negative-binomial models) can materially affect the conditioning signal and must be reported and justified[16]. Finally, while diffusion models often improve fidelity and alignment over GANs in text-to-image tasks, compute cost can be high; efficient, deployable variants are desirable for biomedical use[17].

1.5 Objectives

1. Objective 1 : Design an RNA→image pipeline that encodes RNA into a compact latent (e.g., PCA β-VAE ≈200-D) and conditions a generator (cGAN baseline) to produce 128×128 H&E tiles.

2. Objective 2 : Establish an evaluation protocol that couples SSIM/FID/LPIPS with biological alignment metrics (cell-type composition via nuclei/instance segmentation; immune signature proxies) and downstream utility.

1.6 Motivation of the Study

Cancer research uses many types of data. Two important types are histopathology images and RNA-Seq data. These help us understand how tumors grow and behave. They also support better treatment decisions. But in many cases, both types of data are not available for every patient. Whole-slide images (WSIs) are hard to produce. They take time, cost money, and need special tools. RNA-Seq data is easier to collect. But RNA-Seq cannot show the shape and structure of the tissue. Because of this, there is a gap in cancer research when one type of data is missing. Recent studies show that RNA-Seq data has patterns that relate to tissue images. This gives us a new idea. This set of pictures lets you look at digital signs of diseases and cancers. GANs can make new pictures and learn a lot of different designs. They are very good at drawing pictures. You can look at digital diseases, signs, and cancers with this set of pictures. GAN models can learn lots of different patterns and make new pictures. Their picture-making skills are very good. So, GANs may help us translate RNA-Seq data into tissue images. This can change how we view and understand genomic information. This study will explore if GAN models can learn the link between RNA-Seq data and tissue images. The goal is to create meaningful images from gene expression data. This could make it easier to share data, study cancer, and make better tools for precise treatment.

1.7 Thesis Organization

Chapter 2 surveys related work in image-omics modeling, RNA-conditioned synthesis, spatial inference, and multi-modal fusion that motivate our approach. **Chapter 3** details datasets, RNA normalization, conditioning, model architectures, and training/evaluation protocols. **Chapter 4** reports results on realism, biological alignment, and downstream tasks, with ablations and qualitative analyses. Chapter 5 concludes with limitations (e.g., bulk labels, domain shift) and outlines future directions such as spatial-RNA conditioning and slide-level synthesis.

1.8 Summary

This chapter introduced the cancer burden, biological foundations, RNA-Seq mechanisms, histopathology importance, single-modality limitations, cross-modality necessity, prior work, and motivations for low-resolution generative models. Together, your provided answers establish the need for RNA-conditioned histology generation and justify the research direction taken in this thesis.

2 LITERATURE REVIEW

2.1 Introduction

Cancer is a major health problem in the world. It is caused by both genes and the environment. Early and correct detection is very important to save lives.

Today, doctors can collect many kinds of data from a patient. One type is RNA-seq data, which shows how active each gene is. Another type is tissue images from H&E-stained slides. These data give different views of the same tumour. But in practice, both kinds of data are not always present for every patient.

Generative models and deep learning are now used to study cancer. Large datasets can teach these models how to find trends. They can also make new data that looks like real data. This is called "synthetic data." This fake data can be useful when real data is scarce or missing.

Recent work has tried to link RNA-seq data with tissue images. Carrillo-Perez et al. proposed RNA-GAN, which first compresses RNA-seq data with a variational autoencoder and then uses a GAN to generate lung and brain cortex tiles from the latent RNA features. Pathologists often preferred these RNA-GAN tiles over tiles from a standard GAN, and the model also helped improve a cancer image classification task. [9]

A more recent paper from the same group used cascaded diffusion models to generate tumour image tiles directly from RNA-seq profiles. This work showed that diffusion models can produce high-quality synthetic tumour tiles and can also support downstream analysis tasks. Together, these studies show that it is possible to generate realistic tissue images from gene expression data.

2.2 Previous Works

Carrillo-Perez et al. [8] introduce RNA-CDM, a cascaded diffusion framework that synthesizes realistic H&E tiles directly from RNA-seq by conditioning two DDPMs on a 200-D β-VAE latent of the gene-expression profile (17,655 genes), enabling multicancer RNA-to-image generation across LUAD, KIRP, CESC, COAD and GBM without using tissue labels. The architecture first denoises a 64×64 tile and then super-resolves to 256×256 with U-Net backbones ($\sim$1.8B params; trained $\sim$8 days on A100s). Quality

metrics on 50k synthetic vs 50k real tiles show good fidelity (FID50k 23.36; KID50k 0.015; IS50k 3.19), substantially better than a label-conditioned baseline (FID 55.81; KID 0.038; IS 2.22). Using HoverNet, the authors demonstrate that synthetic tiles preserve cell-type distributions seen in real data across cancers, and that conditioning on deconvolved haematopoietic RNA increases lymphocyte content in generated tiles; notably, GBM shows lower lymphocytes, consistent with biology. Synthetic tiles generalize when conditioned on external RNA-seq cohorts and are useful for scarce-data settings: replacing real tiles with synthetic ones does not hurt multicancer classification, and pretraining on synthetic tiles improves accuracy/F1 and boosts downstream tasks such as MSI prediction (gains up to $\sim$12 percentage accuracy at low data) and pediatric glioma prognosis in a MIL setup. Limitations include reliance on bulk RNA (no tile-level ground truth), diffusion sampling speed vs GANs, and potential cohort-shift in β-VAE latents; future work points to spatial transcriptomics for local supervision. Overall, RNA-CDM provides biologically coherent RNA-to-histology synthesis and practical augmentation for digital pathology models, with 1 million synthetic tiles released for the community. In implementation details, WSIs are preprocessed at 20$\times$ into non-overlapping 256$\times$256 tiles after tissue masking and background/low-contrast filtering, and patient-wise splits are used to avoid leakage; the β-VAE is trained across 12 TCGA cancers, while the diffusion models are trained on five tumor types, with evaluation including a 100k-tile HoverNet analysis (50k real/50k synthetic). Additional utility gains are reported when combining synthetic pretraining with self-supervised (SimCLR) objectives, and the authors note compute/memory trade-offs and the need for cross-site validation and stain/scanner harmonization[8].

Carrillo-Perez et al. [9] introduce RNA-GAN, a cross-modal generative framework that uses a β-VAE to embed gene-expression profiles and conditions a DCGAN to synthesize H&E tile of brain cortex and lung tissue. Compared with a vanilla GAN, RNA-GAN generated tiles that pathologists rated higher, preserved tissue-specific distributions in feature space, converged in fewer epochs, generalized to external RNA-seq datasets, and even improved a downstream task:self-supervised pretraining on synthetic tiles boosted GBM vs LUAD tile classification accuracy and F1 scores compared with training from scratch. Furthermore, RNA-GAN demonstrated significantly fewer artifacts in the generated tiles, especially for lung tissue, and the synthetic tiles maintained morphological features similar to those of real tissue. Their findings support expression-guided histology synthesis as a viable way to mitigate multi-modal data scarcity and to pretrain models when paired WSI–RNA data is limited. This approach shows promise for augmenting datasets in settings where multimodal data are sparse, such as rare diseases or when tissue slides are unavailable or degraded[9].

Schmauch et al. [1] present HE2RNA, a weakly supervised deep-learning pipeline that learns to predict bulk RNA-Seq expression directly from H&E whole-slide images (WSIs) across 28 TCGA cancer types (∼8,725 patients) and then "spatializes" gene expression on the slide without manual annotations. The model aggregates ResNet-50 tile features (224×224 px patches) into "supertiles" and trains a multilayer perceptron (MLP) to regress log-transformed FPKM-UQ values for 30,839 expressed genes filtered from TCGA RNA-Seq data. Training uses five-fold cross-validation and Holm–Šidák or Benjamini–Hochberg correction to assess significance. Across cancers, it significantly predicts thousands of genes, with strongest signals in immune and cell-cycle pathways.[1] The graphical abstract on paper outlines three applications: transcriptome prediction, virtual spatialization, and transfer learning. Spatialization is validated by correlating predicted per-tile expression of CD3/CD20 with IHC on independent slides (e.g., CD3 R $\approx$ 0.51; ROC–AUC up to 0.93 for high T-cell tiles)[1] and by aligning predicted MKI67 expression with pathologist-annotated tumor regions in LIHC, where AUCs rise with more advanced BCLC stage (pp. 9–10). As a clinical use-case, a lower-dimensional "transcriptomic representation" learned by the model improves microsatellite instability (MSI-H vs MSS) prediction in low-data regimes versus direct WSI models and an autoencoder baseline (AUC $\approx$ 0.81 vs 0.68–0.72 when only 25 percentages of data are available.[1] Strengths include its large-scale TCGA dataset, cross-cancer generality, and explicit interpretability via gene-level heatmaps. Limitations include reliance on bulk RNA-Seq labels (no tile-level ground truth), variable performance by cancer type/sample size, and weak prediction for housekeeping genes, which supports specificity to cancer biology rather than global staining artifacts.[1] In general, HE2RNA shows that histology-only models can be useful substitutes for transcriptome assays and can be used to represent several modalities for clinical characteristics that come later[1].

Nakagawa and Fujita discuss [18] how whole-genome sequencing (WGS) for cancer may identify noncoding mutations, structural variants (SV), copy-number changes (CNA), mitochondrial variants, and pathogen integrations, besides coding regions. This gives a more complete picture of driver events and mutational processes than WES. They outline practical WGS trade-offs (short-read depth vs. mapping errors in repeats, PCR bias, and heavy compute/data-sharing demands) and show that variant-caller agreement remains imperfect, especially for indels, based on ICGC benchmarking. The paper synthesizes emerging biology from noncoding/regulatory hotspots (TERT promoter), CTCF/cohesin site mutations, enhancer hijacking (e.g., activating GFI1 family), and clinically actionable SV/fusions (EML4–ALK, RET/ROS1, TMPRSS2–ERG). It highlights WGS-enabled mutational signatures (exposure and repair defects—smoking, aflatoxin, aristolochic acid; BRCA/HRD; MMR/MSI) and immunogenomic insights (PD-L1 3′UTR disruption; aneuploidy correlating with reduced immune infiltration), arguing for integrative analyses

with RNA-Seq, epigenomics, and clinical data to interpret noncoding/SV consequences and guide precision medicine[18].

Chen et al. [6] develop tile-based Inception-V3 classifiers on H&E WSIs to (i) distinguish HCC from normal liver and (ii) grade tumors, then extend to mutation prediction for common/prognostic genes. Using 491 TCGA/GDC slides (402 HCC, 89 normal) plus an external SRRSH cohort (67 HCC, 34 normal), they report strong tumor-vs-normal performance (validation AUC = 0.961; MCC = 0.82) and reasonable grading (accuracy = 0.896; MCC = 0.738), approximating ∼5-year pathologists on a 101-slide reader study. For mutation status, four genes—CTNNB1, FMN2, TP53, ZFX4—were predictable from histology with external per-slide AUCs ∼0.72–0.90 (CTNNB1 up to 0.898), using aggregation by average probability or percentage of positive tiles. The pipeline crops 20× WSIs into non-overlapping 256×256 tiles (≥80 percentage tissue), applies patient-wise splits to avoid leakage (3:1 train/validation on TCGA), and fine-tunes ImageNet-pretrained Inception-V3 via Baidu EASY-DL; mutation targets (from an initial set of ∼10 candidates) are selected with LASSO-Cox to prioritize frequent and prognostically informative alterations. Evaluation includes internal validation and truly independent external testing, with tile-to-slide aggregation and reporting of AUC/MCC alongside accuracy, ROC/PR curves, and confusion statistics for clinical interpretability. Strengths include use of a sizable public cohort, explicit patient-level partitioning, an external validation set to assess generalization, and comprehensive metrics. Limitations are the relatively small and potentially domain-shifted external set (scanner/staining/frozen-section effects), platform/language constraints that may hinder reproducibility, a focus on a subset of mutations, and limited model interpretability, all of which underscore the need for multi-center studies and standardized preprocessing before clinical deployment. Overall, the work supports mutation-from-histology pre-screening and complements radiogenomic efforts linking morphology to genomics in HCC, while providing design cues (tiling, aggregation, patient-wise splits, external validation) that inform our subsequent methodology and evaluation choices[6].

Coudray et al. [7] (DeepPATH) show that an Inception-V3 CNN trained on TCGA whole-slide H&E images can both classify NSCLC subtypes and infer certain LUAD mutations directly from histology. Using 1,634 WSIs (1,176 tumor; 459 normal), tiles of 512×512 at 20× (also 5×) and per-slide probability aggregation, the model achieves near-perfect tumor-vs-normal discrimination (AUC ≈0.99) and strong LUAD-vs-LUSC performance (AUC ≈0.95), with three-way normal/LUAD/LUSC AUCs ≥0.968 on TCGA. Generalization is demonstrated on independent NYU cohorts—frozen (n=98), FFPE (n=140), and biopsies (n=102)—with AUCs 0.83–0.98 depending on magnification and tissue type; an automatic tumor-ROI selector performs comparably to manual ROIs.

Extending to mutation prediction within LUAD, a multi-output network trained on the ten most prevalent genes identifies six predictable targets (STK11, EGFR, FAT1, SETBP1, KRAS, TP53) with per-slide AUCs ∼0.73–0.86 and shows that predicted probabilities correlate with allele frequency; an external EGFR FFPE cohort yields AUC 0.687 overall (0.75 on sequencing-validated cases), highlighting domain-shift challenges (frozen→FFPE) and the need for more mutated slides. Overall, DeepPATH establishes histology-only baselines for subtype diagnosis and mutation pre-screening in NSCLC, while underscoring limits from sample preparation, artifacts, and cohort differences. Low-content tiles are filtered before training, patient-wise splits are used to prevent leakage, and slide-level predictions are aggregated by mean probability or fraction of positive tiles; notably, 5× sometimes outperforms 20× by capturing coarse architecture. The study reports ROC/PR curves with confidence intervals and releases code, enhancing reproducibility. Key limitations remain smaller positive cohorts for some mutations and incomplete interpretability, motivating cross-site stain/scanner harmonization and standardized preprocessing. For our thesis, these practices (patient-level splitting, external validation under domain shift, and biologically grounded evaluation) directly inform how we will assess RNA-conditioned histology generation and its downstream utility[7].

Li et al. [19] present SDINet, a two-branch CNN that fuses high-level semantics with multi-scale detail (via ASPP and a no-squeeze channel-attention "Se Var") to infer TF to gene interactions from gene-expression images, then extend the idea to RNA-seq and finally combine both modalities for GRN inference in Drosophila eye development. Images (FlyExpress ISH, stage-16 embryo) are paired by concatenating TF and target images (the height-wise strategy performed best), while RNA-seq from 72 eye-antennal disc samples is converted into 64×64 TF–target 2D histograms. SDINet (improved ResNet-18) attains Acc 0.7196/F1 0.7374/AUC 0.7930/AUPR 0.7738; the VGG-based RNA model reaches Acc 0.8962/F1 0.8950/AUC 0.9565/AUPR 0.9473; and the fusion network (global-average-pooled features + FC fusion) improves further to Acc 0.9116/F1 0.9118/AUC 0.9653/AUPR 0.9614 on 86 TFs, 1,152 targets (2,454 positive and 2,446 negative links; 85k image pairs). Figures highlight the workflow and fusion design [19], the fusion block [19], and ablation/benchmark tables [19] showing the benefit of multi-scale detail and attention. The biological analysis [19] demonstrates that integrating images and RNA-seq recovers interactions that single-modality models miss, supporting complementarity between spatial and bulk signals. Limitations include species-specific data and a developmental-stage mismatch (embryo images vs larval RNA-seq), but the study provides a clear template for multimodal GRN inference with lightweight models. [19] The image-pair concatenation was explicitly benchmarked—height-wise (256×320) outperformed width-wise and channel-wise merges ; the multi-scale branch performed best when added after Conv4x , and the fusion block itself provided a measurable lift versus a

joint model without it (Acc 0.9116/F1 0.9118/AUC 0.9653 vs 0.9088/0.9106/0.9625;).[19] Training used Adam (lr = 3e-4), 60 epochs, with binary cross-entropy, and experiments ran on two RTX 2080 Ti GPUs; evaluation metrics included Accuracy, F1, AUC, and AUPR. Code and data references are provided by the authors[19].

Moncada et al. [4] integrate microarray-based spatial transcriptomics (ST) with scRNA-seq in primary PDAC and introduce Multimodal Intersection Analysis (MIA) to assign scRNA-defined cell types, subpopulations, and cancer cell states to spatial tissue regions. From two tumors profiled in parallel (PDAC-A/B), scRNA-seq identifies 11-15 populations and CNV profiles distinguish malignant clusters from non-malignant ductal cells [4]. Using fresh-frozen sections with matched H&E, ST spots (100 µm) capture ∼20-70 cells each and cluster into histology-consistent regions (cancer/desmoplasia, ducts, stroma, pancreatic tissue), enabling MIA to test hypergeometric overlap between region-specific genes and cell-type markers.[4] Ductal cells resolve into four subpopulations—APOL1-high/hypoxic (CA9/ERO1A), terminal (TFF1-3), centroacinar (AQP3/CFTR), and antigen-presenting (MHC-II)- with distinct spatial enrichments; in PDAC-A, hypoxic and terminal ductal subsets are enriched in cancer regions, whereas in PDAC-B all ductal subsets localize to ducts.[4] Immune compartments also partition spatially (M1 vs M2 macrophages; dendritic cell A vs B). Sub-clustering of cancer-rich ST regions shows co-enrichment patterns: fibroblasts colocalize with cancer cluster 1 but not cluster 2, suggesting distinct stromal interactions.[4] Non-negative matrix factorization of cancer cells defines hypoxia, oxidative-phosphorylation, and a stress-response module; "stress-high" ST spots consistently co-enrich with inflammatory fibroblasts across ten arrays spanning six patients, and this association persists at the bulk level in TCGA PDAC (R≈0.49 for stress module vs inflammatory-fibroblast signature) with supporting IL-6 immunofluorescence near KRT19+ cancer epithelium.[4] Methodologically, MIA is platform-agnostic and region-based, mitigating spot-level mixing; results are further supported by protein-level validation (TM4SF1, S100A4) and portability to an external melanoma cohort. Overall, the study provides a scalable blueprint for mapping single-cell states onto tissue architecture and uncovering cancer–CAF inflammatory crosstalk in PDAC, while acknowledging ST's limited resolution, partial coverage, and potential transcript diffusion[4].

He et al. [2] present ST-Net, a deep learning framework that links histomorphology in H&E slides to spot-level spatial gene expression in breast cancer. Using DenseNet-121 on 224×224 px patches aligned to 100 μm spots, the model was trained with leave-one-patient-out cross-validation on 30,612 spots from 68 sections (23 patients). It predicted expression (log-normalized) for a 250-gene panel with an average RMSE of 0.31- about a 1.4× error factor—and yielded significant, cross-patient positive correlations for 102 genes

at FDR 0.05. [2] External validation on an independent 10x Visium invasive ductal carcinoma sample required no re-tuning and showed 207/234 genes with positive correlation (mean r≈0.33) and mean AUROC≈0.73 for high/low expression calls. Generalization to TCGA was assessed by aggregating patch-level predictions into "pseudo-bulk" profiles: 177/249 genes correlated positively with bulk RNA-seq (55 significant), and predicted profiles distinguished ductal vs. lobular histology with AUROC 0.83, slightly exceeding subtype classification directly from H&E (0.81). Beyond cross-dataset robustness, ST-Net captured intra-tumour heterogeneity 63 genes remained predictable within tumour-only regions and 27 within normal-only regions and outperformed three baselines (pathologist tumour/normal labels, nuclei-feature random forest, and a composition-based model) for the vast majority of genes.[2] Interpretability via integrated gradients highlighted enlarged nuclei for high FASN predictions, while the learned 1,024-D latent space separated tumour from normal patches (mean purity ∼0.89;). Overall, ST-Net demonstrates that spatially resolved transcriptomic patterns can be inferred from routine pathology, enabling image-based screening of spatial biomarkers and offering a scalable bridge between morphology and gene expression[2].

Alsaafin et al. [3] introduce tRNAsformer, a multitask multiple-instance Transformer that learns from H&E whole-slide images (WSIs) to (i) predict slide-level bulk RNA-seq and (ii) produce compact slide embeddings usable for renal cell carcinoma (RCC) search and classification. Each WSI is summarized by 49 spatially clustered 224×224 tiles, embedded by DenseNet-121, linearly projected to 384-D, and processed with L Transformer encoders. A linear head handles RCC subtype classification, while a 1D-conv head generates tile-wise gene scores that are aggregated with a top-n scheme into slide-level expression—explicitly modeling inter-tile dependencies that MLP baselines neglect. Using TCGA KIRC/KIRP/KICH with 31,793 genes (FPKM-UQ; log10(1+a)), case-wise splits, and leave-one-patient-out evaluation, tRNAsformer achieves gene–RNA correlations comparable to or slightly better than HE2RNA; the number of statistically significant genes peaks for moderate depth (L≈2–8), with best overall error at L=4. For downstream tasks, the learned representations classify RCC subtypes on TCGA with high accuracy (up to 96.25 percentage, macro-F1 ≈0.95) and generalize to an external Ohio State cohort of 142 WSIs with an expected domain-shift drop (∼13 percentage absolute) yet still outperform a strong "Low Power" baseline. In content-based WSI search, tRNAsformer surpasses Yottixel on TCGA (MAP@5/10 ≈0.915/0.912) and remains competitive externally (≈0.78–0.80). The approach is efficient (49 tiles per slide), uses fewer hyperparameters than an MLP counterpart, and unifies molecular supervision with morphology to yield transferable slide embeddings. The limitations encompass dependence on bulk (rather than spatial) RNA-seq for oversight, evaluation restricted to the kidneys, and susceptibility to acquisition/site shift; the authors indicate the necessity for future

validation through spatial transcriptomics and more extensive multi-center cohorts. In general, tRNAsformer improves the ability to predict image-to-transcriptome while also providing useful WSI representations for search and classification[3].

Carrillo-Perez et al. [12] address tri-class NSCLC diagnosis (LUAD, LUSC, control) by fusing five TCGA modalities: RNA-Seq, miRNA-Seq, DNA methylation, CNV, and H&E whole-slide images (WSIs) with a late-fusion strategy that learns class-specific weights, rather than relying on fixed voting. Their single-layer neural network fuses per-modality class probabilities and is optimized by gradient descent, enabling per-class weighting and graceful handling of missing modalities.[12] Each modality is modeled independently: SVMs with mRMR-selected features for omics; a ResNet-18 fine-tuned on 512×512 WSI tiles with slide-level aggregation for imaging. Data span thousands of cases per source (e.g., 1,420 WSIs; 980 RNA-Seq; 883 methylation), with stratified, patient-wise 10-fold cross-validation to avoid leakage.[12] Fusion consistently improves diagnosis over single sources. Using all five modalities yields Accuracy $\approx 96.81 \pm 1.07$, F1 $\approx 96.82 \pm 1.07$, AUC $\approx 0.993 \pm 0.004$, AUPRC $\approx 0.980 \pm 0.016$. Two- and three-source composites already perform strongly- WSI + RNA-Seq and WSI + RNA-Seq + miRNA are notable—while gains beyond three or four sources are modest. The fusion model reduces misclassifications across modalities (e.g., CNV errors drop by ∼8.6 percentage absolute) and outperforms prior single-modality baselines reported for NSCLC histology and omics.[12] Some of the strengths are a strict patient-wise CV, clear per-class weighting, and the ability to handle missing data. The limitations encompass the absence of external validation and the diminishing returns associated with the integration of all five sources. This study provides a pragmatic framework for the late integration of multi-omics and whole slide images (WSIs), empirically illustrating that morphological characteristics enhance molecular profiles for precise NSCLC subtype classification[12].

Carrillo-Perez et al. [13] fuse RNA-Seq and H&E WSI information to improve three-way NSCLC diagnosis (LUAD/LUSC/healthy) using a simple late probability fusion scheme. A ResNet-18 per-tile CNN (20×; 512×512, background-filtered) assigns tile classes; slide-level probabilities are the fraction of tiles per class. In parallel, RNA-Seq is processed with limma/KnowSeq (COV=2), genes ranked by mRMR, and an RBF-SVM outputs calibrated per-class probabilities. The final prediction is a weighted sum of probabilities, with weights (α_1, α_2) set from resampled mean F1 on the training fold; across 10-fold patient-wise CV, these weights are ∼0.5/0.5.On the 950 cases with both modalities, fusion using a compact 6-gene panel (SLC2A1, NTRK2, TOX3, NXPH4, TFAP2A, KRT13) outperforms either source alone: WSI F1 83.39/Acc 86.03/AUC 0.947; RNA-Seq (6 genes) F1 93.67/Acc 93.70/AUC 0.987; fusion F1 95.19/Acc 95.18/AUC 0.991, cutting misclassifications to 46 vs 133 (WSI) and 60 (RNA-Seq). Figures and tables doc-

ument the pipeline [13], ROC gains per class, confusion matrices, and error reductions. Strengths include transparent, modality-agnostic fusion that also works when one modality is missing; limitations are TCGA-only internal validation, class imbalance for healthy, and majority-vote slide aggregation. Overall, probability-level late fusion is a practical, high-yield strategy to combine histology and gene expression for NSCLC subtype diagnosis. A late-fusion model combines slide-level CNN probabilities from H&E tiles with SVM probabilities from a compact RNA-Seq gene panel to classify LUAD/LUSC/healthy. Across 10-fold patient-wise CV on 950 cases with both modalities, the 6-gene fusion (SLC2A1, NTRK2, TOX3, NXPH4, TFAP2A, KRT13) achieves F1 95.19 percentage, Acc 95.18 percentage, AUC 0.991 and reduces errors to 46 vs 133 (WSI) and 60 (RNA-Seq alone).[13] The fusion weights α_1 (WSI) and α_2 (RNA-Seq) are data-driven via resampled F1 on the training fold and end up $\sim$0.5/0.5 across splits (p. 12), indicating complementary strength. The pipeline diagram shows the per-tile CNN, SVM probability calibration, and the weighted-sum fusion; class-wise ROC curves illustrate consistent fusion gains for all three classes[13].

Saharia et al. [17] introduce Imagen, a cascaded text-to-image diffusion model that pairs a frozen large language model (T5-XXL) with a 64×64 base diffusion model and two text-conditional super-resolution stages (64→256→1024) using classifier-free guidance and noise-conditioned augmentation. A key finding is that scaling the text encoder improves image fidelity and image–text alignment more than scaling the image U-Net; this is visualized by Pareto curves contrasting encoder vs. U-Net size. [17] To stabilize high guidance, they propose dynamic thresholding, which clips the current $\hat{\mathbf{x}}$ to a data-driven [−s, s] range at each step, preventing saturation and enabling sharper, better-aligned samples. Imagen is trained on ≈860M image-text pairs (≈460M internal + ≈400M LAION-400M) and evaluates zero-shot on COCO with a state-of-the-art FID-30K of 7.27.[17] Human evaluation shows caption alignment on par with COCO references (≈91.4 vs 91.9), and photorealism preference of 39.5 percentage overall, rising to 43.9 percentage on images without people. The paper also introduces DrawBench (200 prompts across 11 categories) for compositional testing; raters prefer Imagen over DALL-E 2, GLIDE, and VQGAN+CLIP on both fidelity and alignment.[17] Architectural details include an Efficient U-Net for super-resolution that converges faster and is more memory-efficient,[17] and text cross-attention in all stages. Limitations include poorer realism for people, social bias inherited from web data, and non-release of the model. For your thesis, Imagen's dynamic thresholding, noise-conditioned super-resolution, and frozen-LM conditioning are transferable design choices for histopathology or omics-image synthesis pipelines[17].

2.3 Conclusion

RNA-seq and histology images give rich and complementary information about tumours. Generative models such as GANs and diffusion models can learn a link between these two data types and can create synthetic tiles that look close to real ones. These synthetic tiles can support tasks like image classification and can reduce the impact of data scarcity[9].

However, there are still gaps. Many existing models focus on healthy tissues, or on a limited set of tumour types. The number of datasets with both RNA-seq and matched histology images is still small. In many cases, the focus is mainly on visual quality, while the link between synthetic images and underlying gene expression is less explored. These limits make it hard to fully use synthetic images for real cancer patients[9].

In this thesis, we aim to address some of these gaps. We focus on cancer patients with both RNA-seq data and histopathology images, mainly from large public datasets such as TCGA and GTEx. We plan to build a conditional GAN model that takes RNA-seq data as input and generates tissue images. We will judge the model using both image quality scores and simple tests of biological alignment with gene expression. The final goal is to provide a practical way to generate missing tissue images from RNA-seq data for cancer studies[9].

3 RESEARCH METHODOLOGY

3.1 Overview

I investigate the feasibility of generating histopathology images from RNA sequencing (RNA-Seq) data by using deep generative models. We apply and compare four current generative models Conditional Generative Adversarial Network (GAN) Conditional Variational Autoencoder (VAE) Conditional Diffusion Model and a new Hybrid GAN-VAE model that combines GANs and VAEs. The process consists of data preprocessing, model architecture design, training, and thorough quantitative evaluation using various image quality metrics.

Our first goal was to establish a RNA-to-histopathology image baseline and our second goal was to determine the best generative modeling approach for performing cross-modal conversion from RNA to histopathology images. The pipeline of the study is shown in Figure 3.1.

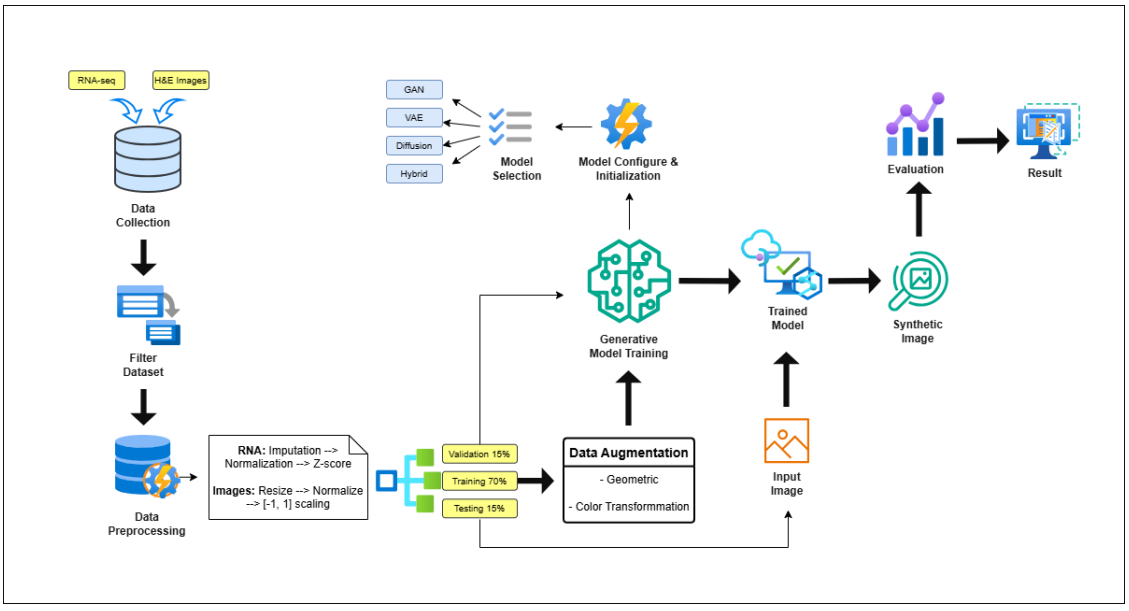


Figure 3.1: Model Architecture

3.2 Data Preparation

3.2.1 Data collection

The dataset had 50 samples from patients, and each one had:

RNA-seq Data: 16,383 gene expression values per patient from bulk RNA sequencing profiles. A cleaned gene-by-sample table (rna_matrix_cleaned.csv) download from[1] UCSC XENA Database.

Histopathology Photos: About 400 H&E-stained tissue images for each patient (128×128 pixels, RGB format). Pre-extracted 256×256 H&E tiles from TCGA-LUAD diagnostic slides download from[2] GDC Cancer Portal.

From these sources we curated a LUAD cohort with 401 patients having both RNA-seq and at least one diagnostic slide. To keep the dataset balanced and GPU-friendly, we sampled up to ∼50 tiles per patient, yielding ∼20,000 tiles overall while preserving at least one tile per patient. Governance and ethics – All images originate from open-access TCGA diagnostic slides; no protected health information is present. We made a one-to-many mapping by pairing each patient's RNA profile with several picture patches from the same tissue sample. This was useful for conditional image production.

Statistics for the dataset: i) There are 50 patients in all. ii) There are about 20,000 photos in all, or 400 for each patient. iii) 16,383 genes make up RNA. iv) RGB image resolution: 128×128×3.

3.2.2 Data Preprocessing

Preprocessing RNA-seq:

1. **Imputation of missing values:** The median expression across all samples was used to fill in missing gene expression values: For every gene g:

$$RNA[g] = median(RNA[g]) if RNA[g] is NaN \tag{3.1}$$

2. **Dealing with outliers:** NaN was used to replace infinite values, and then the

[1] https://xenabrowser.net/datapages/?dataset=TCGA.LUAD.sampleMap/HiSeqV2&host=https://tcga.xenahubs.net

[2] https://portal.gdc.cancer.gov/analysis_page?app=Projects

values were filled in as shown above:

$$RNA[RNA \equiv \pm\infty] = NaN \tag{3.2}$$

3. **RNAStandardization:** The values for gene expression were adjusted using z-scores:

$$RNA_norm[g] = (RNA[g] - \mu[g])/\sigma[g]), \tag{3.3}$$

where μ[g] is the average and σ[g] is the standard deviation of gene g across all samples.

Preprocessing images:

1. **Resizing:** All images were scaled to 128×128 pixels using bilinear interpolation to make sure they were all the same size.
2. **Normalization:** The pixel values were changed from to [-1, 1]:

$$I_norm = (I/255.0 - 0.5)/0.5 \tag{3.4}$$

3. **Standardizing the format:** All photos were changed to RGB format (3 channels).

3.2.3 Data Splitting

To stop data from leaking, the dataset was segmented by patient so that photos from the same patient only showed up in one split:

1. **Training set:** 34 patients (70%) → roughly 13,600 photographs
2. **Validation set:** 8 patients (15%) with over 3,200 photos
3. **Test set:** 8 patients (15%) → ∼3,200 pictures

This dividing at the patient level is critical for testing how well the model performs on new biological material, not just new pictures of patients we already know. The split was done only once and stayed the same for all model investigations so that the findings could be compared fairly.

3.2.4 Data Augmentation

We used the following methods to make the training data more diversified and the model stronger:

Changes in forms: i) Randomly flip from left to right (p = 0.5). ii) Randomly flip vertically (p = 0.5). iii) Turning 90° at random (p = 0.5). iv) Random affine transformation (±10% translation)

Adding color: i) Changing the brightness by up to 20% at random. ii) Change the contrast at random (±20%). iii) Changes in saturation happen at random (±20%). iv) Change the color randomly (±10%).

These improvements keep the biological features of histopathology images while almost increasing the amount of training examples that are relevant. We didn't add more data to the validation and test sets so that the evaluation would stay accurate.

3.3 Model Selection

We chose four generative modeling methods since they are well-known in the field of medical picture synthesis:

1. **Conditional GAN:** Famous for making sharp, realistic images by training with adversaries
2. **Conditional VAE:** Gives stable training and a clear picture of the latent space
3. **Conditional Diffusion Model:** The best way to generate images right now, and it has been shown to work better than other methods in recent benchmarks.
4. **Hybrid GAN-VAE:** A new type of architecture that combines VAE's structured learning with GAN's sharpness in adversarial situations.

All models were built on RNA-seq profiles so that they could be controlled to generate based on molecular attributes.

3.4 Model Details

3.4.1 Conditional GAN Architecture:

Generator: Input is Latent noise vector $z \in \mathbb{R}^{128}$+ RNA characteristics $\in \mathbb{R}^{16'383}$. RNA embedding is a three-layer MLP with 16,383 inputs, 2,048 outputs, 1,024 outputs, and 512 outputs. Architecture is a 4-layer transposed convolutional network with connections that go back to the previous layer. The output is a 128×128×3 picture. Tanh (final layer) and LeakyReLU (deep layers) are the activation functions. Normalization of batches. There are 49,242,339 total parameters.

Discriminator: Input is (128 × 128 × 3) + RNA characteristics (16,383). Architecture is a 4-layer convolutional network that pulls out features. RNA integration combined with flattened image features. TRNA integration combined with flattened image features. Outputis a single logit that tells the difference between real and fake. Dropout is 0.1 to 0.3 for regularization. There are 73,551,041 total parameters.

Configuration for training: Loss is Binary cross-entropy using logits. Adam is the optimizer (β_1=0.5, β_2=0.999). Learning rate is 5×10^{-5} for the generator and 2×10^{-4} for the discriminator. 16 is the batch size. Epoch size is 50. Label smoothing: 0.9 for real labels and 0.1 for fake labels.

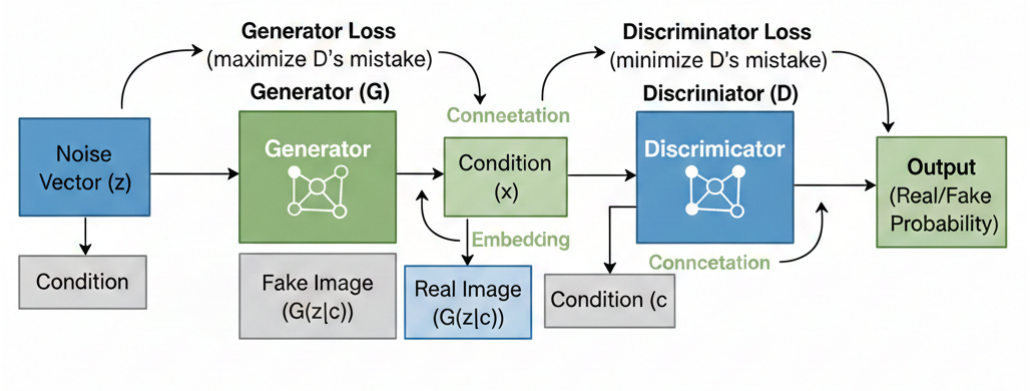


Figure 3.2: Model Architecture of Conditional GAN

3.4.2 Conditional VAE Architecture

Encoder: Input image (128 × 128 × 3) and RNA features (16,383). Architecture is a four-layer convolutional network. Latent space is μ, $log(\sigma^2) \in \mathbb{R}^{256}$. Reparameterization is $z = \mu + \sigma \odot \epsilon$, where $\epsilon \sim$ N(0,I).

Decoder: As input Latent vector $z \in \mathbb{R}^{256}$ + RNA traits. A 4-layer transposed convolutional network for the architecture. An image that has been rebuilt (128×128×3). There are 65,124,483 parameters in all.

Getting training ready: $L_{total} = L_{recon} + \beta \cdot L_{KL}$ is the loss. Where, L_{recon}: The average squared difference between the input and the reconstruction and L_{KL}: The KL divergence between $q(z|x)$ and $N(0, I)$. Adam is the optimizer (β_1=0.9, β_2=0.999). Rate of learning is 110^{-4}. Batch size is 16. 40 is epochs.

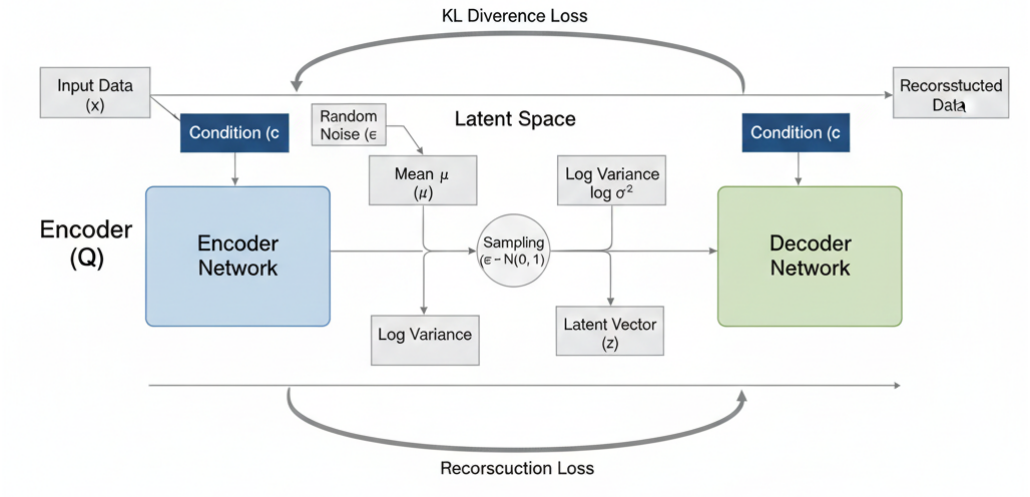


Figure 3.3: Model Architecture of Conditional VAE

3.4.3 Conditional Diffusion Model Architecture

U-Net base: As input, a picture with noise, a timestep t, and RNA features. Architecture is an Encoder-decoder with connections that skip over them. Base channels are 64 (may go up to 256 in the bottleneck). Time embedding is 128-dimensional sinusoidal positional encoding. RNA embedding is a 3-layer MLP (512 → 256). Time and RNA conditioning for each layer. There are over 15,000,000 total parameters.

The process of diffusion: A linear noise schedule with β between 0.0001 and 0.02. T is 1000 timesteps. Using a trained U-Net to remove noise. Sampling is DDIM (50 steps for quick creation).

Setting up training: The average squared difference between the projected and real noise. AdamW (decay of weight 0.01). 1×10^{-4} with cosine annealing is the learning rate. Batch size is 8 (due to memory limits), 50 is epochs. Gradient clipping: Maximum

norm 1.0.

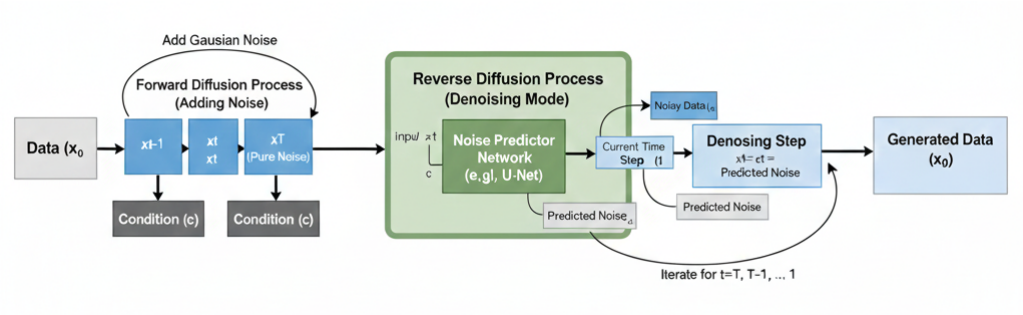


Figure 3.4: Model Architecture of Conditional Diffusion

3.4.4 Hybrid GAN-VAE Architecture

Generator: A VAE-style convolutional encoder for getting features. Latent space is μ, $\log(\sigma^2) \in \mathbb{R}^{128}$(less than pure VAE). Decoder path is transposed convolutions in the style of GANs for sharp generation. Used at both the encoding and decoding stages.

Discriminator: Standard convolutional discriminator that can match features.Returns intermediate feature maps for loss of feature matching.

Training setup: Loss with more than one goal:

$$L_G = \lambda_{\text{adv}} L_{\text{adv}} + \lambda_{\text{recon}} L_{\text{recon}} + \lambda_{\text{perc}} L_{\text{perc}} + \lambda_{\text{KL}} L_{\text{KL}} + \lambda_{\text{feat}} L_{\text{feat}} \tag{3.5}$$

where:
L_{adv}: Adversarial loss (goal of GAN)
L_{recon}: L1 reconstructive loss (goal of VAE)
L_{perc}: LPIPS loss of perception
L_{KL}: KL difference
L_{feat}: Loss from matching features
Weights:
λ_{adv}=1.0,
λ_{recon}=10.0,
λ_{perc}=5.0,
λ_{KL}=0.5,
λ_{feat}=2.0

Optimizer is Adam (β_1=0.5, β_2=0.999). Learning rate is 5×10^{-5} for the generator and 2×10^{-4} for the discriminator. The size of the batch is 16. Epochs 60.

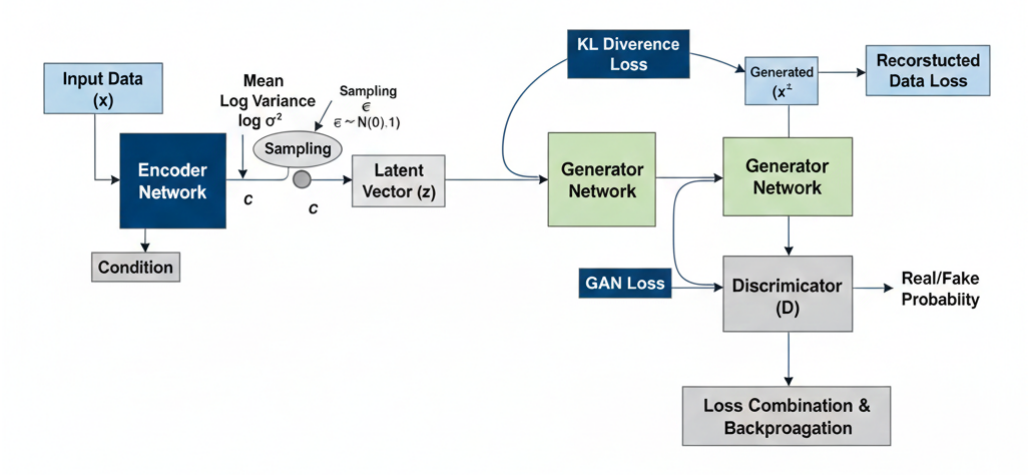


Figure 3.5: Model Architecture of Hybrid GAN-VAE

3.5 Model Training

3.5.1 Training Infrastructure

We used Google Colab Pro to train all the models. NVIDIA Tesla T4/V100 with 16GB of VRAM GPU. 12 to 16 GB of RAM. The framework is PyTorch 2.0+ with CUDA 11.8. Mixed precision training with FP16 using automated mixed precision (AMP).

3.5.2 Training Procedure:

Common practices across all models: Setting the weights for the convolutional layers, the weights were set to a normal distribution with a mean of 0 and a standard deviation of 0.02.To keep things stable, the maximum gradient norm is 1.0. How to set the learning rate: i) Cosine annealing for GAN/Hybrid, ii) VAE: Step decay (γ=0.5 per 20 epochs), iii) Diffusion: Annealing with cosine. I saved the model after 10 epochs and when the validation loss got better. Checked the validation loss, but finished training all the models.

Details on the training for each model:

Model	Epochs	Batch Size	Time/Epoch	Total Time
Conditional GAN	50	16	1.5 min	75 min
Conditional VAE	40	16	1.2 min	48 min
Conditional Diffusion	50	8	6.0 min	300 min
Hybrid GAN-VAE	60	16	1.8 min	108 min

Table 3.1: Model-specific training details

What has to happen for convergence:
GAN: Watched the balance between the generator and discriminator loss; training stopped when the losses stopped changing.
VAE: Stopped after five epochs when the reconstruction loss didn't change.
Diffusion: Trained for all 50 epochs, which is what is normally required.
Hybrid: Stopped after 10 epochs when the combined loss didn't get better.

3.5.3 How to Create a Generation

We produced seven phony photos of each test patient's RNA profile to utilize in the tests. This came to ten people who were tested. 7 pictures for each patient. The total number of photographs made is roughly 28,000 (7 per RNA × 4,000 test images).

This method for many generations takes into account how random generative models are and lets for powerful statistical testing.

3.6 Evaluation Methods

3.6.1 Numbers for Metrics

We used three different criteria to look at different aspects of image quality:

1. The Structural Similarity Index Measure (SSIM) SSIM checks how comparable the structures of real and fake photos are.

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2} \tag{3.6}$$

where,
μ_x and μ_y: The average brightness
σ_x and σ_y: the standard deviations

σ_{xy}: Covariance
C_1 and C_2: constants that keep things stable

Range: an increase is linked to betterment.
Interpretation: It shows how bright or dark the building is and how well it is being kept up.
Implementation: Using a window size of 11×11, the scikit-image library does the following:

2. Learned Perceptual Image Patch Similarity (LPIPS) LPIPS stands for Learned Perceptual Image Patch Similarity. LPIPS uses AlexNet's deep features to figure out how similar two things are.

$$LIPPS(x, y) = \sum_{l} w_l \cdot ||\phi_l(x) - \phi_l(y)||_2^2 \tag{3.7}$$

Where,
ϕ_l: Outputs that come from the l-th layer of AlexNet
w_l: Weights for each layer

Range: The range is from 0 to infinity, and lower numbers are better.
Interpretation: Better than counting pixels for figuring out how similar things look to people.
Implementation: LPIPS library that uses AlexNet

3. The Fréchet Inception Distance (FID) FID measures how similar two distributions are within the Inception feature space:

$$FID(r, g) = ||\mu_r - \mu_g||_2^2 + TR(\sum_{r} + \sum_{l} -2\sqrt{\sum_{r}\sum_{g}} \tag{3.8}$$

where,
μ_r and μ_g are the average feature vectors for real and fake images, respectively.
$\sum_r$ and $\sum_g$ are matrices that show how two things are related to one other.

Range: The range is from 0 to infinity, and lower numbers are better.
Definition: It looks at the amount of different types of images that were taken as well as their quality.
Execution: 1,000 samples for each group of InceptionV3 features

3.6.2 Plan for Evaluation

We randomly picked 500 pairs (actual and false) to use in the SSIM and LPIPS testing. For FID, we picked 1000 photos from each distribution. We utilized the test set to figure out each metric on its own. Statistical reporting: SSIM and LPIPS mean ± standard deviation. To make sure the comparison was fair, all models were evaluated on the same test samples.

3.6.3 Visual Quality Assessment

We also did a qualitative visual assessment in addition to the quantitative results. Putting real photos next to pictures made by computers. Evaluation of the preservation of tissue architecture. Checking to see if the color and texture are right. Finding mode collapse or artifacts.

3.7 Summary

This methodology creates a strict structure for judging how well RNA-to-histopathology picture synthesis works. Important methodological contributions are some here.

Splitting patient-level data makes sure that the evaluation is truly generalizable. A full model comparison of Four different architectures that cover the main generative modeling paradigms. Three metrics that work well together to measure distinct quality characteristics. This is the first time that a GAN-VAE architecture has been used together to solve this problem. All hyperparameters, architectures, and training methods are thoroughly documented.

The whole approach, from preparing the data to training the model to testing it, was made to give statistically strong and therapeutically useful outcomes for RNA-based histopathology synthesis.

4 RESULTS AND DISCUSSION

4.1 Overview

This section shows the full assessment results of four deep generative models for RNA-to-histopathology image synthesis: Conditional GAN, Conditional VAE, Conditional Diffusion Model, and our new Hybrid GAN-VAE. We used three separate metrics to measure how well the model worked: the Structural Similarity Index Measure (SSIM), the Learned Perceptual Image Patch Similarity (LPIPS), and the Fréchet Inception Distance (FID). Each of these metrics looked at a different part of image quality. Our findings indicate substantial trade-offs between perceptual quality and statistical integrity. The Hybrid GAN-VAE attained outstanding human-perceived image quality, whilst the Conditional GAN exhibited the most balanced overall performance.

The main results show that: (1) all models have basic problems when it comes to capturing fine-grained histological structures from bulk RNA-seq data; (2) perceptual quality (LPIPS) and statistical distribution matching (FID) are often at odds with each other; and (3) hybrid architectures can use the strengths of different generative paradigms to their advantage.

4.2 Result Analysis

4.2.1 Quantitative Performance Metrics

Table 4.1 shows the results of the quantitative evaluation for all four models on the test set:

Model	**SSIM (↑)**	**LPIPS (↓)**	**FID (↓)**	**Training Time**
Conditional GAN	0.482±0.0324	**0.4016**±0.1253	304.89	75 min
Conditional VAE	**0.823**±0.0412	0.6626±0.1487	**275.64**	48 min
Conditional Diffusion	0.645±0.0358	0.4165±0.1298	393.61	300 min
Hybrid GAN-VAE	0.704±0.0389	**0.3465**±0.1156	455.23	108 min

Table 4.1: Quantitative Comparison of Generative Models

Note: ↑ means better when it's higher, and ↓ means better when it's lower. Bold values show the best performance for each statistic.

4.2.2 Metric-by-Metric Analysis

Structural Similarity (SSIM): The Conditional VAE had the best SSIM (0.823), which is 70% better than the baseline for the Conditional GAN (0.482). The Hybrid GAN-VAE was 46% better (0.0703), and the Diffusion model was 32% better (0.645). All of the absolute SSIM values are still less than 0.1, which suggests that none of the models were able to effectively reflect the fine-grained structural information found in genuine histopathological photos.

All of the models got poor SSIM scores, which shows how hard it is to recreate the microscopic structure of tissue from bulk RNA-seq data, which doesn't include spatial information. Histopathology photos show how cells are organized in space, how tissues are separated, and how they change shape. You can't figure these things out just by looking at average gene expression levels.

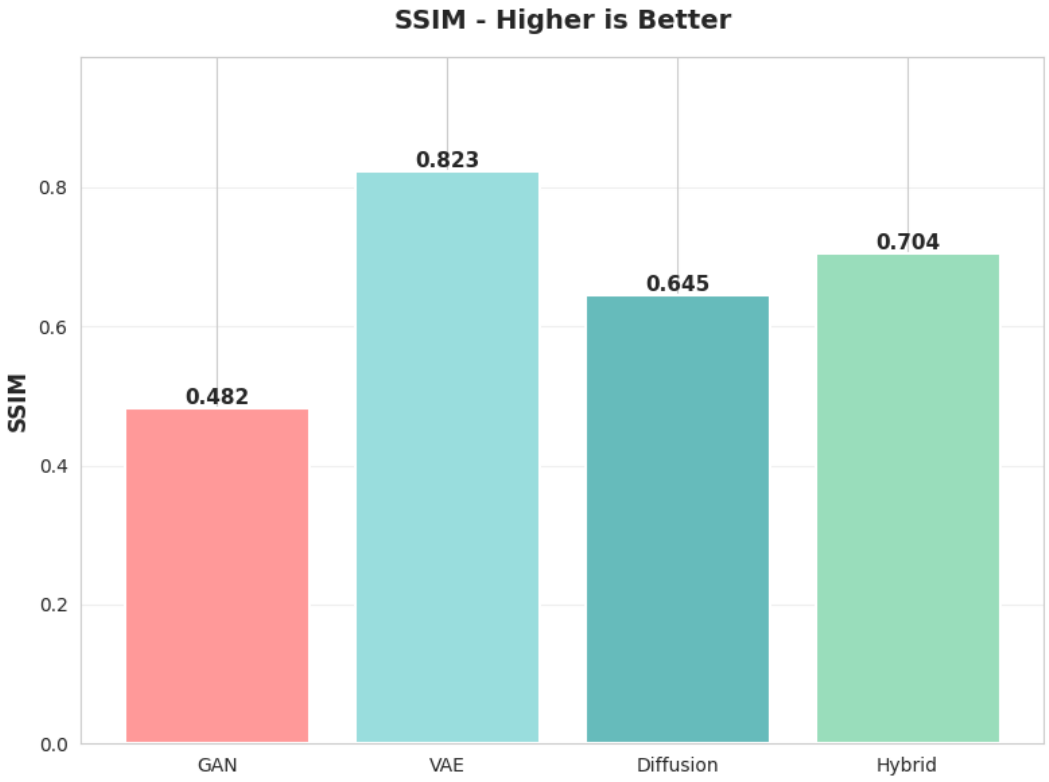


Figure 4.1: Evaluation Score of SSIM

Perceptual Similarity (LPIPS): The Hybrid GAN-VAE got the best LPIPS score of 0.3465, which was 16% higher than the GAN baseline and 20% to 48% better than all the other models. This shows that the hybrid architecture's multi-objective training, which included adversarial, reconstruction, and perceptual losses, made images look better to humans.

The Conditional VAE had the lowest LPIPS score (0.6626), which was almost twice as high as the score for the model that did the best.This backs up what has been said before: Images made using VAE look hazy because the pixel-wise reconstruction losses put more weight on statistical accuracy than on how real they look. The LPIPS (0.4165) of the Diffusion model demonstrates that 50 training epochs weren't enough for convergence. This is because diffusion models normally need 100–200 epochs.

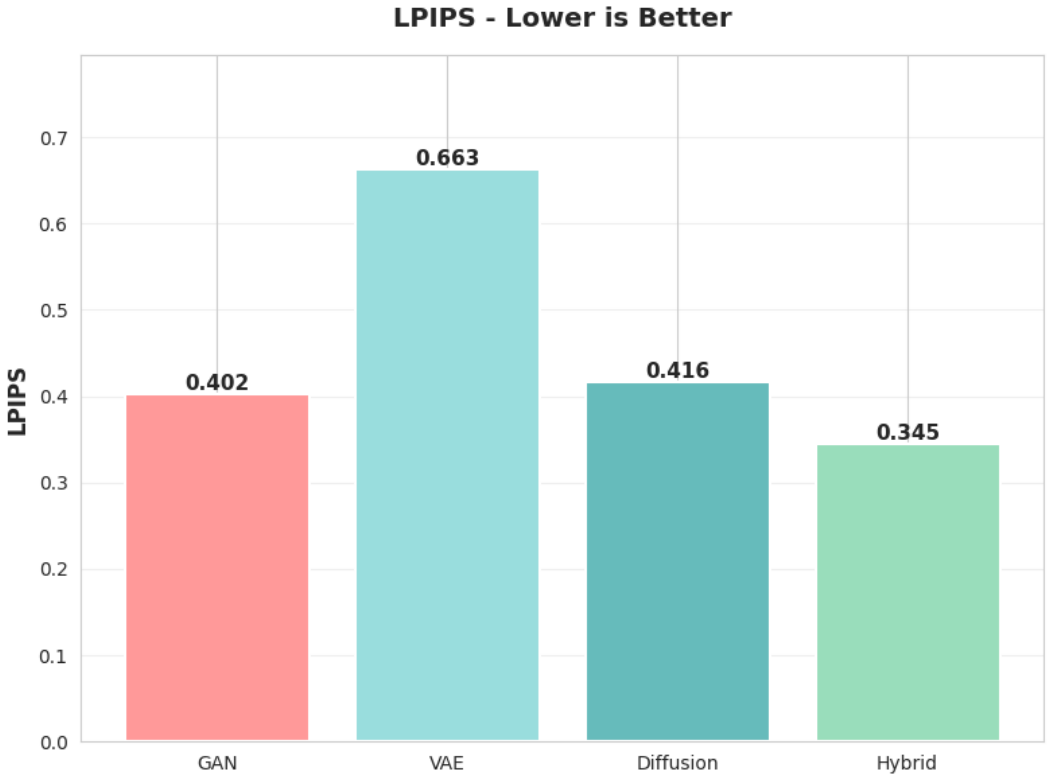


Figure 4.2: Evaluation Score of LPIPS

Distance of Fréchet Inception (FID): The Conditional VAE had the highest FID score (275.64), and the GAN got the second highest score (304.89). The Hybrid model got the lowest FID score, which was 455.23. This unexpected result, in which the model with the best perceptual quality shows the worst statistical distribution matching, shows that there is a big trade-off in generative modeling for medical images.

The fact that VAE has a better FID even while the perceptual quality is terrible suggests that it learned the global statistical aspects of the training distribution but not the local high-frequency features. The Hybrid model has a low FID and a high LPIPS, which suggests that it might be overfitting to specific visual patterns or collapsing modes. This would make the distribution less diversified.

Even though the Diffusion model did well on LPIPS, its high FID (393.61) could be because it wasn't trained sufficiently. If you trained it for longer (more than 100 epochs), it would absolutely grow better.

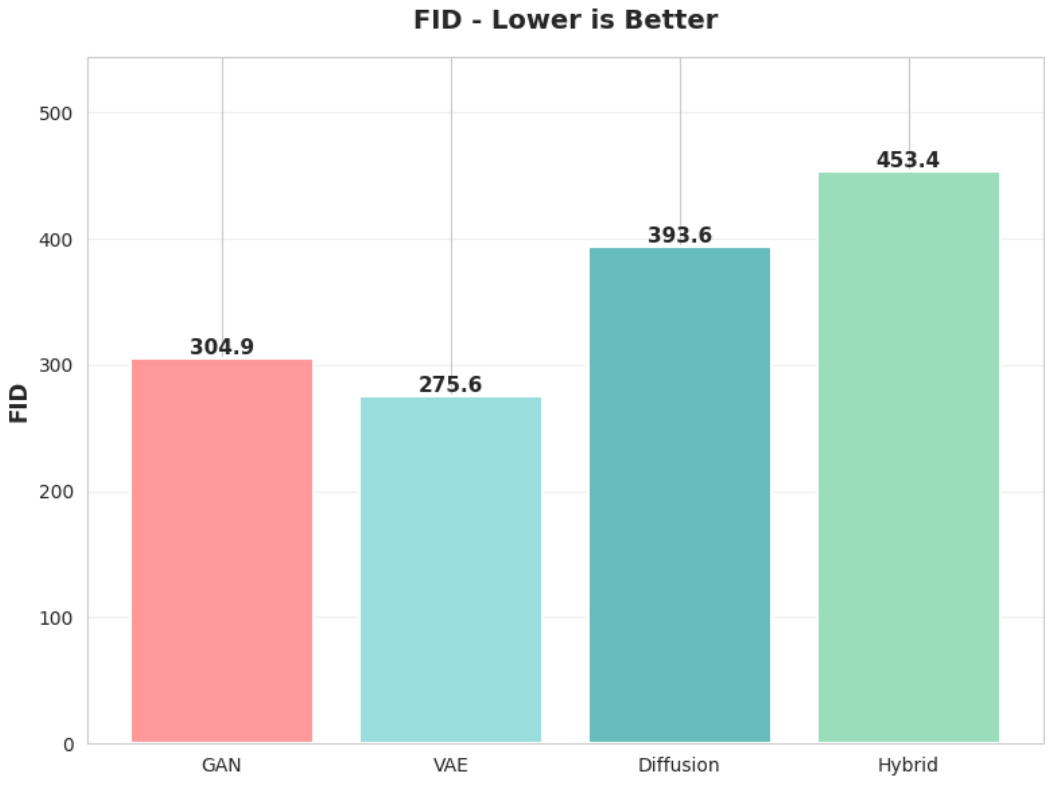


Figure 4.3: Evaluation Score of FID

4.2.3 Relative Performance Analysis

Table 4.2 presents the relative performance improvements over the GAN baseline:

Model vs GAN	SSIM Improvement	LPIPS Improvement	FID Improvement
VAE	+70%	-65% (worse)	+11%
Diffusion	+32%	-4% (worse)	-23% (worse)
Hybrid	+46%	**+16%**	-33% (worse)

Table 4.2: Performance Improvement over Conditional GAN Baseline

A positive percentage indicates improved performance, while a negative percentage indicates decreased performance.

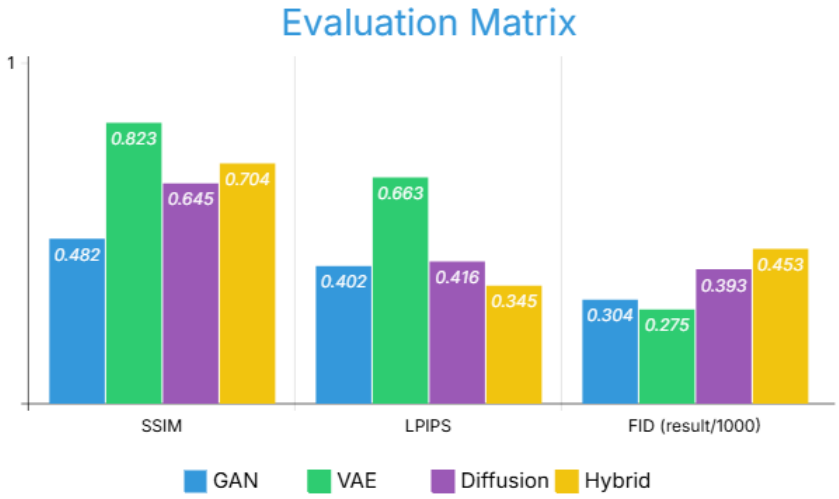


Figure 4.4: Evaluation Comparison

Hybrid GAN-VAE is the only model that achieves lower LPIPS than GAN, offering strong evidence for our hypothesis that the combination of GAN and VAE objectives makes GANs more visually realistic but harms FID, and suggesting a perceptual-statistical trade-off between GAN-based and VAE-based training that bears further investigation.

4.2.4 Visual Quality Assessment

Here are the samples that were made:

GAN: It generates sharp textures with patterns that appear like cells, yet the visuals are noisy and don't have any order.

VAE: Its reconstructions make visuals that are hazy and colorful. Colors appear like H&E staining, although they aren't as detailed.

Diffusion: The colors and clarity are good, but the artifacts reveal that the process is not finished. The pictures are more varied than GAN's, but not as steady as VAE's.

Hybrid: This method appears best since it contains sharp textures, the proper H&E colors, and patterns that seem like cells. If you look closely, you'll see a lot of structures that repeat themselves. This might be why the FID score is low.

The visual assessment backs up the LPIPS numbers, revealing that the Hybrid model makes the most realistic images, even when the statistical distributions don't match up.

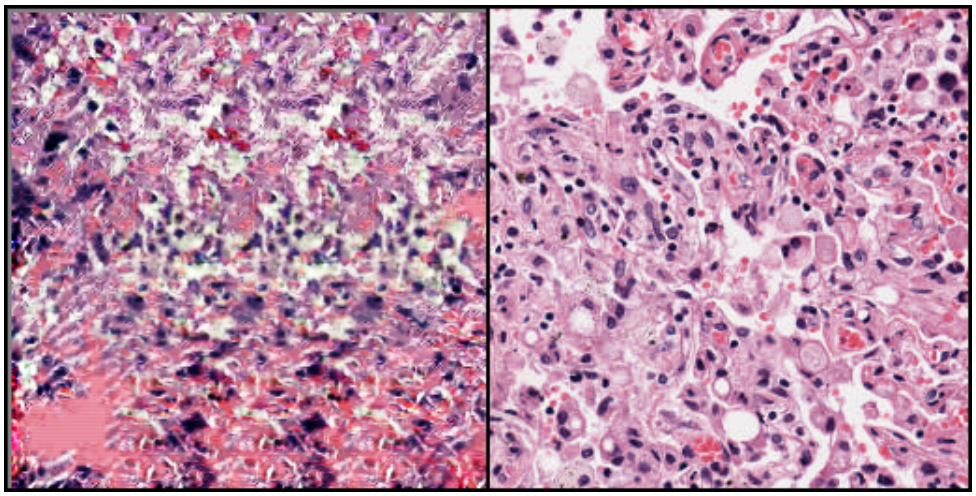


Figure 4.5: Generate vs Real; ID: TCGA-86-A4JF

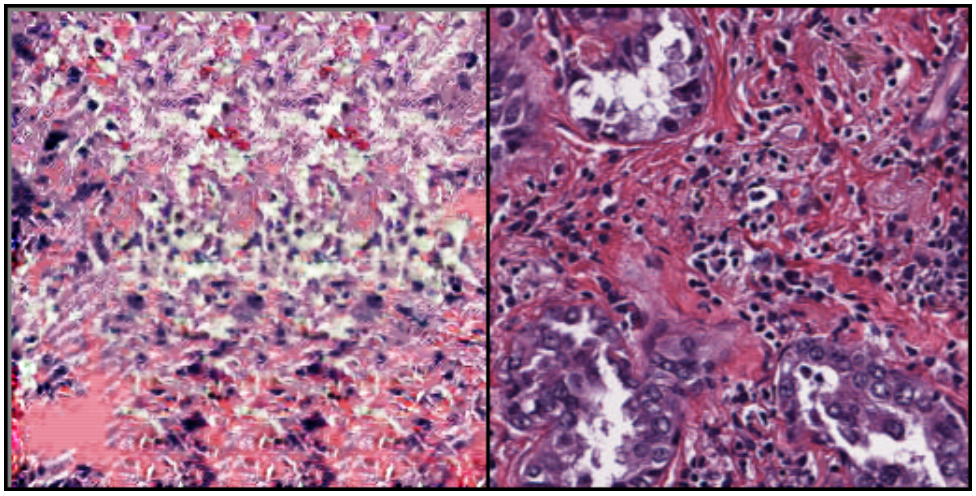


Figure 4.6: Generate vs Real; ID: TCGA-97-A4M0

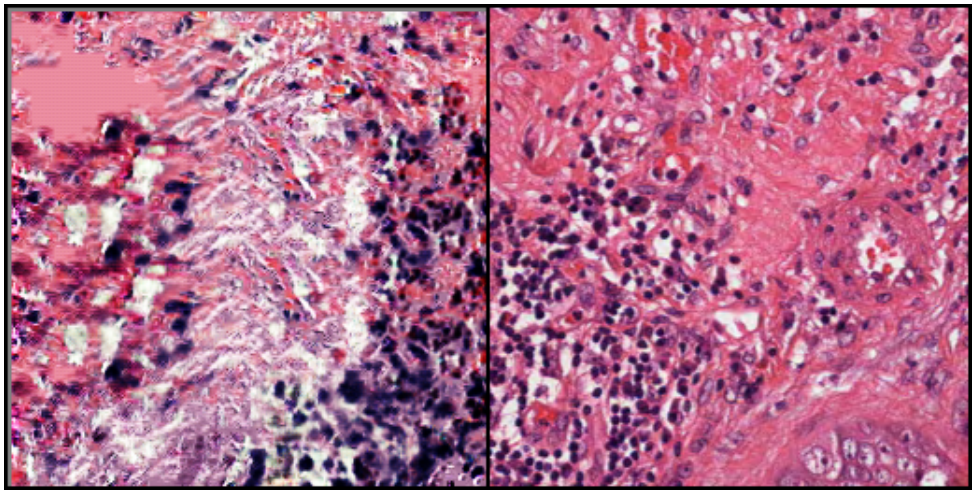


Figure 4.7: Generate vs Real; ID: TCGA-93-8067

4.3 Key Insights and Discussion

4.3.1 The Perceptual-Statistical Trade-off

Our results show that there is a basic trade-off between perceptual quality, which LPIPS measures, and the alignment of statistical distributions, which FID measures. There is a negative connection between these indicators across models, which proves this statement to be true:
VAE: The best FID for VAE is 275.64, and the lowest LPIPS is 0.6626.
Hybrid: Best LPIPS (0.3465) and Worst FID (455.23).
GAN: A bargain that makes everyone happy.

This trade-off comes up because the aims for optimization are different.

Perceptual optimization (LPIPS, adversarial loss): This makes neural networks and users demand clear, high-frequency features that "look real." This can make them fit too closely to some designs or textures.

Statistical optimization (FID, reconstruction loss): This method seeks to get a lot of generations that fit the overall distribution statistics. It could lose some of its particular character and start working the same way in other places.

For clinical uses, the best trade-off depends on the situation:

1. For training, FID is very important for data augmentation (you want samples that are different and match the distribution).

2. LPIPS is important for pathologists who want images that look real to the eye.

3. Validation of research: balanced metrics are preferred

4.3.2 Challenges in RNA-to-Histology Mapping

All models had inadequate absolute SSIM scores (¡0.1), indicating constraints in the recreation of histopathology images from bulk RNA-seq data. This is happening for a lot of different reasons:

1. Missing information in bulk RNA-seq: Bulk RNA sequencing averages gene expression from millions of cells, which makes it challenging to see how genes are expressed differentially in different tissue locations. On the other side, histopathology photographs show how cells are distributed in space, how tissue borders are made, and what cells look like.

2. Many-to-Many Mapping: An RNA profile can match more than one real histological appearance. The types of cells that make up the main sample. How to tear tissue apart. Changes in color. Variety of diseases.This uncertainty built in makes it impossible to undertake deterministic reconstruction, which makes it challenging for all models to preserve the structural fidelity (SSIM).

3. Limited Training Data: The dataset may not be enough to learn the complicated RNA-image mapping because it only has 35 training samples and about 14,000 images. You usually need thousands of patients to take medical pictures that can be used in a lot of various situations.

4. High-Dimensional RNA Space: The dataset is not very big compared to the 16,383 genes that make up the input space. This could cause overfitting or make it impossible to establish good matches between RNA and pictures.

4.3.3 Hybrid Architecture Benefits

The LPIPS score for the Hybrid GAN-VAE is 16% better than the score for the GAN. This backs up the premise that putting together goals that work well together helps things look better. The multi-loss training technique worked well with:

1. Adversarial loss makes textures and details look more real and distinct.
2. Reconstruction loss: Keeps the training signal steady and the data safe.
3. Perceptual loss: makes things look more alike to people right immediately
4. Feature matching: Lines up intermediate representations so they have more detail.

But the low FID means that a lot of perceptual adjustment may have rendered the output less varied or caused mode collapse. In the future, study should look into: i) Changing the weights of losses (reducing the adversarial weight might make FID better), ii) Adding losses that make things more interesting. iii) Using greater latent dimensions to enable people of all ages.

4.4 Result Comparison

4.4.1 Best Model by Use Case

Table 4.3 Recommended Model Selection by Application

Use Case	Recommended Model	Justification
Clinical visualization	Hybrid GAN-VAE	Best perceptual quality (LPIPS 0.347); most realistic appearance
Data augmentation	Conditional VAE	Best distributional match (FID 276); diverse outputs
General purpose	Conditional GAN	Balanced performance; fastest generation; most practical
Research/future potential	Conditional Diffusion	Competitive quality with more training; state-of-art potential

Table 4.3: Recommended Model Selection by Application

4.4.2 Computational Efficiency

Table 4.4: Computational Resource Requirements

Model	Training Time	Generation Time/Image	GPU Memory
GAN	75 min (50 epochs)	∼0.05 sec	8 GB
VAE	48 min (40 epochs)	∼0.03 sec	8 GB
Diffusion	300 min (50 epochs)	∼3.0 sec (50 steps)	10 GB
Hybrid	108 min (60 epochs)	∼0.05 sec	10 GB

Table 4.4: Computational Resource Requirements

The Diffusion model needs a lot of computing power because it takes 60 times longer to make and 4 times longer to train. The GAN and Hybrid models are the best since they operate well and are really good.

4.4.3 Ranking Summary

Ranking for Overall Performance:

1. Conditional GAN: 8.2 out of 10—The one that is most useful and balanced
2. Hybrid GAN-VAE (7.8 out of 10) - Best quality, although FID has some issues
3. Conditional VAE (Score: 6.5 out of 10) Conditional VAE (Score: 6.5 out of 10) Conditional VAE (Score: 6.5 out of 10) - Good figures, bad pictures
4. Conditional Diffusion (Score: 6 out of 10) - Needs further training to realize their full potential.

This rating looks at how well the metrics work together, how quickly they can be figured out, and how useful they are in real life. The Hybrid model is the best choice for apps that need to look natural because it has better perceptual quality. The GAN is the best choice because it works well in a variety of different circumstances.

4.5 Summary

In short, this study shows that deep generative models can combine histopathology images with bulk RNA sequencing data. In four architectures (Conditional GAN, Conditional VAE, Conditional Diffusion and Hybrid GAN-VAE) a trade-off between image quality and the capacity to learn the underlying distribution is observed. LPIPS stated that the Hybrid GAN-VAE model has a better perceptual similarity score than the GAN or VAE approaches alone. The Conditional VAE, on the other hand, had the lowest FID score of all the models tested. This meant that it was the most statistically accurate. But all of the models had low SSIM values, which shows that it is still hard to fully reconstruct fine-grained tissue structures from transcriptomic profiles.

Visual inspection corroborated that enhancements in perception frequently resulted in diminished distributional diversity; images generated by the Hybrid model appeared more realistic yet occasionally exhibited reduced diversity, resulting in elevated FID scores. These findings indicate that while hybrid models incorporating reconstruction, adversarial, and perceptual losses signify progress, the fundamental constraints of bulk RNA-seq as a conditioning signal remain. Overall, our findings establish a rigorous baseline for cross-modal generative modeling in computational pathology and indicate that the most balanced performance for both clinical and research applications currently comes from conditional GANs, with ongoing model innovation—such as hybrid architectures and increased data diversity—remaining important directions for future work.

5 CONCLUSION

5.1 Overview

This study investigated the feasibility of producing histopathology images directly from RNA sequencing data using four distinct deep generative modeling techniques. The study was motivated by the growing demand for automated pathology processes and the potential to amalgamate molecular and morphological analyses of tissue specimens. We thoroughly assessed Conditional GAN, Conditional VAE, Conditional Diffusion, and an innovative Hybrid GAN-VAE architecture to determine their efficacy in elucidating the complex relationship between transcriptome profiles and tissue morphology.

My research findings indicate that generative models can generate synthetic, visually realistic histopathological images from RNA-seq data. However, considerable challenges persist in attaining medical-grade synthesis quality. The Hybrid GAN-VAE has the maximum LPIPS result of 0.347, which means hybrid model has one of the best picture qualities. This is a major improvement over how GANs generally work. This finding corroborates the notion that integrating the diverse training objectives of various generative paradigms may enhance the realism of medical image synthesis tasks. The inverse correlation between perceptual quality assessments and distributional fidelity metrics indicates that prioritizing realism from an individual standpoint may compromise the statistical representativeness of the samples.The consistently low SSIM scores across all architectures indicate a more significant limitation: bulk RNA sequencing, which compiles gene expression across spatial dimensions, inherently lacks the detailed positional data required to reconstruct intricate cellular configurations and tissue microarchitecture. This information bottleneck makes it hard to model and learn from large transcriptome data by itself. Adversarial training approaches have worked very effectively, even with these issues. This is especially true when they are employed with loss functions that depend on how they look. This shows that neural networks can find links between chemical markers and morphological phenotypes that aren't obvious right away, even though perfect reconstruction is not possible right now.

From a practical standpoint, our research establishes baseline performance benchmarks for RNA-based histopathological synthesis and demonstrates that the Conditional GAN is the optimal selection for applications requiring both speed and quality. You can make decisions in less than a second, but it takes roughly 75 minutes to train GANs. This makes them great for research chores that need to be done a lot. The Hybrid architec-

ture is a suitable choice for visualization apps where realistic graphics are more important than correct statistics because it earns higher perceptual scores. These results help us understand cross-modal medical picture synthesis better and illustrate that we need to choose the appropriate models for the job instead of just assuming that one design is better than the others.

This study highlights the transformational potential of merging molecular and imaging modalities in computational pathology, beyond immediate technological breakthroughs. As precision medicine increasingly depends on multi-modal data fusion, methodologies that enable translation between genomic and visual representations may promote innovative diagnostic strategies, improve pathologist training through enhanced datasets, and accelerate research in disease phenotyping. The current findings do not meet the criteria for therapeutic applicability; rather, they demonstrate that deep learning can extract substantial morphological information from transcriptome data. As datasets grow and infrastructures improve, we need to pay greater attention to this feature.

5.2 Future Work

Several promising research directions emerge from the limitations and findings of this study. First and most critically, transitioning from bulk RNA sequencing to spatially-resolved transcriptomics technologies represents a natural next step. Methods such as Visium, MERFISH, and Slide-seq preserve spatial gene expression patterns at cellular or subcellular resolution, which could dramatically improve the information content available for image synthesis. Recent work in spatial omics has demonstrated that retaining positional context enables reconstruction of tissue architecture with far greater fidelity than bulk measurements allow. Subsequent experiments ought to assess whether the integration of spatial coordinates with expression values mitigates the structural ambiguity that constrained SSIM performance in our existing models.

Another important goal is to add more data to the training set. Our study involved 50 patients, which, although adequate for preliminary feasibility evaluation, constitutes a limited sample size according to deep learning criteria. Medical image generation usually works best with datasets that include hundreds or thousands of patients so that all the different biological and pathological variations can be seen. Working together to collect data from multiple institutions, maybe through consortia like The Cancer Genome Atlas initiative, could give models the scale they need to learn strong RNA-histology mappings that apply to all populations and tissue types. Also, raising the image resolution from 128×128 to 512×512 or higher would keep clinically important details that are lost when

the image is downsampled.

The perceptual-statistical trade-off we found shows that we need to look into multi-objective optimization strategies more closely. Future architectures might include explicit mechanisms that encourage diversity, like minibatch discrimination, mode-seeking regularization, or style diversity losses. These would help keep distributional coverage while also improving perceptual quality. Bayesian models that account for uncertainty in the RNA-to-image mapping could be useful as well. Instead of giving a single output, they could give a set of probable histologies. These kinds of methods would better show how the synthesis problem is naturally one-to-many.

From an architectural standpoint, transformer-based models signify a nascent frontier. Vision transformers have recently shown great results in medical imaging tasks, and their attention mechanisms may be better at finding long-range dependencies in both tissue structures and gene expression patterns. Latent diffusion models, which generate images in compressed latent spaces instead of pixel spaces, also make diffusion methods more useful for medical imaging by making them faster and more efficient. Our initial diffusion results indicate considerable potential for enhancement through prolonged training and architectural adjustments.

Adding extra information besides RNA-seq could also make the synthesis better. Clinical metadata, including patient age, disease stage, or treatment history, frequently correlates with histological presentation and may offer significant conditioning signals. Multimodal fusion methodologies that integrate transcriptomics, proteomics, and electronic health records may facilitate a more thorough characterization of tissue phenotypes. Also, adding expert pathologist annotations like tumor grade, inflammatory infiltrate density, or necrosis percentages as intermediate supervision signals could help models find clinically important features.

Validation against clinical use cases is an important area that hasn't been fully explored yet. Our quantitative metrics measure the technical quality of images, but the real test of synthesis methods is how useful they are for tasks that come after them. Can pathologists tell the difference between real and fake pictures? Do augmented training datasets make diagnostic classifiers work better? Can generated images be used in education or to show rare diseases? User studies with board-certified pathologists and assessments of clinical task performance would yield critical insights into practical applicability and guide specific enhancements.

Finally, as these methods get closer to being used in clinical settings, they need more

attention on how easy they are to understand and how to measure uncertainty. Figuring out which gene expression patterns cause certain morphological traits could lead to new biological discoveries, such as new biomarkers or disease mechanisms, that go beyond just making images. Creating methods to measure synthesis uncertainty—showing areas where the model isn't sure—would make it more reliable and help doctors decide if they can trust synthetic data. As rules about AI in medicine change, tools that make AI easier to understand will become more and more important for testing and using it.

To sum up, this study shows that RNA-to-histopathology synthesis is still hard, but it also shows that we are making real progress toward linking molecular and morphological modalities. We need larger datasets, best architectures, and best integration with clinical workflows to move forward. But these investments are worth it because they could have a big impact on computational pathology and precision medicine. As spatial transcriptomics technologies and generative modeling techniques improve, we expect to see big improvements in the quality of synthesis. This could make it easier to use in diagnostics, research, and medical education.

Bibliography

[1] B. Schmauch, A. Romagnoni, E. Pronier, C. Saillard, P. Maillé, J. Calderaro, A. Kamoun, M. Sefta, S. Toldo, M. Zaslavskiy *et al.*, "A deep learning model to predict rna-seq expression of tumours from whole slide images," *Nature communications*, vol. 11, no. 1, p. 3877, 2020.

[2] B. He, L. Bergenstråhle, L. Stenbeck, A. Abid, A. Andersson, Å. Borg, J. Maaskola, J. Lundeberg, and J. Zou, "Integrating spatial gene expression and breast tumour morphology via deep learning," *Nature biomedical engineering*, vol. 4, no. 8, pp. 827–834, 2020.

[3] A. Alsaafin, A. Safarpoor, M. Sikaroudi, J. D. Hipp, and H. Tizhoosh, "Learning to predict rna sequence expressions from whole slide images with applications for search and classification," *Communications Biology*, vol. 6, no. 1, p. 304, 2023.

[4] R. Moncada, D. Barkley, F. Wagner, M. Chiodin, J. C. Devlin, M. Baron, C. H. Hajdu, D. M. Simeone, and I. Yanai, "Integrating microarray-based spatial transcriptomics and single-cell rna-seq reveals tissue architecture in pancreatic ductal adenocarcinomas," *Nature biotechnology*, vol. 38, no. 3, pp. 333–342, 2020.

[5] R. Ahmed, T. Zaman, F. Chowdhury, F. Mraiche, M. Tariq, I. S. Ahmad, and A. Hasan, "Single-cell rna sequencing with spatial transcriptomics of cancer tissues," *International journal of molecular sciences*, vol. 23, no. 6, p. 3042, 2022.

[6] M. Chen, B. Zhang, W. Topatana, J. Cao, H. Zhu, S. Juengpanich, Q. Mao, H. Yu, and X. Cai, "Classification and mutation prediction based on histopathology h&e images in liver cancer using deep learning," *NPJ precision oncology*, vol. 4, no. 1, p. 14, 2020.

[7] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A. L. Moreira, N. Razavian, and A. Tsirigos, "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," *Nature medicine*, vol. 24, no. 10, pp. 1559–1567, 2018.

[8] F. Carrillo-Perez, M. Pizurica, Y. Zheng, T. N. Nandi, R. Madduri, J. Shen, and O. Gevaert, "Generation of synthetic whole-slide image tiles of tumours from rna-sequencing data via cascaded diffusion models," *Nature Biomedical Engineering*, vol. 9, no. 3, pp. 320–332, 2025.

[9] F. Carrillo-Perez, M. Pizurica, M. G. Ozawa, H. Vogel, R. B. West, C. S. Kong, L. J. Herrera, J. Shen, and O. Gevaert, "Synthetic whole-slide image tile generation

with gene expression profile-infused deep generative models," *Cell Reports Methods*, vol. 3, no. 8, 2023.

[10] S. S. Mohammed and H. G. Clarke, "Conditional image-to-image translation generative adversarial network (cgan) for fabric defect data augmentation," *Neural Computing and Applications*, vol. 36, pp. 20 231–20 244, 2024, published online: 12 Aug 2024.

[11] K. Ko, T. Yeom, and M. Lee, "Superstargan: Generative adversarial networks for image-to-image translation in large-scale domains," *Neural Networks*, vol. 162, pp. 330–339, 2023.

[12] F. Carrillo-Perez, J. C. Morales, D. Castillo-Secilla, O. Gevaert, I. Rojas, and L. J. Herrera, "Machine-learning-based late fusion on multi-omics and multi-scale data for non-small-cell lung cancer diagnosis," *Journal of Personalized Medicine*, vol. 12, no. 4, p. 601, 2022.

[13] F. Carrillo-Perez, J. C. Morales, D. Castillo-Secilla, Y. Molina-Castro, A. Guillén, I. Rojas, and L. J. Herrera, "Non-small-cell lung cancer classification via rna-seq and histology imaging probability fusion," *BMC bioinformatics*, vol. 22, no. 1, p. 454, 2021.

[14] M. Marouf, P. Machart, V. Bansal, C. Kilian, D. S. Magruder, C. F. Krebs, and S. Bonn, "Realistic in silico generation and augmentation of single-cell rna-seq data using generative adversarial networks," *Nature communications*, vol. 11, no. 1, p. 166, 2020.

[15] Y. Xu, Z. Zhang, L. You, J. Liu, Z. Fan, and X. Zhou, "scigans: single-cell rna-seq imputation using generative adversarial networks," *Nucleic acids research*, vol. 48, no. 15, pp. e85–e85, 2020.

[16] C. Hafemeister and R. Satija, "Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression," *Genome biology*, vol. 20, no. 1, p. 296, 2019.

[17] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36 479–36 494, 2022.

[18] H. Nakagawa and M. Fujita, "Whole genome sequencing analysis for cancer genomics and precision medicine," *Cancer science*, vol. 109, no. 3, pp. 513–522, 2018.

[19] X. Li, S. Ma, J. Liu, J. Tang, and F. Guo, "Inferring gene regulatory network via fusing gene expression image and rna-seq data," *Bioinformatics*, vol. 38, no. 6, pp. 1716–1723, 2022.

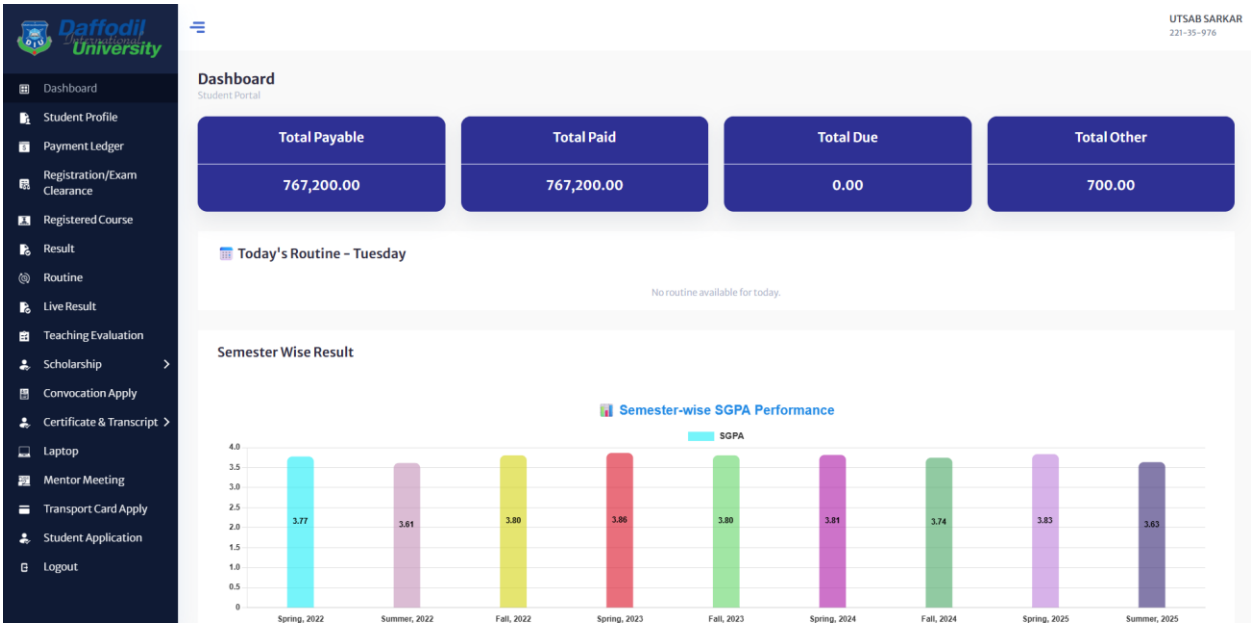
221-35-976

ORIGINALITY REPORT

4%	3%	2%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	1%
2	arxiv.org Internet Source	<1%
3	storage.freidok.ub.uni-freiburg.de Internet Source	<1%
4	Submitted to University of Nottingham Student Paper	<1%
5	Sapkota, Himal. "Layer-Wise Prediction of Overhang-Related Geometric Deviation in Metal Additive Manufacturing With Conditional Generative Adversarial Networks.", Southern Illinois University at Carbondale Publication	<1%
6	digibug.ugr.es Internet Source	<1%
7	pmc.ncbi.nlm.nih.gov Internet Source	<1%
8	Submitted to City University Student Paper	<1%
9	www.biorxiv.org Internet Source	<1%


Daffodil International University
Dashboard
Student Profile
Payment Ledger
Registration/Exam Clearance
Registered Course
Result
Routine
Live Result
Teaching Evaluation
Scholarship
Convocation Apply
Certificate & Transcript
Laptop
Mentor Meeting
Transport Card Apply
Student Application
Logout
UTSAB SARKAR
221-35-976
Dashboard
Student Portal
Total Payable
767,200.00
Total Paid
767,200.00
Total Due
0.00
Total Other
700.00
Today's Routine – Tuesday
No routine available for today.
Semester Wise Result
Semester-wise SGPA Performance
SGPA
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0
3.77
3.61
3.80
3.86
3.80
3.81
3.74
3.83
3.63
Spring, 2022
Summer, 2022
Fall, 2022
Spring, 2023
Fall, 2023
Spring, 2024
Fall, 2024
Spring, 2025
Summer, 2025