UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE FÍSICA


# Synthesizing High B-Value Diffusion-Weighted Imaging of the Prostate using Diffusion Models

Pedro Rodrigues Parracho

**Mestrado em Engenharia Biomédica e Biofísica**

Dissertação orientada por:
Prof. Nuno Matela
Dr. José Almeida

2025

“THE ESSENCE OF INTELLIGENCE IS THE ABILITY
TO LEARN FROM EXPERIENCE”

**Frank Rosenblatt**

# Agradecimentos

Esta dissertação representa o culminar de 5 anos de um percurso académico tão exigente como enriquecedor. Agradeço desde já ao leitor, que corajosamente, se dispõe a ler este trabalho.

Começo por agradecer ao professor Nuno Matela, não só pela orientação contínua neste projeto, mas por todos os ensinamentos transmitidos ao longo desta jornada e pelas aulas lecionadas com entusiasmo e dedicação que me despertaram tanta curiosidade e interesse nesta área.

Agradeço também ao Dr. Nickolas Papanikolaou e a toda a equipa da CCIG da Fundação Champalimaud, que me fizeram sentir bem-vindo num ambiente profissional de investigação. Um agradecimento especial ao Dr. José Almeida que me acompanhou e orientou durante todo este processo com constante disponibilidade e muita paciência.

Aos meus pais pelo amor e apoio incondicional, por serem os mentores que me ajudam a tomar decisões corretas, mas especialmente por me darem espaço para tomar decisões erradas. Um grande obrigado por me concederem um ambiente onde sempre me senti seguro e onde tinha todas condições para concluir este percurso. Aos meus irmãos mais velhos André e Tiago por me tolerarem com paciência quando estou stressado e ouvirem com atenção quando estou entusiasmado. São os meus maiores exemplos. Agradeço também ao meu primo João pela ajuda e mais do que isso pela motivação que me transmite todos os dias ao atravessar o seu doutoramento, que é um percurso académico muito mais comprido, difícil, interminável e comprido do que o meu. A tua persistência inspirou-me.

Deixo expresso a minha enorme gratidão aos meus amigos Tomás, Ricardo, Augusto, Miguel, Rafa, João, Jaime e Filipa, por me fazerem rir quando estou desmotivado, por não me perguntarem "como vai a tese?" e por me mostrarem que existe uma outra vida para além da académica e que estas vidas não são antónimas, mas sim complementares. Deram-me a motivação que precisava para encarar este desafio e sem isso certamente teria desistido. Um obrigado especial ao meu colega Rafael por todas as tardes que passou a trabalhar ao meu lado, por ajudar a tornar um processo entediante e difícil numa jornada entusiasmante e divertida. Não a iria querer fazer doutra forma.

Um grande obrigado aos meus amigos e colegas da FCUL, Afonso, Miguel, Hugo, Carol, Sofia, Catarina, Sara e Joana, por me terem acompanhado durante estes 5 anos de estudo árduo. Sem vocês este curso teria sido muito aborrecido.

E por fim aos meus padrinhos Gui e Rafa, por me terem acolhido e feito sentir menos sozinho numa cidade tão diferente da minha.

# Abstract

Prostate cancer (PCa) remains among the most prevalent and deadly cancers among men worldwide, with rising incidence rates placing increasing pressure on healthcare systems. Magnetic Resonance Imaging (MRI), and more specifically, Diffusion-Weighted Imaging (DWI), has proven to be an ideal diagnostic tool in the early detection of PCa. However, acquisition of high b-value DWI, which is critical in detecting diffusion-restricted lesions, is challenging due to the prolonged scan times and susceptibility to noise and motion artifacts.

This dissertation explores the clinical practicability of generating high b-value prostate DWI from low b-value prostate DWI inputs using generative artificial intelligence (AI) models with the aim of decreasing acquisition times and artifact susceptibility.

Three generative models were tested: Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and Denoising Diffusion Probabilistic Models (DDPMs). Each model was conditioned to take any low b-value DWI and generate a resulting DWI at any requested high b-value.

Quantitative evaluation was conducted with image similarity metrics, comparing the generated DWI with the real acquired DWI. Diagnostic performance was also quantified with a deep learning lesion segmentation model in an attempt to determine diagnostic preservation in the generated DWI.

The DDPM architecture was the most effective generative architecture, with the best performance in quantitative scores, in addition to high visual realism in maintaining diffusion effects. The generated images also preserved the majority of the diagnostic accuracy of the original high b-value scans, with preservation of lesion detection rates and localisation of lesions.

The results verify that diffusion models are the most adequate architecture for low-to-high b-value DWI generation and can retain diagnostic performance.

This study proved that generative AI can significantly enhance MRI workflows, reducing scan times, and facilitating scalable PCa screening programs, especially in resource-constrained environments, paving the way for more accessible diagnostic pathways.

**Keywords:** DDPM; GAN; VAE; generative AI; Low-to-high b-value DWI generation

# Resumo

Atualmente, o cancro da próstata (PCa) representa uma das neoplasias mais prevalentes e mortais entre homens, com uma incidência que tem vindo a aumentar e que coloca uma pressão crescente nos sistemas de saúde. As ressonâncias magnéticas (MRI), em particular imagem ponderada por difusão (DWI), tem-se revelado essencial na deteção precoce de neoplasias clinicamente significativas.

Esta técnica baseia-se na medição dos movimentos microscópicos das moléculas de água nos tecidos, permitindo identificar alterações nas estruturas celulares. As DWI com b-values elevados são especialmente importantes porque amplificam os efeitos de difusão, tornando mais evidente a restrição do movimento de moléculas de água em regiões onde há elevada densidade celular, como em lesões tumorais. Isto porque as neoplasias malignas tendem a apresentar uma estrutura celular mais compacta e desorganizada, o que dificulta a difusão de água e se traduz num sinal mais intenso nas DWI de alto b-value.

Contudo, a aquisição de DWI com b-values elevados envolve tempos de aquisição prolongados, maior suscetibilidade a ruído, artefactos de movimento e limitações logísticas, incluindo custos elevados de equipamento, manutenção e profissionais especializados, o que dificulta significativamente a sua aplicação em larga escala.

Esta dissertação propõe uma abordagem baseada em modelos de inteligência artificial generativos para gerar DWI de alto b-value a partir de DWI de baixo b-value, que são adquiridos significativamente mais rápido, porem não obtêm o mesmo nível de sensibilidade de efeitos de difusão. Permitindo assim, reduzir o tempo de aquisição de DWI de b-value alto e aumentar a acessibilidade ao diagnóstico com MRI. Para alcançar este objetivo foram exploradas três arquiteturas generativas distintas: Autoencoders Variacionais (VAE), Redes Adversárias Generativas (GAN) e Modelos de Difusão Probabilísticos de Denoising (DDPM). Os modelos generativos foram concebidos para produzir imagens de alto b-value condicionadas no caso clínico específico, recebendo apenas um DWI de baixo b-value e informação sobre o b-value alvo. Durante o treino, cada par (DWI b-baixo DWI b-alto) permite ao modelo aprender a relação entre a anatomia observável no b-baixo e os padrões de difusão que se manifestam quando o contraste de difusão é aumentado. Os modelos não se limitam a copiar texturas gerais, mas aprendem a extrair características como contornos anatómicos, artefactos de movimento e até padrões que não são visíveis ao olho humano, utilizando estas representações como guia para gerar DWI de alto b-value. O condicionamento é implementado através de uma imagem de baixo b-value que é fornecida ao gerador em cada etapa da geração da imagem sintética por concatenação de canais e o b-value alvo é incorporado como vetor de contexto que guia o processo de geração. Esta arquitetura garante que a síntese reflete a anatomia e características específicas ao caso clínico, enquanto ajusta o contraste de difusão para reproduzir a supressão de sinal em tecido saudável e o destaque em áreas com difusão restrita, conforme o b-value alvo.

Numa primeira fase, foi realizada uma análise comparativa entre os três modelos, em que se utilizou a base de dados PROSTATEx. Esta análise permitiu uma avaliação direta da qualidade das imagens geradas por cada arquitetura.

Foi realizada uma avaliação quantitativa através de métricas de semelhança de imagem que foi complementada com uma análise visual. Os resultados demonstraram que o DDPM superou significativamente os outros modelos, com valores superiores em todas as métricas quantitativas e com imagens geradas com maior fidelidade, quando comparadas com as imagens geradas pelos outros modelos. As imagens produzidas pelo DDPM apresentam elevada semelhança com as imagens reais, sem distorções visuais, boa definição de estruturas anatómicas da próstata e preservação eficaz dos efeitos de difusão característicos das DWI de alto b-value. Em contraste, as imagens geradas pelas

arquiteturas VAE e GAN revelaram métricas inferiores, distorções visuais, estruturas anatómicas mal delineadas, ausência de efeitos de difusão e presença de artefactos como o padrão de "checkboard".

Com base nestes resultados, o DDPM foi selecionado para uma segunda fase de treino, otimização e validação, desta vez com a base de dados ProCAncer-I, que apresenta uma maior complexidade devido à sua maior diversidade de b-values (incluindo valores mais elevados) e à heterogeneidade das imagens DWI provenientes de diferentes centros. A otimização do DDPM foi realizada de forma iterativa e sequencial, ajustando hiperparâmetros para maximizar o desempenho do modelo. Na fase de validação, o DDPM manteve bons resultados em todas as métricas quantitativas e na avaliação visual, mesmo em transformações complexas entre b-values muito baixos para muito elevados, sem comprometer a definição anatómica ou os efeitos de difusão.

Adicionalmente, foram gerados mapas de calor que analisaram a relação entre a distância de transformação dos b-values (distância entre o b-value da DWI de input e o b-value da DWI gerada) e a qualidade das imagens geradas. Observou-se que transformações com intervalos maiores de b-value resultavam em menor semelhança com as imagens reais, enquanto transformações entre b-values mais próximos mantinham maior fidelidade. Este fenómeno pode ser explicado pelo facto de as DWI de b-value baixo apresentarem menor sensibilidade aos efeitos de difusão, captando também efeitos de perfusão e apresentando um menor contraste entre tecidos. Por outro lado, as DWI de b-value elevado são altamente sensíveis à restrição de difusão. Assim, quando o modelo tenta gerar uma imagem de b-value elevado a partir de uma imagem de b-value muito baixo, enfrenta maior dificuldade em inferir os padrões de difusão mais subtis que não estão presentes no input, o que compromete a fidelidade da geração e conduz a resultados menos precisos.

Por fim, procedeu-se a uma análise de desempenho diagnóstico utilizando um modelo de segmentação de lesões a fim de avaliar se as DWIs geradas pelo DDPM preservam a informação patológica presente nas DWIs adquiridas. O ensaio comparou as segmentações obtidas a partir do modelo treinado com imagens adquiridas (ACQ) com as segmentações obtidas a partir do modelo treinado com imagens geradas (SYN) e com as máscaras de referência anotadas por radiologistas, usando como métricas principais o coeficiente de Dice e uma métrica binária de deteção baseada num limiar prático de Dice ($\geq$0,1) para avaliar a localização aproximada da lesão. Os resultados mostraram que, embora haja uma redução média do Dice nas segmentações produzidas sobre imagens sintéticas, a maior parte das lesões presentes nas DWIs adquiridas foi corretamente preservada nos exemplos gerados, com a taxa de deteção global a manter-se em níveis próximos aos do conjunto adquirido (a maioria das lesões detectadas pelo modelo treinado em imagens adquiridas também foram identificadas quando o modelo foi treinado em imagens geradas).

Estes resultados reforçam o potencial clínico dos modelos de difusão. O DDPM mostrou capacidade de capturar e transferir grande parte das dinâmicas de difusão relevantes para a deteção de lesões da próstata a partir de imagens de baixo b-value, preservando a localização e a visibilidade das lesões na maioria dos casos. Na prática, isto sugere que a síntese de DWI de alto b-value pode acelerar protocolos de MRI ao reduzir a necessidade de aquisições longas, diminuir a suscetibilidade a artefactos de movimento e tornar o exame mais tolerável e eficiente, sem sacrificar a capacidade de identificação de lesões. O que demonstra que este estudo se alinha com as iniciativas europeias de integração de IA em imagiologia, estas conclusões apontam para um caminho viável de implementação clínica, desde que acompanhado por validação adicional multicêntrica e uma avaliação por radiologistas.

**Palavras-chave:** DDPM; GAN; VAE; AI generative; geração de DWI de baixo-para-alto b-value.

# Table of Contents

# List of Figures

# List of Tables

# Acronyms

**ADC** Apparent Diffusion Coefficient

**ADCVCGAN** Attention Dual Contrast Vision Transformer CycleGAN

**AE** Autoencoders

**AI** Artificial Intelligence

**ANOVA** Analysis of Variance

**CBAM** Convolutional Block Attention Module

**CNN** Convolutional Neural Network

**CNR** Contrast-to-Noise Ratio

**DDPM** Denoising Diffusion Probabilistic Model

**DWI** Diffusion Weighted Imaging

**FID** Fréchet Inception Distance

**FSIM** Feature Similarity Index Measure

**GAN** Generative Adversarial Network

**HSD** Honest Significant Difference

**ML** Machine Learning

**MRI** Magnetic Resonance Imaging

**MS-SSIM** Multi-Scale Structural Similarity Index Measure

**MSD** Mean Squared Difference

**MSE** Mean Squared Error

**NN** Neural Network

**PCa** Prostate Cancer

**PI-RADS** Prostate Imaging Reporting and Data System

**PSNR** Peak Signal-to-Noise Ratio

**RF** Radiofrequency

**ReLU** Rectified Linear Unit

**SNR** Signal-to-Noise Ratio

**SSIM** Structural Similarity Index Measure

**TE** Echo Time

**TR** Repetition Time

**VAE** Variational Autoencoder

**VAE-GAN** Variational Autoencoder Generative Adversarial Network

**mpMRI** multiparametric MRI

# 1. Introduction

This dissertation describes the development of a well-performing generative model capable of taking low b-value DWI images as inputs and generating higher b-value DWI images, maintaining anatomical details and diagnostic performance of the original high b-value images.

## 1.1. Context and motivation

Today's world is characterized by rapid scientific and technological progress, yet profound disparities in access to high-quality healthcare persist. In 2015 the United Nations adopted the 2030 Agenda for Sustainable Development, a global commitment framed around 17 Sustainable Development Goals (SDGs) (displayed in Figure 1.1) that aim to mitigate these healthcare disparities, eradicate poverty and ensure prosperity and well-being for everyone [1].


Figure 1.1: UN Department of Economic and Social Affairs: sustainable development goals (from [1]).

Prostate cancer (PCa) exemplifies why this commitment matters, remaining the second most commonly diagnosed cancer in men in 2020, with about 1.4 million new cases and 375 000 deaths, and with its incidence expected to double to roughly 2.3 million new cases by 2040 based on an increasing demographic as well as an aging population, placing additional burdens on health care systems as well as increasing barriers and inequalities in the access for advanced screening and treatment technologies [2] [3] [4]. In many European countries, PCa is the most frequently diagnosed cancer in men and a leading cause of cancer death, highlighting the urgent need for practical, scalable early-detection strategies that reduce harm from overdiagnosis while identifying clinically significant disease early enough for effective treatment [5].

Multiparametric MRI (mpMRI), and in particular diffusion-weighted imaging (DWI), has emerged as a promising tool to improve prostate cancer screening by increasing sensitivity for clinically significant lesions and reducing unnecessary biopsies compared with the current screening strategies using prostate-specific antigen testing [6]. Yet the routine deployment of mpMRI at a widespread scale is constrained by long scan times (generally requiring 30 to 40 minutes for each scan), susceptibility to motion and noise, and substantial costs in equipment, infrastructure and specialized personnel, which

together limit equitable access across regions and population groups [7]. Acknowledging these challenges, initiatives like the EU4Health program's call for proposals aim to incorporate artificial intelligence (AI) into MRI workflows to improve efficiency, with a direct focus on piloting AI-driven solutions that allow for a gradual implementation of MRI as a standard protocol for PCa screening [8].

This dissertation presents itself within this context by proposing a generative AI approach in order to reduce acquisition times in DWI, generating diagnostically useful high b-value DWI, which are critical for identifying diffusion-restricted lesions that are indicative of PCa, from short, low b-value acquisitions using a Denoising Diffusion Probabilistic Model (DDPM). Furthermore, the proposed DDPM would allow for the generation of DWIs at any specified high b-value from any singular low b-value, thus providing a versatile tool for clinicians to investigate optimal b-values for lesion detection, a procedure that is intrinsically iterative and patient-specific, enabling real-time adjustments to enhance diagnostic accuracy.

By enabling the production of high-quality, high b-value images without prolonged scanning, this study targets the core aims of SDG 3, “Ensure healthy lives and promote well-being for all at all ages” with earlier detection, timelier referral and more appropriate treatment pathways that improve health outcomes and reduce avoidable morbidity and mortality. At the same time, the research directly engages SDG 9, “Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation,” by developing a model that can be integrated with existing MRI workflows to reduce per-exam time and operational cost, thereby promoting resilient and scalable diagnostic infrastructure. The approach also speaks to SDG 10, “Reduce inequalities,” because reducing acquisition time and resource requirements lowers the economic and logistical barriers that currently restrict access to high-quality prostate imaging, making mpMRI more affordable and faster helps ensure that advanced diagnostics are not limited to wealthier regions or well-resourced institutions but can reach underserved populations as well. There are additional synergies with SDG 12 on responsible consumption and production, insofar as shorter and more efficient imaging protocols reduce energy use, scanner occupation time and redundant repeat imaging without compromising diagnostic quality.

## 1.2. Objectives

The primary objective of this Master's dissertation is to assess the clinical viability of employing generative AI models to reduce prostate MRI acquisition times by generating high b-value diffusion-weighted images (DWIs) from low b-value inputs, investigating architectures such as Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and Denoising Diffusion Probabilistic Models (DDPMs) for this task, and enabling the generation of any high b-value DWI from any low b-value DWI. In addition to this main objective, the following secondary objectives are defined:

1. Research on generative networks, including VAEs, GANs, and DDPMs, for medical image synthesis.
2. Review of current literature to determine the state of the art in low-to-high b-value DWI generation.
3. Establish an adequate training and validation framework for the selected models.
4. Identify the most suitable architecture among the evaluated models for DWI generation.
5. Optimize the selected model to enhance its performance in the generation task.
6. Validate the optimized model on an independent external dataset through diagnostic performance assessment and evaluate its clinical applicability.

## 1.3. Structure

This dissertation is organized into five chapters, each contributing to a comprehensive exploration of the subject matter. The structure of this dissertation is as follows:

1. Introduction: Introduces the motivation and context of the study, research objectives, the challenges addressed, and the proposed solution.
2. Theoretical Background: Presents the foundational concepts necessary for the development and understanding of this study and reviews the current state of the art knowledge on the topic.
3. Methodology: Details the experimental framework, including the datasets used, data preprocessing steps, the architectures explored and validation methods.
4. Results and Discussion: Presents the outcomes of the model training and validation, including image quality metrics, diagnostic performance and discusses the implications of the findings.
5. Conclusions and Future Work: Summarizes the key contributions of the study and outlines potential future research directions.

# 2. Theoretical Background

## 2.1. Magnetic Resonance Imaging (MRI)

Magnetic resonance imaging (MRI) is a medical imaging technique used to generate detailed images of the anatomy and physiological processes inside the body [9]. MRI relies on the principle that atomic nuclei have an inherent magnetic moment and interact with external magnetic fields in a predictable manner [10]. Given the fact that water constitutes a significant portion of the human body, the hydrogen nuclei of water molecules are used as a target nucleus of choice for MRI [11]. When placed in an external magnetic field of strength $B_0$, the hydrogen protons align either with (parallel) or against (antiparallel) this field. This orientation generates a net magnetization parallel to $B_0$, with an excess of protons in the lower-energy parallel alignment (represented in Figure 2.1) [12].

Instead of remaining stationary, these protons experience a rotary motion called precession, around the axis of the external field at a particular frequency. This frequency, known as the Larmor frequency, is directly proportional to the intensity of the external magnetic field and is given by the Larmor equation:

$$\omega_0 = \gamma B_0 \tag{2.1}$$

Where $\omega_0$ is the precession frequency, $B_0$ is the external magnetic field magnitude, and $\gamma$ is the gyromagnetic ratio, a constant that establishes the relation between the magnetic moment and angular momentum of the proton [13].

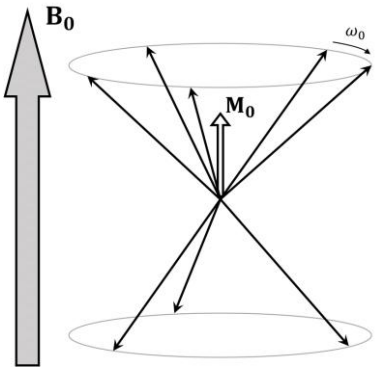


Figure 2.1: Longitudinal magnetization: When an external static magnetic field $B_0$ (oriented along the z-axis, grey vector) is applied, hydrogen protons align either with (parallel) or against (antiparallel) $B_0$. The parallel alignment is the lower energy state, resulting in a slight excess of protons aligned with $B_0$ (black arrows). The magnetic moments of precessing protons cancel out in the transverse plane, but along the longitudinal direction, the excess of parallel protons produces a net magnetization vector $M_0$ (hollow vector) aligned with $B_0$. The protons precess at the Larmor angular frequency $\omega_0$ (curved arrow). (from [116])

To produce an MRI signal, a second radiofrequency oscillating magnetic field, $B_1$, is introduced perpendicularly to $B_0$ at the Larmor frequency. The radiofrequency (RF) pulse perturbs the equilibrium of the system by tilting the net magnetization away from its alignment with $B_0$ and into the transverse plane. This generates a detectable transverse component of magnetization, which causes a current in the receiver coil to oscillate, an effect called free induction decay [14]. After the RF pulse is turned off, the protons slowly return to their equilibrium with the applied field during the process of relaxation (see Figure 2.2) [15].

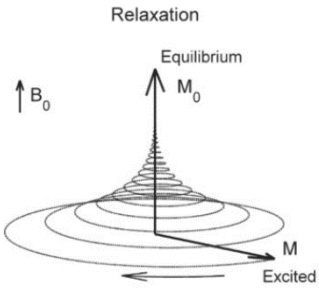


Figure 2.2: Recovery of magnetization following a 90° RF pulse: $B_0$ is the external static magnetic field along the z-axis, aligning hydrogen protons to produce a longitudinal magnetization $M_0$. Initially, protons precess in phase, forming a coherent $M$. Subsequently, dephasing occurs, causing $M$ to decrease as protons spread out, while $M_0$ begins to recover along $B_0$ (from [117]).

There are two ways that relaxation happens. The first, longitudinal relaxation, is responsible for the recovery of net magnetization in the direction of $B_0$. It is affected by interactions of protons with their molecular environment and is quantified by the $T_1$ relaxation time. The second, transverse relaxation, accounts for the loss of coherence between the spinning protons due to their interactions with themselves. The loss of coherence or the dephasing leads to a decay in the transverse magnetization and is accounted for by the $T_2$ relaxation time. $T_1$ and $T_2$ differ among tissue types and thus enable contrast in MRI [16].

Besides quantifying tissue relaxation properties, MRI also has to quantify the spatial origin of signals to generate an image. This is achieved by magnetic field gradients, namely space-varying inhomogeneities of the magnetic field strength. Gradients applied along the x, y, and z directions make the Larmor frequency of the protons depend on position so that position can be encoded accurately [17].

MRI contrast is controlled by the balance of the repetition time (TR), the period between consecutive RF pulses, and the echo time (TE), the period between the RF pulse and the echo signal. Low TR and TE values produce $T_1$-weighted images, which highlight variations in longitudinal relaxation, with water appearing dark and fat tissue appearing bright [18].

Conversely, long TR and TE values produce $T_2$-weighted images, where water-containing structures appear bright because of their long transverse relaxation times. Proton density-weighted images are generated when both TR and TE are long, which primarily represent the density of hydrogen protons within tissues (see Figure 2.3) [18] [19].

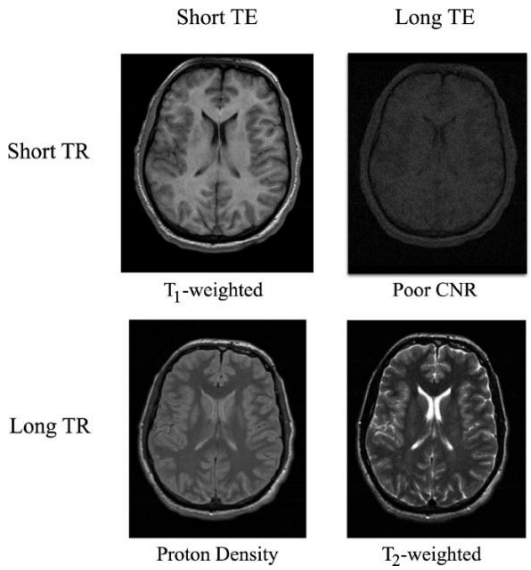


Figure 2.3: MRI contrast influenced by the balance of the TR and TE: For T1-weighted images (top left, short TE and short TR), cerebrospinal fluid appears dark, and white matter is brighter than grey matter. For proton density images (bottom left, short TE and long TR), the contrast highlights proton density with moderate grey and white matter differentiation. For poor contrast-to-noise ratio (CNR) images (top right, long TE and short TR), there is a reduced tissue differentiation. For T2-weighted images (bottom right, long TE and long TR), cerebrospinal fluid appears bright, and white matter is darker than grey matter. (from [118])

## 2.2. Diffusion-Weighted Imaging (DWI)

### 2.2.1 Basic Principles of DWI

Diffusion-Weighted Imaging (DWI) is an MRI modality that captures the microscopic movement of water molecules in biological tissue [20]. This movement, known as Brownian motion, is influenced by tissue type, as cellular structures such as membranes and organelles, and the composition of extracellular space hinder water diffusion to varying degrees.

In conventional MRI, hydrogen protons in water molecules are aligned using an external magnetic field and perturbed using radiofrequency pulses to generate a detectable signal. In DWI, some additional magnetic field gradients are applied to specifically detect the movement of these protons [21]. Initially, a gradient pulse is applied, this pulse will dephase the hydrogen protons in water molecules causing them to precess at different rates depending on their position and scatter their spins in various directions. To rephase the protons, a second gradient pulse of opposite polarity is applied after a brief delay. For protons that remain stationary, this second pulse reverses the dephasing effect, allowing their spins to return to the original phase and produce a strong signal. However, if water molecules diffuse during the delay, the protons move to new positions, experiencing a different magnetic field during the second gradient. This movement prevents the spins from being correctly rephased, as the second pulse cannot align them with their original phase, resulting in partial signal attenuation in the MRI image [22] [23]. The signal attenuation is directly proportional to the extent of water diffusion and hence DWI can distinguish between tissues with varying cellular densities.

The DWI images are a result of the way water diffusion affects signal attenuation. When water molecules can move freely, for instance, in cerebrospinal fluid, diffusion is unrestricted. This causes intense attenuation of the original signal, which appears dark on DWI images [24]. Free movement of water molecules inhibits adequate spin rephasing, resulting in large signal loss. Conversely, in tissues where there are diffusion restrictions such as brain parenchyma, tumors, acute ischemic stroke areas,

structural barriers like cell membranes and other structural components that block the flow of water molecules, the signal attenuation is reduced, leading to higher signal intensity and a bright appearance on DWI [25].

DWI is a valuable diagnostic technique due to its sensitivity to microscopic structural changes that are imperceptible on routine MRI imaging. It is especially helpful in the early diagnosis of strokes, being capable of detecting ischemic events before routine imaging methods [26]. It is also essential in oncology, for instance, in tumor characterization, where it can differentiate between malignant and benign tissue based on water diffusion changes, since in oncological lesions the diffusion is restricted due to the uncontrolled proliferation of tumor cells and denser cellular areas, characteristic of this type of lesions. In pathologies including multiple sclerosis, DWI can give information on white matter integrity by identifying structural damage. And is also useful in the detection of infection and inflammation since water restriction in regions of active cellular activities like abscesses, can be highlighted on DWI images [27].

### 2.2.2 The Role of the b-Value in DWI Images

The b-value of Diffusion-Weighted Imaging (DWI) is a parameter that measures the sensitivity of the scan to water molecular diffusion. It is determined based on factors such as the strength, length, and timing of the applied magnetic field gradients during imaging [27]. The b-value is described mathematically as:

$$b = \gamma^2 G^2 \delta^2 (\Delta - \delta / 3) \tag{2.2}$$

Where $\gamma$ is the hydrogen gyromagnetic ratio, $G$ is the amplitude of the diffusion-sensitizing gradient, $\delta$ is the length of the gradient pulses, and $\Delta$ is the time interval between the gradient pulses. These parameters can be modulated to control the amount of diffusion influence on the DWI image.
Lower b-values, typically between 0-100 s/mm², use weaker gradients or shorter gradient durations, which produce images with minimal diffusion weighting. These images mainly reflect tissue structure and have a high signal-to-noise ratio (SNR), similar to conventional T2-weighted MRI. Because diffusion effects are minimal at low b-values, low b-value images retain fine anatomy and thus are suitable for viewing normal tissue morphology. However, they are less effective at identifying areas where water diffusion is restricted, e.g., situations like ischemic strokes or tumors [28].

Conversely, higher b-values, typically between 800-3000 s/mm², use stronger gradients and longer durations, which enhance water diffusion sensitivity. Consequently, tissues with limited diffusion appear bright because of limited molecular movement, while regions with free diffusion have decreased brightness due to the severe signal attenuation (see Figure 2.4, where several b-values are displayed).This feature renders high b-value imaging essential for distinguishing between disease tissues and normal anatomical tissues [29].

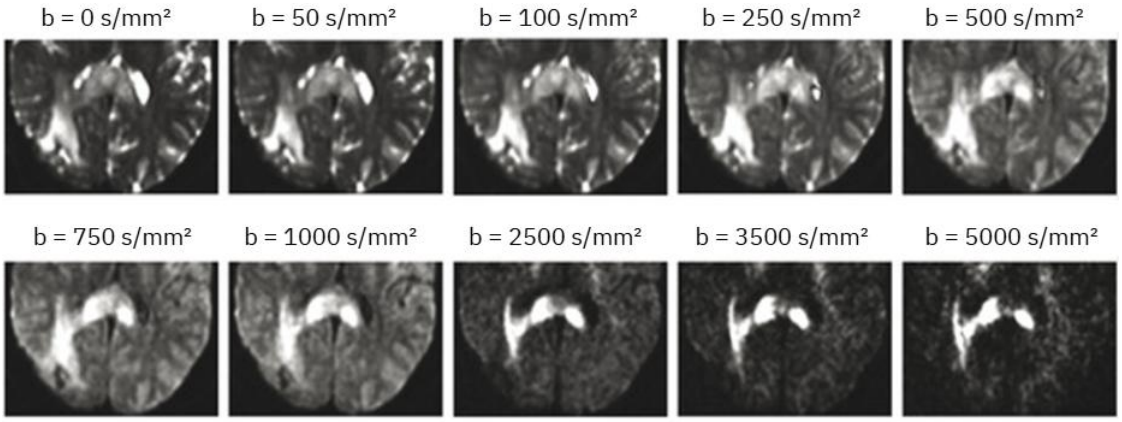


Figure 2.4: Diffusion-weighted images (DWIs) of a glioblastoma at various b-values: 0, 50, 100, 250, 500, 750, 1000, 2500, 3500, 5000 s/mm². As the b-value increases, the signal intensity of the cerebrospinal fluid and normal brain tissue decreases, enhancing the visibility of the glioblastoma due to its restricted diffusion, which appears brighter at higher b-values. (from [120])

However, increasing the b-value can lead to several issues that can impact image quality as well as clinical image usability. Its primary limitation is a substantial SNR loss [30] [31]. Additionally, as high-b-value images require longer acquisition times, this can lead to patient discomfort as well as greater chances of motion artifacts resulting in blurring and distortions that can cover subtle details [32]. This is especially challenging in patients who cannot be kept still, for example, those in pain, with cognitive impairment, or children. In addition to this, geometric distortions due to the strong diffusion gradients can reduce anatomical accuracy [32] [33].

Therefore, choosing the right b-value is critical to the optimization of image contrast, signal integrity, and diagnostic performance, with most applications requiring personalized diffusion-weighting strategies.

### 2.2.3 Diffusion-Weighted Imaging (DWI) in Cancer Detection

Diffusion-Weighted Imaging offers information regarding cellular density and cellular integrity, which are key indicators of malignancy. This is because malignant tissue is denser and has an altered anatomy, impeding water diffusion more than normal tissue, thus resulting in distinct signal patterns in the images of DWI [34].

Tumors have increased cellularity with reduced extracellular space and altered permeability of the membranes, all of which hinder the diffusion of water [35] [36]. Such restricted diffusion is indicated as an intense signal on DWI and thus differentiates malignant lesions from normal surrounding tissue. Unlike T1 or T2-weighted MRI, which gives primarily anatomical information, DWI gives functional information about the behaviour of the tissue and therefore is a helpful imaging method to identify and delineate neoplasms.

DWI is also able to detect subtle pathologic changes, allowing diagnosis to take place in an earlier stage, which is critical in the cancer prognosis [37]. DWI is broadly applied in the clinical treatment of most forms of cancer, such as the prostate, the breast, the liver, and cancers of the brain, due to its diagnostic capability [38] [39] [40] [41]. In prostate cancer, it is an indispensable component of mpMRI, supported by protocols including the Prostate Imaging Reporting and Data System (PI-RADS) [42]. The use of DWI greatly enhances the specificity and the sensitivity of the detection of the cancer of the prostate, especially of the peripheral zone, in which most of the cancers are present [43] [44] [45]. For instance, in a study by Syer et al [46], the combination of high b-value DWI and T2 weighted imaging considerably improved the diagnostic accuracy of prostate cancer.

DWI can also be used to quantify the apparent diffusion coefficient (ADC), a numerical measurement of how freely water molecules move within tissue, calculated from two or more DWIs with different b-values to provide a measure of molecular diffusion independent of other MRI signal influences, such as the T2 shine-through effect. The resulting ADC value is inversely related to the degree of diffusion restriction. In cancer diagnostics, ADC maps help detect and differentiate between benign and malignant lesions by highlighting areas of restricted diffusion, which appear as lower ADC values compared to normal tissue. Studies have shown that ADC measurements significantly improve diagnostic accuracy, with meta-analyses reporting increased sensitivity and specificity for prostate cancer detection [47].

Despite the high level of anatomical detail that can be obtained with low b-value DWI (e.g., b = 0-100 s/mm²), high b-value imaging (e.g., b = 1000-2000 s/mm² or more) is particularly valuable for cancer detection due to its enhanced sensitivity to restricted water motion. The malignant lesions stand out more as a result of selective suppression of the background signal, with the consequent improvement of lesion conspicuity (see Figure 2.5 (b), where a high b-value DWI is used to identify a lesion) [48]. Prostate cancer imaging with PI-RADS v2.1 recommends imaging with a minimum of one high b-value of ≥1400 s/mm² for the best detection of lesions with the aid of DWI [49].

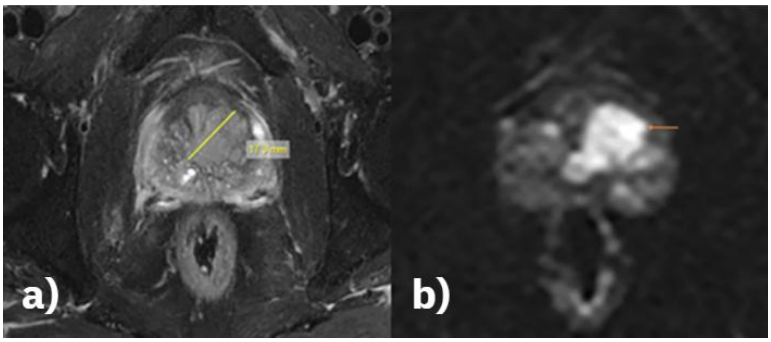


Figure 2.5: A patient diagnosed with prostatic acinar adenocarcinoma: (a) Prostate Axial T2: fat suppression sequence shows a lesion with defined boundaries. (b) Prostate DWI at a b-value of 1500 s/mm²: enhances the lesion's visibility compared to T2, demonstrating reduced signal-to-noise ratio (SNR) but improved CNR, enhancing cancer visibility for improved detection and characterization in prostate MRI. (from [121])

However, the practical issues of high b-value imaging, such as longer scan times and added susceptibility to motion artifacts, pose a particular challenge. These considerations emphasize the need for new approaches, such as machine learning-based synthesis, to enable the generation of high b-value images from faster, less noisy low b-value images [50].

## 2.3. Machine Learning

Machine learning (ML) is a subfield of artificial intelligence that allows programs to learn from data and enhance their performance on a certain task without explicitly being programmed for each step [51]. ML consists of teaching algorithms to find patterns, make predictions, or make decisions through learning from large data sets. Because ML models can adapt and improve with experience, they are viable solutions for complex problems where rule-based coding is not feasible, and patterns can be learned from the data. ML is best suited to solve problems involving large, complex, or noisy data sets, making it especially useful in fields like medical imaging [52].

In order to train an ML model, a large amount of data is given to an algorithm; the algorithm then adjusts its parameters internally to minimize error or maximize accuracy. After the model is trained, it is able to generalize to new, unseen data and provide solutions.

### 2.3.1 Neural Networks

The neural network (NN) is one of the most important concepts of ML due to its simplicity and efficiency. It can be described as a computational model inspired by the function and structure of the human brain [53]. Neural networks consist of artificial "neurons" stacked upon one another that receive input data, transform it through linear and non-linear operations, and generate an output.

Every neuron of a NN is provided with inputs (e.g., numeric values of a dataset) and each input learns a numerical coefficient (weight). These weighted inputs are summed, and a bias term is added to shift the result. In a final step, an activation function such as sigmoid, ReLU (Rectified Linear Unit), or tanh is used to introduce non-linearity into the model and determine the output of the neuron [54]. The output of a neuron $y$ can be represented as:

$$y = f\left(\sum_{i=1}^{n} w_i \cdot x_i + b\right) \tag{2.3}$$

Here, $x_i$ are inputs, $w_i$ are weights, $b$ is bias, and $f$ is the activation function (this equation can be visually interpreted as represented in Figure 2.6). For example, the ReLU function outputs zero for negative results and the sum itself for positive ones, allowing the network to filter and encode intricate patterns [55].

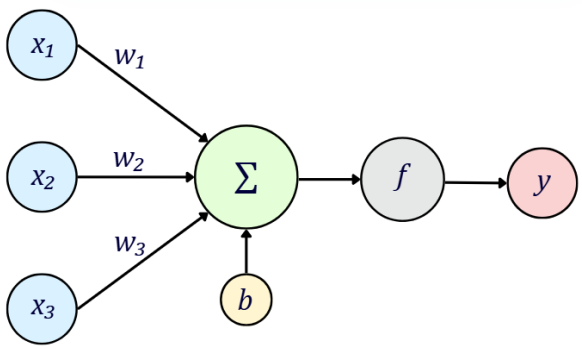


Figure 2.6: Single neuron model in a NN: processing of inputs to generate an output. Inputs $x_1$, $x_2$, $x_3$ (blue circles) are assigned weights $w_1$,$w_2$, $w_3$, respectively, and their weighted sum Σ (green circle) is computed. A bias term $b$ (yellow circle) is added to the sum, and the result is passed through an activation function $f$ (grey circle) to produce the output $y$ (red circle). The process is described by Equation 2.3.

The core architecture includes an input layer, the point at which data enters, hidden layers, where data processing takes place, and the output layer, where outputs are generated (this process is illustrated in Figure 2.7) [56]. Each neuron applies a mathematical operation to its inputs, based on learned weights, and passes the output to enable the network to model complex patterns in the data. In a multi-layer NN, the output of one layer becomes the input of the next, where the depth (number of layers) and the width (number of neurons per layer) control the network's ability to learn sophisticated patterns between data [57].

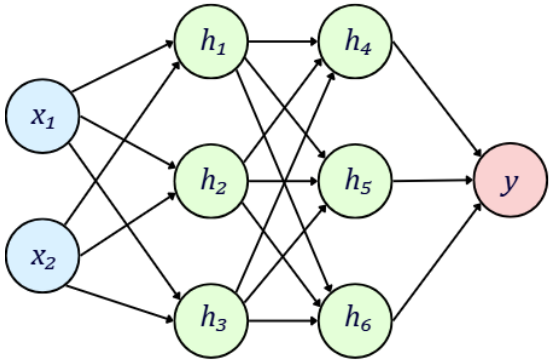


Figure 2.7: Model of a NN with two hidden layers: The input layer (blue circles, $x_1$, $x_2$) receives the data and passes it to the first hidden layer (green circles, $h_1$, $h_2$, $h_3$). The data is then processed through weighted connections to the second hidden layer (green circles, $h_4$, $h_5$, $h_6$), and finally the output layer (red circle, $y$) produces the predicted result.

#### 2.3.1.1. Training a Neural Network

Training a neural network is an iterative process, where each step can be understood as a two-part process: a forward propagation process and a backward propagation process. In the forward propagation phase, the network is activated by calculating the inputs of the initial layer, which propagate via the hidden layers through weighted connections until the output layer produces prediction outputs [58]. A cost function, such as the mean squared error, measures the performance of the algorithm by comparing the predictions of the model and real values:

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(y_{pred,i} - y_{true,i}\right)^2 \tag{2.4}$$

Where $J(\theta)$ is the cost function, $y_{pred}$ is the predicted output, $y_{true}$ is the true value, and $n$ is the number of examples.

In the backward pass, the parameters $\theta$, e.g., the weights and biases, are updated to minimize the cost function. This process is optimized with a gradient-based method: the derivative of the cost function with respect to each weight and bias is calculated in order to determine in which direction the cost function decreases the most [59]. The weights are then updated by an optimization algorithm such as gradient descent:

$$w_i \Leftarrow w_i - \eta \cdot \frac{\partial J}{\partial w_i} \tag{2.5}$$

Where $\eta$ is the learning rate, which determines the step size given the gradient and $\frac{\partial J}{\partial w_i}$ is the partial derivative of the cost function $J$ with respect to the weight $w_i$. When the whole training dataset goes forward and backward through the network once, which is referred to as an epoch, the predictions of the model are improved. Depending on the model and data in question, a large number of epochs may be necessary in order to reduce the cost function.

For the NN to accurately predict the data, it is necessary to optimize it. One of the most important aspects of the optimization is the selection of hyperparameters. The learning rate is one of the most critical hyperparameters. If the learning rate is too high the model overshoots the minimum and if it is

too low, training takes too long and can easily get stuck on a local minimum [60]. To help the model converge, optimization algorithms such as the Adam optimizer have been invented. Adam employs adaptive learning rate methods, calculating individual rates for each parameter from estimates of the first and second moments of the gradients. This makes it computationally effective and simpler to set up, incrementally adjusting $\theta$ until the cost is minimized and accurate predictions are made [61].

Overall, the goal of the training process is to generate a model that performs well not just on the training data but also on new inputs, like a test set - an ability referred to as generalization [62]. The main challenge is minimizing both the training error together with the test error. Underfitting happens when the training error is too high, and the network is not able to represent the data accurately. This usually occurs because the model's complexity is too low and can be resolved by adding more hidden layers to the NN [63]. Overfitting happens when the difference between the training error and the test error is too high, the network memorizes the training data, along with its noise and outliers, instead of learning the underlying general trends [64]. This can be understood through the bias-variance trade-off: bias refers to the expected error in predictions, while variance indicates how much the model's predictions vary with changes in the data. An ideal model achieves minimal bias and minimal variance. Overfitting often occurs when a network is too complex for the data it is trying to learn (e.g., has too many neurons or layers, or is trained for too many epochs), as increasing complexity to reduce bias (e.g., by adding more parameters) typically leads to higher variance, making the model overly sensitive to the training data [65].

Overfitting results in poor generalization to new data, lowering the practical use of the NN. It can be addressed by methods such as increasing the size of the training data, implementing regularization methods (e.g. weight decay) that focus on minimizing test error by penalizing larger weights, dropout regularization that randomly removes neurons during training, or early stopping that consists of stopping training when performance levels drop or the model has converged [66] [67].

### 2.3.2 Convolutional Neural Network

Convolutional Neural Networks (CNNs) utilize the concepts of the NN and adapt them to process structured grid-like data, such as images. In contrast to the conventional NN, which process inputs as flat vectors, CNNs take advantage of convolutional layers to extract spatial relationships in image data, making them highly effective for image classification, segmentation, and generation tasks [68]. As is illustrated in Figure 2.8 the convolutional layers employ filters or kernels that slide over the input with a given stride (number of pixels by which the filter moves across the input image) and calculate feature maps capturing local patterns, i.e., edges, textures, and shapes, minimizing the requirement for manual feature engineering.

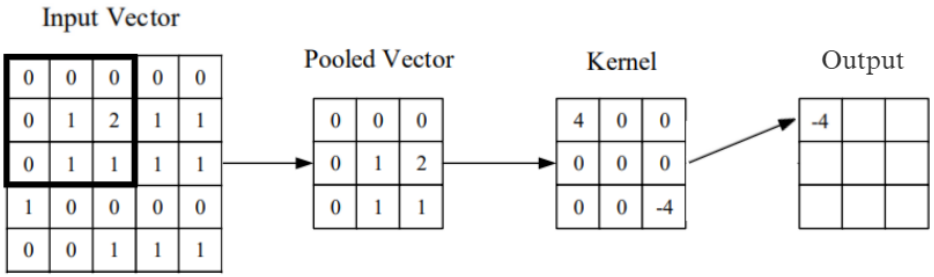


Figure 2.8: Representation of a convolutional layer. The kernel slides through the input vector, where each element of the pooled vector is multiplied by the elements in the kernel and summed, then added to the corresponding location in the output. (from [68])

The structure of a CNN for supervised learning consists of convolutional layers followed by activation functions, pooling layers to downsample and decrease computational complexity, and fully connected layers for final classification or regression, as illustrated in Figure 2.9. As noted, the network is trained to minimize a cost function through backpropagation, where gradients are calculated to modify filter weights. This trains the network, enabling it to efficiently generalize, focusing on spatial hierarchies, although careful tuning of hyperparameters such as filter size and stride is needed to manage bias and variance [69].

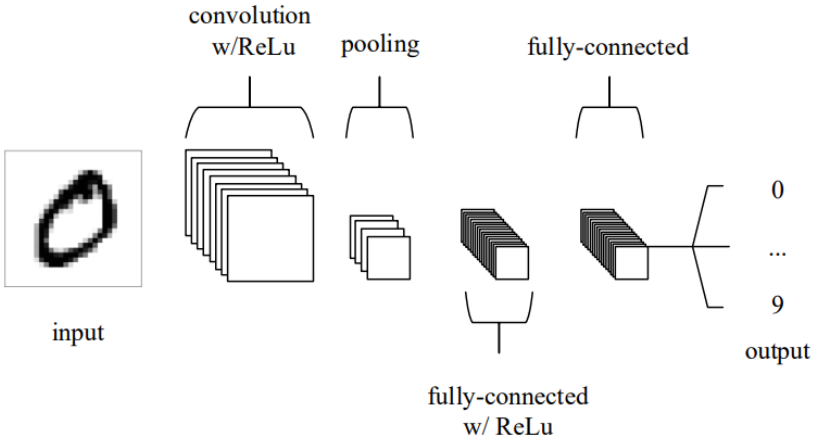


Figure 2.9: Representation of CNN architecture comprised of convolution layers, max-pooling, and fully-connected layers. (from [68])

### 2.3.3 U-Net

The U-Net model is a specialized CNN configuration specifically designed for image-specific tasks such as segmentation and image generation. Introduced by Ronneberger et al. in 2015 [70], U-Net is currently widely utilized for medical imaging due to its ability to use limited training sets while producing high-resolution outputs. Its unique architecture, which follows the shape of a "U," involves a contracting path to capture context and a symmetric expanding path to allow accurate reconstruction. Figure 2.10 exhibits a visual representation of a typical U-Net architecture.

The encoder path, or the contracting path, consists of repeated applications of convolutional layers and max-pooling steps, which iteratively reduce the spatial dimensions while increasing the number of feature channels. This downsampling captures context within the image by learning hierarchical features. Every convolutional block is normally composed of two convolutions, each succeeded by an activation function, and a max-pooling for dimensionality reduction. The expanding path, or the decoder, replicates this using upsampling layers, e.g., transposed convolutions or up-convolutions, that expand the spatial resolution but decrease the number of channels. Upsampling is supplemented by concatenation with matching feature maps from the contracting path, made possible by skip connections. These skip connections link the encoder and decoder at the same resolution levels (e.g., 128 to 128, 256 to 256), allowing high-resolution information from the encoder to be passed to the decoder, enabling the network to maintain spatial context, mitigate the loss of information due to max-pooling and thus reconstruct the image accurately.

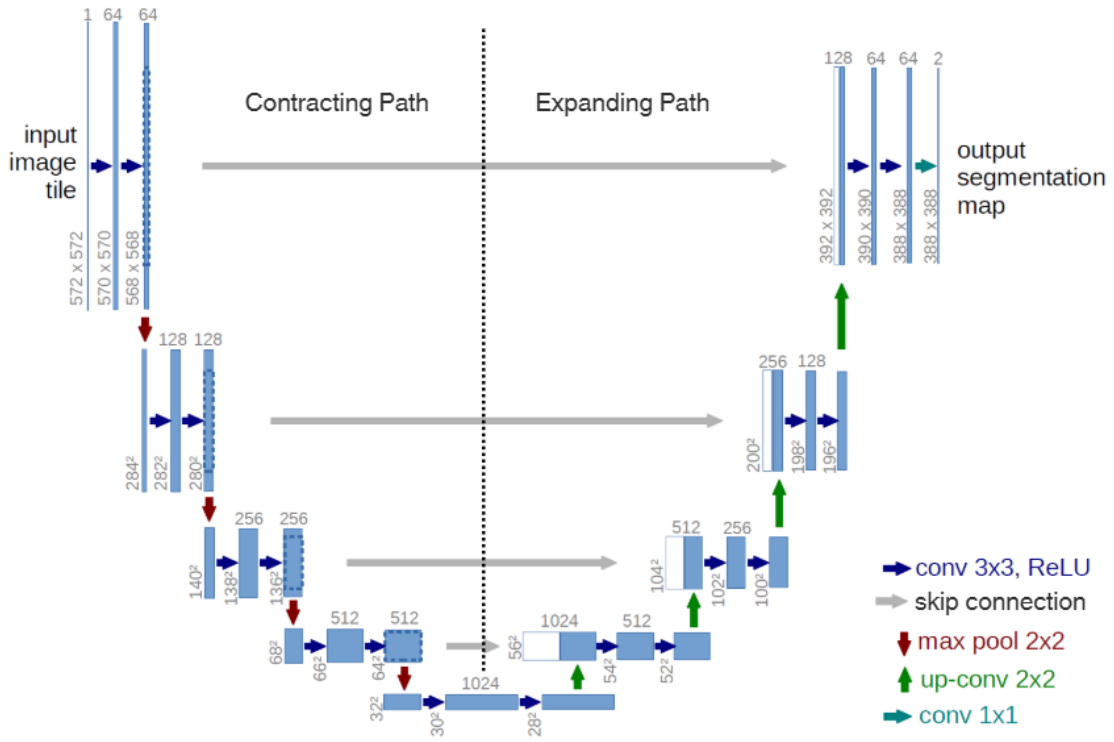


Figure 2.10: Representation of U-net architecture. featuring a contracting path with downsampling layers (e.g., 64, 128, 256, 512, 1024 channels) for feature extraction, an expanding path with upsampling layers for reconstruction, and skip connections linking corresponding resolution levels. Example with 572x572 in the highest resolution and 32x32 pixels in the lowest resolution. Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The resolution size is provided at the lower left edge of the box. White boxes represent copied feature maps provided by the skip connections. The arrows denote the different operations. (from [70])

Forward propagation in U-Net involves passing the input image to the contracting path to represent the features and then to the expanding path to reconstruct the features from the encoded form to a segmented or generated output. Backward propagation then refines the network's parameters in order to decrease the cost function to a minimum.

### 2.3.4 Generative Models

Generative models are ML algorithms that learn the underlying structure and distribution of data in an attempt to generate new similar samples [71]. Unlike discriminative models, where the inputs are discriminated or classified, generative models aim at learning the process of generating the data. Through the learning of the training data's structure, they are able to generate new outputs, e.g., images or text, equivalent to the original dataset [71]. This capability makes them useful for data synthesis or augmentation tasks, setting the stage for even more sophisticated neural network-based methods [72]. In the next sections, we will focus on the most relevant generative models, namely Autoencoders (AEs), GANs, and DDPMs.

### 2.3.4.1. Autoencoders

The autoencoders (AE) are some of the earliest generative models to be developed. An AE is an unsupervised learning model that learns a representation of data in latent space by trying to reconstruct its input into its output [69]. It has a symmetric neural network structure with an encoder that transforms the input into a latent space and a decoder that reconstructs it, as is depicted in Figure 2.11 [69] [73]. Typically, the input layer and the output layer have the same size, but the space has lower dimensions, which capture key features from the data to reconstruct.

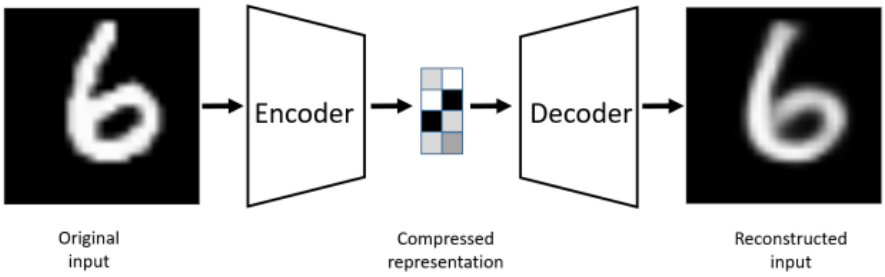


Figure 2.11: Structure of an AE: The original input is encoded, and the data is compressed into a latent space. Then the decoding from the latent space, which contains the important features, allows for the reconstruction of the original image. (from [122])

The AE is deterministic: it will always generate the same output for a given input. A variational autoencoder (VAE), its probabilistic counterpart, generates parameters of a probability distribution instead of an explicit reconstruction, allowing it to sample new data [74]. VAEs are regularized AEs, avoiding overfitting and making the latent space adequate for generation tasks [75]. AEs are applied in data compression and feature extraction, and VAEs generalize this to sampling new data [76].

Autoencoders are a simple solution to generative modelling but are limited by poor quality and diversity in the generated output and by the dependence on the consistency of reconstruction [77]. These restrictive factors led the search for more advanced solutions, such as GANs, which circumvent these limitations through competitive learning dynamics to perform generative tasks more effectively.

### 2.3.4.2. Generative Adversarial Networks

Generative adversarial Networks (GAN) are a significant advancement in generative techniques. GANs learn data patterns through the employment of two adversarial networks. This enables the generation of highly realistic results, which surpass the previous methods of creating data [78].

GANs consist of two main components: the generator that produces artificial data and the discriminator that verifies it (displayed in Figure 2.12). The generator begins with a random noise vector (e.g., uniform or Gaussian distribution) and generates new data that resembles real samples. The discriminator judges the probability that a sample belongs to real data rather than artificial data generated by the generator [78] [79].

Training is an adversarial process: the discriminator learns to distinguish real samples from fabricated ones, and the generator learns to generate more realistic samples, making it harder for the discriminator to distinguish between the artificial data and the real data. The parameters of the networks are adjusted through backpropagation as the discriminator learns through the analysis of the patterns of both real and fabricated data and the generator learns based on the discriminator's output. The gradients are propagated from the discriminator to the generator to maximize the generator's ability to produce realistic images (i.e., prompting the discriminator to classify them as 1 = real). This process is repeated

until a point known as the Nash equilibrium, where one network cannot improve without the other network changing as well [80].

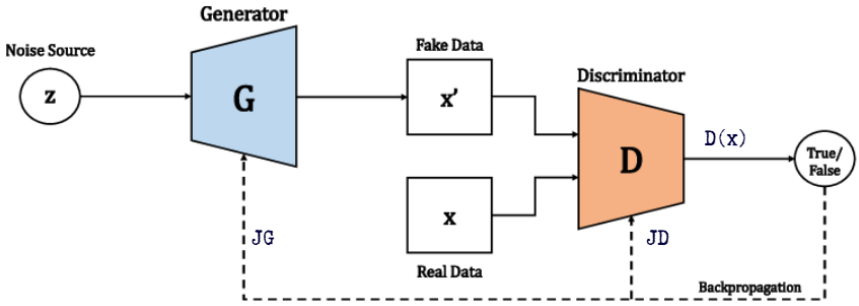


Figure 2.12: The architecture of a GAN. Real samples $x$, are drawn from the training dataset, while fake samples $x'$ are generated by the generator $G$. The generator $G$ takes a noise vector $z$ (a fixed-length random vector following a uniform or Gaussian distribution) as input and produces synthetic data $x'$ This process maps $z$ to a latent space representing the data distribution after training. The discriminator $D$ evaluates the likelihood of an input being real ($x$, assigned a value near 1) or fake ($x'$, assigned a value near 0). The discriminator is optimized using the loss function $JD$, while the generator uses the loss function $JG$. During training, the discriminator updates its weights using both real and fake samples, whereas the generator adjusts its parameters solely through backpropagation from the fake outputs. (from [123])

GANs create realistic data and can be used to generate text or pictures. GANs are effective when high quality is required, leveraging the adversarial process between the two models to improve their output more so than generative models like autoencoders [81].

However, GANs are restricted by a series of limitations that may render their practical application challenging. A key issue is training instability, wherein the adversarial process may fall out of balance, either the discriminator overwhelming the generator, resulting in poor generation, or the generator overpowering the discriminator, resulting in mode collapse [82]. Mode collapse occurs whenever the generator produces a limited set of outputs and is not able to model the full variety of the actual data distribution [83]. Furthermore, GANs can struggle to efficiently produce high-resolution outputs because of a compounded increase in instability and computation demands. These drawbacks have motivated research into other generation methods, including diffusion models, which balance some of the shortcomings of GANs while introducing novel advantages [84].

#### 2.3.4.1. Denoising Probabilistic Diffusion Models

Denoising diffusion probabilistic models (DDPM) are a class of generative models inspired by thermodynamics and stochastic processes, presenting an innovative method for data creation [85]. Unlike earlier techniques such as GANs, which rely on competitive networks, diffusion models generate data by simulating a process that gradually adds noise to real samples and learns to reverse it [86]. This begins with some initial data, like an image, and progressively adds Gaussian noise in several steps until it is transformed into pure noise in a process called forward diffusion. Then a neural network learns to invert the process, reconstructing the data step-by-step to learn its underlying distribution in a process called reverse diffusion or denoising process. Diffusion models are based on probability theory, using Markov chains and score-based learning as their building blocks. In a Markov chain, each step only depends on the last one, enabling a controllable process of noise addition and removal. Score-based learning guides the reversal by estimating how to direct the noisy data back towards its original form [86] [87]. Through this optimization, diffusion models produce samples that closely capture the patterns of the real data.

#### 2.3.4.1.1. The Diffusion Process and Reverse Generation

The forward diffusion process starts with the input data, denoted $x_0$, e.g. an unaltered image. It incrementally adds Gaussian noise in a series of timesteps, generating a chain of increasingly noisy versions. The first image with added noise is $x_1$ and noise is added until the image reaches pure noise $x_T$. Mathematically, this is denoted by the equation:

$$q(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I\right) \tag{2.6}$$

Here, $x_t$ is the data at step $t$ with some noise, and $x_{t-1}$ is the data at the previous step. The term $\beta_t$ is a noise scheduler that determines how much noise is added to the image at each step $t$. While using the identity matrix $I$ to define the covariance structure of the added noise, ensuring that the noise applied to each pixel is independent of the others. The mean of this Gaussian distribution, $\sqrt{1-\beta_t}x_{t-1}$, slightly contracts the previous data while the variance $\beta_t I$ adds the noise. As t approaches the final timestep $T$, $\beta_t$ tends to increase, and towards the end, $x_T$ is nearly equivalent to pure Gaussian noise, with no trace of the original structure left.

The reverse diffusion process does the opposite, attempting to recover $x_0$ from the noisy $x_T$. Every backward step, from $x_t$ to $x_{t-1}$, is learned by a neural network, progressively removing the noise. This can be expressed as:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1};\ \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\right) \tag{2.7}$$

Here, $x_{t-1}$ is the predicted data at the previous timestep, conditioned on the current noisy $x_t$ at timestep $t$. The network, parameterized by $\theta$, predicts $\mu_\theta(x_t, t)$, the mean of the less-noisy data, and $\Sigma_\theta(x_t, t)$, the variance, which represents the uncertainty of the prediction. This process is displayed in Figure 2.13.

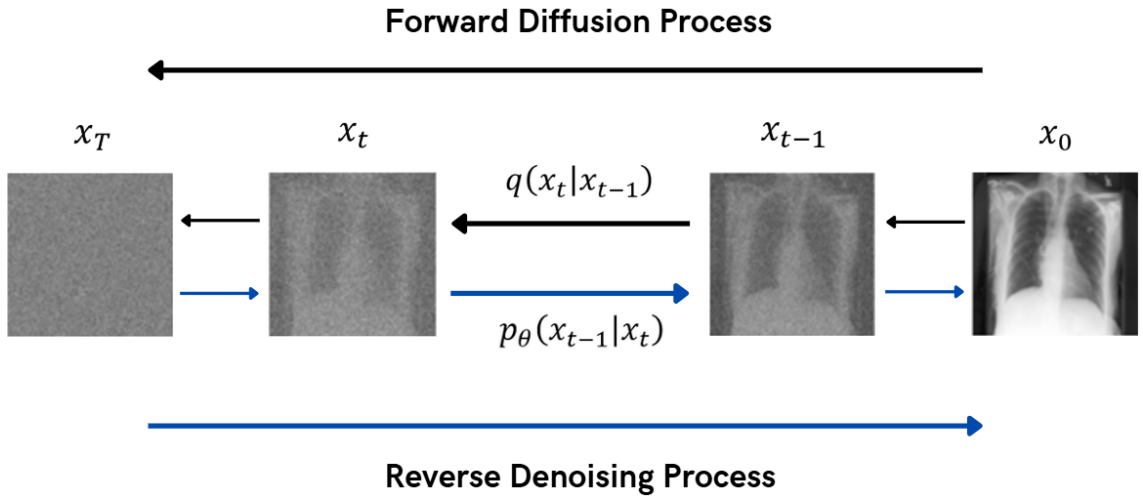


Figure 2.13: Diffusion model process for image generation: The forward diffusion process (top) gradually adds noise to the original image $x_0$, transitioning through intermediate states $x_{t-1}$ and $x_t$ with probability $q(x_t|x_{t-1})$, until reaching a noise distribution $x_T$. The reverse denoising process (bottom) starts from $x_T$ and iteratively removes noise using the learned probability $p_\theta(x_{t-1}|x_t)$, reconstructing the original image $x_0$, as exemplified with a chest radiograph.

Typically, the variance is simplified or fixed, and the mean is the main target of prediction since it drives the step-by-step denoising process [86]. The network is trained by matching its noise predictions with the real noise added in the forward process, refining its parameters $\theta$ incrementally. To generate new data, the model starts with random noise $x_T$ and applies this reverse process for every $T$ step until it ends at a clean sample $x_0$ that contains the features of the training data. This generation is done using the learned ability of the network to transform unstructured noise into structured data, creating new samples with the statistical features of the original dataset. During these sequential processes, the noise gradually turns into meaningful patterns under the guidance of the model's learned data distribution [88] [89].

Diffusion models produce diverse outputs, capturing a broad range of patterns in the training data because the reverse process gradually, explores the distribution in the data step-by-step. They have a high stability in training and a clear optimization objective: to minimize the difference between the noise added and predicted.

Diffusion models can also perform conditional generation, where the reverse process is guided by specific inputs, such as an image or text, to shape the generated output [90]. For example, feeding a low-resolution or noisy input image enables the model to denoise or upscale it to generate a high-quality version that maintains the structural attributes of the input, such as edges or anatomical structures in medical images [91]. To incorporate any other conditional data, like text prompts or class labels, this data can be included in the model, usually by concatenation with either the noisy input $x_t$ or the timestep $t$ in the neural network architecture, so that the model can condition on the denoising process and sample based on those conditions.

Additionally, this type of conditioning can be effectively achieved using a cross-attention mechanism, a technique widely adopted in transformer-based architectures. In this type of mechanism, the model attends to the conditional input (e.g., a low b-value DWI or target b-value) by computing attention scores that weigh the relevance of the condition across the noisy input features at each denoising step. The cross-attention process involves projecting the conditional data into a query space and aligning it with keys and values derived from the noisy input, allowing the model to dynamically focus on relevant features. This enables the generation process to be guided precisely, ensuring that the output (e.g., a high b-value DWI) reflects the structural and diffusion characteristics specified by the condition.

These models excel in generating high-quality outputs. The iterative process of the model guarantees that the final generated samples, such as images, have fine details and realistic textures that closely resemble the real data on which they are trained, rendering diffusion models especially useful for medical image synthesis, an area in which anatomical consistency and accuracy are essential. These models have been utilized to create realistic MRI or CT scans, recreating anatomical features with high accuracy, which can be helpful for augmenting datasets or simulating cases where actual data is limited [92] [93].

## 2.4. Quality Assessment of Generated Images: Metrics

As explored, generative models represent an effective tool for creating high-quality images, making them strongly indicated for medical image generation. However, the clinical utility of the generated images depends on the extent to which they are able to replicate real images, and therefore need thorough evaluation metrics by which their quality can be determined.

In order to quantitatively measure the quality of the generated images, it is necessary to consider three key concepts of image comparison analysis: pixel-level coherence, perceptual coherence, and semantic coherence. Pixel-level coherence focuses on measuring the exact correspondence of individual

pixel values between the generated and real images, which can be quantified using metrics such as mean squared difference (MSD) and peak signal-to-noise ratio (PSNR). Perceptual coherence evaluates how visually similar the images appear to observers, focusing on the overall appearance and structural resemblance, assessed by metrics such as the structural similarity index measure (SSIM) and feature similarity index measure (FSIM). Semantic coherence assesses the higher-level context, ensuring the generated images preserve the intended content, evaluated for instances by the metric Fréchet inception distance (FID).

It is important to use this diverse set of metrics because pixel-focused metrics alone do not account for perceptual differences. In areas such as medical image generation, where the main concern is how the image is perceived rather than exact pixel matches, it is recommended to combine pixel-focused metrics with perceptual and semantic metrics to ensure the preservation of diagnostic integrity and clinical relevance.

## 2.4.1 Pixel-level Metrics

### 2.4.1.1. Mean Squared Difference (MSD)

Mean Squared Difference (MSD), also known as Mean Squared Error (MSE), is a simple measure for the comparison of two images based on the average squared difference between pixel intensities. MSD measures the numerical difference between a target image and a synthesized image based on pixel-wise fidelity. MSD is widely used in image processing, where accurate pixel matching is prioritized [94]. A lower MSD means greater similarity, making it a straightforward measure of error. However, MSD does not account for human perception, two images with equal MSD can be visually different due to its sensitivity to pixel intensity shifts and lack of structural awareness [95].

The MSD is calculated as follows for two images $I_{gen}$ (generated) and $I_{real}$ (real), each of size $M \times N$ pixels:

$$MSD = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} \left( I_{gen}(i,j) - I_{real}(\mathrm{i},j) \right)^2 \tag{2.6}$$

Here, $I_{gen}(i,j)$ and $I_{real}(\mathrm{i},j)$ are the pixel intensities at position $(\mathrm{i},j)$. The result is averaged over all existing pixels, providing a direct measure of pixel-level discrepancy.

### 2.4.1.2. Peak Signal-to-Noise Ratio (PSNR)

The Peak Signal-to-Noise Ratio (PSNR) is a simple metric that assesses the quality of an image by measuring the ratio between the maximum pixel intensity and the error or noise of an image, where higher values represent higher quality and less noise, and is expressed in decibels [96]. It is widely applied in image compression and medical imaging as an efficient tool to measure the level of noise [97].

However, PSNR is only a measurement of quality. In applications where the image features are relevant, it is recommended to combine PSNR with metrics like SSIM or FSIM [98]. The PSNR is calculated as:

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX^2}{MSE}\right) \tag{2.7}$$

Here, $MAX$ is the maximum possible pixel value (e.g., 255 for 8-bit images) and $MSE$ is mean squared error.

## 2.4.2 Perceptual Coherence Metrics

### 2.4.2.1. Structural Similarity Index Metric (SSIM)

The Structural Similarity Index Measure (SSIM) measures image similarity based on luminance, contrast, and structural features with the aim of simulating human visual perception. SSIM captures how humans perceive images based on structural patterns instead of pixel-wise discrepancy [99]. This metric is extensively used in image quality assessment in areas such as medical image synthesis, where preserving the perceived structure is more important than the exact pixel matching [100]. SSIM ranges from -1 to 1, with 1 indicating identical images. SSIM is calculated as:

$$SSIM(I_{gen}, I_{real}) = \frac{(2\mu_{gen}\,\mu_{real} + C_1)(2\sigma_{gen,real} + C_2)}{(\mu_{gen}^2 + \mu_{real}^2 + C_1)(\sigma_{gen}^2 + \sigma_{real}^2 + C_2)} \tag{2.8}$$

Here, $\mu_{gen}$ and $\mu_{real}$ are mean intensities and $\sigma_{gen}$, $\sigma_{real}$ are standard deviations, $\sigma_{gen,real}$ is the covariance, and $C_1$ and $C_2$ are constants to avoid division by zero. A high SSIM reflects perceptual similarity.

### 2.4.2.2. Feature Similarity Index Metric (FSIM)

The Feature Similarity Index Measure (FSIM) focuses on analysing low-level features like phase congruency and gradient magnitude, which quantify concepts like textures and edges [101]. FSIM is an extension of SSIM that highlights perceptually significant features and is therefore more adequate for tasks involving fine details, e.g., image enhancement or medical image creation [101] [102]. It ranges from 0 to 1, where a score of 1 indicates a perfect similarity between features. This metric is able to capture textural differences effectively, but can be computationally demanding due to feature extraction. The FSIM is calculated as:

$$FSIM = \frac{\sum_{i,j} S_L(\mathrm{i}, j) \cdot PC_m(\mathrm{i}, j)}{\sum_{i,j} PC_m(\mathrm{i}, j)} \tag{2.9}$$

Here, $S_L(\mathrm{i}, j)$is the similarity at position $(\mathrm{i}, j)$ based on phase congruency $PC$ and gradient magnitude, and $PC_m(\mathrm{i}, j)$ weighs perceptually significant regions.

### 2.4.3 Semantic Coherence Metrics

#### 2.4.3.1. Fréchet Inception Distance (FID)

The Fréchet Inception Distance (FID), measures the similarity between the vector feature distributions of two image sets, typically extracted using a pre-trained neural network that has learned to identify meaningful features from a large dataset (originally the Inception v3 network). Theoretically, FID assesses the distance between these feature sets by assuming they are drawn from multivariate normal distributions, evaluating high-level feature similarities, such as shapes and textures rather than relying solely on raw pixel comparisons. This approach provides a robust metric for comparing statistical properties of deep features extracted by the network.

However, the choice of the pre-trained network and the training data used significantly influences the features it is capable of extracting. For medical images, for example, the network must be pre-trained on a medical imaging dataset, such as one containing radiographs or MRIs, as it would be better optimized to identify domain-specific features like anatomical structures or pathological patterns, which significantly differ from those in natural images.

FID is widely used in fields like generative modelling in order to assess how well generated images match the overall distribution of real images [103]. A lower FID represents a better alignment between distributions and high realism of the synthetic images. FID is calculated as:

$$FID = \left\|\mu_{gen} - \mu_{real}\right\|_2^2 + \mathrm{Tr}\left(\Sigma_{gen} + \Sigma_{real} - 2\left(\Sigma_{gen}\Sigma_{real}\right)^{1/2}\right) \tag{2.10}$$

Here, $\mu_{gen}$, $\mu_{real}$ are the mean feature vectors, and $\Sigma_{gen}$, $\Sigma_{real}$ are the covariance matrices of features extracted from the generated and real images. The trace (Tr) of the sum of these matrices quantifies the extent to which the distributions diverge, providing a comprehensive measure of dissimilarity between the two sets.

## 2.5. Literature Review: Generative Models for High B-value DWI Synthesis

As noted in section 2.2.3, DWI is particularly useful in the detection and characterization of prostate cancer, due to its ability to measure water molecule motion within tissue, which can be correlated with cellular density and malignancy. High b-value DWI (for example, b = 1500 s/mm² or greater) distinguishes cancerous from non-cancerous tissues clearly by enhancing diffusion effects and therefore, improving diagnostic accuracy and lesion visibility. However, scanning for high b-value DWI in the clinical setting has some drawbacks, like increased scan duration, increased susceptibility to motion artifacts, and reduced SNR, potentially compromising image quality and patient comfort. These limitations have led the research on generative models to produce high b-value DWI from lower b-value images (e.g., b = 800-1000 s/mm²).

### 2.5.1 GAN-Based Methods

In a promising study, Hu *et al*. (2021) [104] suggested using Generative Adversarial Networks (GANs) to generate high b-value DWI (b = 1500 s/mm²) of the prostate from routine scans (b = 800 or 1000 s/mm²). Their method focused on preserving anatomical and pathological details by learning the mapping between low and high b-value images. The study involved 395 patients and to validate their results quantitative metrics and quality and diagnostic scores annotated by radiologists were utilized. Quantitatively, the synthesized images were compared to DWI obtained at b = 1500 s/mm² using MSD, PSNR, SSIM, and FSIM, with successful results on all metrics. For image quality, two 3-year-experienced radiologists independently scored the DWI sets (acquired and generated) on a five-point Likert scale for benign tissue suppression, anatomic distortion, artifacts, and overall quality, and found the generated sets to be superior on all aspects.

Diagnostic performance was assessed by four radiologists (two residents and two with more than 20 years of experience) who scored prostate cancer likelihood on a five-point scale, validating the reliability of generated images with AUC 0.89, showing great clinical potential for prostate cancer imaging.

In another research, Rezaeijo *et al*. (2022) [105] investigated the possibility of using GAN-based models, including CycleGAN, Pix2Pix, and DC²Anet, for synthetic DWI generation of the prostate with varying b-values. Through their research, it was demonstrated that the models could accurately generate synthetic DWI images with b-values of 400 and 800 s/mm² from acquired DWI with a b-value of 50 s/mm².

The research involved 170 patients with prostate cancer. The synthetic images generated were validated with five metrics: mean absolute error, root mean squared error, Pearson's correlation coefficient, PSNR, and SSIM. From the assessed models, DC²Anet exhibited the highest accuracy in comparison with CycleGAN and Pix2Pix. The generated images closely resembled real high b-value DWI, maintaining structural fidelity and diagnostic quality, demonstrating the potential of these GAN-based methods.

Hu e*t al*. (2024) [106] proposed another study that suggested an unsupervised learning model called Attention Dual Contrast Vision Transformer CycleGAN (ADCVCGAN) to improve image quality and shorten scanning time in gastric DWI. The model produced high b-value DWI (b = 1200 s/mm²) using lower b-value acquisitions (b = 800 s/mm²). The ADCVCGAN employed attention mechanisms with a Convolutional Block Attention Module (CBAM) and vision transformers in a U-Net framework to improve feature extraction and refinement of details to produce synthetic images of the same quality as directly acquired high b-value images.

The study included 200 patients with stomach cancer. The quality of the created DWI was checked using different measurements, like PSNR, SSIM, FSIM, MSE, weighted peak signal-to-noise ratio, and weighted mean squared error. The generated DWI showed high scores in all these measurements when compared to the images directly acquired with high b-values. For image quality, two radiologists with 8 and 10 years of abdominal MRI experience independently scored acquired and generated sets of DWI in blinded sessions at 2-week intervals using a five-point Likert scale for lesion visibility, anatomical detail, distortion, and overall quality, with the generated DWI set scoring similarly to the acquired DWI set.

To assess consistency, two radiologists identified the largest nodule area in both DWI sets, three times, one month apart for each marking. It was observed from the results that the repeatability and consistency were alike for both sets, proving the generated DWI sets to be consistent in making clinical decisions.

The vision transformer and attention mechanism improved feature extraction and refinement of the details, resulting in synthetic images with the quality of acquired high b-value images. Although the study was focused on gastric cancer, the methodology could be adapted for prostate DWI synthesis.

However, these methods presented some limitations, particularly Hu *et al*. (2021) [104] primarily noted the restriction to DWI sets with b-values ranging from 800 to 1000 s/mm², limiting the model's applicability to higher or lower b-values. Rezaeijo *et al*. (2022) [105] highlighted the lack of multi-center validation as a key constraint, compounded by their use of a narrow b-value range (50, 400, and 800 s/mm²), which falls short of the higher b-values (e.g., >1500 s/mm²) critical for prostate cancer assessment. Hu e*t al*. (2024) [106] emphasized the model's reliance on a single b-value pair (800 to 1200 s/mm²) and potential variability in DWI and ADC values due to equipment and population differences, underscoring the need for large-scale, multi-center validation to ensure generalizability.

The reviewed articles are constrained by their b-value transformations, often employing a starting b-value that is excessively high or a target b-value that falls short of the desired high range. This restricts the potential benefits of these studies, as the optimal goal is to significantly reduce acquisition times while generating images with very high b-values, which are the most challenging to acquire. Moreover, in real-world clinical settings, where a wide variety of protocols are utilized, a single solution may not suffice to produce consistently useful results across diverse imaging conditions.

### 2.5.2 Diffusion Model-Based Approaches

Diffusion models have shown significant potential in medical image generation. Although there is limited research on the application of diffusion models for the generation of high b-value diffusion-weighted imaging (DWI) in prostate cancer, the prospect for these models in this context is evident.

Several studies in more general medical imaging areas have shown the power of diffusion models for image generation, denoising, and super-resolution, which can benefit future research on prostate DWI synthesis.

Khader *et al*. (2022) [107] developed DDPM for the generation of high-quality 3D medical images like MRI and CT scans. This research highlights the ability of DDPMs in creating realistic and high-detailed anatomical images through a process of progressively adding and removing noise.

The model was tested with quantitative and qualitative approaches, in which two radiologists scored the synthesized images for realism, anatomical accuracy, and slice consistency. The results indicated that synthetic images met high-quality standards appropriate for medical use. The model guaranteed the preservation of fine-grained anatomical details, which is important for medical use.

While this study was not DWI-specific, the technique could be used as a foundation for generating high b-value prostate DWI using the same denoising and noise-adding processes. The fact that the method can be used to generate high-quality 3D volumes makes it particularly relevant to volumetric prostate imaging, where tissue integrity is important to preserve.

Friedrich *et al*. (2024) [108] proposed a novel wavelet-based diffusion model for synthesizing high-resolution 3D medical images. This research applied a diffusion model on wavelet decomposed images. By decomposing the images into multiple frequency bands, the model was capable of effectively scaling 3D diffusion models to high resolutions maintaining both global and fine-grained structures, making it adequate for high-resolution imaging tasks.

The research was tested on two public datasets: the BraTS 2023 Adult Glioma dataset for MRI of the brain and the LIDC-IDRI dataset for CT of the lungs. Quantitative analysis utilized metrics including the FID for calculating image fidelity and the multi-scale structural similarity index measure (MS-SSIM) for determining image diversity. Results showed that the wavelet-based diffusion model achieved state-

of-the-art performance, with low FID scores reflecting realistic generation of images and low MS-SSIM scores indicating high diversity among generated samples.

This approach can be relevant to high b-value DWI reconstruction because the wavelet approach helps decrease the image noise that typically occurs with higher b-values. In DWI, the bigger the b-value, the lower the signal-to-noise ratio. Wavelet-based models can rectify this by maintaining high-frequency information, which is crucial for correct prostate lesion detection. The fact that the model can generate high-resolution 3D images qualifies it as a strong contender for extension to prostate DWI synthesis, where anatomical fidelity is critical.

In another study, Dorjsembe *et al.* (2024) [109] explored conditional diffusion models for semantic 3D brain MRI image synthesis. This model employs semantic labels as conditional inputs to direct the generation process so that the generated images are anatomically consistent and correctly delineate the structures of interest. In this research a dataset of 3D brain MRI scans was employed to train and test the Med-DDPM model. Quantitative assessment was conducted with the Dice score on tumor segmentation tasks, which measures the intersection of predicted and ground-truth segmentations.

Med-DDPM model obtained a similar dice score among synthetic images and real images. The paper showed how conditional diffusion models can generate high-quality MRI images very similar to actual brain scans.

Applying this approach to high b-value DWI generation could involve using anatomical information, such as the location of the prostate or other organs, as conditional inputs to guide the synthesis of high b-value images from low b-value images. This targeted approach would enhance anatomical coherence between the original low b-value scans and the generated high b-value outputs, thereby improving the preservation of pathological details and boosting the diagnostic quality of the resulting images.

# 3. Methodology

The primary objective of this chapter is to present the research process. It begins by detailing the materials employed, including hardware, datasets, models, and training techniques. Subsequently, it offers a comprehensive exposition of the methodologies applied for model validation and statistical analysis of results.

## 3.1. Objective of the Study

The objective of this study is to develop a generative model capable of taking low b-value DWI images as inputs and generating the corresponding high b-value DWI images, maintaining anatomical details and diagnostic performance of the original high b-value images. The models explored for this task include DDPMs, GANs, and VAEs.

## 3.2. Hardware and Software

The models were trained on a computer provided by the Champalimaud Foundation, equipped with two NVIDIA TITAN RTX GPUs, which provided the necessary computational power for processing the PROSTATEx and ProCAncer-I datasets. The software environment utilized Python 3.13, along with PyTorch (version 2.3.0) for deep learning and MONAI (version 1.3.0) for medical imaging tasks. The radiology images were stored in DICOM and NIfTI formats, processed using libraries such as pydicom and nibabel for compatibility with the MONAI pipeline.

## 3.3. Data Description

In this study, two databases were utilized to develop and evaluate the proposed models. The larger dataset, ProCAncer-I, was used to train the main diffusion model, offering a substantial and varied source of information to guide the learning process. In addition, the PROSTATEx database, which is a smaller dataset, was utilized in the initial stages of model development and subsequently for validation purposes, since it is an external dataset that is not analyzed by the model in the training process, providing an independent benchmark to assess the generalizability and performance of the trained models.

### 3.3.1 PROSTATEx Database

The PROSTATEx database was developed as a benchmark for prostate MRI analysis and includes more than 10,000 curated prostate MRI examinations aimed at validating modern AI algorithms and radiologist performance in the detection of clinically significant prostate cancer. Created in collaboration with an international, multi-disciplinary scientific advisory board of 16 experts in AI, radiology, and urology. PROSTATEx comprises prostate MRI examinations acquired on a Siemens 3T MR scanner and encompasses multiple imaging modalities such as T2-weighted, proton density-weighted, dynamic contrast-enhanced, DWI and ADC. However, for the purposes of this study, only the DWI, ADC and T2 images were used.

The PROSTATEx dataset includes 345 studies of prostate DWI, each consisting of 19 slices, and acquired using three distinct b-values: 50 s/mm², 400 s/mm², and 800 s/mm², resulting in a total of 19,665 images. The DWI images were obtained with a single-shot echo planar imaging sequence incorporating diffusion-encoding gradients, and an ADC map was then computed by the scanner software for each case. In addition, 205 of these studies include radiologist segmentations, delineating regions where lesions are present.

The PROSTATEx dataset was selected due to its relatively small size and high quality. Making it ideal for experimentation with various models in order to identify the most suitable approach before scaling up to a larger and more diverse dataset: the ProCAncer-I.

### 3.3.2 ProCAncer-I Database

The ProCAncer-I database is a large-scale repository developed to aid in prostate cancer research. ProCAncer-I is an EU-funded initiative that collaborates with 20 partners, including leading clinical centers and AI experts. The dataset contains more than 17,000 anonymized mpMRI cases obtained from various clinical sites across Europe, resulting in a diverse set of imaging data for prostate cancer analysis. It encompasses several types of mpMRI images, such as T2-weighted, dynamic contrast-enhanced, DWI and ADC sequences. However, only the T2, ADC and DWI were used in this study.

The DWI in the ProCAncer-I database consisted of 8731 studies. The data utilized in this study had a cut-off of acquisition until 31 of March of 2024 and any data acquired after that was excluded. Additionally, studies that contained lesion segmentation annotations were not used to possibilitate a future validation. The remaining 5,114 studies that were used for training, encompassed a total of 473,232 images. Each study represents a patient's prostate DWI acquisition, and each image reflects the individual slices within these studies, capturing a range of b-values used to measure diffusion. To ensure the analysis relied solely on scanner-acquired data, all files corresponding to synthetic DWIs images generated through post-processing rather than direct acquisition were excluded, leaving 459,636 images for analysis.

The characteristics of the DWI dataset are summarized through several distributions. There are 35 unique b-values present in the dataset, ranging from 0 s/mm² to 2200 s/mm² and the corresponding distribution is represented in Figure 3.1 (a), showing the top 10 b-values with 0 s/mm² and 1500 s/mm² being the most frequent. The manufacturers of the MRI machines used to acquire the images differ depending on the clinic of origin and are depicted in Figure 3.1 (b), with Philips contributing with the majority of the files, followed by GE Medical Systems. The pixel spacings of the DWI images reflect their spatial resolution and are shown in Figure 3.1 (c), where the most common spacing is around 1.50 mm x 1.50 mm. Lastly, the images also differ in size, as shown by Figure 3.1 (d), where the most common size is 256×256 pixels, followed by 224×224 and 160×128 pixels.

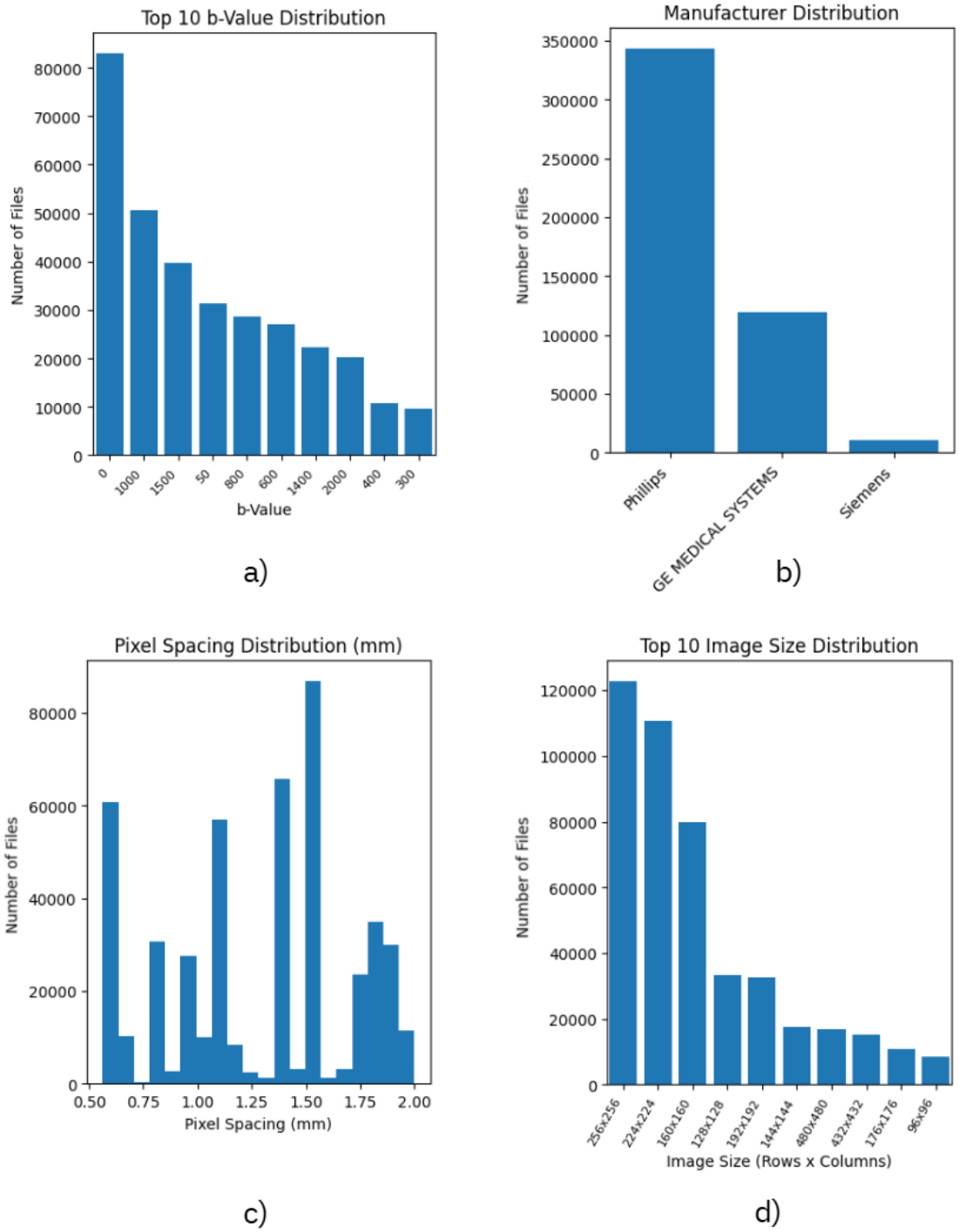


Figure 3.1: Histograms of DWI slices in the ProCAncer-I database. a) Histogram of top 10 b-values distribution. b) Histogram of manufacturer distribution. c) Histogram of pixel spacing distribution. d) Histogram of top 10 image size distribution.

### 3.3.3 Data Processing and Preparation

Preprocessing the DWI images in the PROSTATEx and ProCAncer-I datasets consisted in a set of operations to standardize the images for model training and validation.

The initial step was the normalization of pixel intensity values, specific to each image. For both datasets, the minimum and maximum pixel values of each DWI image were registered, and the intensities were subsequently scaled to the range [0, 1], in order to ensure consistency across images.

Subsequently, to preserve the spatial resolution between all of the images in the dataset and allow the model to learn spatial features persistently, the pixel spacing was adjusted to the median value across each dataset. For the PROSTATEx dataset, the pixel spacing was normalized to 2.00 mm x 2.00 mm, while in the ProCAncer-I dataset the spacing was adjusted to 1.38 mm x 1.38 mm.

Then, each image underwent a central spatial crop to a uniform size of 96x96 pixels in order to focus computational resources on the most relevant anatomical regions. This cropping targeted the central area of the pelvis, containing the prostate, which is the primary region of interest for lesion detection and therefore, generation tasks, effectively discarding peripheral regions that contain less diagnostically significant information. By concentrating on this core region, the models were optimized to learn patterns and transformations relevant to prostate pathology, avoiding the inefficiency of processing the entire DWI image.

The main step of the data preparation involved pairing the DWI images to support the training of models capable of learning the transformations between different diffusion weightings. For both the PROSTATEx and ProCAncer-I datasets, DWI studies were acquired across multiple b-values, with each case providing a series of slices at these diffusion levels. The images were paired in a low-to-high b--value manner, generating all possible combinations of low and high b-value pairs within each study (Figure 3.2 illustrates this pairing process). This approach maximized the utilization of the available DWI data, enabling the models to capture the nuanced patterns and structural changes associated with varying diffusion weightings. As a result, the ProCAncer-I dataset yielded 475,525 pairs, while the PROSTATEx dataset produced 19,665 pairs. These pairs served as the main foundation for training the models to predict and generate high b-value images from their lower b-value counterparts.

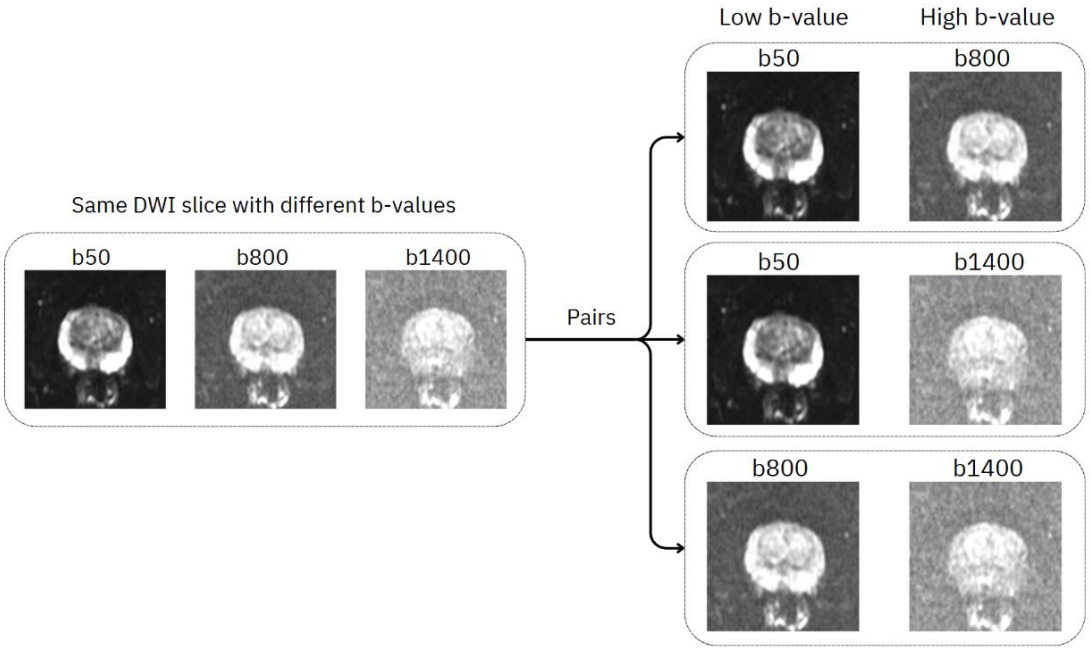


Figure 3.2: Example of low-to-high b-value pairs from the ProCAncer-I dataset, illustrating a single DWI slice acquired at three different diffusion weightings: 50 s/mm², 800 s/mm², and 1400 s/mm². These images form three distinct low-to-high b-value pairs (b50-b800, b50-b1400, and b800-b1400).

## 3.4. Generative Models

This study experiments with VAE, GAN, and DDPM to generate high b-value DWIs from low b-value inputs. These models were selected due to their capabilities in generative tasks: VAEs for their structured latent space and reconstruction potential, GANs for their ability to produce realistic images through adversarial training, and DDPMs for their stability and high-quality synthesis, making them well-suited to preserve anatomical details and diagnostic accuracy in medical imaging.

### 3.4.1 Variational Autoencoder

The VAE used to generate high b-value DWI images is based on the MONAI framework. The encoder part of the VAE has several convolutional layers with channel sizes progressing from 128 to 256 to 384, hierarchically. It includes residual blocks to improve gradient flow and stability of learning with a single block per level of resolution. Attention mechanisms are applied specifically in the deepest layer, with self-attention applied to focus on key spatial features and improve the model's capacity for learning important anatomical features. The decoder part of the VAE mirrors this structure in reverse, starting with a convolutional layer that upsamples from the latent space to 384 channels, followed by multiple residual blocks and attention blocks to refine features. It includes upsampling layers to increase resolution, reducing channel sizes from 384 to 256 and then to 128, with final convolutions to output a single-channel image, ensuring detailed reconstruction. Each convolutional layer utilizes LeakyReLU as the activation function. The model incorporates L1 loss to enable direct pixel comparison with a weight of 1, perceptual loss to promote quality with a weight of 0.001, and Kullback-Leibler divergence loss, that approximates the latent space to a normal distribution with mean of 0 and variance of 1, assigned a weight of $1 \times 10^{-6}$.

Conditioning is achieved by feeding the encoder the pre-processed input low b-value image and through a conditioning label that is incorporated into the encoder and into the decoder, providing information about the input b-value and the target b-value. The input b-value and target b-value are extracted from each individual pair and guide the model to generate the specific high b-value image starting from the given low b-value, ensuring the transformation preserves anatomical and diagnostic accuracy. Figure 3.3 displays the conditioned generation process of the VAE.

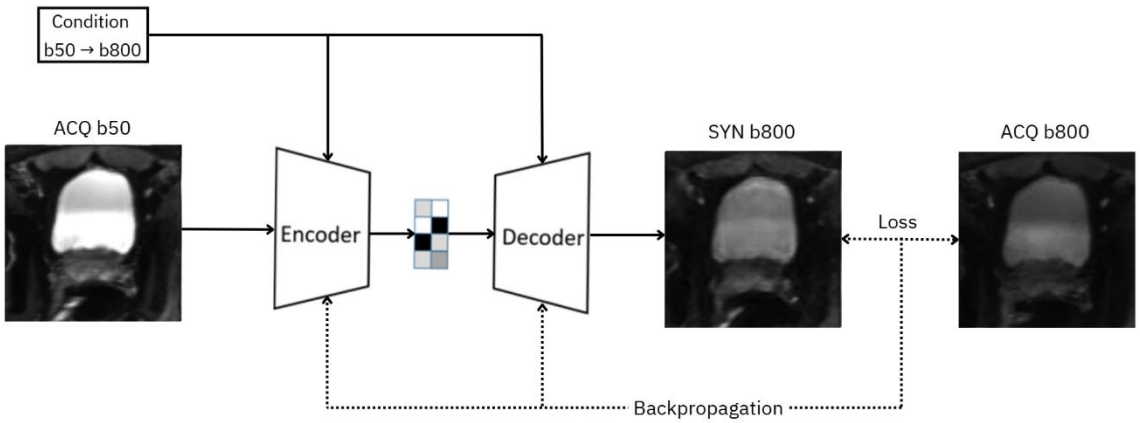


Figure 3.3: Schematic representation of the VAE process for generating high b-value images. The encoder receives a low b-value image (ACQ b50) and a conditioning label, which indicates the input b-value and target b-value (e.g., b50-b800). Subsequently the encoder encodes the input image and label into a latent space. This conditioning is also integrated into the decoder, guiding the reconstruction of the high b-value image. The decoder then generates an image (SYN b800) based on the latent space and this image is compared to its corresponding real high b-value image (ACQ b800), with the resulting loss being calculated and with subsequent backpropagation to refine both the encoder and decoder.

### 3.4.2 Variational Autoencoder - Generative Adversarial Network

The GAN used to synthesize high b-value DWI images is based on the MONAI framework. The model can be better described as a VAE-GAN (Variational Autoencoder Generative Adversarial Network) since it features a VAE that acts as a generator and a discriminator that helps guide the generation through an adversarial process. The encoder consists of two downsampling stages using convolutional layers with 256 and 512 channels, supported by two residual layers per stage to enhance learning stability. The decoder is symmetrical with two upsampling stages, reducing channels back to 512 and then to the output's single channel, using residual connections for refinement. The discriminator, that assesses whether the images belong to the acquired set or the synthetic set, is composed of three layers of 64 channels each, using 4x4 kernels. Each convolutional layer utilizes LeakyReLU as the activation function. The model uses L1 loss for pixel accuracy with a weight of 1, perceptual loss for quality assigned a weight of 0.001, and adversarial loss for realism with a weight of 0.01. Additionally, a quantization loss, inherent to the VAE component, is included with its default contribution of 1 to regulate the latent space discretization.

The conditioning process consists of feeding the encoder the pre-processed input low b-value image and through a conditioning label that is incorporated into the encoder, decoder, and discriminator, providing information about the input b-value and the target b-value throughout the generation and discrimination process. Once more, the input b-value and target b-value are extracted from each individual pair and help guide the model to generate the specific high b-value image starting from the given low b-value. Additionally, the conditioning label will help the discriminator distinguish between real and generated images based on the intended b-value transformation. Figure 3.4 offers a visual representation of this conditioning process.

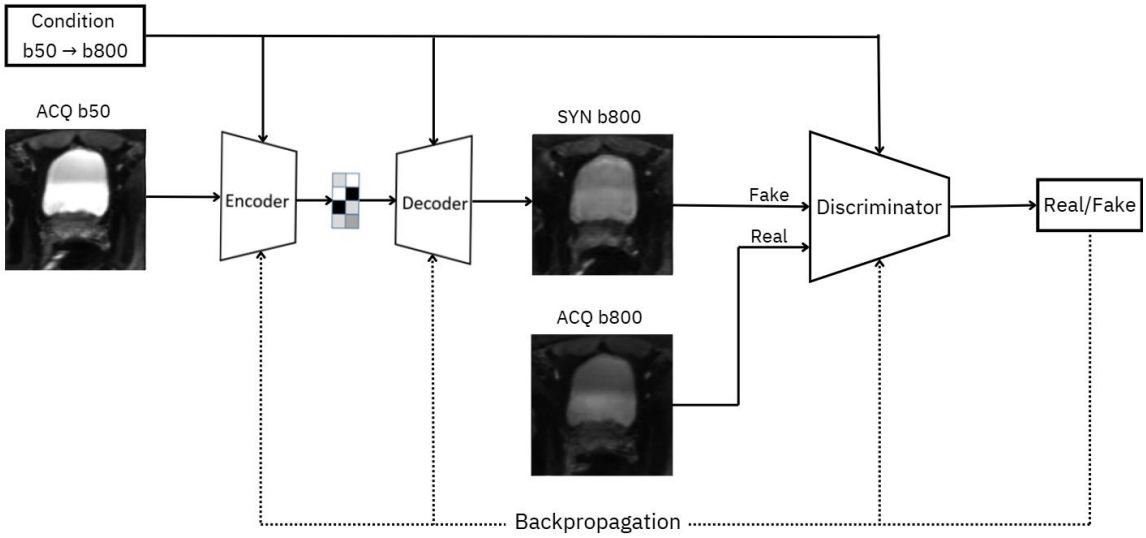


Figure 3.4: Schematic overview of the VAE-GAN generation process for high b-value images. The encoder receives a low b-value input image (ACQ b50) along with a conditioning label, which contains the input b-value and target b-value (e.g., b50-b800), and encodes these inputs into a latent space. This conditioning is also integrated into the decoder and discriminator. The decoder then generates an image (SYN b50) that alongside its corresponding real high b-value image (ACQ b800), is evaluated by the discriminator, which outputs a real/fake classification. The resulting feedback loop, guided by backpropagation, refines the encoder, decoder, and discriminator.

### 3.4.3 Denoising Diffusion Probabilistic Model

The Denoising Diffusion Probabilistic Model (DDPM) utilized for generating high b-value DWI images is constructed with the MONAI framework, utilizing a U-Net architecture as its base structure. The U-Net is comprised of an encoder path that progressively downsamples the input through convolutional layers, reducing spatial dimensions while increasing channel depths, followed by a decoder path that upsamples the feature maps to restore the original resolution. The architecture incorporates multiple resolution levels, with the number of channels set to start at 128 and doubling to 256 and then 512, where each level includes a single residual block in order to enhance gradient flow and stabilize the training. Attention mechanisms are strategically applied at the middle and deepest layers levels (corresponding to 256 and 512 channels) to focus on critical spatial dependencies, utilizing one attention head per level in order to improve the model's ability to capture fine anatomical details. Additionally, the model features multi-head attention with 256 channels to enhance the feature representation by allowing multiple attention heads to simultaneously capture diverse spatial dependencies and contextual relationships, such as anatomical details and diffusion patterns, thereby enriching the feature maps for more accurate denoising. The loss function is based on the mean squared error equation and is computed between the noise predicted and the actual noise added during the diffusion process. This loss drives the model to accurately predict the noise at each timestep, allowing for a gradual denoising from a noisy state to a clear high b-value DWI image. The training and sampling process is supported by a cosine scheduler wherein $\beta_1 = 10^{-4}$ and $\beta_T = 0.02$, which operates with 1000 timesteps in its Markov chain.

The conditioning of this DDPM is designed to guide the generation process using both image and b-value data. The training begins with the introduction of controlled random noise into the target high b-value image, a step facilitated by the scheduler, which adds Gaussian noise to the clean high b-value

images based on randomly selected timesteps from a total of 1000 timesteps, creating a noisy image that represents an intermediate state in the diffusion process. This noisy image is then concatenated along the channel dimension with the corresponding low b-value image, forming a combined input that helps the model by providing anatomical context from the low b-value image, which the model uses as a structural guide to refine the noisy data during the denoising process. The model predicts the noise added to this noisy image using its U-Net architecture, and the loss is computed, driving the training and allowing the model to learn the reverse diffusion process over the multiple epochs. Additionally, the input b-value and target b-value are embedded as a context vector. This context is integrated into the model's attention layers, where the cross-attention mechanism correlates the b-value information with spatial features across the image. This operation enables the model to adapt the denoising process to the specific b-value transformation, making sure the generated high b-value image accurately reflects the intended diffusion properties.

The sampling process, which occurs after training, reverses the diffusion process to generate new synthetic high b-value images. It starts with an input image of pure random noise, which the model iteratively denoises over the same 1000 timesteps. At each timestep, the trained model predicts the noise present in the current noisy sample (t) and uses the scheduler to compute the previous, less noisy state (t-1), gradually refining the image. The low b-value image is used as a conditional anchor, which is concatenated with the evolving noisy sample at each step, providing consistent anatomical context. Additionally, the embedded context vector, containing the input and target b-values, is fed into the attention layers. This iterative refinement continues until the model produces a final high b-value image that aligns with the target diffusion characteristics and low b-value image anatomy. This process is visually represented in Figure 3.5.

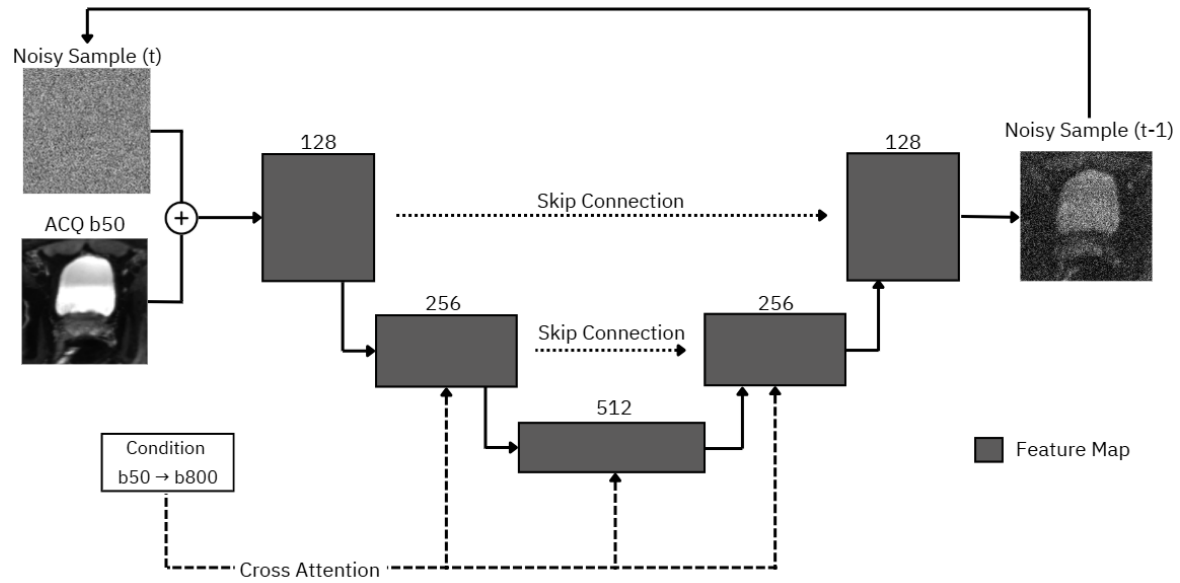


Figure 3.5: Schematic Illustration of the sampling process in the diffusion model for generating high b-value images. The process begins with pure random noise concatenated with a low b-value input image (ACQ b50) along the channel dimension, serving as the initial input to a U-Net architecture. The U-Net features a down-sampling path and an up-sampling path (512, 256, 128 channels), with skip connections linking the 128 and 256 layers between down-sampling and up-sampling stages. Cross-attention layers at the 256 and 512 channels integrate a conditioning context vector, embedding the input and target b-values to guide the denoising process. At each of the 1000 timesteps, the U-Net generates a high b-value image that is one step less noisy than the input, with the output fed back as the new input noise, concatenated again with the low b-value image. This iterative refinement, driven by the scheduler and anatomical context from the low b-value image, continues until a fully clear high b-value image is produced.

## 3.5. Training Models

The training process was conducted in two stages. Initially, all models, including the VAE, GAN, and DDPM were trained on the smaller PROSTATEx dataset, with the objective of probing for the most adequate model for the high b-value DWI generation task. After this initial training, the model that demonstrated the best performance was selected to be trained on the larger ProCAncer-I database. Each model was individually optimized for the specific characteristics of its respective dataset to ensure a satisfactory generalization capability and accuracy.

For the PROSTATEx dataset, which comprises 345 cases, 15% of the data, equivalent to 52 cases, was allocated for validation. This partitioning resulted in 16,701 low-to-high b-value pairs of DWI slices for training and 2,964 low-to-high b-value pairs of DWI slices for validation. For the ProCAncer-I database, which consists of 5,114 cases, 5% of the data, equivalent to 256 cases, was utilized for validation, resulting in a total of 450,991 low-to-high b-value pairs of DWI slices for training and 24,534 low-to-high b-value pairs for validation.

The training procedure of all models spanned 150 epochs, using the Adam optimizer with an initial learning rate of 0.0001. A cosine annealing learning rate scheduler was used and complemented by a 10-epoch warm-up period. In the warm-up period, the learning rate started from a value much smaller than the initial rate and increased gradually to the original 0.0001, in order to reduce early overfitting. Additionally, a batch size of 16 was used throughout the training.

## 3.6. Validation

The validation of the models was conducted to comprehensively assess their performance and the realism of the generated high b-value DWI images. The generated DWI were evaluated quantitatively with metrics and qualitatively with diagnostic performance analysis measured by a deep learning segmentation model.

### 3.6.1 Metrics

A diverse set of metrics was utilized to compare the generated high b-value DWI images with their real counterparts. In this procedure all the synthetic images were generated from the low b-value images of the validation partition of each database, ensuring that none of the training data contaminated the validation results. The metrics employed include MSD, PSNR, SSIM, FSIM and FID.

The MSD was selected as a validation metric since it provides a sensitive indicator of pixel-level discrepancies and helps detect gross errors, ensuring that the models accurately replicate the intensity patterns critical for DWI analysis, where subtle variations can change lesion visibility and thus impact diagnostic interpretation. The PSNR was also employed once it quantifies the relative quality of the generated images, with higher values indicating a better fidelity to the real images. This metric is particularly helpful in evaluating the signal preservation in high b-value DWI images, where maintaining the signal integrity is essential for distinguishing tissue characteristics, especially in regions with low signal-to-noise ratio, which are frequent in high b-value DWI. Additionally, a perceptual metric such as the SSIM was essential for evaluating discrepancies between the generated and real DWI, since it considers structural details like lesion boundaries, which are essential for diagnosis. This metric is complemented by the FSIM which focuses on low-level features like edges, vital for tissue delineation. Finally, FID was incorporated to assess the overall distribution similarity between generated and real high b-value DWI images, capturing higher-level statistical differences, including texture and structural

coherence. The FID utilized a network pre-trained on the Radimagenet database, a large database of medical images that includes several modalities such as CT and MRI.

To statistically validate the performance differences within the three models, a one-way Analysis of Variance (ANOVA) was computed for each metric with a $p < 0.05$, followed by a post-hoc analysis using Tukey's Honest Significant Difference (HSD) to determine which specific models exhibited statistically significant differences.

For the model trained on the ProCAncer-I dataset, besides the general evaluation metrics that were computed for all of the generated validation images, a more specific evaluation was done by computing all of the metrics for specific b-value transformation intervals and forming heatmaps for each of these metrics. This interval-specific analysis provides a nuanced evaluation of the model's generation accuracy based on input and target b-values.

### 3.6.2 Diagnostic Performance Analysis

Recognizing that metrics alone are insufficient to assess the diagnostic utility of the generated medical images, a diagnostic performance analysis was conducted, comparing the synthetic images to their real counterparts, in order to evaluate the generated images' ability to retain diagnostic information and ensure quality. This process was carried out only on the images generated by the diffusion model trained on the ProCAncer-I database, with the assistance a deep learning segmentation model.

#### 3.6.2.1. Segmentation Model Assessment

In order to evaluate the model’s ability to preserve lesion visibility and diagnostic capability in the generated high b-value DWI images a deep learning segmentation model was tasked with marking lesion regions in a real set of images and in their counterpart generated set of images. The objective was to evaluate if the lesion representation is the same in both sets, allowing for a direct comparison of diagnostic performance between the acquired and generated sets.

This validation process involved training a lesion segmentation model on the partition of the PROSTATEx dataset with radiologists’ segmentations, which was not used to train the generative model, hence acting as an independent external validation set. Utilizing an external validation dataset (sourced from a different database rather than a subset of the training data) helps evaluate the model's generalizability because it tests the network's performance on data with potentially different acquisition protocols, patient populations, or imaging characteristics, which more closely mimics real-world clinical scenarios. Additionally, using an external validation dataset reinforces the credibility of the results and offers a more unbiased assessment of the generative model’s ability to preserve the diagnostic.

The segmentation model used the nnU-Netv2 framework and was trained to identify prostate lesions using three modalities: T2-weighted images, high b-value DWI images (with the highest b-value in PROSTATEx being b800), and the corresponding ADC images, calculated from the low b-value (b50) and high b-value DWI pairs.

The training process used the standard nnU-Net protocol and spanned 1000 epochs with a 5-fold cross-validation. The segmentation model employed a 3D full-resolution U-Net architecture with seven stages, featuring channel sizes of 32, 64, 128, 256, 320, 320, and 320 across the stages. The network utilized LeakyReLU as the activation function and training incorporated a combination of Dice loss and cross-entropy loss to maximize segmentation accuracy. The model processed data with a batch size of 2, with a median image size of 21×384×383 voxels and spacing of 3.0×0.5×0.5 mm.

Following this training, the model was switched to inference mode, where it was first tested with real T2 images, real high b-value DWI, and real ADC images to establish a baseline for lesion identification. Subsequently, the model was evaluated using the same real T2 images paired with generated high b-value DWI images and corresponding ADC images, computed from the original real low b-value DWI and the generated high b-value DWI. The goal was to determine if the segmentation model identified lesions in a similar manner in both real and generated image sets, thereby indicating whether the generative model preserved lesion characteristics.

The results of the segmentation were measured using the Dice score, which is a metric that assesses the overlap between the predicted and ground truth lesion masks. These scores were visualized in box plots that compare real and generated cases. Additionally, a two-sample t-test was conducted with a significance level of $\alpha = 0.05$ to statistically evaluate whether significant differences existed between the lesion segmentation performance on real DWI versus generated DWI.

# 4. Results and Discussion

This chapter presents the results of the study and discusses its findings. It begins with the evaluation of the VAE, GAN and DDPM models trained on the PROSTATEx database, focusing on the visual evaluation and obtained metric scores of the generated images. Subsequently, it examines the performance of the DDPM trained on the ProCAncer-I database, assessed through visual evaluation, quantitative metrics and a lesion location performance evaluation carried out by a deep learning-based segmentation model.

## 4.1. Models Trained on PROSTATEx Database

The VAE, GAN and DDPM were trained on the PROSTATEx database. These models were specifically optimized for this database and the presented results reflect the best performance achieved by each model over the 150 epochs of training. The validation images were generated using the low b-value images from the validation partition as inputs and each generated high b-value DWI was compared against its acquired high b-value DWI counterpart.

### 4.1.1 Visual Assessment of Generated DWI

Figure 4.1 presents the synthetic high b-value (b=800 s/mm²) images generated by the VAE, GAN and DDPM, relative to the acquired low b-value (b=50 s/mm²) and high b-value (b=800 s/mm²) DWIs acquired in three randomly selected slices of prostate studies. The images, produced by the DDPM, demonstrate significantly better visual resemblance to ground truth acquired high b-value DWIs, with improved performance in terms of perceptual quality, clearer anatomical details and precise diffusion effects representation. Specifically, the DDPM-generated images present minimal anatomical distortion, as evident in the central region within the prostate, in which the gland's contours, internal structures, and signal intensity distribution significantly resemble ground truth acquired ones. The brighter regions within the prostate, corresponding to areas of restricted diffusion, are accurately depicted in the DDPM outputs and retain the observed ground truth high contrast and textural accuracy. The DDPM also accurately models suppression of the signal in surrounding tissue, producing cleaner boundaries and fewer artifacts, thus providing improved enhancement in lesions' visualization and overall image clarity.

In contrast, both the VAE and GAN models struggle to replicate these features successfully, producing noisier images with higher levels of undesirable signal. The VAE exhibits higher distortion and noise than the GAN, with blurry edges and less similarity to the real DWIs, resulting in a loss of textural details and overall deterioration in perceptual quality. Even though the GAN performs somewhat better than the VAE in preserving the overall shape of the prostate, it still shows significant artifacts, such as inconsistent signal intensity and incomplete modelling of diffusion effects, in which brighter regions within the prostate are muted or uneven. Neither the VAE nor the GAN accurately represents the signal suppression within normal tissues, producing images that are visually different from the original acquired high b-value DWIs (ACQ b800).

Furthermore, both the GAN and VAE exhibit visually noticeable checkerboard artifacts, appearing as grid-like patterns or distortions due to improper upsampling or deconvolution processes when generating images. These artifacts arise primarily from transposed convolutions (deconvolutions) used within the VAE or GAN generator or decoder structure, where there is an uneven overlap when the stride is not a divisor of the kernel, leading to some output positions receiving more influence from input

points than others, which causes a checkerboarding effect that compounds across layers. This issue is inherent to the deconvolution operation, appearing even in untrained networks or during feature visualization [110].

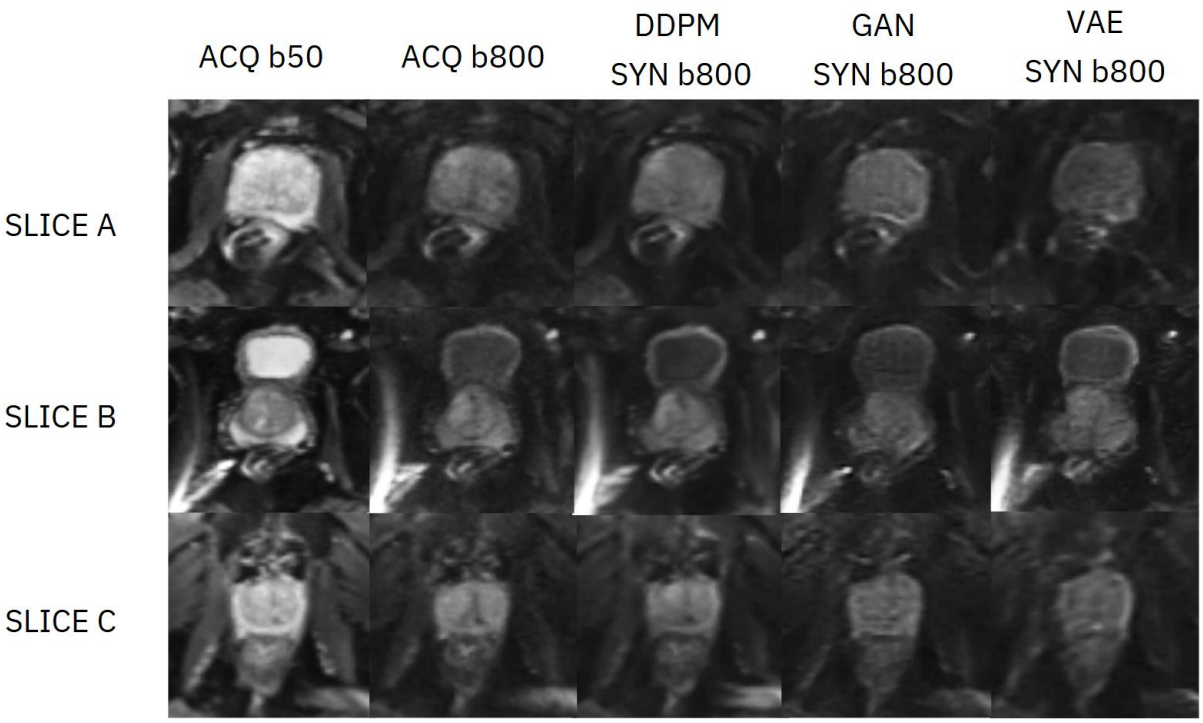


Figure 4.1: Comparative visualization of DWI for three randomly selected prostate slices (SLICE A, B and C) of the validation set of the PROSTATEx database. The sequence includes the acquired low b-value DWI (ACQ b50), the acquired high b-value DWI (ACQ b800), and the synthetic high b-value DWIs generated by the DDPM (DDPM SYN b800), GAN (GAN SYN b800), and VAE (VAE SYN b800).

### 4.1.2 Quantitative Evaluation of Generated DWI

The quantitative evaluation results, reported in Table 4.1, are aligned with the visual evaluations and highlight the better performance of the DDPM on all assessed metrics. The DDPM achieved the lowest MSD (0.0006) and FID (24.835) and achieved the highest PSNR (34.168), SSIM (0.902) and FSIM (0.920), the GAN ranked second on all metrics, while the VAE performed the worst with the highest MSD (0.0015) and FID (28.163) and lowest PSNR (29.053), SSIM (0.846) and FSIM (0.874). These results corroborate the visual assessment conclusions, where the DDPM consistently outperformed both the GAN and VAE.

Table 4.1: Performance metrics (MSD, PSNR, SSIM, FSIM and FID) comparing the generated high b-value DWIs from the VAE, GAN, and DDPM against the acquired high b-value DWIs from the validation set.

| | **MSD ↓** | **FID ↓** | **PSNR (dB)↑** | **SSIM ↑** | **FSIM ↑** |
|---|---|---|---|---|---|
| **VAE** | 0.0015 | 28.163 | 29.053 | 0.846 | 0.874 |
| **GAN** | 0.0011 | 26.671 | 31.251 | 0.861 | 0.898 |
| **DDPM** | 0.0006 | 24.835 | 34.164 | 0.902 | 0.920 |

Additionally, in Figure 4.2, violin plots are presented for the MSD, PSNR, SSIM, and FSIM performance that support this conclusion by demonstrating that the DDPM exhibits a larger proportion of its distribution concentrated in lower scores for MSD and higher scores for SSIM, FSIM, and PSNR, with reduced variability across all metrics, demonstrating its robustness and consistency in generating high-quality high b-value DWIs compared to the other models that present more instability in their generation. The GAN comes in second, showing a distribution with lower scores for MSD and higher scores for PSNR, SSIM and FSIM when compared to the VAE's distributions. The GAN's and VAE's variability is similar in almost all metrics, except for MSD (Figure 4.2 a)), where it presents significantly less variability.

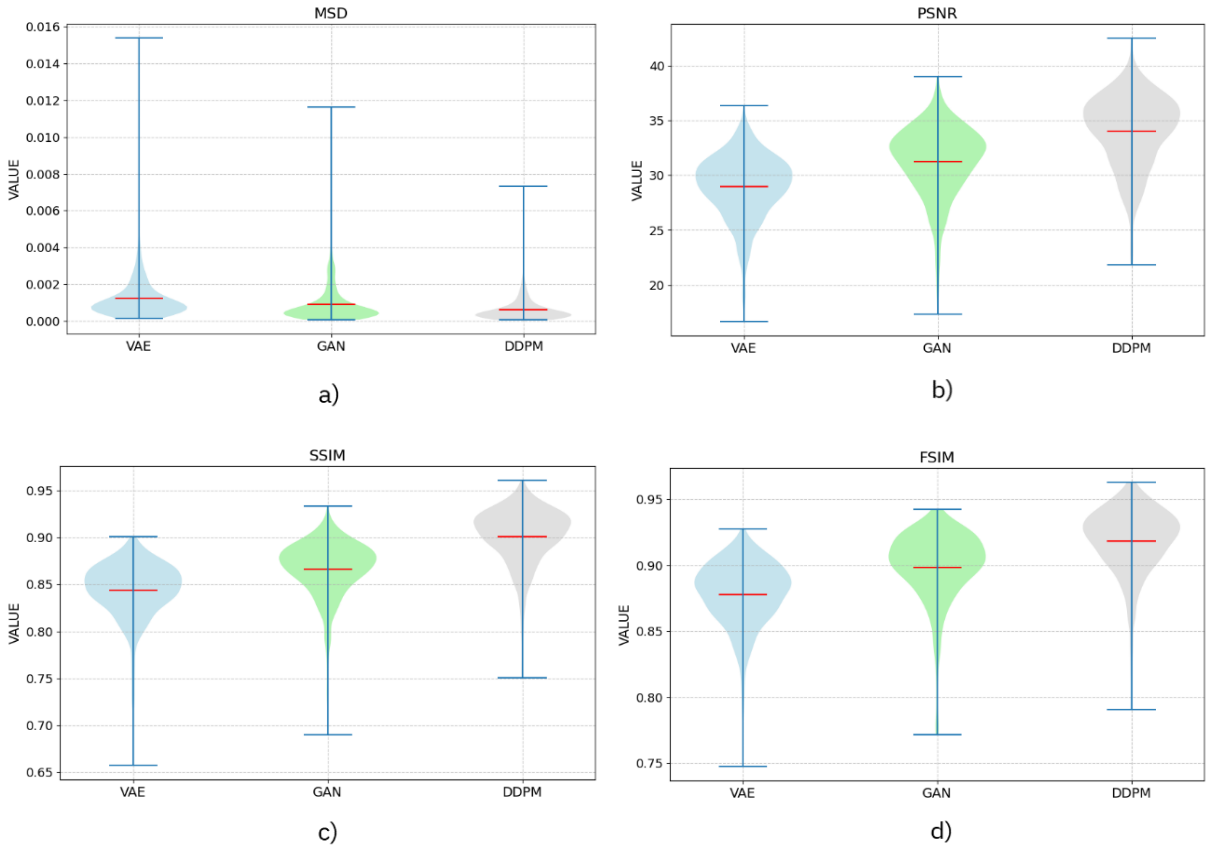


Figure 4.2: Violin plots illustrating the distribution of the performance metrics by slice: a) MSD, b) PSNR, c) SSIM and d) FSIM of the VAE, GAN, and DDPM generated DWI slices compared against the acquired DWI slices in the validation partition of the PROSTATEx database.

To statistically validate the performance differences identified in the violin plots (Figure 4.2) and in the metrics table (Table 4.1), a one-way ANOVA was computed for each metric across the three models using the individual metric scores from the set of validation images. The ANOVA presented significant differences in mean scores for all metrics ($p < 0.05$), meaning that the three model performances are not equivalent. Post-hoc analysis using Tukey's HSD concluded that DDPM consistently outperformed both the VAE and GAN for each measure with significantly lower MSD and FID and significantly higher PSNR, SSIM, and FSIM ($p < 0.05$), as can be observed in Table 4.2.

Table 4.2: Post-hoc Tukey HSD test results comparing the performance of VAE, GAN, and DDPM models across MSD, PSNR, SSIM, and FSIM metrics for generated slices, with a significance level of $\alpha = 0.05$. Mean differences (Mean Diff) and associated p-values are reported, where $p < 0.05$ indicates a statistically significant difference between models.

| Difference | MSD | | PSNR | | SSIM | | FSIM | |
|---|---|---|---|---|---|---|---|---|
| | **Mean Diff** | **P-value** | **Mean Diff** | **P-value** | **Mean Diff** | **P-value** | **Mean Diff** | **P-value** |
| **VAE -DDPM** | $-9\times10^{-4}$ | $5.26\times10^{-4}$ | 5.097 | 0.007 | 0.056 | 0.002 | 0.046 | 0.011 |
| **VAE -GAN** | $-4\times10^{-4}$ | 0.458 | 2.308 | 0.226 | 0.015 | 0.331 | 0.022 | 0.367 |
| **GAN -DDPM** | $-5\times10^{-4}$ | 0.023 | 2.789 | 0.019 | 0.041 | 0.037 | 0.024 | 0.043 |

The DDPM exhibited superior performance among the three models for generating high b-value DWIs from low b-value inputs. Both visual assessment and quantitative metrics support the fact that the DDPM generates images that are visually more similar to true acquired high b-value DWIs, accurately capturing the diffusion effects and tissue characteristics, particularly in the central prostate region. This superior performance is therefore likely attributed to the architectural strengths of the DDPM, such as the U-Net architecture, under which critical features and detailed prostate tissue information are extracted and preserved. The employment of skip connections within the U-Net also minimizes information loss across layers, preserving fine anatomical details throughout the generative process. Furthermore, the attention mechanism enables the model to selectively focus on some parts of the image and this improves the model's understanding of the global structure and emphasizes significant details like lesion boundaries or tissue contrasts. The gradual denoising process that is inherent to diffusion models also plays an important role, since it systematically removes the intrinsic noise within DWI images, resulting in clearer outputs.

On the other hand, the suboptimal performance of the VAE and GAN could be attributed to intrinsic limitations within their architectures. Because of the VAE's use of a variational inference method and its focus on approximating a latent distribution, the generated images appear overly smoothed, causing the loss of fine textural details and increased noise, as seen in the sampled high b-value DWIs in Figure 4.1. This smoothing effect likely contributes to the elevated MSD and reduced SSIM, FSIM, and PSNR ratings. Likewise, the GAN's adversarial training, while effective for generating plausible images, lacks the preservation of diffusion effects or lesion-specific features, often resulting in anatomical distortions and inconsistent signal suppression. Both of these models lack a designed denoising process and, therefore, are more prone to noise and unable to properly replicate the detailed diffusion characteristics observed within practical high b-value DWIs.

The DDPM's superior performance in generating high-fidelity medical images aligns with findings from recent literature. For instance, Mueller-Franzes *et al.* (2023) [111] conducted a multimodal comparison of DDPMs and GANs on medical datasets, which, consistently with this study, demonstrated that DDPMs outperform GANs in terms of stability and ability to capture complex distributions without suffering from training instability. It is also important to note that the analysis highlighted visual artifacts in GAN-generated images, such as artificial grid patterns, which we similarly observed in our GAN architecture, manifesting as checkerboard distortions due to deconvolution processes. Similarly, Wang *et al.* (2024) [112] in their systematic review of generative AI for medical images highlighted that diffusion models, just as verified in this study, achieve superior results in PSNR, SSIM, and FID compared to GANs when generating medical images, attributing this to their denoising process that approximates the true data distribution from initial noise, enabling better preservation of fine details and structural integrity, observations that closely parallel our findings, where the DDPM's

focus on gradual refinement lead to superior preservation of diffusion effects and anatomical structures in the synthesized high b-value DWIs.

## 4.2. DDPM Trained on ProCAncer-I Database

Having identified the DDPM as the most promising method of generating high b-value DWIs from low b-value inputs based on its superior performance on the PROSTATEx database, this model was now trained and tested on the ProCAncer-I database. This larger and more diverse dataset presents a wider variety of b-values, and thus a more challenging training process, where there is a need to condition the model with a more diverse set of pairs of input and target b-values.

### 4.2.1 DDPM Hyper-parameter Optimization

To fit the model to this diverse data, a targeted hyperparameter optimization strategy was implemented sequentially, focusing on key parameters deemed to be more influential: initial learning rate, batch size, channel architecture, diffusion timesteps, and noise timesteps scheduler. The configurations that were experimented with in this optimization routine are listed in Table 4.3 and tested in terms of metrics, presented in Table 4.4. These metrics were calculated on the validation set of the ProCAncer-I dataset and they were used as a base for enhancing the DDPM's performance and ensuring its effectiveness across an expanded range of imaging conditions.

Table 4.3: Hyperparameter configurations explored during the optimization of the DDPM for the ProCAncer-I database.

| Trial ID | Learning Rate | Batch Size | Channels Architecture | Diffusion Timesteps | Noise Scheduler |
|---|---|---|---|---|---|
| 1 | 0.0001 | 16 | [64, 128, 256] | 1000 | linear |
| 2 | 0.0001 | 16 | [128, 256, 512] | 1000 | linear |
| 3 | 0.0001 | 16 | [64, 128, 256, 512] | 1000 | linear |
| 4 | 0.0005 | 16 | [128, 256, 512] | 1000 | linear |
| 5 | 0.001 | 16 | [128, 256, 512] | 1000 | linear |
| 6 | 0.00005 | 16 | [128, 256, 512] | 1000 | linear |
| 7 | 0.0001 | 16 | [128, 256, 512] | 750 | linear |
| 8 | 0.0001 | 16 | [128, 256, 512] | 1250 | linear |
| 9 | 0.0001 | 16 | [128, 256, 512] | 1500 | linear |
| 10 | 0.0001 | 16 | [128, 256, 512] | 750 | cosine |
| 11 | 0.0001 | 16 | [128, 256, 512] | 1000 | cosine |
| 12 | 0.0001 | 16 | [128, 256, 512] | 1250 | cosine |
| 13 | 0.0001 | 16 | [128, 256, 512] | 1500 | cosine |
| 14 | 0.0001 | 16 | [64, 128, 256] | 1000 | cosine |
| 15 | 0.0001 | 16 | [64, 128, 256, 512] | 1000 | cosine |
| 16 | 0.0001 | 16 | [64, 128, 256, 512] | 1500 | linear |
| 17 | 0.0001 | 8 | [128, 256, 512] | 1000 | linear |

Trial 1 in Table 4.3 establishes the baseline configuration and is used as the standard network for comparison. The initial priority considered was to optimize the channels architecture, an important factor governing the capacity of the model to extract important features. Trial 2 with a configuration of

[128, 256, 512] channels performed best compared to both the less elaborate architecture of Trial 1 ([64, 128, 256]) and the complex structure of Trial 3 ([64, 128, 256, 512]). The reason Trial 1 performed worst stems from its lack of depth to effectively capture the diffusion patterns in the ProCAncer-I dataset. The excessive depth and complexity of Trial 3 may have led to overfitting.

Then the initial learning rate was tested in Trials 4 (0.0005), 5 (0.001) and 6 (0.00005), with Trial 2's rate yielding higher metric results. This implies an initial learning rate of 0.0001, allowing a stable gradient update and avoids overshooting optimal weights, while a lower learning rate, like 0.00005 may get stuck in a local minimum.

Subsequent experiments targeted the diffusion timesteps in Trials 7 (750), 8 (1250), and 9 (1500), with the 1000 timesteps from Trial 2 proving to be optimal. The 750 timesteps may have been insufficient to fully resolve the intricate diffusion processes, leading to incomplete feature modeling, while the higher timesteps (1250 and 1500) probably added unnecessary computational effort and led to overfitting to noise and a decline in performance metrics by overexposure to stochastic perturbations.

The noise scheduler was then experimented with, Trials 10-13 investigating a cosine scheduler on varied timesteps (750, 1000, 1250, 1500). Unexpectedly, these configurations underperformed in relation to the linear scheduler, possibly because the cosine scheduler's non-linear noise decay disrupted the model's ability to maintain consistent diffusion gradients, particularly on a heterogeneously distributed b-value dataset. Attempts to correct this in Trials 14 and 15 through the use of a cosine scheduler with other architectures ([64, 128, 256] and [64, 128, 256, 512] respectively) also resulted in inferior outcomes, which reinforces the linear scheduler's suitability in this database. Attempt 16 tested the strategy of using a more complex architecture paired with a higher number of diffusion steps, however, this configuration yielded inferior results to Trial 2.

The batch size was also altered, with Trial 17 (batch size of 8) underperforming when compared to Trial 2, likely because of less stable gradients when smaller batch sizes are utilized.

Table 4.4: Performance metrics for the hyperparameter configurations explored during the optimization of the DDPM for the ProCAncer-I database.

| Trial ID | MSD ↓ | FID ↓ | PSNR (dB) ↑ | SSIM ↑ | FSIM ↑ |
|---|---|---|---|---|---|
| 1 | 0.0038 | 43.658 | 26.211 | 0.731 | 0.816 |
| **2** | **0.0032** | **40.047** | **27.363** | **0.764** | **0.834** |
| 3 | 0.0034 | 40.524 | 27.214 | 0.753 | 0.827 |
| 4 | 0.0036 | 41.211 | 26.828 | 0.731 | 0.801 |
| 5 | 0.0039 | 45.604 | 24.715 | 0.718 | 0.805 |
| 6 | 0.0035 | 41.132 | 26.676 | 0.746 | 0.822 |
| 7 | 0.0037 | 43.542 | 26.433 | 0.723 | 0.814 |
| 8 | 0.0033 | 40.315 | 27.421 | 0.758 | 0.829 |
| 9 | 0.0034 | 40.757 | 27.288 | 0.751 | 0.820 |
| 10 | 0.0038 | 43.601 | 26.126 | 0.717 | 0.809 |
| 11 | 0.0034 | 40.139 | 27.344 | 0.755 | 0.825 |
| 12 | 0.0034 | 40.318 | 27.413 | 0.753 | 0.821 |
| 13 | 0.0035 | 41.031 | 26.996 | 0.746 | 0.811 |
| 14 | 0.0037 | 43.491 | 26.198 | 0.729 | 0.818 |
| 15 | 0.0034 | 40.613 | 27.266 | 0.756 | 0.824 |
| 16 | 0.0036 | 41.120 | 26.866 | 0.737 | 0.807 |
| 17 | 0.0033 | 40.102 | 27.235 | 0.760 | 0.831 |

Overall, Trial 2 proved to be the setup that exhibited the best performance in all metrics considered. This setup of [128, 256, 512] channel architecture, 0.0001 learning rate, 1000 diffusion steps, linear noise schedule, and batch size 16 emerged as the best setup to implement DDPM on the ProCAncer-I database.

### 4.2.2 Visual Assessment of Generated DWI

A detailed visual evaluation was performed on the validation images generated by the optimized DDPM trained on the ProCAncer-I database. This assessment used acquired low b-value DWI sets as input to synthesize corresponding high b-value DWI sets, which were then compared against their acquired high b-value DWI counterparts.

Figure 4.3 presents exemplary DWI images generated by the model (SYN) with different b-value transformations and provides a general assessment of the performance of the model. The DDPM captured effectively the overall prostate anatomy, including its borders and structural integrity, with no anatomic distortions evident across the tested transformations. The images also preserve the diffusion effects as evidenced in Slices 1, 2 and 4 of Figure 4.3, where the accurate representation of brighter regions indicative of restricted diffusion, such as those associated with prostate lesions, across all b-value ranges. This preservation implies that the model effectively captured the underlying tissue diffusion features upon which diagnostic accuracy is deeply dependent on.

Additionally, SYN images show an apparent reduction in noise levels in comparison to their respective high b-value acquired DWI (ACQ), resulting in sharper contours and enhanced visibility of anatomical details. This denoising feature enhances the interpretability of the images at the clinical level.

However, as is visible in slices 3 and 5 of Figure 4.3, it is clear that differences in background pixel intensity exist between the SYN and ACQ high b-value sets. This discrepancy may be attributed to the ProCAncer-I database's high variability in acquisition protocols, which might hinder the accuracy of the model in predicting uniform background intensity. The variety in b-values and acquisition conditions likely introduces complexity that the current optimization may not fully resolve, leading to localized differences in non-prostate areas.

Furthermore, images with larger input-target b-value differences, like the 0 s/mm² - 2000 s/mm² transformations (visible in Slice 5 of Figure 4.3), display slight distortions in diffusion effects, as would be expected low b-value images (e.g., 0 - 200 s/mm²) are more sensitive to perfusion effects, heavily influenced by the anatomy of vascular structures and the rapid movement of water in capillary vessels, which masks the subtle diffusion effects at the micro-tissue level critical for identifying tissue abnormalities. In contrast, high b-value images (e.g., 1800 s/mm² or 2000 s/mm²) are more attuned to the spatial scale of cellular structures, strongly correlating with cellularity and revealing detailed diffusion patterns, such as restricted water movement in tumors, which remain concealed in low b-value DWIs [113]. The DDPM, despite its consistent performance, may struggle to extrapolate these diffusion patterns over such wide ranges, resulting in minor artifacts. Despite this, the DDPM retains the general ability to capture diffusion trends, ensuring that key features like lesion contrast and tissue suppression remain discernible.

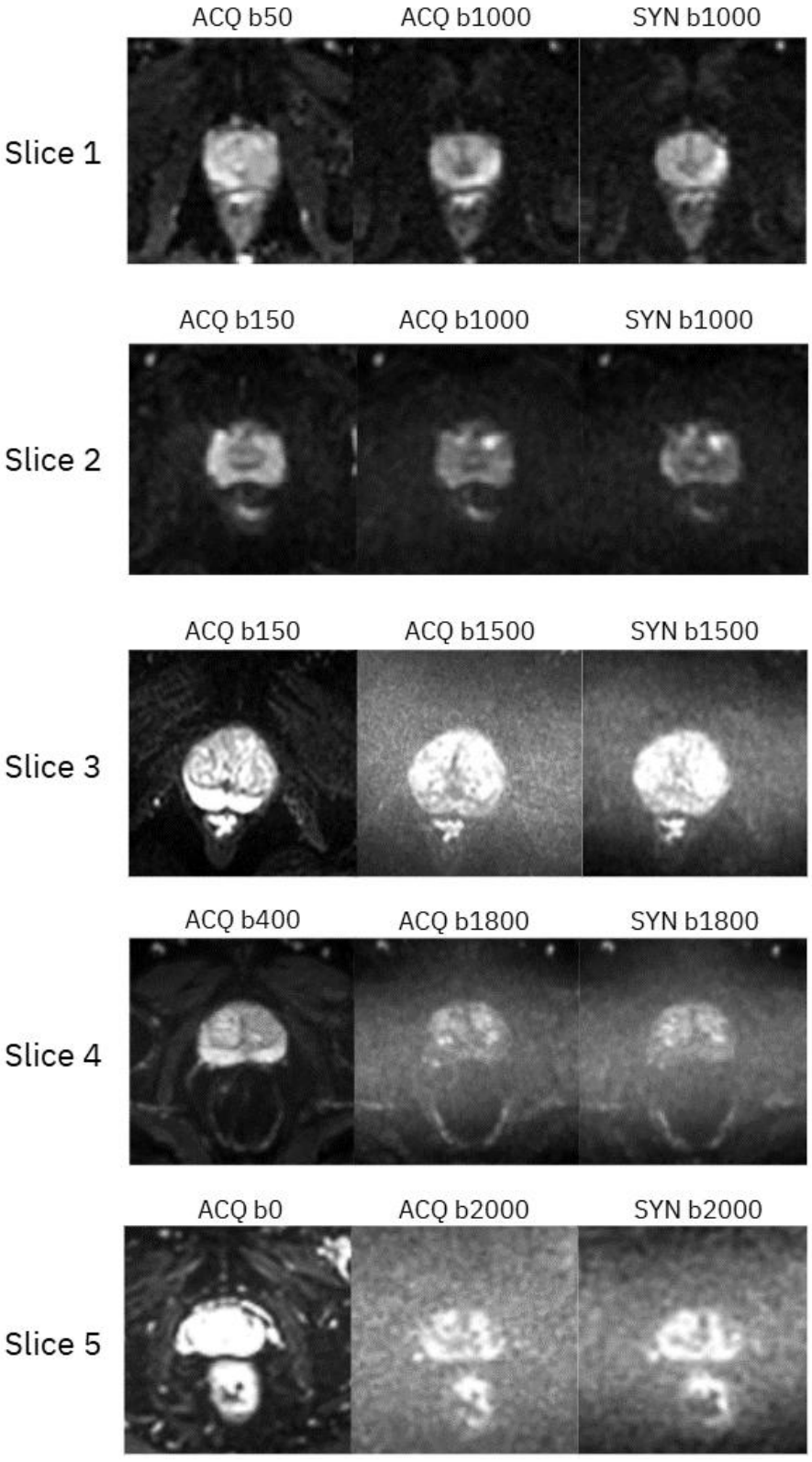


Figure 4.3: Comparative visualization of DWI across four randomly selected prostate study slices from the ProCAncer-I database. Each row displays the acquired low b-value DWI (input), the acquired high b-value DWI (target), and the synthetic high b-value DWI generated by the optimized DDPM (SYN), showcasing the model's performance across varying b-value transformations.

These findings are aligned with recent literature, particularly with Hung *et al.* (2023) [114], who introduced a conditional diffusion model for medical image generation that demonstrated that DDPMs

excel in generating high-fidelity images with realistic textures and intricate details. This aligns with the results of this study, where the DDPM accurately captured the complex diffusion patterns and fine tissue textures of the prostate across a wide range of b-values. Furthermore, Hung *et al*. highlighted the efficiency of DDPMs in denoising medical images, a feature mirrored in this study, where the generated high b-value DWIs exhibited reduced noise levels compared to acquired counterparts.

In addition to the previous visual assessment of the DDPM-synthesized validation images on different b-value pairs, a further evaluation was performed to investigate the ability of the model to generate a series of b-value transformations from a single acquired slice of DWI. As shown in Figure 4.4, this evaluation involves a single slice acquired from a particular study of the validation set of the ProCAncer-I database measured at b-values 0 s/mm², 50 s/mm², 200 s/mm², 400 s/mm², 1000 s/mm², and 2000 s/mm². From only the b-value 0 s/mm² DWI as an input, the DDPM effectively generated corresponding DWI images for 50 s/mm², 200 s/mm², 400 s/mm², 1000 s/mm², and 2000 s/mm², which were then compared to their acquired versions.

The synthesized images prove highly similar to the acquired DWI set, successfully maintaining the prostate's anatomical configuration and contours without inducing visible distortions. This fidelity is evident throughout all of the b-value transformations. The DDPM effectively retains the inherent tissue properties despite the broad b-value transformation spectrum. Once more, the generated images exhibit significantly reduced noise levels when compared to the acquired set, an effect more noticeable in higher b-values, such as the DWI with 1000 s/mm², and 2000 s/mm² in Figure 4.4. This improvement in image clarity is represented by sharper edges and can be explained by the iterative denoising process, which progressively refines the image during generation, mitigating the inherent noise typical in high b-value DWI acquisitions.

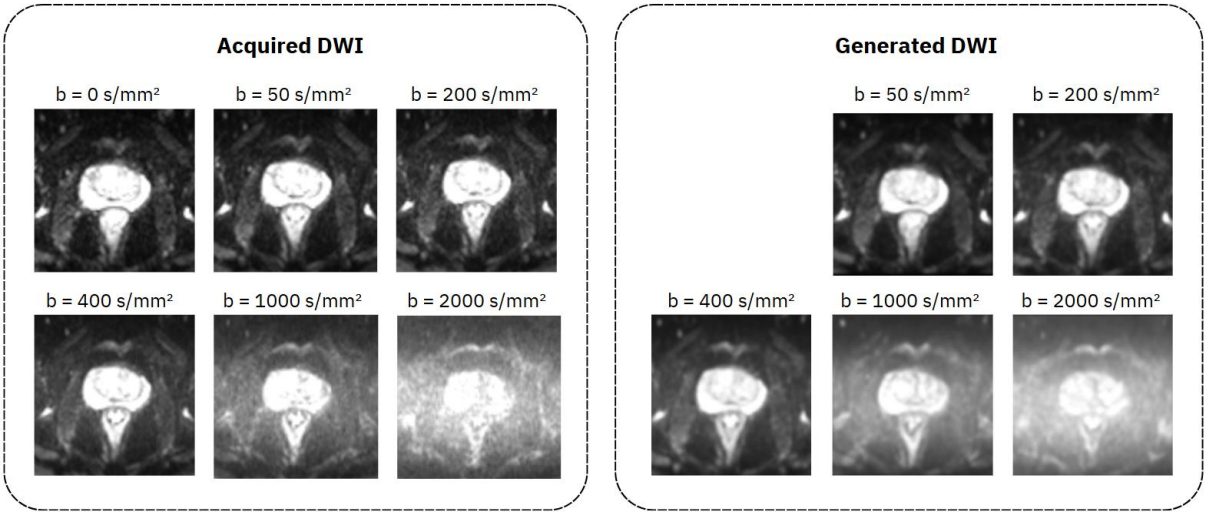


Figure 4.4: Comparative DWI for a single prostate study slice from the ProCAncer-I database validation set. The image presents six acquired DWIs with b-values of 0 s/mm², 50 s/mm², 200 s/mm², 400 s/mm², 1000 s/mm², and 2000 s/mm², alongside the corresponding synthetic DWIs generated by the DDPM trained on ProCAncer-I. The synthetic images (50 s/mm², 200 s/mm², 400 s/mm², 1000 s/mm², and 2000 s/mm²) were produced using only the acquired DWI at 0 s/mm² as input.

The versatility of this model is evident by its ability to handle both small and large b-value transformations (e.g., 0 s/mm² to 50 s/mm² compared to 0 s/mm² to 2000 s/mm²) with low information loss. For smaller steps, for instance, 0 s/mm² to 200 s/mm², high fidelity is preserved because of the proximity of diffusion characteristics, but bigger steps, such as 0 s/mm² to 2000 s/mm², still retain essential features despite higher extrapolation difficulty from subtle diffusion effects. This flexibility represents a valuable asset, as optimal b-value selection in DWI imaging is often dynamic and depends on various factors, including tissue type, lesion characteristics, and clinical objectives, such as balancing

signal-to-noise ratio with diffusion sensitivity. Having a tool that can generate any high b-value DWI at will from a single low b-value input enables clinicians to iteratively explore and visualize multiple b-values without extending acquisition time, facilitating a more adaptive diagnostic process where the most informative b-value for lesion identification can be determined post-acquisition.

This dynamic capability sets the proposed architecture apart from previous models mentioned in Section 2.5, such as the generative low-to-high b-value DWI GAN models described in Hu *et al.* (2021) [104], which were limited to generating a static high b-value of 1500 s/mm² from a fixed input of 800 s/mm², or the model in Hu *et al.* (2024) [106], which could only produce a b-value of 1200 s/mm² from an 800 s/mm² input. Unlike these fixed-transformation approaches, the current architecture supports generation in a dynamic manner, accommodating any available low b-value input to generate any desired optimal high b-value output, thereby offering greater clinical versatility. This capability not only accelerates MRI workflows by minimizing the need for prolonged scans but also enhances diagnostic precision in scenarios where identifying the ideal b-value is iterative, such as in real-time lesion characterization or personalized imaging protocols, contributing to a more efficient and accessible prostate cancer screening protocol.

### 4.2.3 B-value Interval based Assessment of Metrics

The range of b-value transformation, defined by the difference between the input b-value and target b-value, plays a significant role in determining the quality of synthetic DWIs generated using the DDPM. To better understand how the model performs in all of the b-value ranges, a comparative analysis was conducted by computing metrics based on the initial input b-value and target b-value of the generation. The validation set is balanced by intervals of b-value transformation, selected for their clinical relevance in diagnostic imaging (e.g., capturing diffusion effects critical for lesion detection), theoretical significance in understanding diffusion processes, and alignment with the b-values available in the training database, ensuring no interval is underrepresented, allowing for the creation of heatmaps for the metrics: MSD, PSNR, SSIM, FSIM and FID, represented in Figure 4.5. The heatmaps provide a visual explanation of the variation of the model's efficiency with the magnitude of b-value differences.

Overall, the heatmaps show that the metrics have better scores when the distance between input and target b-values is shorter. This trend is evidently observed in the MSD heatmap (Figure 4.5 a)), where the range with the smallest b-value difference (500-800 to 1000-1299 s/mm²) performs the best, with an MSD of 0.0022, reflecting minimal error. Performance deteriorates as the input b-value decreases or the target b-value increases, with the largest b-value range (0-199 to 1500-2000 s/mm²) recording the highest MSD score of 0.0044. This is also corroborated by the PSNR heatmap (Figure 4.5 b)), with better scores for lower target b-values (e.g., 1000-1299 s/mm²), as high noise levels associated with higher b-values (e.g., 1500-2000 s/mm²) tend to lower PSNR, consistent with the metric's sensitivity to noise.

This tendency is corroborated by the FID heatmap (Figure 4.5 e)), in which worse scores are presented in the higher target b-values, matching increased perceptual dissimilarity as the model fails to reproduce fine details of high b-value DWIs. The SSIM and FSIM heatmaps (Figure 4.5 c) and d) respectively), also reflect this trend with higher scores for lower intervals. The exception occurs in the 200-499 to 1500-2000 s/mm² interval (SSIM: 0.698, FSIM: 0.801), which performs worse than the 0-199 to 1500-2000 s/mm² interval (SSIM: 0.713, FSIM: 0.805), even with the greater b-value difference of the latter. This unexpected result can be attributed to an imbalance in the training data since the range 200-499 to 1500-2000 s/mm² is less prevalent than other b-values, with the model having fewer examples to learn the specific diffusion patterns of this interval, thus impairing its generalization.

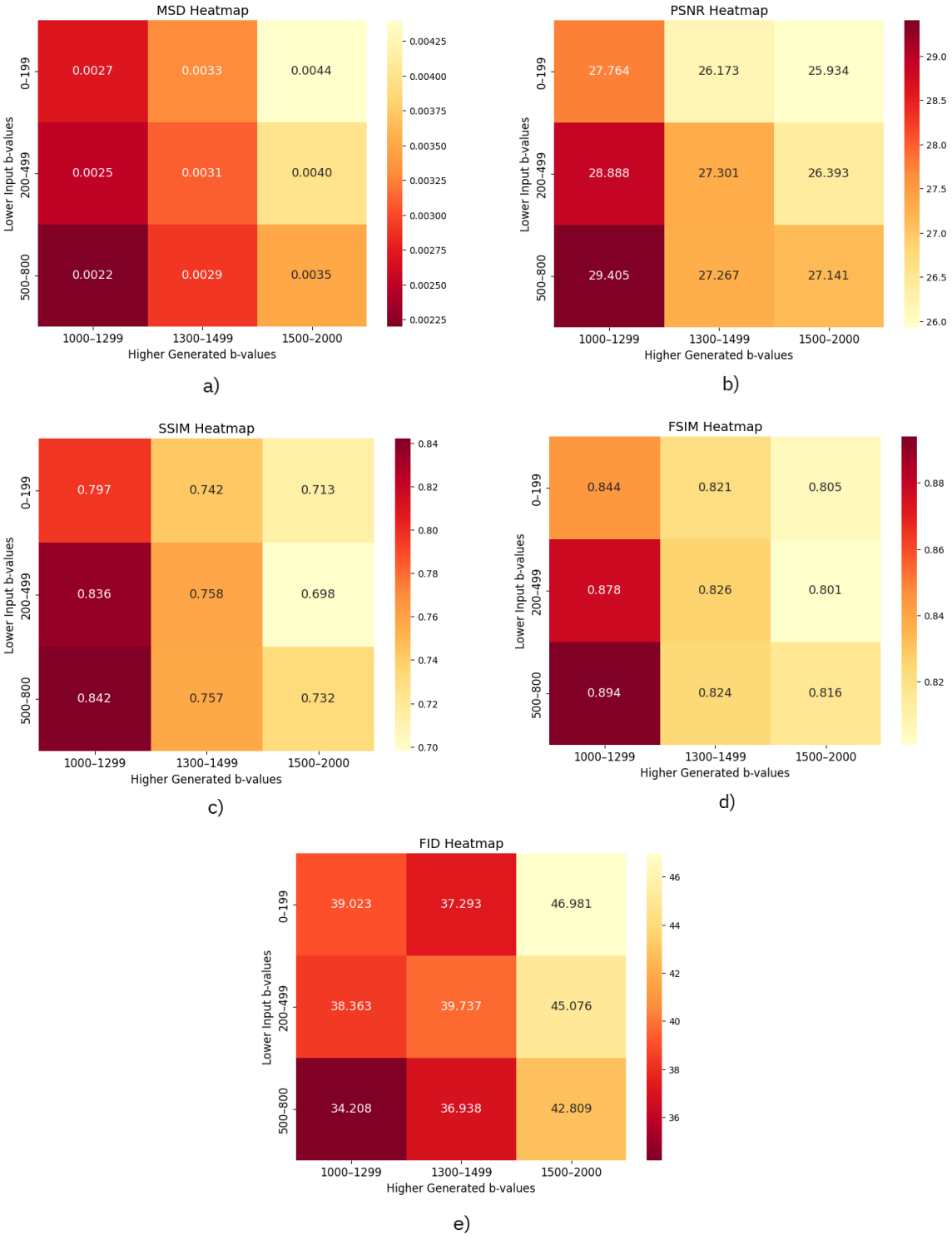


Figure 4.5: Heatmap of performance metrics for the DWI slices generated by the DDPM trained on the ProCAncer-I database, analyzed across varying b-value transformation intervals. The heatmaps include the metrics: (a) MSD, (b) PSNR, (c) SSIM, (d) FSIM, and (e) FID. The y-axis denotes intervals of input low b-value ranges, while the x-axis represents intervals of target high b-value ranges.

As expected, generated high b-value DWIs are more similar to their acquired high b-value DWI counterpart when the distance between input and target b-values is shorter. This effect is expected because smaller b-value differences mean a greater similarity between the input and target DWIs, simplifying the generation process and enhancing fidelity. Larger intervals, where the input and target images have large differences in diffusion properties, are more complex to model and lead to worse

scores across all metrics. This can be explained by the fact that DWIs at higher b-values are more sensitive to diffusion effects, outlining areas of restricted diffusion as bright regions. Low b-value inputs, by contrast, are less sensitive to these subtle patterns of diffusion, and do not capture the contrast that is present at higher b-values, which makes it more difficult for the model to accurately predict the target image in very-low to very-high b-value generations.

This is consistent with current literature on the different diffusion effects captured by DWI at variable b-values, as highlighted in Woo *et al.* (2018) [115], who carried out a meta-analysis comparing b-values for the detection of prostate cancer. This analysis underscores that DWI images acquired with high b-values suppress the signal intensity of normal or benign prostate tissue, giving enhanced visibility to the tumor by accentuating restricted diffusion, while images acquired with low b-values retain higher intensity due to perfusion, making them less capable of capturing subtle diffusion patterns and thus present very different visual appearance than high b-value DWI. This divergence corroborates this dissertation's findings, where the DDPM exhibits greater difficulty in generating high b-value DWIs from distant low b-value inputs (e.g., 0-199 to 1500-2000 s/mm²), resulting in inferior scores in metrics and minor distortions. The difference between these b-value ranges, as shown in the diffusion model's performance, reflects the difficulty of extrapolating complex diffusion effects from very low b-value DWI.

### 4.2.4 Lesion Detection Assessment

To assess the clinical utility of the high b-value DWI generated by the DDPM, an evaluation of diagnostic accuracy was conducted. Two deep learning lesion segmentation models based on the nnU-Netv2 architecture were established: the first (ACQ model) was trained on the PROSTATEx data, including T2-weighted images, acquired high b-value DWIs (b=800 s/mm²), and corresponding ADC maps (derived from b=50 and b=800 s/mm² pairs). The second model (SYN model) was trained with the same T2-weighted images but with the generated high b-value DWIs and synthesized ADC maps, computed by deriving from the acquired low b-value DWIs (b=50 s/mm²) and corresponding generated high b-value DWIs. This dual-model scheme allowed for a direct comparison of lesion identification capabilities and aimed to ascertain whether the generated images preserve the diagnostic utility and diffusion characteristics of acquired DWIs.

The diagnostic performance test began with a visual assessment of the lesion segmentations masks overlapping the T2-weighted images, presented in Figure 4.6. In each image, the radiologists' annotations (ground truth lesion masks) are depicted in yellow, the ACQ model's predicted lesion masks are in red, and the SYN model's predicted masks are in blue. For the cases a) and b) of Figure 4.6 the ACQ and SYN models both correctly localize lesions with high correlation to the ground truth locations, however the masks are not completely overlapping, indicating some boundary discrepancies. The partial overlap is however reassuring, suggesting that both models correctly capture lesion locations. Conversely, example c) in Figure 4.6 shows divergence, where both models detect portions of a lesion but fail to identify two distinct lesions present in the ground truth, showing cases where the two models differ in lesion interpretation. Example d) of Figure 4.6 also shows limitations with the ACQ model inaccurately marks a non-lesion area and the SYN model does not identify a lesion at all, underscoring challenges in lesion detection for the two methods. This visual assessment indicates that although the DDPM tends to maintain diffusion effects and lesion visibility similar to acquired images, there are instances where inaccuracies in modeling arise. Significantly, the ACQ model also exhibits limitations, which may imply that certain errors may stem from the segmentation model constraints rather than the generated images alone.

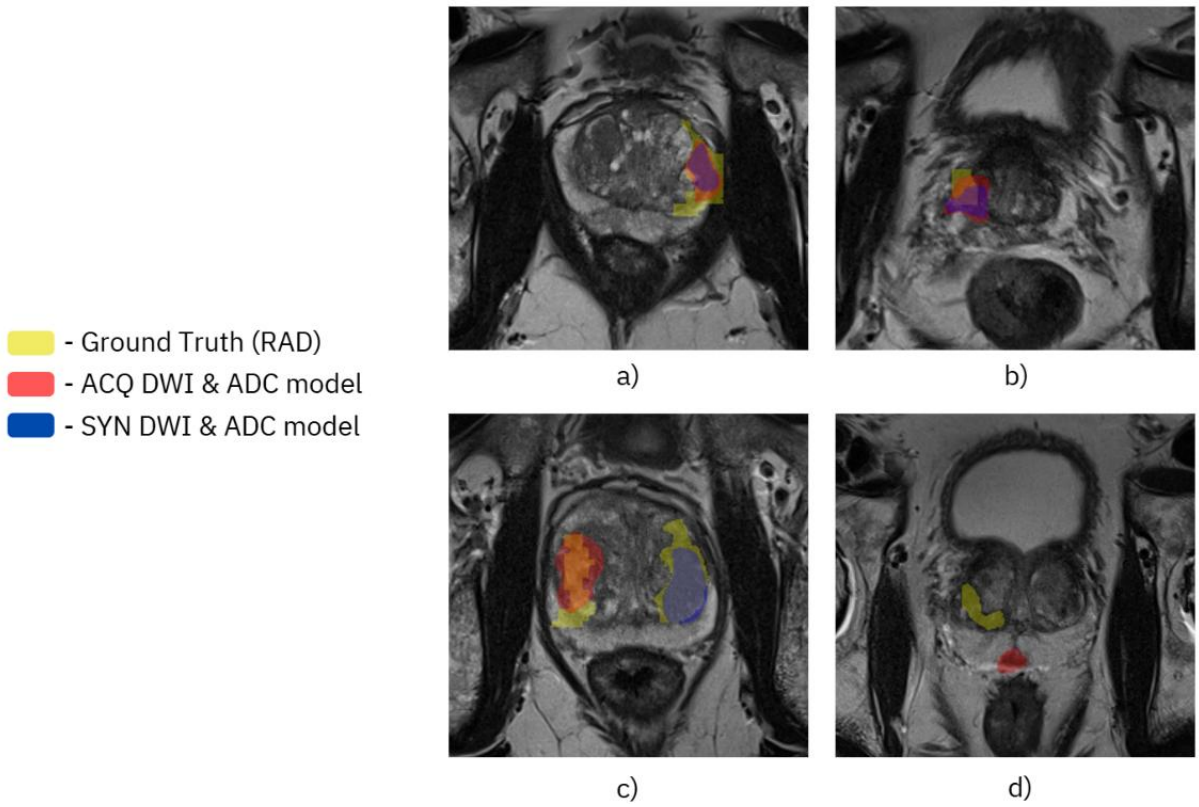


Figure 4.6: Visual assessment of lesion segmentation performance on four representative T2-weighted images (a, b, c, d) from the PROSTATEx database. The images display overlapping lesion masks: yellow for radiologist-annotated ground truth (RAD), red for predictions from the ACQ model trained on acquired high b-value DWIs (b=800 s/mm²) and corresponding ADC maps, and blue for predictions from the SYN model trained on DDPM-generated high b-value DWIs and synthesized ADC maps.

Quantitative validation is given in Table 4.5 and Figure 4.7, whereby the Dice similarity coefficient was used to quantitatively determine overlap between predicted segmentations and ground truth masks. Dice score of the ACQ model relative to radiologist-defined ground truth (RAD) is 0.248 and that of SYN model relative to RAD is a lower 0.193, indicating a moderate decline in precision of identifying lesions with generated images. Yet the Dice score of ACQ relative to SYN (0.402) indicates a high overlap in the number of lesions each detects, which is a measure of consistency in localization of lesions.

Table 4.5: Dice similarity coefficients comparing lesion segmentation performance across models, derived from the PROSTATEx validation set. Dice scores for the ACQ model versus radiologist-defined ground truth (RAD), the SYN model versus RAD, and the ACQ model versus SYN model.

| | **RAD vs ACQ** | **RAD vs SYN** | **ACQ vs SYN** |
|---|---|---|---|
| **Dice** | 0.248 | 0.193 | 0.402 |

Boxplots in Figure 4.7 further clarify this, such that RAD versus ACQ model distribution has a higher median Dice score and a wider interquartile range compared to RAD versus SYN to substantiate a slight decline in performance of SYN model. By contrast, ACQ versus SYN model displays a higher median Dice score, which substantiates overlap of their predicted results across validation set.

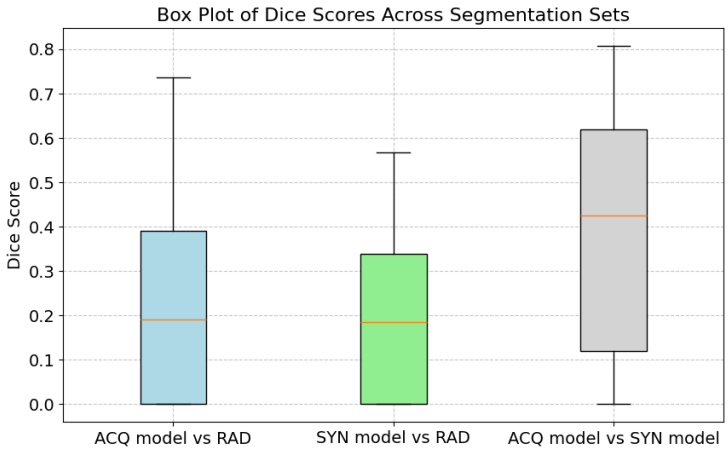


Figure 4.7: Boxplots of the distribution of Dice similarity coefficients for lesion segmentation performance on the PROSTATEx validation set. The plots compare the radiologist-defined ground truth (RAD) versus the ACQ model, RAD versus the SYN model, and the ACQ model versus the SYN model.

In addition, a two-sample t-test was used to provide a statistical comparison of Dice similarity measures of the SYN model and RAD and of the ACQ model and RAD. When conducted at a significance of $\alpha = 0.05$, it identified no significant difference between groups ($p > 0.05$) and consequently suggested that performance in the detection of lesions by the SYN and the ACQ model could not be considered statistically different.

Despite observing a mean Dice score degradation of 0.055 from the ACQ versus RAD to SYN versus RAD comparison, the overall diagnostic performance of the DWIs generated by the DDPM is largely preserved. This is especially remarkable considering that the DDPM was only trained on the ProCAncer-I database and yet generalized effectively to an external testing set of the PROSTATEx database in a different patient set and acquisition protocol. Preserving lesion visibility and diffusion information even with a slight performance drop reflects the DDPM's potential for clinical applicability and provides a promising pathway to reduce MRI acquisition time without losing detection integrity.

To further complement the lesion detection test, another analysis was carried out by creating a Dice similarity coefficient threshold of 0.1 to analyze the detection accuracy of lesions between the ACQ and SYN models. This value was selected as a practical indicator of lesion location accuracy, where a Dice score of more than 0.1 implies that a model was successful in localizing the approximate position of a lesion, even if the precise boundaries are not perfectly delineated, a measure relevant in clinical scenarios where lesion localization is prioritized rather than precise contouring. On the PROSTATEx validation set of 29 cases, the ACQ model identified 14 lesions (48.3% of available cases) with a Dice score above 0.1, while the SYN model detected 11 lesions (37.9% of available cases), with a decrease in detection accuracy of 10.4%.

This decrease in detection rate from 48.3% to 37.9% implies that the DDPM-generated high b-value DWIs, could fail to adequately reproduce subtle diffusion features and microstructural information important for lesion detection, making them diverge slightly from their acquired counterparts. In some cases, the diffusion effects of the generated high b-value DWI are incomplete and only the broader diffusion patterns are reproduced, and in a small number of cases the model misinterprets the diffusion effects and cannot render the intricate diffusion characteristics. The decline could also indicate limitations in the adaptability of the DDPM in generalizing across different datasets with diverse patient cohorts and imaging protocols.

In spite of the decline of 14%, the SYN model still detects more than two-thirds of the lesions detected by the ACQ model, demonstrating preservation of detection sensitivity. This preservation suggests that the DDPM effectively replicates the diffusion effects within prostate tissue across most

generated high b-value DWIs. It demonstrates an understanding of the tissue microstructure and diffusion dynamics, being able to infer regions of restricted diffusion, crucial for lesion identification, even when such effects are not apparent in the corresponding low b-value DWIs.

These findings are aligned with recent studies, mentioned previously in section 3.5, which explore the generation of high b-value DWIs using GANs, although with methodological differences from this study. For instance, Hu *et al.* (2021) [104] focused on prostate cancer using a GAN-based framework to generate 1500 s/mm² b-value DWIs from intermediate 800 s/mm² b-value images. Subsequently, the study relied on four radiologists who were blinded to patient details, to evaluate their confidence in detecting prostate cancer in the high b-value DWI. The analysis found that the generated images enabled cancer detection performance on par with the real high b-value images. The generated images even increased accuracy for less experienced radiologists, which suggests that the generation of high b-values can reduce noise and enhance lesion conspicuity to aid diagnosis.

Similarly, Hu *et al.* (2024) [106] focusing on gastric cancer imaging, developed an unsupervised GAN framework to generate high 1200 s/mm² b-value DWIs from intermediate 800 s/mm² b-value DWI. This study then relied on two experienced radiologists to conduct blinded assessments across criteria such as lesion visibility and location. The results showed that the generated images achieved detection ratings equivalent to the real high b-value images, which demonstrated that generative models are capable of capturing and recreating diffusion effects.

These studies corroborate the current results by affirming that generated high b-value DWIs can maintain diagnostic accuracy relative to acquired images, as evidenced by preserved lesion visibility and detection ratings. However, they differ in some aspects. The referenced works relied on radiologist evaluations, which provide a human-centric, bias-free assessment of clinical utility, avoiding the potential limitations of automated segmentation models like nnU-Netv2. In the present study, the ACQ model's baseline Dice score (mean 0.248 versus radiologist ground truth) already indicates inherent constraints in deep learning-based lesion detection, even on acquired DWIs, limitations that are amplified in the SYN model but not solely attributable to the generated images. By contrast, the radiologist assessments in the prior studies eliminate such model-specific biases, offering a purer measure of image fidelity. Additionally, the aforementioned studies generated fixed high b-values from specific intermediate b-values, limiting flexibility to predefined protocols. The DDPM, however, enables dynamic generation of arbitrary higher b-values conditioned on user-specified targets, using lower b-values DWI as inputs, allowing for broader applicability across varying clinical needs and protocols. This dynamic process, combined with training the DDPM on the ProCAncer-I dataset with images acquired from three scanners compared to the two used in Hu *et al.* (2021) [104] studies, and testing its generalization capability on the external PROSTATEx dataset, extends the diagnostic preservation seen in the referenced works to more heterogeneous, real-world scenarios. Furthermore, the DDPM's training across a wider variety of magnetic field strengths (1.5T and 3.0T versus only 3.0T in Hu *et al.* (2021) [104]), enhances its robustness and adaptability, positioning this framework as superior for diverse clinical applications.

Nonetheless, the DDPM's ability to yield SYN model performance with only a slight Dice degradation (0.055) and detection rate drop (10.4%) mirrors the high authenticity observed in those GAN-based syntheses, where quality ratings and detection confidence showed equivalence or superiority.

These correlations reinforce the clinical potential of diffusion-based models, by proving that generated high b-value DWIs can replicate diagnostic information, reducing scan times while retaining lesion detection. The slight performance drops in the current automated evaluation, when viewed alongside the radiologist-validated equivalence in prior GAN studies, suggest that DDPM-generated images could achieve even closer parity in human assessments, paving the way for integration into multiparametric MRI workflows for prostate cancer diagnosis.

# 5. Conclusions and Future Work

## 5.1. Conclusions

This dissertation has been designed to address the challenges in PCa screening by using generative models of AI to generate high b-value DWI from low b-value images and optimize MRI protocols for efficiency and accessibility. PCa remains the second most diagnosed cancer in men worldwide, underscoring the need for efficient early detection strategies that will reduce mortality while minimizing overdiagnosis and unnecessary intervention. MRI with DWI has emerged as the most efficient non-invasive screening tool, having higher sensitivity to patterns indicative of malignancy than any other screening tool. However, its widespread use has been constrained by its elevated cost, long acquisition time, and susceptibility to motion artifacts particularly for high b-value DWI that enhances tumoral contrast. This study sought to address this gap by using generative AI techniques to reduce DWI acquisition time.

Firstly, a comprehensive review of generative networks covering VAEs, GANs, and DDPMs was conducted to explore their uses in the generation of medical images. This was complemented by a detailed literature analysis to determine the state of the art in the generation of low-to-high b-value DWI with the latest advances in the simulation of realistic diffusion features while compensating for noise and variability inherent in MRI images. Afterwards, an adequate training and validation setup was established where the models were conditioned to utilize the low b-value DWI images as inputs while the target b-value was embedded to guide the generation of the corresponding high b-value equivalents.

Among the assessed architectures, the DDPM demonstrated the best performance for the generation of low-to-high b-value DWI, as evaluated using quantitative measures such as MSD, PSNR, SSIM, FSIM, FID, alongside visual evaluations that verified preserved anatomical and diffusion information. Conversely, the VAEs and GANs presented limitations in terms of handling the complex noise and diffusion patterns of high b-value DWI and generating artifact-free outputs, leading to the selection of DDPM as the optimal model. The DDPM model was then optimized to further boost its generative performance, showing strong metric performance and high visual similarity to the original images while retaining critical diffusion effects such as the restricted motion of the water molecules in tumoral regions. Notably, the difficulty of generation increased when the input and target b-value distances increased, resulting in worse metric performance accompanied by the presence of minor distortions in a few cases.

Validation of the optimized DDPM on an independent external dataset, including diagnostic performance assessment using a segmentation model, confirmed its potential and clinical applicability. For the majority of instances, the generated images preserved diagnostic efficacy, with lesion locations identified in the same regions as ground-truth high b-value sequences and diffusion effects consistently maintained across generation and ensured no significant loss of sensitivity or specificity in the detection of tumors.

In summary, the DDPM approach successfully generates high b-value DWI from low b-value acquisitions and provides a promising route towards accelerating MRI protocols by reducing scan time and thus related expenses.

However, this study is not without limitations. The ProCAncer-I database, which the DDPM was trained on, while large and multi-centered, presents an imbalance in the distribution of b-value DWI acquisitions, in that some of the b-value intervals are underrepresented. This can impact model learning toward more common ranges and impact generalization to underrepresented ranges. Additionally, a more severe limitation lies in the lesion segmentation models used to evaluate the diagnostic accuracy of generated high b-value DWI images, which showed suboptimal localization accuracy even for the

model trained on acquired images. This constraint arises due to inherent limitations in the nnU-Net v2 architecture and, more importantly, due to the scale of training data with a limited number of DWI scans annotated with lesion masks by radiologists, thus limiting the capability of the segmentation model and the resultant diagnostic preservation analysis. Initially, the intent was to complement this with a qualitative evaluation by radiologists reviewing the generated DWI for lesion detectability and image quality, however due to limitations on radiologists' availability, this validation was not possible.

## 5.2. Future Work

Despite the advancements achieved with the development of this dissertation, there are still some limitations and aspects that warrant further investigation in future work to expand on the conclusions drawn.

The precision of the DDPM is directly related to the training database, and although the ProCAncer-I database employed provided a robust and multi-centered foundation, its imbalance in b-value distributions limited the model's exposure to specific ranges. Future work can enhance b-value generation diversity and quality by adding or merging other databases with a broad set of b-values, possibly allowing generation beyond the ranges explored in this study, like ultra-high b-values (>2000 s/mm²) and enabling on-demand arbitrary b-value generation. This technique can also be extended to non-prostate DWI applications, including other cancer or tissue imaging reliant on high b-value sequences for diagnosis, e.g., breast or brain DWI, hence broadening its clinical use.

Due to time constraints, optimization of the DDPM within this work was restricted to baseline configurations; future research can investigate other architectures, i.e., with other attention mechanisms or hybrid diffusion-transformer models, and systematically tune hyperparameters like learning rates, diffusion noise schedules, or diffusion timesteps.

Moreover, the diagnostic performance analysis based on an automatic segmentation model with inherent inaccuracies should be complemented in future work by independent expert radiologist assessments. These assessments might qualitatively score generated DWI sets on key criteria like benign prostate tissue signal suppression, anatomic distortion, artifacts, global image quality, and lesion visibility and localization, resulting in a more clinically applicable assessment of diagnostic preservation.

# 6. References


[1] Department of Economic and Social Affairs, United Nations, “Sustainable Development.”, 2025.

[2] Culp, M. B., Soerjomataram, I., Efstathiou, J. A., Bray, F., & Jemal, A., “Recent Global Patterns in Prostate Cancer Incidence and Mortality Rates,” European urology, vol. 77, no. 1, pp. 38-52, 2020.

[3] Bergengren, O., Pekala, K. R., Matsoukas, K., Fainberg, J., Mungovan, S. F., Bratt, O., Bray, F., Brawley, O., Luckenbaugh, A. N., Mucci, L., Morgan, T. M., & Carlsson, S. V., “2022 Update on Prostate Cancer Epidemiology and Risk Factors-A Systematic Review,” European urology, vol. 84, no. 2, pp. 191-206, 2023.

[4] Kimura, T., Sato, S., Takahashi, H., & Egawa, S., “Global trends of latent prostate cancer in autopsy studies,” Cancers, vol. 13, no. 2, p. 359, 2021.

[5] European Commission, “Prostate cancer burden in EU-27,” 2021.

[6] Fazekas, T., Shim, S. R., Basile, G., Baboudjian, M., Kói, T., Przydacz, M., Abufaraj, M., Ploussard, G., Kasivisvanathan, V., Rivas, J. G., Gandaglia, G., Szarvas, T., Schoots, I. G., van den Bergh, R. C. N., Leapman, M. S., Nyirády, P., & Shariat, S. F., “Magnetic Resonance Imaging in Prostate Cancer Screening: A Systematic Review and Meta-Analysis,” JAMA oncology, vol. 10, no. 6, pp. 745-754, 2024.

[7] JAMA Netw Open, vol. 1, no. 2, p. e180219, 2018., Boesen, L., Nørgaard, N., Løgager, V., et al., “Assessment of the Diagnostic Accuracy of Biparametric Magnetic Resonance Imaging for Prostate Cancer in Biopsy-Naive Men: The Biparametric MRI for Detection of Prostate Cancer (BIDOC) Study".

[8] EU4Health work programme, “Commission Initiative on Prostate Cancer,” Brussels, 2025.

[9] R. R. Edelman, “The history of MR imaging as seen through the pages of Radiology,” Radiology, vol. 273, no. 2S, pp. 181-200, 2014.

[10] R. A. Pooley, “AAPM/RSNA physics tutorial for residents: Fundamental principles of MR imaging,” RadioGraphics, vol. 25, no. 4, pp. 1087-1099, 2005.

[11] M. K. Patralekh, S. Singh, S. K. Sharma, and S. Khandelwal, “Basics of magnetic resonance imaging, 2016.

[12] G. Katt, S. A. Ara, A. Shireen, “Magnetic resonance imaging (MRI) - A review,” International Journal of Dental Clinics, vol. 3, no. 1, pp. 65-70, 2011.

[13] M. H. Levitt, “Spin dynamics: Basics of nuclear magnetic resonance,” John Wiley & Sons, vol. 90, no. 2, pp. 20-25, 2008.

[14] R. Bitar, G. Leung, R. Perng, S. Tadros, A. R. Moody, J. Sarrazin, T. P. Roberts, “MR pulse sequences: What every radiologist wants to know but is afraid to ask,” RadioGraphics, vol. 26, no. 2, pp. 513-537, 2006.

[15] J. P. Ridgway, “Cardiovascular magnetic resonance physics for clinicians: Part I,” Journal of Cardiovascular Magnetic Resonance, vol. 12, no. 71, pp. 10-17, 2010..

[16] G. B. Chavhan, P. S. Babyn, B. G. Jankharia, H.-L. M. Cheng, M. M. Shroff, “Steady-state MR imaging sequences: Physics, classification, and clinical applications,” RadioGraphics, vol. 28, no. 4, pp. 1147-1160, 2008..

[17] D. B. Twieg, “The k-trajectory formulation of the NMR imaging process with applications in analysis and synthesis of imaging methods,” Medical Physics, vol. 10, no. 5, pp. 610-621, 1983..

[18] D.B. Plewes, W. Kucharczyk, “Physics of MRI: a primer,” Journal of Magnetic Resonance Imaging, vol. 35, no. 5, pp. 1038-1054, 2012..

[19] B. A. Jung, M. Weigel, “Spin echo magnetic resonance imaging,” Journal of Magnetic Resonance Imaging, vol. 37, no. 4, pp. 805-817, 2013..

[20] D. Le Bihan, “Diffusion MRI: What it can tell us about tissue microstructure,” NMR in Biomedicine, vol. 268, no. 2, pp. 1103, 2013..

[21] R. Bammer, “Basic principles of diffusion-weighted imaging,” European Journal of Radiology, vol. 45, no. 3, pp. 169-184, 2003..

[22] R. Luypaert, S. Boujraf, S. Sourbron, M. Osteaux, “Diffusion and perfusion MRI: basic physics,” European Journal of Radiology, vol. 38, no. 1, pp. 19-27, 2001..

[23] S.J. Holdsworth, R. Bammer, “Magnetic resonance imaging techniques: fMRI, DWI, and PWI,” Seminars in Neurology, vol. 28, no. 4, pp. 395-406, 2008..

[24] D. M. Koh, D. J. Collins, “Diffusion-weighted MRI in the body: Applications and challenges in oncology,” American Journal of Roentgenology, vol. 188, no. 6, pp. 1622-1635, 2007..

[25] G. W. Albers, “Diffusion-weighted MRI for evaluation of acute stroke,” Neurology, vol. 51, no. 3, pp. 47-49, 1998..

[26] S. Warach, J. Gaa, B. Siewert, P. Wielopolski, R. R. Edelman, “Acute human stroke studied by whole brain echo planar diffusion-weighted magnetic resonance imaging,” Annals of Neurology, vol. 37, no. 2, pp. 231-241, 1995..

[27] A. Buisson, A. Joubert, P. F. Montoriol, D. D. Ines, C. Hordonneau, B. Pereira, J. M. Garcier, G. Bommelaer, V. Petitcolin, “Diffusion-weighted magnetic resonance imaging for detecting and assessing ileal inflammation in Crohn’s disease,” Alimentary, Pharmacology & Therapeutics, vol. 37, no. 5, pp. 537-545, 2013..

[28] L. Tang, X. J. Zhou, “Diffusion MRI of cancer: From low to high b-values,” Journal of Magnetic Resonance Imaging, vol. 49, no. 1, pp. 23-40, 2019..

[29] F.P.A. Pereira, G. Martins, E. Figueiredo, M.N.A. Domingues, R.C. Domingues, L.M.B. da Fonseca, E.L. Gasparetto, “Assessment of breast lesions with diffusion-weighted MRI: comparing the use of different b values,” American Journal of Roentgenology, vol. 1.

[30] K. B. Grant, H. K. Agarwal, J. H. Shih, M. Bernardo, Y. Pang, D. Daar, M. J. Merino, B. J. Wood, P. A. Pinto, P. L. Choyke, B. Turkbey, "Comparison of calculated and acquired high b value diffusion-weighted imaging in prostate cancer," Abdominal Imaging,, vol. 40, no. 3, pp. 578-586, 2015.

[31] A. B. Rosenkrantz, L. Mannelli, X. Kong, B. E. Niver, D. S. Berkman, J. S. Babb, J. Melamed, S. S. Taneja, "Prostate cancer: utility of fusion of T2-weighted and high b-value diffusion-weighted images for peripheral zone tumor detection and localization,", Journal of Magnetic Resonance Imaging: JMRI, vol. 34, no. 1, pp. 95-100, 2011..

[32] D. Le Bihan, C. Poupon, A. Amadon, F. Lethimonnier, "Artifacts and pitfalls in diffusion MRI," Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine, vol. 24, no. 3, pp. 478-488, 2006..

[33] E.B. Hutchinson, A.V. Avram, M.O. Irfanoglu, C.G. Koay, A.S. Barnett, M.E. Komlosh, E. Özarslan, S.C. Schwerin, S.L. Juliano, C. Pierpaoli, "Analysis of the effects of noise, DWI sampling, and value of assumed parameters in diffusion MRI models," Magnetic, Resonance in Medicine, vol. 78, no. 5, pp. 1767-1780, 2017..

[34] B. Zelhof, M. Pickles, G. Liney, P. Gibbs, G. Rodrigues, S. Kraus, L. Turnbull, "Correlation of diffusion-weighted magnetic resonance data with cellularity in prostate cancer," BJU International, vol. 103, no. 7, pp. 883-888, 2009..

[35] R. L. Gullo, S. C. Partridge, H. J. Shin, B. Sunitha, S. B. Thakur, K. Pinker, "Update on DWI for breast cancer diagnosis and treatment monitoring," American Journal of Roentgenology, vol. 222, no. 1, pp. e2329933, 2024..

[36] N.S. White, C.R. McDonald, N. Farid, J. Kuperman, D. Karow, N.M. Schenker-Ahmed, H. Bartsch, R. Rakow-Penner, D. Holland, A. Shabaik, A. Bjørnerud, "Diffusion-weighted imaging in cancer: physical foundations and applications of restriction spectrum, imaging," Cancer Research, vol. 74, no. 17, pp. 4638-4652, 2014..

[37] J.M. Baehring, W.L. Bi, S. Bannykh, J.M. Piepmeier, R. K. Fulbright, "Diffusion MRI in the early diagnosis of malignant glioma," Journal of Neuro-Oncology, vol. 82, pp. 221-225, 2007..

[38] C.H. Tan, W. Wei, V. Johnson, V. Kundra, "Diffusion-weighted MRI in the detection of prostate cancer: meta-analysis," American Journal of Roentgenology, vol. 199, no. 4, pp. 822-829, 2012..

[39] H. J. Meyer, M. Martin, T. Denecke, "DWI of the Breast-Possibilities and Limitations," RöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren, vol. 194, no. 9, pp. 966-974, 2022..

[40] S. Altmayer, L.M. Armelin, J.S. Pereira, L.V. Carvalho, J. Tse, P. Balthazar, M.Z. Francisco, G. Watte, B. Hochhegger, "MRI with DWI improves detection of liver metastasis and selection of surgical candidates with pancreatic cancer: a systematic review, and meta-analysis," European Radiology, vol. 34, no. 1, pp. 106-114, 2024..

[41] E.R. Gerstner, A.G. Sorensen, "Diffusion and diffusion tensor imaging in brain cancer," Seminars in Radiation Oncology, vol. 21, no. 2, pp. 141-146, 2011..

[42] E. Hassanzadeh, D.I. Glazer, R.M. Dunne, F.M. Fennessy, M.G. Harisinghani, C.M. Tempany, “Prostate imaging reporting and data system version 2 (PI-RADS v2): a pictorial review,” Abdominal Radiology, vol. 42, no. 1, pp. 278-289, 2017..

[43] X. Yuan, L. Guo, W. Du, F. Mo, M. Liu, “Diagnostic accuracy of DWI in patients with ovarian cancer: a meta-analysis,” Medicine, vol. 96, no. 19, pp. e6659, 2017..

[44] A.B. Rosenkrantz, L. Mannelli, X. Kong, B.E. Niver, D.S. Berkman, J.S. Babb, S.S. Taneja, “Prostate cancer: Utility of fusion of T2-weighted and high b-value diffusion-weighted images for peripheral zone tumor detection and localization,” Journal of, Magnetic Resonance Imaging, vol. 34, no. 1, pp. 95-100, 2011..

[45] W. Isaacs, A. De Marzo, W.G. Nelson, “Focus on prostate cancer,” Cancer Cell, vol. 2, no. 2, pp. 113-116, 2002..

[46] T. J. Syer, K. C. Godley, D. Cameron, P. N. Malcolm, “The diagnostic accuracy of high b-value diffusion- and T2-weighted imaging for the detection of prostate cancer: a meta-analysis.” Abdominal Radiology, vol. 43, no. 7, pp. 1787-1797, 2018.

[47] M. Scialpi, E. Martorana, P Scialpi, A. D’Andrea, F. M. Mancioli, M. Mignogna, A. Di Blasi, F. Trippa, “MRI apparent diffusion coefficient (ADC): A biomarker for prostate cancer after radiation therapy.” Turkish Journal of Urology, vol 47, no. 6, pp. 448,, 2021..

[48] S. Woo, C.H. Suh, S.Y. Kim, J.Y. Cho, S.H. Kim, “Head-to-head comparison between high-and standard-b-value DWI for detecting prostate cancer: a systematic review and meta-analysis,” American Journal of Roentgenology, vol. 210, no. 1, pp. 91-100, 2018..

[49] T. Ullrich, L. Schimmöller, “Perspective: a critical assessment of PI-RADS 2.1,” Abdominal Radiology, vol. 45, no. 12, pp. 3961-3968, 2020..

[50] L. Hu, D.W. Zhou, Y.F. Zha, L. Li, H. He, W.H. Xu, L. Qian, Y.K. Zhang, C.X. Fu, H. Hu, J.G. Zhao, “Synthesizing high-b-value diffusion-weighted imaging of the prostate using generative adversarial networks,” Radiology: Artificial Intelligence, vol. 3, no, . 5, pp. e200237, 2021..

[51] T.M. Mitchell, “Machine learning: A guide to current research,” Machine Learning, vol. 1, no. 1, pp. 65-95, 1980..

[52] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A.W.M. van der Laak, B. van Ginneken, C.I. Sánchez, “A survey on deep learning in medical image analysis,” Medical Image Analysis, vol. 42, no. 1, pp. 60-88, 2017..

[53] D.E. Rumelhart, G.E. Hinton, R.J. Williams, “Learning representations by back-propagating errors,” Nature, vol. 323, no. 6088, pp. 533-536, 1986..

[54] V. Nair, G.E. Hinton, “Rectified linear units improve restricted Boltzmann machines,” Proceedings of the International Conference on Machine Learning, vol. 27, no. 1, pp. 807-814, 2010..

[55] Y. LeCun, Y. Bengio, G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436-444, 2015..

[56] X. Glorot, A. Bordes, Y. Bengio, “Deep sparse rectifier neural networks,” Proceedings of the International Conference on Artificial Intelligence and Statistics, vol. 15, no. 1, pp. 315-323, 2011..

[57] K. He, X. Zhang, S. Ren, J. Sun, “Deep residual learning for image recognition,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, no. 1, pp. 770-778, 2016..

[58] D.E. Rumelhart, B. Widrow, M.A. Lehr, “The basic ideas in neural networks,” Communications of the ACM, vol. 37, no. 3, pp. 87-93, 1994..

[59] L. Bottou, “Large-scale machine learning with stochastic gradient descent,” Proceedings of COMPSTAT, vol. 19, no. 1, pp. 177-186, 2010..

[60] C.P. Igiri, O.U. Anyama, A.I. Silas, “Effect of learning rate on artificial neural network in machine learning,” 2015..

[61] D.P. Kingma, J. Ba, “Adam: A method for stochastic optimization,” International Conference on Learning Representations, vol. 1, no. 1, pp. 1-15, 2015..

[62] R. Novak, Y. Bahri, D. A. Abolafia, J. Pennington, J. Sohl-Dickstein, “Sensitivity and generalization in neural networks: an empirical study,” arXiv preprint arXiv:1802.08760, 2018..

[63] A. D. Gavrilov, A. Jordache, M. Vasdani, J. Deng, “Preventing model overfitting and underfitting in convolutional neural networks,” International Journal of Software Science and Computational Intelligence (IJSSCI), vol. 10, no. 4, pp. 19-28, 2018..

[64] M. M. Bejani, M. Ghatee, “A systematic review on overfitting control in shallow and deep neural networks,” Artificial Intelligence Review, vol. 54, no. 8, pp. 6391-6438, 2021..

[65] S. Lawrence, C. L. Giles, A. C. Tsoi, “Lessons in neural network training: Overfitting may be harder than expected,” Aaai/iaai, pp. 540-545, 1997..

[66] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929-1958, 2014..

[67] C. F. G. D. Santos, J. P. Papa, “Avoiding overfitting: A survey on regularization methods for convolutional neural networks,” ACM Computing Surveys (Csur), vol. 54, no. 10, pp. 1-25, 2022..

[68] K. O'Shea, R. Nash, “An introduction to convolutional neural networks,” 2015..

[69] P. Baldi, “Autoencoders, unsupervised learning, and deep architectures,” Proceedings of the International Conference on Machine Learning, vol. 27, no. 1, pp. 37-50, 2012..

[70] O. Ronneberger, P. Fischer, T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in International Conference on Medical image computing and computer-assisted intervention, Cham: Springer International Publishing, pp. 234-241, 2015.

[71] A. Oussidi, A. Elhassouny, “Deep generative models: Survey,” International Conference on Intelligent Systems and Computer Vision (ISCV), vol. 1, no. 1, pp. 1-8, 2018..

[72] R. Salakhutdinov, “Learning deep generative models,” Annual Review of Statistics and Its Application, vol. 2, no. 1, pp. 361-385, 2015..

[73] W.H.L. Pinaya, S. Vieira, R. Garcia-Dias, A. Mechelli, “Autoencoders,” Machine Learning, vol. 1, no. 1, pp. 193-208, 2020..

[74] D.P. Kingma, M. Welling, “An introduction to variational autoencoders,” Foundations and Trends in Machine Learning, vol. 12, no. 4, pp. 307-392, 2019..

[75] C. Doersch, “Tutorial on variational autoencoders,” arXiv preprint arXiv:1606.05908, vol. 1, no. 1, pp. 1-23, 2016..

[76] W. Harvey, S. Naderiparizi, F. Wood, “Conditional image generation by conditioning variational auto-encoders,” arXiv preprint arXiv:2102.12037, vol. 1, no. 1, pp. 1-19, 2021..

[77] R. Wei, C. Garcia, A. El-Sayed, V. Peterson, A. Mahmood, “Variations in variational autoencoders—a comparative evaluation,” IEEE Access, vol. 8, no. 1, pp. 153651-153670, 2020..

[78] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, “Generative adversarial networks,” Communications of the ACM, vol. 63, no. 11, pp. 139-144, 2020..

[79] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, A.A. Bharath, “Generative adversarial networks: An overview,” IEEE Signal Processing Magazine, vol. 35, no. 1, pp. 53-65, 2018..

[80] F. Farnia, A. Ozdaglar, “Do GANs always have Nash equilibria?,” International Conference on Machine Learning, vol. 1, no. 1, pp. 3029-3039, 2020..

[81] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, “Improved techniques for training GANs,” Advances in Neural Information Processing Systems, vol. 29, no. 1, pp. 2234-2242, 2016..

[82] L. Mescheder, A. Geiger, S. Nowozin, “Which training methods for GANs do actually converge?,” International Conference on Machine Learning, vol. 80, no. 1, pp. 3481-3490, 2018..

[83] Y. Kossale, M. Airaj, A. Darouichi, “Mode collapse in generative adversarial networks: An overview,” International Conference on Optimization and Applications (ICOA), vol. 1, no. 1, pp. 1-6, 2022..

[84] P. Dhariwal, A. Nichol, “Diffusion models beat GANs on image synthesis,” Advances in Neural Information Processing Systems, vol. 34, no. 1, pp. 8780-8794, 2021..

[85] J. Sohl-Dickstein, E.A. Weiss, N. Maheswaranathan, S. Ganguli, “Deep unsupervised learning using nonequilibrium thermodynamics,” International Conference on Machine Learning, vol. 1, no. 1, pp. 2256-2265, 2015..

[86] J. Ho, A. Jain, P. Abbeel, “Denoising diffusion probabilistic models,” Advances in Neural Information Processing Systems, vol. 33, no. 1, pp. 6840-6851, 2020..

[87] F.A. Croitoru, V. Hondru, R.T. Ionescu, M. Shah, “Diffusion models in vision: A survey,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 9, pp. 10850-10869, 2023..

[88] C. Luo, “Understanding diffusion models: A unified perspective,” arXiv preprint arXiv:2208.11970, 2022..

[89] M. Fuest, P. Ma, M. Gui, J. Schusterbauer, V.T. Hu, B. Ommer, “Diffusion models and representation learning: A survey,” arXiv preprint arXiv:2407.00783, 2024..

[90] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, M. Norouzi, “Palette: Image-to-image diffusion models,” ACM SIGGRAPH Conference Proceedings, vol. 1, no. 1, pp. 1-10, 2022..

[91] A.L.Y. Hung, K. Zhao, H. Zheng, R. Yan, S.S. Raman, D. Terzopoulos, K. Sung, “Med-cDiff: Conditional medical image generation with diffusion models,” Bioengineering, vol. 10, no. 11, pp. 1258, 2023..

[92] G. Webber, A.J. Reader, “Diffusion models for medical image reconstruction,” BJR|Artificial Intelligence, vol. 1, no. 1, 2024..

[93] A. Kazerouni, E.K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, D. Merhof, “Diffusion models for medical image analysis: A comprehensive survey,” arXiv preprint arXiv:2211.07804, 2022..

[94] Z. Wang, A.C. Bovik, “Mean squared error: Love it or leave it? A new look at signal fidelity measures,” IEEE Signal Processing Magazine, vol. 26, no. 1, pp. 98-117, 2009..

[95] A.K. Moorthy, Z. Wang, A.C. Bovik, “Visual perception and quality assessment,” Optical and Digital Image Processing: Fundamentals and Applications, pp. 419-439, 2011..

[96] A. Hore, D. Ziou, “Image quality metrics: PSNR vs. SSIM,” International Conference on Pattern Recognition, pp. 2366-2369, 2010..

[97] A.J. Dinu, R. Ganesan, A.A. Kebede, B. Veerasamy, “Performance analysis and comparison of medical image compression techniques,” International Conference on Control, Instrumentation, Communication and Computational Technologies, pp. 738-745, 2016..

[98] U. Sara, M. Akter, M.S. Uddin, “Image quality assessment through FSIM, SSIM, MSE and PSNR—a comparative study,” Journal of Computer and Communications, vol. 7, no. 3, pp. 8-18, 2019..

[99] I. Bakurov, M. Buzzelli, R. Schettini, M. Castelli, L. Vanneschi, “Structural similarity index (SSIM) revisited: A data-driven approach,” Expert Systems with Applications, vol. 189, pp. 116087, 2022..

[100] D. Gourdeau, S. Duchesne, L. Archambault, “On the proper use of structural similarity for the robust evaluation of medical image synthesis models,” Medical Physics, vol. 49, no. 4, pp. 2462-2474, 2022..

[101] L. Zhang, L. Zhang, X. Mou, D. Zhang, “FSIM: A feature similarity index for image quality assessment,” IEEE Transactions on Image Processing, vol. 20, no. 8, pp. 2378-2386, 2011..

[102] A. Breger, C. Karner, I. Selby, J. Gröhl, S. Dittmer, E. Lilley, J. Babar, J. Beckford, T.R. Else, T.J. Sadler, S. Shahipasand, “A study on the adequacy of common IQA measures for medical images,” International Conference on Medical Imaging and Computer, -Aided Diagnosis, pp. 451-462, 2024..

[103] Y. Yu, W. Zhang, Y. Deng, “Fréchet inception distance (FID) for evaluating GANs,” China University of Mining Technology Beijing Graduate School, vol. 3, no. 11, 2021..

[104] L. Hu, D. Zhou, Y. Zha, L. Li, H. He, W. Xu, L. Qian, Y. Zhang, C. Fu, H. Hu, J. Zhao, “Synthesizing high b-value diffusion-weighted imaging of the prostate using generative adversarial networks,” Radiology: Artificial Intelligence, 2021..

[105] S.M. Rezaeijo, H. Entezari Zarch, H. Mojtahedi, N. Chegeni, A. Danyaei, “Feasibility study of synthetic DW-MR images with different b values compared with real DW-MR images: Quantitative assessment of three models based-deep learning including CycleGAN,, Pix2PiX, and DC²Anet,” Applied Magnetic Resonance, vol. 53, no. 12, 2022..

[106] C. Hu, C. Bian, N. Cao, H. Zhou, B. Guo, “Synthesizing high b-value diffusion-weighted imaging of gastric cancer using an improved vision transformer CycleGAN,” Bioengineering, vol. 11, no. 8, 2024..

[107] F. Khader, G. Mueller-Franzes, S. Tayebi Arasteh, T. Han, C. Haarburger, M. Schulze-Hagen, P. Schad, S. Engelhardt, B. Baessler, S. Foersch, J. Stegmaier, C. Kuhl, S. Nebelung, J.N. Kather, D. Truhn, “Medical diffusion: Denoising diffusion probabilistic, models for 3D medical image generation,” 2022..

[108] P. Friedrich, J. Wolleb, F. Bieder, A. Durrer, P.C. Cattin, “WDM: 3D wavelet diffusion models for high-resolution medical image synthesis,” 2024..

[109] Z. Dorjsembe, H.-K. Pao, S. Odonchimed, F. Xiao, “Conditional diffusion models for semantic 3D brain MRI synthesis,” 2024..

[110] A. Odena, V. Dumoulin, and C. Olah, “Deconvolution and checkerboard artifacts,” Distill, vol. 1, no. 10, p. e3, 2016..

[111] G. Müller-Franzes, J. M. Niehues, F. Khader, et al., “A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis,” Scientific Reports, vol. 13, p. 12098, 2023..

[112] H. Wang, “Comparative Analysis of GANs and Diffusion Models in Image Generation,” Highlights in Science, Engineering and Technology, vol. 120, pp. 59-66, 2024..

[113] L. Tang and X. J. Zhou, “Diffusion MRI of cancer: From low to high b-values,” Journal of Magnetic Resonance Imaging, vol. 49, no. 1, pp. 23-40, 2019..

[114] A. L. Y. Hung, K. Zhao, H. Zheng, R. Yan, S. S. Raman, D. Terzopoulos, K. Sung, “Med-cDiff: Conditional Medical Image Generation with Diffusion Models,” Bioengineering (Basel, Switzerland), vol. 10, no. 11, pp. 1258, 2023..

[115] 2018., S. Woo, C. H. Suh, S. Y. Kim, J. Y. Cho, and S. H. Kim, “Head-To-Head Comparison Between High- and Standard-b-Value DWI for Detecting Prostate Cancer: A Systematic Review and Meta-Analysis,” American Journal of Roentgenology, vol. 210, no. 1, pp. 91-100, .

[116] R. Tomi-Tricot, “Clinical applications of the kT-points method to homogenise spin excitation in 3T MRI,” PhD diss., Université Paris Saclay (COmUE), 2018..

[117] P.V. Prasad, P. Storey, “Magnetic Resonance Imaging,” Molecular Biomethods Handbook, Humana Press, 2008..

[118] M. Chappell, “Generating and enhancing contrast,” Principles of Medical Imaging for Engineers, Springer, Cham, 2019..

[119] R. S. Abou khadrah, H.H. Imam, “Multiple b values of diffusion-weighted magnetic resonance imaging in evaluation of solid head and neck masses,” Egypt Journal of Radiology and Nuclear Medicine, vol. 50, no. 54, 2019..

[120] B.M. Ellingson, B. Yoo, W.B. Pope, “8 Diffusion-Weighted Imaging for Gliomas,” RadiologyKey, 2016..

[121] B. Goveas, W. Dkhar, R. Kadavigere, et al., “Enhancing prostate cancer detection: The role of b-value and apparent diffusion coefficient in DWI,” F1000Research, vol. 14, pp. 155, 2025..

[122] D. Bank, N. Koenigstein, R. Giryes, “Autoencoders,” Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook, pp. 353-374, 2023..

[123] N.A. Letizia, A.M. Tonello, “Segmented generative networks: Data generation in the uniform probability space,” IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 3, pp. 1338-1347, 2020..