UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Computer Science Curriculum

Glib Manaiev

Medical Image Classification with Limited Data

Master's Thesis (30 ECTS)

Supervisors: Dmytro Fishman, PhD
Joonas Ariva, MSc

Tartu 2024

Medical Image Classification with Limited Data

Abstract: Advancements in computational methods have greatly influenced medical imaging, facilitating the development of advanced diagnostic tools. One of the many tasks in this field is classifying images to determine whether they contain disease. This task is challenging because of the scarcity of annotated medical data, as it is harder to annotate because it requires an expert.

Lately, the problem of limited annotations has often been addressed by a group of learning approaches that utilize unannotated data, known as unsupervised learning. Typically, models are pretrained on an artificial task that exploits the properties of images, rather than their annotations, and then fine-tuned on annotated data. Despite the recent success of these methods, they remain minimally explored in the field of medical imaging, particularly in medical image classification.

This thesis investigates the effectiveness of various unsupervised pretraining approaches in enhancing the classification of medical images, specifically focusing on kidney tumor classification from CT (computed tomography) scans, which represents a distinct challenge within medical image classification. In our experiments, these methods do not significantly improve model performance, but offer insights into the limitations and possibilities of unsupervised learning in this area. Contrary to prior expectations about the transformative impact of unsupervised pretraining, the benefits appear dependent on specific contexts and tasks. This work illustrates the complexity of enhancing model performance in this field, emphasizing the need for a comprehensive approach to tackling these challenges.

Keywords:
deep learning in medical imaging, unsupervised learning, self-supervised learning, data augmentation, explainable artificial intelligence, image classification

CERCS: T111 - Imaging, image processing; P176 - Artificial intelligence

Meditsiiniliste piltide klassifitseerimine piiratud andmetega

Lühikokkuvõte: Arvutuslike meetodite edu on märkimisväärselt mõjutanud meditsiinilist kujutamist ning on aidanud kaasa piltdiagnostika vahendite arengule. Üheks keskseks väljakutseks selles valdkonnas on haiguste tuvastamine piltidelt. Selle ülesande teeb keeruliseks märgendatud andmete vähesus, mis on osalt põhjustatud asjaolust, et andmeid saavad märgendada vaid valdkonna eksperdid.

Viimasel ajal on andmete nappuse probleemi proovitud lahendada juhendamata õppe meetoditega, kus kasutatakse mudelite treenimiseks märgendamata andmeid. Tavaliselt selliste metoodikate puhul eeltreenitakse mudel tehislikel ülesannetel, mille puhul kasutatakse ära piltide omadusi ja struktuuri märgenduste asemel. Seejärel peenhäälestatakse mudel väikesel märgendatud andmestikul. Vaatamata nende meetodite hiljutisele edule, on neid meditsiinilise kujutamise valdkonnas vähe uuritud, eriti meditsiiniliste piltide klassifitseerimisel.

Magistritöös uuritakse erinevate juhendatamata eeltreenimismeetodite tõhusust meditsiiniliste piltide klassifitseerimisel, keskendudes eelkõige neerukasvajate tuvastamisele kompuutertomograafia piltidelt. Töö tulemused näitavad, et sellised meetodid ei paranda oluliselt mudelite täpsust. Samuti annavad tulemused ülevaate juhendamata õppe meetodite piirangutest ja võimalustest selles valdkonnas. Vastupidiselt eelnevatele ootustele juhendamata meetoditele, paistab et selliste meetoditest saadav kasu sõltub tugevalt konkreetsetest kontekstidest ja ülesannetest. Käesolev teadustöö ilmestab mudelite täpsuse parandamise keerukust meditsiinilise kujutamise valdkonnas ning rõhutab vajadust tervikliku lähenemisviisi järele nende väljakutsetega toime tulemiseks.

Märksõnad:
sügav õppimine, tehisnärvivõrgud, süvaõpe meditsiinilises kuvamises, järelevalveta õpe, iseõppiv õpe, andmete augmentatsioon, seletatav tehisintellekt, pildi klassifikatsioon

CERCS: T111 - Pilditehnika; P176 - Tehisintellekt

Contents

1 Introduction

Kidney cancer represents a significant global health challenge, diagnosed in over 430,000 individuals annually and responsible for approximately 180,000 deaths each year [1]. This statistic highlights the critical need for advancements in diagnostic methods that can improve early detection and treatment outcomes. Medical imaging plays a crucial role in the diagnosis and management of kidney cancer, providing essential insights that inform clinical decision-making.

Using deep learning in medical image analysis is challenging due to the limited availability of annotated medical data. Annotating medical data is more expensive than other tasks because it requires high precision and can typically only be done by individuals with medical education. Additionally, obtaining medical images, especially those containing diseases, is difficult due to privacy concerns and regulatory restrictions.

These challenges make this a suitable area for applying unsupervised learning, which enables models to learn useful representations from data without the need for annotations. Unsupervised learning generates synthetic annotations that pre-train the model, which is then fine-tuned using the annotated part of the dataset. While unsupervised learning has been popular and effective outside the medical image classification domain, its application in this field remains less explored.

This thesis investigates various unsupervised pretraining methods to evaluate their effectiveness in enhancing the classification of CT scans. In addition to exploring unsupervised learning, data augmentation techniques were used to artificially enlarge the dataset, proving beneficial in improving model accuracy.

The findings presented in this thesis provide a grounded perspective on their effectiveness. The results indicate that while unsupervised pretraining can offer some benefits, its advantages are highly context-dependent and may not be as transformative as previously anticipated.

In the subsequent sections, the thesis elaborates on various aspects of the study. Background section provides a detailed overview of the background technologies and methodologies fundamental to this research. Methods section outlines the dataset and experimental setup used to evaluate the unsupervised pretraining techniques. In Results section, a thorough analysis of the experimental results is presented, showcasing the impact and efficacy of the applied methods. Conclusion section summarizes obtained results. Finally, Discussion section talks about broader implications of the findings and proposes potential directions for future research, aiming to further the application of deep learning in medical image analysis.

2 Background

This section introduces fundamental neural network architectures and their pivotal role in medical image analysis. Starting with an overview of key architectures like ResNet and U-Net, which are essential for tasks such as image segmentation, the discussion progresses to Self-Supervised Learning (SSL). SSL and its specific approaches are explored for their ability to utilize unlabeled data, addressing the notable shortage of annotated medical datasets. This section also covers data augmentation strategies, another critical method to overcome the limitations of scarce training data. Finally, it delves into explainable AI techniques, focusing on their importance in ensuring the reliability and transparency of automated medical image analysis. This structured approach lays a comprehensive foundation for the detailed experimental investigations that follow.

2.1 Neural Network Architectures in Machine Learning

Neural networks, particularly Convolutional Neural Networks (CNNs), form the backbone of many modern machine learning applications, notably in fields requiring pattern recognition such as computer vision and medical image analysis. These computational models are highly effective in recognizing complex patterns and making intelligent decisions from vast amounts of data.

2.1.1 Convolutional Neural Networks (CNNs)

CNNs are specialized neural networks designed to process data with a grid-like topology, such as images. Their capability to capture spatial hierarchies in data stems from their unique architecture, particularly the convolutional layers, which apply filters to the input data. These filters move across the image, detecting local patterns such as edges, textures, and shapes at various levels of abstraction. By stacking multiple layers, CNNs can learn complex representations, starting from simple features in the initial layers to more complex structures in the deeper layers. CNNs are thus well suited for capturing important features such as edges and textures, which are crucial for tasks such as image classification, segmentation, and object recognition. Pooling layers, often interspersed between convolutional layers, further help by reducing the spatial dimensions, thereby enhancing the network's ability to recognize patterns irrespective of their position within the image.

The significant advancement in CNNs was marked by the introduction of AlexNet in 2012, which utilized deep layers, ReLU activations, and dropout to reduce overfitting, substantially outperforming existing models in the ImageNet competition [2]. Building on this, the ResNet architecture innovated with residual blocks that incorporate skip connections, facilitating the training of even deeper networks by enabling more effective

gradient flow [3]. These developments have established deep CNNs as foundational in advancing computer vision.

2.1.2 U-Net for Medical Image Segmentation

U-Net is a CNN variant specifically designed for biomedical image segmentation. Its architecture features a symmetric structure with a contracting path to capture context and an expansive path to enable precise localization. U-Net is particularly noted for its efficiency in training with limited data—a common challenge in medical imaging—leveraging extensive data augmentation to maximize the utility of available annotated samples [4].

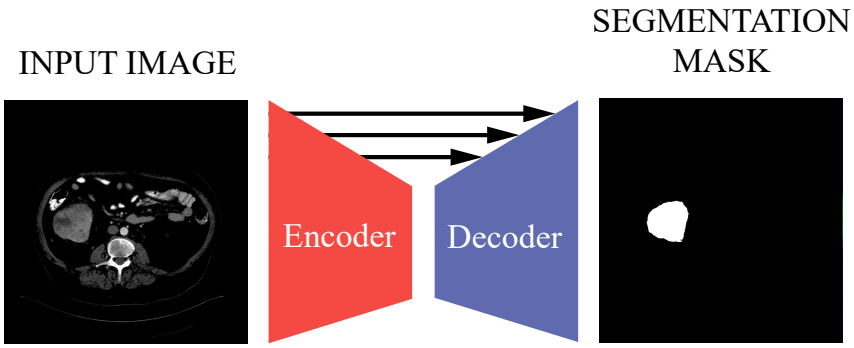


Figure 1. Simplified architecture of the U-Net. The network consists of three main components: the encoder, the decoder, and the skip connections. The encoder, shown in scarlet, is a series of convolutional layers that progressively capture increasingly complex features of the input image while downsampling the spatial dimensions. The decoder, illustrated in purple, is responsible for upsampling the encoded features to reconstruct the image's spatial dimensions. It gradually refines and reconstructs the output, aiming to restore the original image resolution. The skip connections, indicated by arrows, link corresponding layers of the encoder and decoder, allowing the transfer of high-resolution features directly from the encoder to the decoder. This mechanism helps retain detailed spatial information that is crucial for accurate image segmentation.

U-Net's design allows it to excel in capturing both local and global contextual information, which is essential for detailed segmentation tasks in medical imaging. This has led to significant improvements in automated segmentation of medical images such as CT scans, MRI, and microscopy images, and has spurred numerous adaptations and enhancements in the domain.

2.2 Pretext Tasks & Transfer learning

In the advancement of machine learning applications for medical imaging, particularly CT image classification, the advent of self-supervised learning (SSL) presents a paradigm shift. SSL, a branch of unsupervised learning, leverages the vast amounts of unlabeled data to learn robust representations without the need for human annotation, thus addressing the scarcity of annotated medical datasets [5]. This approach is embodied in the formulation of pretext tasks, which are designed to learn predictive features from the data itself, fostering the development of models that can predict certain properties or patterns inherent in the data.

Pretext tasks are ingeniously formulated such that the learning process leverages inherent data characteristics to predict certain properties, patterns, or parts of the data. These tasks can be predictive [6, 7], where the model might predict the missing part of an image or the next frame in a sequence, generative [8, 9], where the model reconstructs or generates new data points from the learned distribution, or contrastive [10, 11], focusing on differentiating between similar and dissimilar pairs of data samples. Each of these approaches teaches the model to understand and encode vital features of the data, crucial for the subsequent task-specific applications.

The efficacy of SSL and pretext tasks in medical imaging has been demonstrated in various studies. Models pretrained on large-scale unlabeled datasets have shown remarkable success in disease detection, segmentation, and classification tasks, demonstrating the potential of SSL to mitigate the challenges posed by limited labeled datasets in medical imaging [12, 13].

Through the strategic implementation of pretext tasks within the self-supervised learning framework, researchers can pretrain models to learn complex, generalizable features from expansive, unlabeled datasets. These features, when refined through fine-tuning, has proven to sometimes noticeably improve medical image analysis tasks [5].

Following the pretext task phase is the transfer learning process, encompassing pre-training and fine-tuning stages. Initially, models are pretrained on large-scale unlabeled datasets through the selected pretext tasks. This pretraining endows the model with a broad understanding of the data's features. Subsequently, the model undergoes fine-tuning on a smaller, task-specific labeled dataset. This fine-tuning process adjusts the pretrained model's parameters to optimize performance for the specific CT image classification task at hand. The transition from generalized pretraining to targeted fine-tuning exemplifies the core of transfer learning, leveraging learned representations from one task to enhance performance on another.

2.2.1 Predictive Approaches

The predictive self-supervised learning approach is centered on developing robust representations from unlabeled data. This process involves assigning each image, or a

portion thereof, a pseudo-label. These labels are directly derived from the data, for instance, by utilizing structural information of the image. Subsequently, this task is approached as either a classification or regression challenge, where the objective is to accurately predict these pseudo-labels. The effectiveness and strength of the features learned during this pretraining phase heavily rely on the strategy used for generating pseudo-labels, especially considering the specific downstream tasks at hand. A variety of predictive pretraining strategies have been implemented in the domains of medical image classification and segmentation; this section will further explore several such methods in greater detail.

Rotation Prediction was initially introduced by Gidaris et al. [14] as a means to acquire visual representations through self-supervised learning. This task involves training a convolutional model to identify the geometric rotation applied to an input image, effectively framing this challenge as a straightforward classification task. Specifically, images are rotated with steps of 90 degrees, resulting in possible rotations of [0°, 90°, 180°, 270°], as can be seen in Figure 2. The idea of rotation prediction is the connection between the model's capacity to discern the applied rotation and its proficiency in identifying key features within the image. The model must learn the types and orientations of objects relative to the rotation, mirroring the human process of recognizing rotated objects. For example, understanding a body slice orientation after a 90° rotation involves recognizing the arrangement of all of its organs: kidneys, backbone, etc. Consequently, rotation prediction facilitates the learning of semantic features by teaching the model to understand image orientations.

In the context of medical imaging, rotation prediction has been explored by various studies [15, 16], with mixed outcomes ranging from no noticeable performance gains—or even decreases—to notable improvements. These varied results highight the challenge of applying rotation prediction in medical settings, where the orientation of anatomical structures may not always provide clear discriminative signals for learning.

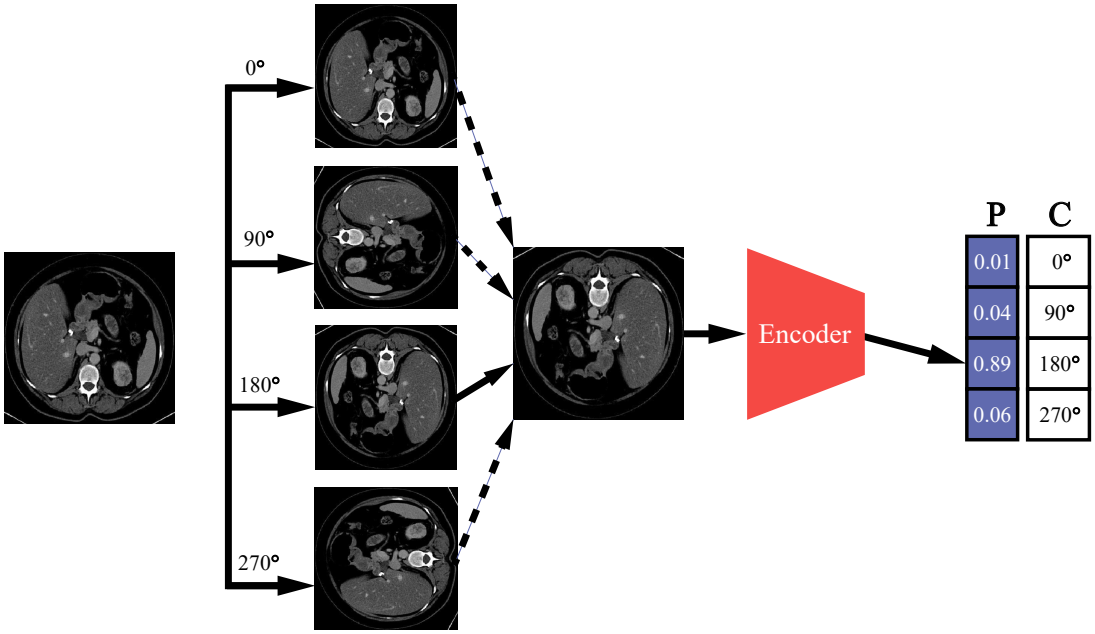


Figure 2. Illustration of Rotation Prediction Task. An image is rotated by 0°, 90°, 180°, and 270°, and the CNN is tasked to predict the rotation angle. P are probabilities of predicted classes C. Image adapted from [17].

Relative position prediction was devised by Doersch et al. [6] as an innovative self-supervised learning approach to cultivate visual representations from unlabeled images. This method trains a model to anticipate the relative position of a second image patch with respect to a first, randomly selected patch from a large collection of unlabeled images. Such training necessitates the CNN to discern the appearance of objects and their parts, promoting the learning of meaningful visual features. The fundamental premise is that successfully predicting the relative positions of patches within images compels the model to understand and recognize the compositional structure of objects and their spatial relationships, as shown in Figure 3. For instance, discerning that a patch depicting a section of a kidney is above a patch showing a tumor, all without any additional contextual information, indicates a sophisticated understanding of the anatomical features and spatial configurations pertinent to medical imaging, such as CT scans. This method effectively converts the unsupervised challenge of learning from unlabeled images into a supervised problem by utilizing the inherent spatial context within images as a self-supervisory signal.

In the field of medical image analysis, the relative position prediction prediction approach as well as similar approaches that were later derived from it have found several applications [18, 19]. Despite showing improvement, leveraging spatial relationships as self-supervision in medical images, while promising, often results in less noticeable improvements.

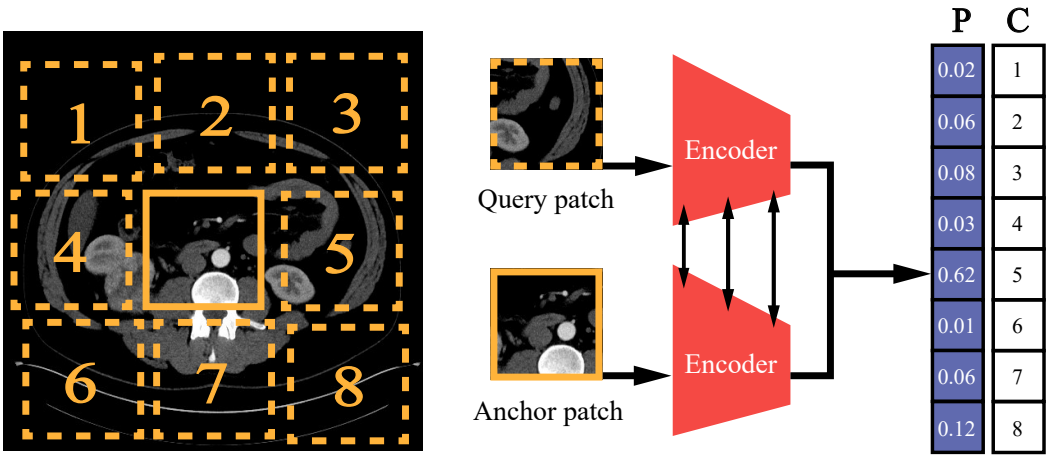


Figure 3. Illustration of self-supervised learning by relative position prediction task. (left): An image is divided into nine patches where the central patch (the one without number) represents the anchor patch and the remaining eight patches (delineated in dashed yellow lines) represent the query patches. (right): a training example that consists of an anchor patch and query patch is passed to a late-fusion convolutional model which shares weights between the two branches to predict the position of the query patch with respect to the anchor patch. P are probabilities of predicted classes C. Image adopted from [17].

2.2.2 Generative Approaches

Generative approaches in self-supervised learning have gained significant traction due to their capacity to model complex data distributions. Central to these methods is the principle of learning to generate or reconstruct instances that mirror the training dataset, without requiring explicit labels. These methods typically involve training generative models, such as Autoencoders [20], Variational Autoencoders (VAEs) [21] or Generative Adversarial Networks (GANs) [22], to recreate the input data. Models are trained to capture the distribution of the data within a latent space, which is construed to retain the most salient features of the data. The quality of these learned features is often contingent upon the generative model's ability to accurately reconstruct the input data while also preserving the richness of the latent representation.

Early explorations in generative self-supervised learning predominantly utilized autoencoders, where the encoder component maps inputs into a latent space and the decoder reconstructs the input from this latent representation [23]. The optimization of these models is driven by the fidelity of the reconstruction to the original input. Later

advancements saw the introduction of VAEs, which added a probabilistic twist to the encoding process, enabling the generation of new data instances by sampling from the learned distribution [21].

The advent of GANs introduced a novel competitive dynamic to the training process, where a generator network competes against a discriminator network that judges the authenticity of the generated images [22]. The discriminator, in the process of distinguishing real from fake, acquires a nuanced understanding of the characteristics of the data. Below, we will discuss various generative methods and their applications to medical image analysis.

Image Inpainting employed as a self-supervised learning strategy, proposed by Pathak et al. [8], involves training models to accurately predict and reconstruct missing or damaged sections of images, as shown on Figure 4. This method requires the model to interpret the remaining parts of an image and use this context to fill in the gaps, effectively teaching the model to understand and replicate the underlying data structures and relationships. By compelling the model to restore lost image parts, it not only learns to identify visual patterns and intricate details essential for whole image interpretation but also enhances its ability to generalize from limited data.

This approach has proven to be very efficient in the field of medical image segmentation, allowing the extraction of the most useful context from usually limited data [24, 25]. However, in the domain of medical image classification, inpainting approaches are a less popular choice than for segmentation-related tasks[19].

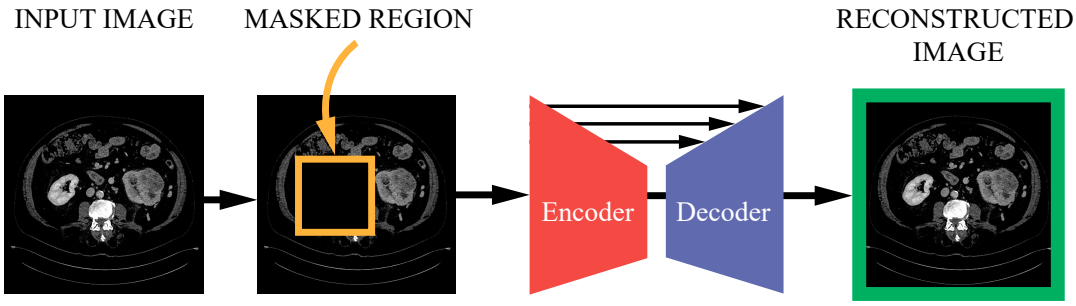


Figure 4. Schematic representation of usual inpainting framework. Part of the input image is removed by masking it with black pixels (outlined in orange). The context encoder is then tasked with reconstructing the occluded segment, thereby learning to infer missing content based on the surrounding visual context, as can be seen from the rightmost image.

The more advanced approach, which shares the same principle, is called Model Genesis. It is a self-supervised learning framework introduced by Zhou et al. [26], specifically designed for 3D medical imaging. Building upon the concept of context

restoration, Model Genesis employs four unique distortion operations to train models: non-linear transformations using the Bézier transformation function, local pixel shuffling, in-painting (similar to the context encoder method), and out-painting, which is essentially the inverse of in-painting, meaning that the outside pixels are masked out. Notably, each input volume is subjected to the first two operations followed by either of the last two operations, but not both. This structured approach to applying transformations enables the generative model to effectively learn from and restore the distorted images back to their original context. The illustrations of the distortion operation, as well as the framework, can be seen on Figure 5. Model Genesis has been extensively evaluated across six downstream tasks, demonstrating its efficacy in both segmentation and classification tasks within medical image analysis. The framework's ability to generalize from intrinsic image features learned through these self-supervised tasks makes it particularly valuable for enhancing performance on specific medical imaging applications. Even though the framework was mainly devised for 3D applications, it has been shown to provide a slight performance boost when used on 2D slices of medical images, namely slices of lung CT scans.

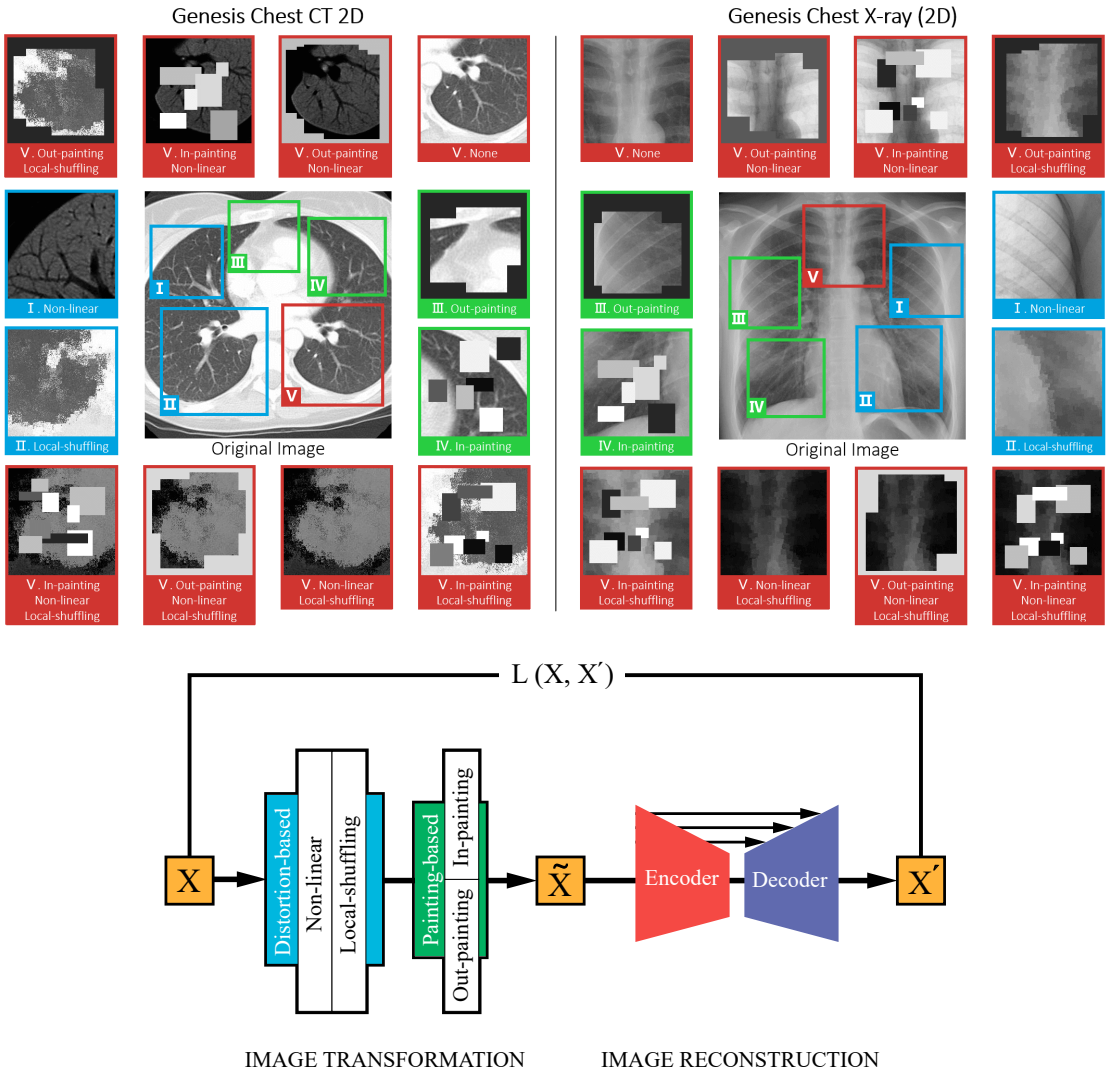


Figure 5. Schematic representation of the Model Genesis transformations and framework. The upper image showcases the diverse transformations employed by the Model Genesis framework for 2D medical imaging. Each panel represents a transformation applied to an original medical image: I) Non-linear transformation, II) Local pixel shuffling, III) Out-painting, and IV) In-painting. Notably, transformations I and II are applied consistently, while either III or IV is selected for a given image, as in-painting and out-painting are mutually exclusive. This framework's encoder-decoder architecture, shown on the lower image, is trained to revert these transformed patches to their original state, thereby learning to reconstruct and understand complex medical image structures without manual labeling. The process is demonstrated on both Genesis Chest CT and Chest X-ray (2D) images, highlighting the framework's adaptability to different medical imaging modalities. Images were taken from the original paper [26].

2.2.3 Contrastive Approaches

The contrastive approach in self-supervised learning emphasizes the differentiation between representations of data samples, often by bringing closer the features of similar (positive) pairs while pushing apart those of dissimilar (negative) pairs. This method relies on constructing pairs or sets of data points where some inherent relationship exists, such as different augmentations of the same image or semantically related images, to teach the model what features are essential for identifying similarities or differences. The core of this approach is a contrastive loss function, which quantifies the degree to which the learned representations adhere to these relational expectations. By optimizing this loss, the model learns to encode rich, discriminative features without the need for explicit labeling, making it particularly adept for tasks where labeled data is scarce but where capturing the nuances between data samples is crucial. This section delves into the details of some popular contrastive approaches and their usage in medical image analysis.

Simple Framework for Contrastive Learning of Representations (SimCLR) [10] utilizes data augmentation to generate multiple views of the same image, thereby creating pairs of correlated samples. These samples are then transformed through operations such as cropping and color distortion to produce augmented images that, while visually distinct, share the same semantic information. These images are encoded into representations via a CNN and a projection head, facilitating the application of a contrastive loss function, as shown in Figure 6.

Central to SimCLR's effectiveness is the contrastive loss function, particularly the normalized temperature-scaled cross-entropy loss (NT-Xent loss), which is formulated to minimize the distance between representations of positive pairs (i.e., different augmentations of the same image) while maximizing the distance between those of negative pairs (i.e., different images). The loss for a pair of positive samples i, j is given by:

$$\mathcal{L}_{\text{NT-Xent}} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)} \tag{1}$$

where z_i and z_j denote the representations of two augmented views of the same image, $\text{sim}(u, v)$ represents the cosine similarity between vectors u and v, N is the batch size of distinct images, τ is a temperature scaling parameter, and $\mathbb{1}_{[k \neq i]}$ is an indicator function that is 1 if $k \neq i$. Through optimizing this loss, SimCLR trains to capture the semantic essence of images, showcasing its utility in downstream tasks such as image classification with minimal labeled data.

SimCLR has been applied across various medical imaging contexts with varying degrees of success [27, 28]. When integrated with additional methodologies [28], it demonstrates a significant enhancement in performance. However, as a standalone approach, SimCLR does not exhibit substantial improvements when compared to models trained with full supervision.

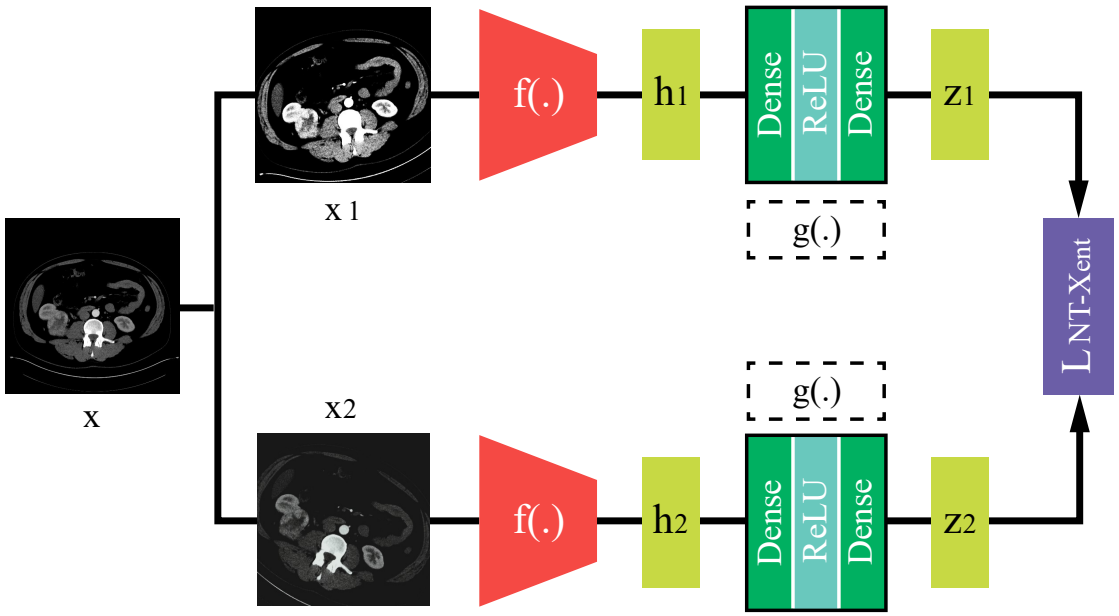


Figure 6. Schematic representation of SimCLR framework. An input image x undergoes a series of transformations to produce two correlated images, x_1 and x_2, which serve as positive pairs. Each image is processed through a feature extractor $f(\cdot)$ to obtain representations h_1 and h_2. These are then passed through a projection network $g(\cdot)$, composed of dense layers with ReLU activation, resulting in embeddings z_1 and z_2. The model is trained using the NT-Xent loss to minimize the distance between these embeddings, thereby encouraging the network to learn invariant features from augmented versions of the same image.

Contrastive Predictive Coding (CPC), proposed by van den Oord et al. in 2018 [29], offers a method for unsupervised learning of representations from diverse data forms, including images, text, and audio. Unlike generative models that predict future data samples from their context, CPC aims to create a compact representation that enhances the mutual information between the context (C) and the target (X). This approach allows for the learning of representations that overlook low-level details of the input data, focusing instead on capturing more abstract, informative features.

The CPC framework comprises three main components: an encoder network, an autoregressive network, and the InfoNCE loss function, as can be seen from the Figure 7. The encoder network converts input data into a latent variable (Z_t), capturing essential features of the input. The autoregressive network then uses these latent variables to produce a context (C_t) for predicting future latent variables. The key to CPC's effectiveness is the InfoNCE loss, derived from the Noise-Contrastive Estimation (NCE) [30] loss, which encourages the model to distinguish between the true future samples and randomly selected negative samples. The InfoNCE loss is expressed as:

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E}\left[\log \frac{\exp(\text{sim}(c, x^+))}{\sum_{x^-} \exp(\text{sim}(c, x^-))}\right] \tag{2}$$

Here, c indicates the context, x^+ is a positive sample that follows the context, and x^- are negative samples chosen randomly from the dataset. The function $\text{sim}(c, x)$ measures the similarity between the context and each sample, usually by a dot product. Through the optimization of the InfoNCE loss, CPC learns to produce representations that predict future states of the input, enriching the model's understanding and processing of sequential data.

Building upon the original CPC framework, CPCv2 [31] introduces several key improvements over its predecessor: it incorporates a larger and more effective set of data augmentations, utilizes a deeper and wider architecture for the encoder network, and employs a modified contrastive loss function that facilitates learning from a greater number of negative samples, significantly improving the quality of the learned representations.

CPC has seen limited application in the field of medical imaging. However, its adaptation for analyzing 3D CT scans [32] demonstrated some improvements over baseline models trained from scratch. The same study also compared this to a 2D version of CPC, which underperformed relative to the baseline. It's important to note that this decrease in performance may be attributed to the difference in dimensionality, as the baseline model was designed for 3D data.

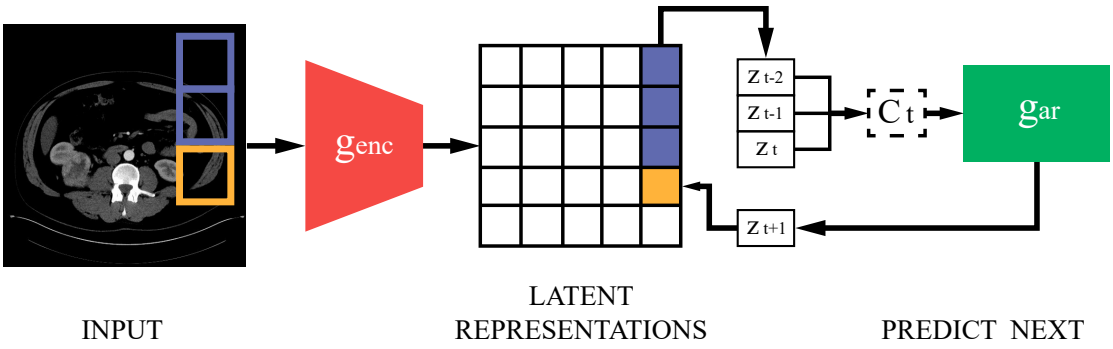


Figure 7. Schematic representation of the Contrastive Predictive Coding (CPC) framework. An input image is divided into a grid and is processed through an encoder network (g_{enc}), resulting in a grid of latent representations. A subset of these representations (highlighted in yellow and purple) is selected to act as the context (C_t) for the autoregressive model (g_{ar}). This model uses the context to predict future latent variables (z_{t+1}), aiming to maximize the mutual information between context and predictions. Training involves a contrastive loss that discriminates between true future representations and a set of negative samples.

2.3 Data Augmentation

Data augmentation technique is often used to artificially expand the training dataset and introduce a variety of conditions under which a medical event might be imaged. Techniques often used include geometric transformations like rotations and flips, which can mimic the various orientations of patients during imaging, and intensity variations that account for different machine calibrations or contrast levels.

Specific to medical imaging, augmentation techniques must be applied with particular care. Unlike natural images, slight alterations in medical images can change the diagnosis or obscure critical features. For example, random noise or color shifts that are commonly used in augmenting natural images may not be suitable for medical datasets, as they could obscure subtle but clinically significant features. Thus, augmentation strategies are often tailored to the specific medical imaging modality and analysis task, ensuring that the augmented data remains representative and useful for clinical purposes. Properly implemented, these techniques not only improve model robustness but also enhance performance in diagnostic tasks, ensuring that models trained on augmented datasets are better equipped to handle real-world variability in medical images [33].

2.4 Explainable AI

Explainable AI (XAI) has become an essential component in the field of medical imaging, aiming to make the decision-making processes of AI models transparent and understandable to human users. With the increasing integration of AI in diagnostic tools, it is crucial that practitioners can interpret and trust the recommendations provided by these systems.

In the context of medical imaging, XAI offers the potential to demystify the outputs of complex models, particularly deep learning architectures, which often operate as 'black boxes'. Given that medical decisions can have profound implications on patient care, the ability to explain and validate these decisions is not merely a matter of academic interest but a clinical necessity. Explainable models can illuminate the features within medical images that are most influential in a model's predictions, thus providing clinicians with valuable insights that may corroborate their expertise or reveal new diagnostic patterns.

In this thesis, XAI was specifically utilized to delve into and compare the performance of various models, looking for explanations behind their differing effectiveness.

2.4.1 Class Activation Mapping

Class Activation Mapping (CAM) was introduced by Zhou et al. in 2016 [34] as a method to identify regions in an image that influence a model's decision for a given task. It operates by mapping the activations of specific layers in a convolutional neural network (CNN) to the output predictions, highlighting important areas in the image that contribute to the final decision.

Building on CAM, Gradient-weighted Class Activation Mapping (GradCAM) utilizes the gradients of any target output flowing back into the final convolutional layers to produce a heatmap [35]. This method refines the visual explanation by indicating where the CNN is focusing its attention for each class in a classification task.

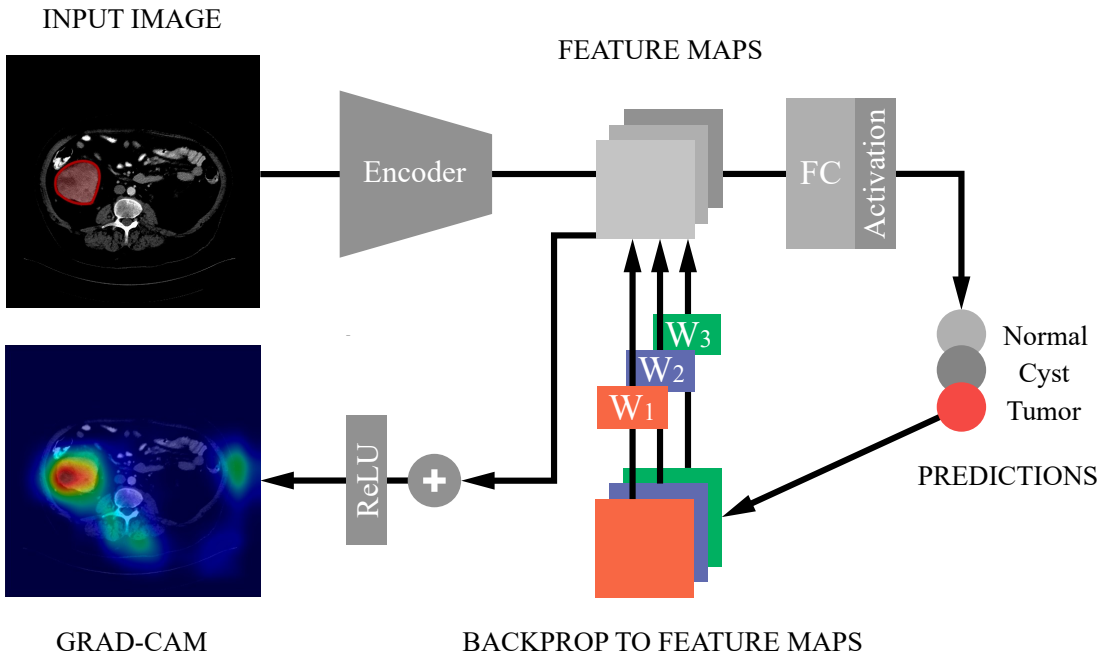


Figure 8. Illustration of the GradCAM process within a CNN. Starting with an input image, such as a CT slice, the image is forward propagated through the CNN to extract feature maps that capture essential visual patterns relevant to the task. For all classes of interest *(Normal, Cyst, Tumor)*, the model computes a raw score by passing feature maps through a fully connected (FC) layer. The gradients are set to zero for all classes except the target class, which is set to 1. This targeted signal is backpropagated to the rectified convolutional feature maps, emphasizing areas crucial for decision-making. These maps are then weighted (*W1, W2, W3*) based on gradient information flowing back from the output layer. The weighted feature maps are combined and passed through a ReLU function to produce a coarse Grad-CAM localization, highlighting where the model focuses while making the decision. [35]

3 Methods

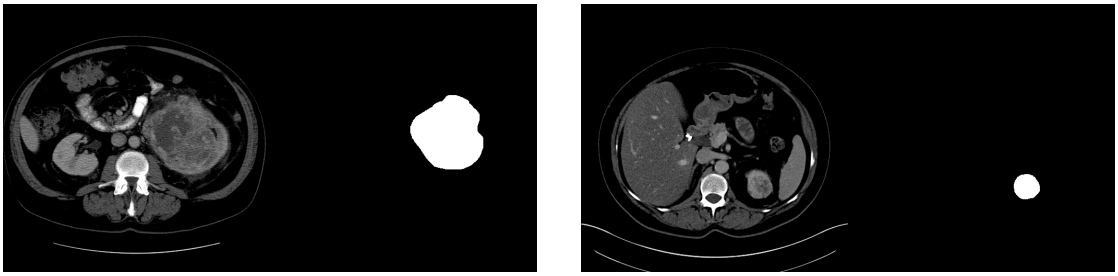
This section provides a comprehensive overview of the methodologies employed in this study. It begins with a description of the dataset used, followed by the rationale for selecting the baseline model and details of its implementation. Next, the section outlines each pretext task tested, including the data augmentation strategies applied. Finally, it details the evaluation procedures and the metrics used to assess model performance.

3.1 Dataset

The primary dataset utilized for all experiments conducted in this thesis was the 2023 Kidney and Kidney Tumor Segmentation Challenge dataset (abbreviated as KiTS23) [36]. The KiTS23 competition challenges teams to create the most effective system for the automatic semantic segmentation of kidneys, renal tumors, and renal cysts, marking the third iteration of the challenge following the 2019 and 2021 competitions. The dataset comprises 599 cases, divided into 489 training and 110 test cases. Each case represents a three-dimensional CT scan of the kidneys, which for the purpose of this study, was processed into transverse (axial, perpendicular to the spinal column) 2D slices to facilitate the experiments on 2D image data.

The dataset annotates three types of segmentation: kidney (encompassing all kidney parenchyma and non-fat tissue within the renal hilum), tumor (kidney masses suspected to be malignant pre-operatively), and cyst (kidney masses identified as cysts either radiologically or pathologically, when available). For the scope of this research, only tumor segments were considered. Given the variable sizes of tumors across slices, the study focused on slices featuring sufficiently large tumors, establishing a minimum area threshold of 1000 pixels for inclusion. This threshold not only facilitated empirical analysis but also allowed for visual identification of tumors in slices, simplifying the experimental results' interpretation as depicted in Figure 9. Additionally, it was assumed to support the models that did the inpainting-related tasks, as bigger tumors are easier to encode and not to miss during the reconstruction. This will be described in more details in the corresponding section.

To mitigate model bias and ensure a balanced training dataset, slices were selected to maintain a 2:1:1 ratio among three categories: slices with tumors, slices without tumors but with kidneys, and slices with neither. The compiled dataset for experimentation included 11,980 slices for training and 3,149 slices for testing.



(a) (b)

Figure 9. Tumor sizes scale: (a) On the right is a slice that contains a big tumor; on the left is its annotation mask, where white pixels signify the tumor. (b) Slice with a tumor of smaller size and its annotation.

All images were uniformly resized to 512x512 pixels for consistency and underwent normalization tailored to highlight tumor intensities, diverging from conventional methods that uniformly treat all pixels. This approach, inspired by the nnUnet framework [37], adjusts the contrast based on the specific intensity range of tumors. Generally, CT scans encompass a wide range of Housefield Units (HU), from -1000 HU (air) to approximately 3000 HU (dense bone). However, the intensity values relevant to kidneys and kidney tumors are much narrower. Typically, soft tissues like kidneys are observed within the -150 to 250 HU range, while renal tumors might exhibit slightly different intensities due to their composition and the presence of contrast agents. This specificity in HU ranges for kidneys and tumors underlines the necessity of targeted normalization to effectively contrast-enhance these areas within the broader HU spectrum of CT scans. This approach was informed by analyzing tumor-associated pixel intensities across the dataset to determine the 0.5th and 99.5th percentile values, mean (μ), and standard deviation (σ). The initial normalization stage, aimed at adjusting image contrast to emphasize tumor areas within the extensive HU range, is represented by the following equation:

$$I_{norm} = \frac{clip(I, [P_{0.5}, P_{99.5}]) - \mu}{\sigma}, \tag{3}$$

where I_{norm} is the normalized image, I is the original image, and $P_{0.5}$ and $P_{99.5}$ denote the 5th and 99.5th percentile values, respectively. The subsequent stage, not included in the nnUnet's normalization method, applies min-max normalization to adjust intensity values to a [0,1] range, essential for the inpainting tasks:

$$I_{final} = \frac{I_{norm} - I_{min}}{I_{max} - I_{min}}, \tag{4}$$

This two-step normalization process, especially with the addition of the second stage not found in the nnUnet framework, is crucial for maintaining consistency across experiments by focusing on renal tumor intensities.

3.2 Architecture and setup

The cornerstone model for all experiments conducted within this thesis was ResNet18 [3], chosen for its balance between accuracy and computational efficiency. Despite experimenting with larger models from the same family, such as ResNet34 and ResNet50, their increased complexity did not translate to significantly better performance in tasks related to classification and inpainting. This observation underscores the thesis's objective: to explore the performance impact of various methodologies rather than seeking the pinnacle of model accuracy. The ResNet18 architecture is notable for its four residual blocks, designed to downscale the input image progressively, culminating in a fully connected layer followed by a sigmoid activation for binary classification tasks.

The utilization of weights pretrained on the ImageNet dataset [38] proved instrumental. ImageNet, a large-scale dataset designed for object recognition in natural images, encompasses over 14 million images categorized into thousands of classes. Despite the absence of medical imagery within ImageNet, initializing ResNet18 with these pretrained weights significantly enhanced its performance. This effect shows the transferability of learned features from natural to medical image contexts.

3.3 Baseline classfier training

The baseline training for the binary classification task was conducted using a modified version of the torchvision ResNet18 model. The dimensions of the fully connected (FC) layer were adjusted from $[256, 1000]$ to $[256, 1]$ with subsequent Sigmoid to suit binary classification. The model was trained over 30 epochs with a batch size of 64, utilizing the Adam optimizer with L2 regularization. The Binary Cross Entropy (BCE) loss, suitable for binary classification tasks, was the chosen loss function and is defined as:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)], \tag{5}$$

where N is the number of samples, y_i is the ground truth label, and $\hat{y}_i$ is the predicted probability for the i-th sample. This loss function is designed to measure the discrepancy between the predicted probabilities and the actual binary labels.

Additionally, several basic data augmentation techniques were applied, including rotation, horizontal and vertical flipping, slight color distortions, and cropping. However, these augmentations did not yield a positive impact, likely due to the uniformity of CT scan images. The consistency in patient positioning and organ locations within scans renders such transformations less meaningful.

3.4 Pretraining Approaches

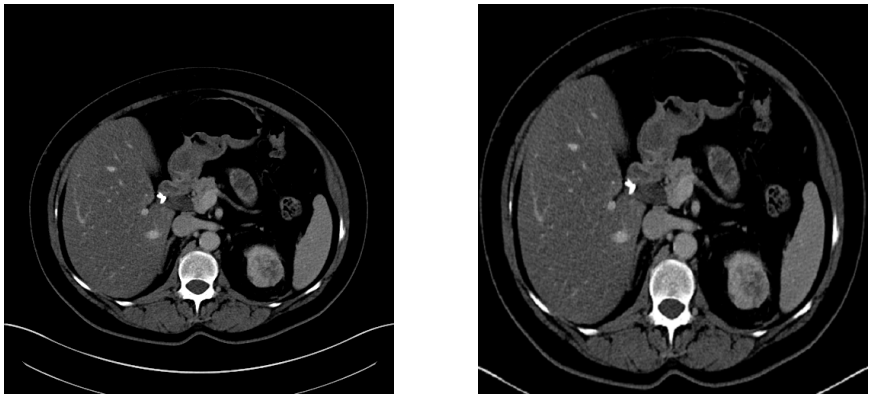
The pretraining methods examined include predictive, generative, and contrastive approaches, each with unique strategies to derive synthetic annotations and improve feature learning. The subsequent subsections provide detailed descriptions of the implementation and evaluation of these pretraining approaches.

3.4.1 Predictive Pretraining

The Rotation classification pretraining was the predictive method of choice for the experiments conducted in this thesis. The model was trained on the original training set. The dataloader of the model was modified to rotate input images by an arbitrary angle from $[0, 90, 180, 270]$ degrees and was assigned the corresponding label from $[0, 1, 2, 3]$ as was shown on Figure 2. All the other preprocessing was identical to the one of the baseline training. The model backbone architecture was left intact, and the final FC layer of the classifier head was changed in dimensions from $[512, 1]$ to $[512, 4]$ to adjust to the 4-class classification task. The optimizer, learning rate, training length, and other parameters were unchanged compared to the baseline training. During the fine-tuning stage, the pretrained backbone was not frozen, so the whole model was trained for the same amount of epochs and with the same hyperparameters as the baseline model.

3.4.2 Generative Pretraining

Inpainting Pretraining tasks were devised with a focus on feature learning. The first method involved blanking out a 20% by 20% square section of the input image at random. The second method, aiming to target the kidney region more precisely, used the same size square but placed it within an area predetermined from dataset analysis to most likely contain kidneys across various slices. Images were processed to eliminate any black pixels outside the body contour for standardization, as indicated in Figure 10. The third method involved swapping regions from the right to the left kidney, training the model to reconstruct the original image and thus encouraging the encoder to concentrate on these specific areas. The kidney regions were determined by looking through a big portion of the dataset. They were chosen to be big to contain kidneys with high confidence.



(a) Original image (b) After crop and resize

Figure 10. Image processing to ensure uniformity. The image was first treated with morphological operations to remove all the pixels outside of the body, that include random noise and parts of the CT machine. Then, the image was cropped to a rectangle that contains all the non-zero values and resized to the original size.

The original ResNet backbone was employed as an encoder in a U-Net architecture, with skip connections and upsampling convolutions suitable for the inpainting tasks. Training spanned over 50 epochs, using a ReduceLROnPlateau scheduler from the PyTorch library, which reduces the learning rate once learning stagnates. This approach allows more detailed learning over time, which is less sensitive to the risk of overfitting in a pretraining context. Two loss functions, $\mathcal{L}_{\text{MSE}}$ for Mean Squared Error and $\mathcal{L}_{\text{dice}}$ for Dice loss, were used, the latter of which is typically employed for binary segmentation but here adapted for continuous-valued images in the range $[0, 1]$. The losses are mathematically formulated as:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2, \tag{6}$$

$$\mathcal{L}_{\text{dice}} = 1 - \frac{2 \times |\hat{Y} \cap Y|}{|\hat{Y}| + |Y|}, \tag{7}$$

where $\hat{Y}$ and Y represent the predicted output and the ground truth, respectively. Results, including visualizations and model predictions, are detailed in Section 4.

Model Genesis was also chosen for use in this work. The method followed was the same as described by Zhou et al. in their paper [26] and the code they shared on GitHub. Like in the original work, the MSE loss was used for training. The main change made was using a U-Net with a ResNet-18 encoder for the model, the same type that was used in the inpainting experiments. This kept the model's setup consistent across this study. Training was done for 300 epochs, which is the same length as what was done in Model Genesis.

As in all other experiments, during the fine-tuning stage, the pretrained backbone, extracted from U-Net, was not frozen, so the whole model was trained for the same amount of epochs and with the same hyperparameters as the baseline model.

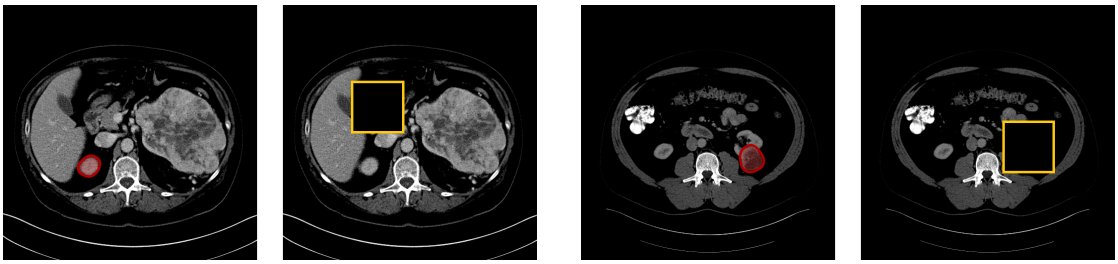
3.4.3 Contrastive Pretraining

SimCLR was selected as one of the two contrastive learning methods evaluated in this study. The SimCLR framework was shown on Figure 6. The implementation was based on a PyTorch version of SimCLR, which can be found in the repository [39]. Unlike the typical use of a standard ResNet-based feature extractor with layer-wise normalization, this experiment utilized a ResNet18 backbone equipped with batch normalization to simplify the fine-tuning process. Additionally, the conventional data augmentation pipeline, notably Color Jitter, was modified for medical imaging applications to reduce the intensity of color adjustments, thus minimizing the risk of overly distorting the medical images.

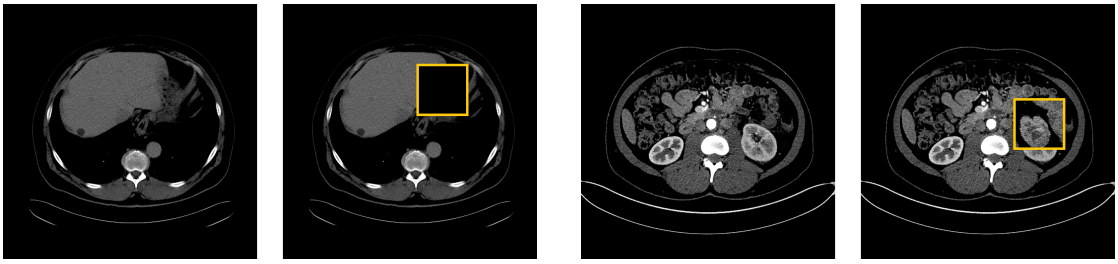
Contrastive Predictive Coding (CPC), shown on Figure 7, the second contrastive learning method tested in this study, drew on an existing PyTorch implementation found online [40]. The code underwent several adjustments to make it compatible with a ResNet18 backbone used throughout these experiments.

3.5 Data Augmentation

In this work, two fundamentally different data augmentation approaches were employed: Supervised and Unsupervised. Supervised Augmentation utilized the tumor annotations that were available. The idea was inspired by the Copy-Paste Data augmentation [41] technique, which was originally proposed for instance segmentation. The main concept of this method involves copying instances from one image and pasting them into another, thereby creating additional ground truth instances in a completely different context. Building on this, the following approach was implemented: on the slices containing a tumor, there was a 50% chance to randomly select a 20% by 20% square to be removed in such a way that the tumor remains unaffected. Additionally, with a 35% probability, the tumor was covered by a 20% by 20% black square, but this was only done if the tumor was relatively small (area less than 2500 pixels). This restriction was imposed because large tumors often cause deformations in surrounding organs and covering such tumors could still leave visible deformations, potentially leading the model to unlearn that deformations can also be an indicator of a tumor. An example of this augmentation is shown in Figure 11.



(a) An image where a tumor is present (marked with red). On the left pair of images, a random 20% by 20% non-tumor region is removed; on the right, the tumor is covered by a black square. Removed regions are marked with yellow.



(b) An image without any tumor. On the left pair of images, a random 20% by 20% non-tumor region is removed; on the right, a tumor region (marked in yellow) is introduced to an image.

Figure 11. Supervised data augmentation examples.

Unsupervised Augmentation does not rely on dataset annotations. Similar to the approach described for inpainting, this method utilizes data-specific prior information. Regions most likely to contain kidneys—and potentially tumors—were identified by examining numerous dataset slices. The augmentation process operates as follows: first, areas presumed to contain kidneys are copied from the image. Then, three rectangles, randomly chosen and varying in size from 0% to 50% of the image's dimensions, are removed. Finally, the initially copied regions, containing the kidneys, are pasted back into the image. This ensures that the key kidney regions remain unaltered. The steps of this process are illustrated in Figure 12.

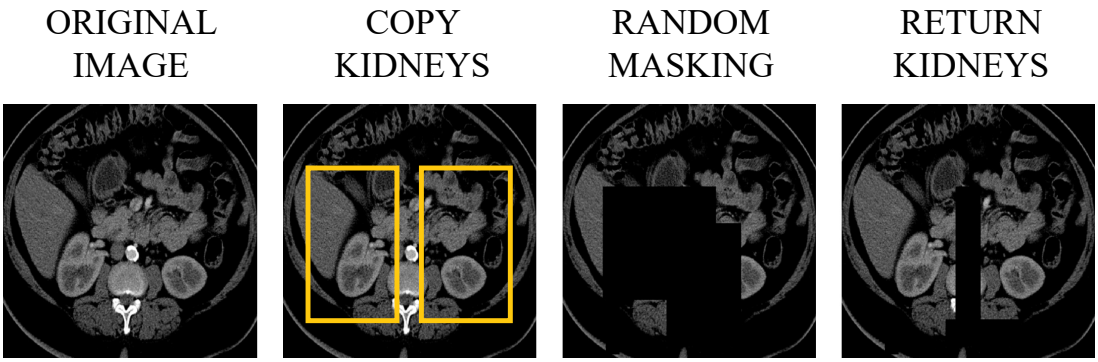


Figure 12. Unsupervised Augmentation Pipeline. Initial steps involve copying regions potentially containing kidneys (yellow rectangles) from the original image. Subsequently, three randomly selected and sized areas of the image are blacked out. The copied regions are then pasted back to preserve the integrity of the kidney areas, ensuring they are not impacted by the random removals.

3.6 Evaluation

The performance of the trained models was compared based on the highest achieved Accuracy and F1 scores on a testing set. The Accuracy reflects the proportion of correct predictions, both true positives and true negatives, in relation to the total number of predictions:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}, \tag{8}$$

where TP, FP, FN and TN are the numbers of True Positive, False Positive, False Negative and True Negative predictions, respectively. The F1 Score is particularly insightful when dealing with class imbalances, as it provides a balance between the precision and recall in a single metric:

$$F1_{score} = \frac{2 \times TP}{2 \times TP + FP + FN}, \tag{9}$$

3.7 Implementation Details

All methods and models were developed using Python 3.8.3 [42], PyTorch 2.2.0 [43], and CUDA version 12.1 [44]. The experiments were conducted on the High-Performance Computing (HPC) facilities at the University of Tartu, utilizing the same NVIDIA Tesla V100 GPU equipped with 32GB of VRAM. The text of this thesis was written using the Overleaf editor [45]. For brainstorming ideas, straightening, and debugging certain portions of the code, ChatGPT version 3.5 [46] was employed. The grammatical accuracy of the thesis was ensured through the use of Grammarly [47] and ChatGPT [46].

4 Results

The model trainings were carried out as detailed in Methods. Table 1 lays out the outcomes when different strategies were used to initialize the baseline model. Results from fine-tuning, after employing predictive, generative, and contrastive tasks as pretexts, are compared against the baseline in Table 2. Furthermore, Table 3 displays a comparison in performance between the baseline model's straightforward training and both supervised and unsupervised data augmentation strategies.

4.1 Model Initialization Comparison

In this study, two different methods of weight initialization were explored and their performance compared: random weight initialization and initialization with pretrained weights from the ImageNet dataset. Random initialization assigns weights in the model randomly at the start of training. In contrast, using pretrained ImageNet weights involves starting with weights that the model learned from a broad and diverse set of non-medical images in the ImageNet database. Despite the non-medical nature of the ImageNet dataset, the model initialized with these pretrained weights achieved a 0.20 higher accuracy and 0.23 higher F1 score, as detailed in Table 1. Thus, the model initialized from ImageNet weights was chosen as a baseline for this study.

Initial weights	Accuracy	F1
Random	0.69	0.66
Pretrained on ImageNet	0.89	0.89

Table 1. Comparison of the results of the model with different initialization weights.

4.2 Predictive Pretraining

The training for the rotation prediction task itself achieved an accuracy of 99.8%. However, when using the weights derived from this task to fine-tune the tumor classifier model, the performance deteriorated compared to the baseline. As detailed in Table 2, there was a 3% drop in accuracy, which was consistently observed across both classes. This uniform decrease in performance suggests that the decline was not due to bias in model learning.

4.3 Generative Pretraining

As mentioned earlier, three distinct inpainting tasks were implemented in this work: inpainting of randomly removed squares, inpainting squares removed from kidney regions,

and switching kidney regions. Illustrations of these image perturbation methods and the results from the trained models are displayed in Figure 13. The models managed to achieve satisfactory results, performing inpainting with considerable precision and quality, despite the models' constrained size.

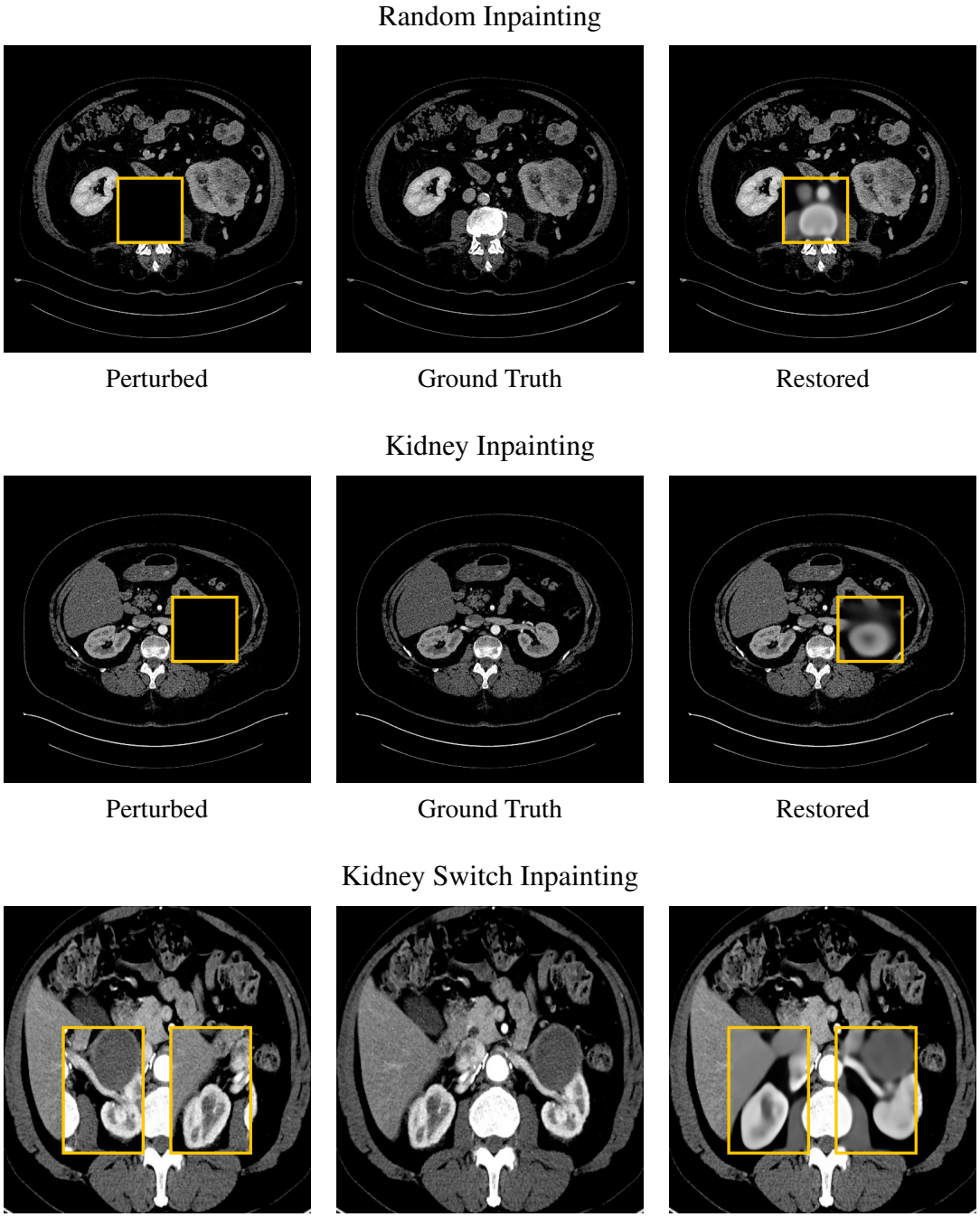


Figure 13. Inpainting Methods and Results. From left to right: Perturbed image, ground truth annotation, and the restored image by the trained model.

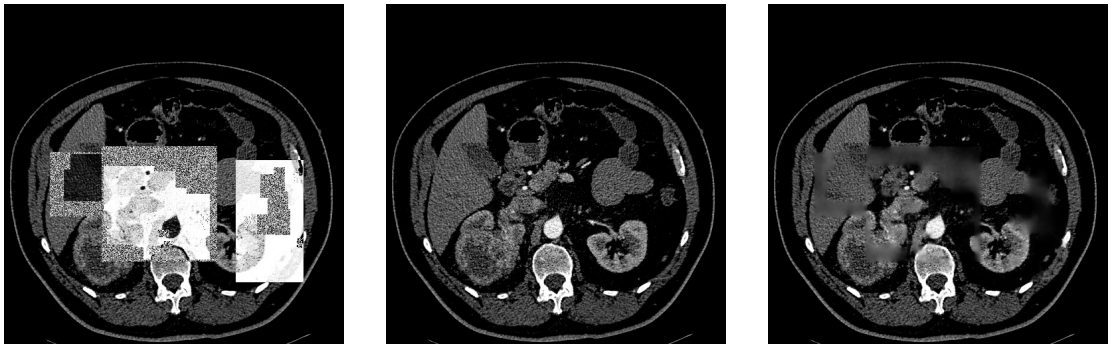
As discussed previously in Generative Pretraining section, the inpainting was conducted using two different loss functions for comparison: MSE and Dice. Although Dice loss is typically used for segmentation tasks with binary values, it proved more effective than MSE for these inpainting tasks, leading to less blurry outcomes as shown in Figure 14. The MSE tended to produce blurrier results due to its averaging effect on pixel values.

MSE loss Ground truth Dice loss

Figure 14. Comparison of MSE and Dice Loss Functions. The left image was inpainted using MSE loss, which resulted in blurriness, whereas the right image was treated with Dice loss, showing clearer and more precise inpainting results. The inpainted regions are highlighted in yellow. The original image is shown in the center for reference.

Despite achieving very good results in the pretext tasks, the classification models fine-tuned from weights obtained through random inpainting, kidney inpainting, and kidney switch inpainting performed 13%, 10%, and 11% worse in terms of accuracy, compared to the baseline model, respectively.

Although the image perturbations used in Model Genesis were more severe than those in the inpainting experiments described above, the model was still able to recover the initial images with good quality, as illustrated in Figure 15. Despite achieving very good results in the pretext task, the classification model, fine-tuned from the backbone extracted from the U-Net used in Model Genesis, achieved accuracy and F1 scores that were 5% lower than those of the baseline model.



Perturbed Ground Truth Restored

Figure 15. Example of Model Genesis Perturbations and the restored results by the trained model. The input image is shown in the center, the perturbed image on the left, and the restored image on the right.

4.4 Contrastive Pretraining

Assessing the SimCLR training was challenging because the only available metric was the loss. However, since the loss consistently decreased, it is reasonable to infer that the training was effective. Despite this, the classification model fine-tuned from the SimCLR-trained weights showed slightly lower accuracy than the baseline model, as detailed in Table 2.

Like SimCLR, the loss for Contrastive Predictive Coding decreased throughout the training, suggesting that the training process was effective. Nonetheless, the model fine-tuned from this training performed 11% worse in terms of accuracy and F1 score compared to the baseline. This indicates that while the loss metrics suggested efficient training, they did not directly translate to improved predictive performance on the target tasks.

Method	Accuracy	F1
Baseline	0.89	0.89
Rotation Prediction	0.87	0.86
Inpainting (Random)	0.76	0.74
Inpainting (Kidney)	0.79	0.76
Inpainting (Switch)	0.78	0.79
Model Genesis	0.84	0.82
SimCLR	0.87	0.86
CPC	0.78	0.78

Table 2. Comparative performance metrics of classification models fine-tuned from various pretraining methods. Accuracy and F1 score are calculated based on the binary classification outcomes for each method.

4.5 Data Augmentation

The supervised augmentation approach showed a 0.02 increase in accuracy and a 0.01 decrease in F1 score compared to the baseline model.

The unsupervised augmentation approach demonstrated a 0.03 increase in accuracy, while the F1 score remained unchanged.

Method	Accuracy	F1
Baseline	0.89	0.89
Supervised Augmentation	0.91	0.88
Unsupervised Augmentation	0.92	0.89

Table 3. Comparison of performance metrics between the baseline model and models trained with supervised and unsupervised data augmentation techniques.

4.6 Activation Visualization

Explainable AI approaches were utilized to investigate why none of the pretraining strategies, particularly inpainting, yielded positive results as initially expected. GradCAM was employed to illuminate the differences between models, revealing what activates them and investigating any potential biases that could have led to a drop in model performance. As depicted on Figure 16, Figure 17 and Figure 18, the activation maps generated by GradCAM clearly show that the models are focusing on tumors and kidneys in cases of positive predictions, which serves as a sensible validation of model behavior.

Despite the clear focus on relevant anatomical features, the activation regions and underlying reasons for model responses were consistent across different models, with

no consistent erroneous activations observed during an extensive review of the results. Consequently, GradCAM was applied to only a limited set of models, as it did not provide further insights into the differences between models.

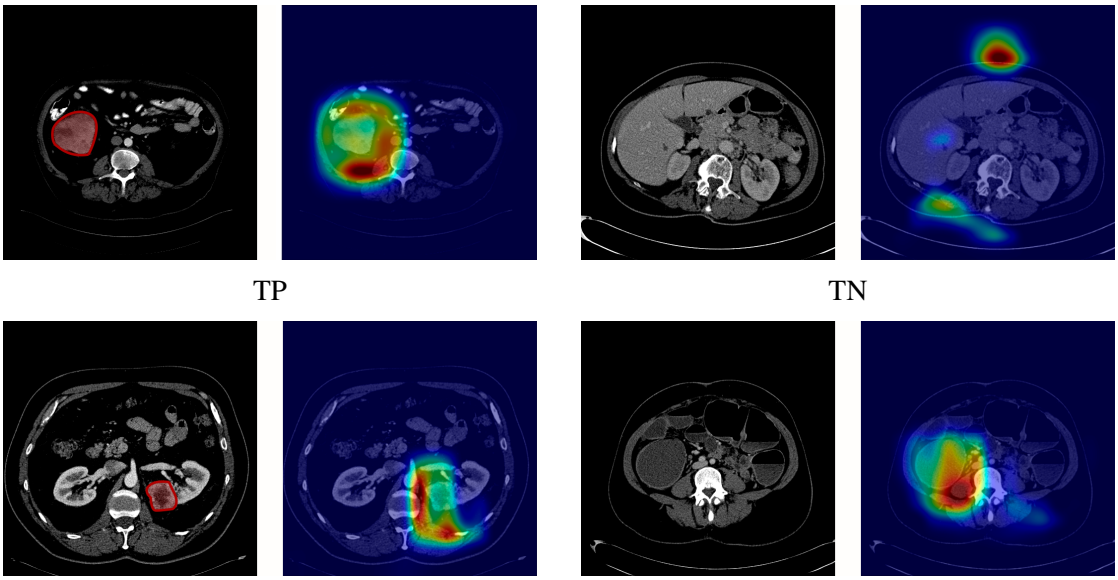


Figure 16. GradCAM Visualizations for the Baseline Model. Each pair consists of the original input image on the left, where tumors, if present, are highlighted in red, and the corresponding activation map on the right, which uses a color gradient from blue (least activation) to red (most activation). Captions below each image identify the case as TP (True Positive), FP (False Positive), FN (False Negative) and TN (True Negative).

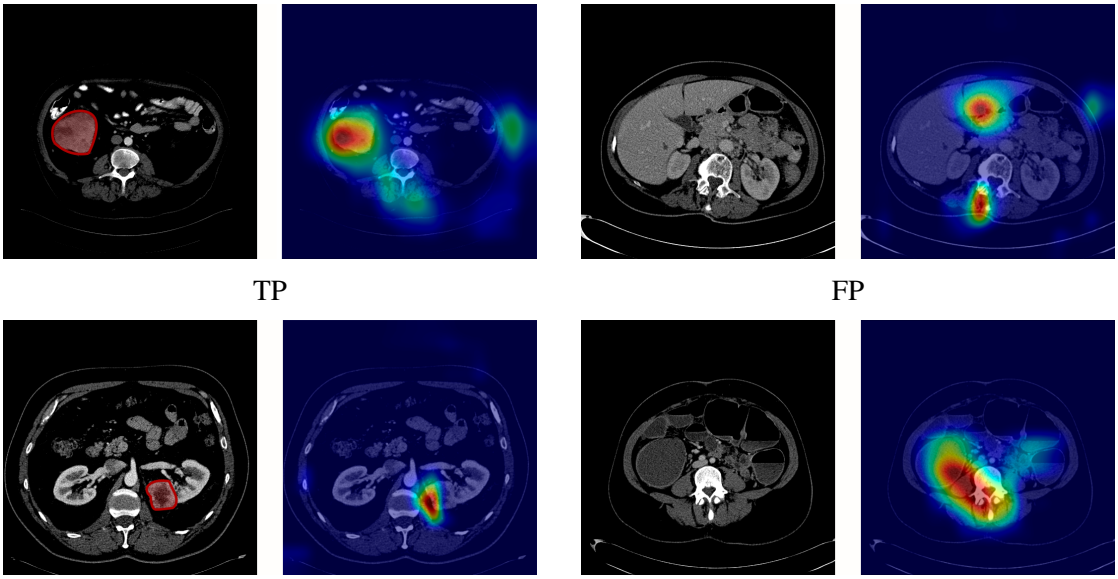


Figure 17. GradCAM Visualizations for the model fine-tuned from the Kidney Inpainting pretraining.

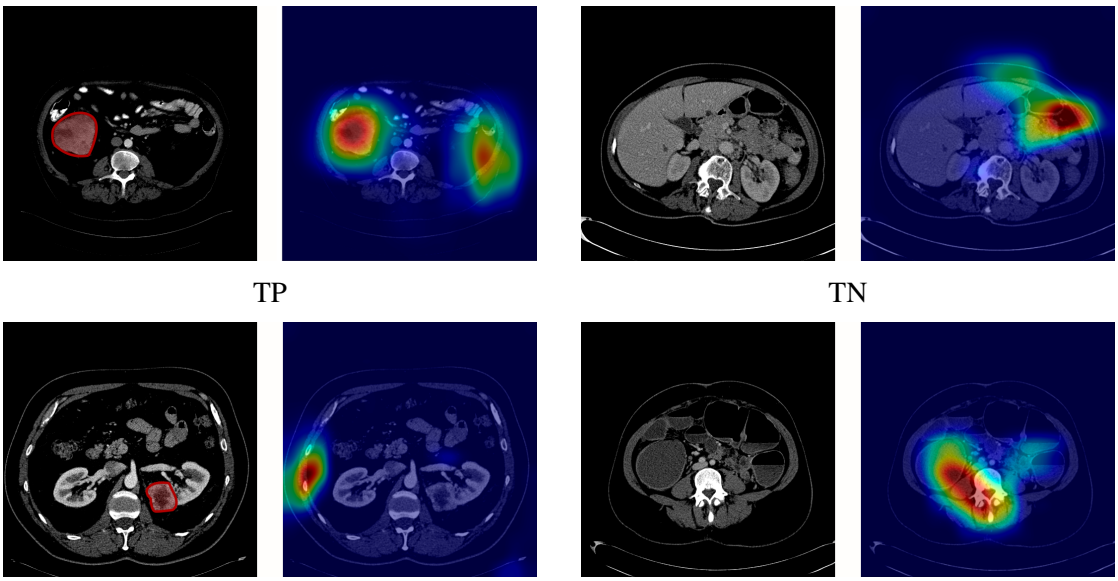


Figure 18. GradCAM Visualizations for the model fine-tuned from the Kidney Switch Inpainting pretraining.

5 Discussion

The results reveal that the pretraining methods explored did not enhance model performance significantly. Specifically, methods like Rotation Prediction and Inpainting not only failed to improve performance but actively deteriorated it. Inpainting, for instance, led to a sharp decline in model accuracy. Conversely, while Model Genesis and SimCLR also did not improve performance, they were less detrimental compared to the more direct approaches like Inpainting and CPC. This distinction highlights the importance of the context and specificity of pretraining tasks in medical imaging: even seemingly similar tasks like Inpainting and Model Genesis resulted in different levels of efficacy.

These results underscore the importance of the context and specificity of pretraining tasks in medical imaging. For instance, despite both employing image perturbation strategies, Inpainting and Model Genesis yielded divergent efficiencies, suggesting that the nature of the pretext tasks needs to be closely aligned with the ultimate diagnostic tasks. Similarly, the disparity in performance between CPC and SimCLR, both contrastive learning methods, further illustrates that subtle differences in task design can significantly affect outcomes.

An intriguing aspect of the study was the baseline model's superior performance, which was pretrained on ImageNet. Despite being a non-medical dataset, this approach outperformed other strategies, indicating that universal features learned from broader contexts might be more beneficial than inaccurately targeted features learned directly from medical data. This observation suggests that the quality of features learned is more critical than the dataset specificity, pointing to the potential superiority of generic pretraining over bespoke but less effective approaches.

Data augmentation methods, though not the primary focus of the thesis, showed consistent improvements in model performance. Both supervised and unsupervised augmentation techniques proved effective, enhancing the robustness and generalization of the models, which contrasts with the limited success of pretraining strategies. The unsupervised data augmentation success can be attributed in part to the utilization of spatial prior information. By manually identifying regions more likely to contain kidneys through an examination of numerous CT slices, the augmentation process was tailored to enhance the model's exposure to relevant variations. This strategy of leveraging specific prior information extracted by humans helped improve model generalization. However, the ultimate goal in machine learning is to develop models that can autonomously identify and utilize such prior information without extensive human intervention. The progress made with unsupervised augmentation suggests that embedding domain-specific knowledge into the learning process is beneficial, but future research should focus on methods that enable models to extract and apply this knowledge independently.

In conclusion, while unsupervised pretraining approaches like Model Genesis and SimCLR showed some promise, none surpassed the efficacy of the baseline model pretrained on ImageNet. This thesis contributes to the ongoing discourse in medical

image analysis by demonstrating that while pretraining can be beneficial, its success is highly contingent on the careful design of pretext tasks that are well-suited to the specific challenges of medical imaging. Future work should focus on exploring more sophisticated data augmentation techniques and machine learning approaches that can more effectively harness the rich, yet often underutilized, information present in medical images. This approach may provide a more fruitful avenue for achieving significant advancements in the field of medical diagnostics.

6 Conclusion

This thesis explored the effectiveness of various unsupervised pretraining approaches to improve the classification of medical images, focusing on CT scans for kidney tumor detection. Despite the potential benefits, their application in medical image classification remains minimally explored.

Multiple unsupervised pretraining methods were implemented and evaluated for their impact on model performance. The findings indicate that these methods did not enhance model performance. While unsupervised pretraining was expected to provide significant improvements, the results suggest its benefits are task-limited and context-dependent.

The study also demonstrated the positive effects of data augmentation techniques on model accuracy and robustness. Both supervised and unsupervised augmentation strategies showed improvements.

In summary, this thesis contributes to understanding the use of unsupervised pretraining methods in medical image classification. It highlights the challenges of enhancing model performance in this area and the need for continued research to address the limitations of sparsely annotated medical data.

7 Acknowledgements

I am very, very grateful to my supervisors, Dmytro Fishman and Joonas Ariva, for their guidance and help throughout this sometimes rocky research journey. I would also like to thank Dima and Daria for lending me their eyes and helping with advice. Additionally, thanks to the members of the BCV Lab, particularly Dmytro Shvetsov and Illia Tsiporenko, for their valuable recommendations.

References

[1] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, 2012.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015.

[5] Saeed Shurrab and Rehab Duwairi. Self-supervised learning methods and applications in medical imaging analysis: a survey. *PeerJ Computer Science*, 8, 2022.

[6] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised Visual Representation Learning by Context Prediction. 5 2015.

[7] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks. 6 2014.

[8] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context Encoders: Feature Learning by Inpainting. 4 2016.

[9] Richard Zhang, Phillip Isola, and Alexei A. Efros. Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction. 11 2016.

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. 2 2020.

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. Technical report.

[12] Shengfeng Liu, Yi Wang, Xin Yang, Baiying Lei, Li Liu, Shawn Xiang Li, Dong Ni, and Tianfu Wang. Deep Learning in Medical Ultrasound Analysis: A Review. *Engineering*, 5(2):261–275, 4 2019.

[13] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. 7 2018.

[14] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised Representation Learning by Predicting Image Rotations. 3 2018.

[15] Nicolas Ewen and Naimul Khan. ONLINE UNSUPERVISED LEARNING FOR DOMAIN SHIFT IN COVID-19 CT SCAN DATASETS. Technical report.

[16] Yujie Zhu Zhu. Self-supervised Learning for Small Shot COVID-19 Classification. In *ACM International Conference Proceeding Series*, pages 36–40. Association for Computing Machinery, 6 2021.

[17] S. Albelwi. Survey on self-supervised learning: Auxiliary pretext tasks and contrastive learning methods in imaging. *Entropy (Basel, Switzerland)*, 24(4):551, 2022.

[18] Yen Nhi Truong Vu, Trevor Tsue, Jason Su, and Sadanand Singh. An improved mammography malignancy model with self-supervised learning. In *Medical Imaging 2021: Computer-Aided Diagnosis*, volume 11597, pages 210–216. SPIE, 2021.

[19] Shih Cheng Huang, Anuj Pareek, Malte Jensen, Matthew P. Lungren, Serena Yeung, and Akshay S. Chaudhari. Self-supervised learning for medical image classification: a systematic review and implementation guidelines, 12 2023.

[20] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 1 2015.

[21] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. 12 2013.

[22] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. 6 2014.

[23] Yoshua Bengio. Deep Learning of Representations for Unsupervised and Transfer Learning. Technical report, 2012.

[24] Songping He, Yi Zou, Bin Li, Fangyu Peng, Xia Lu, Hui Guo, Xin Tan, and Yanyan Chen. An image inpainting-based data augmentation method for improved sclerosed glomerular identification performance with the segmentation model EfficientNetB3-Unet. *Scientific Reports*, 14(1), 12 2024.

[25] Jeffrey Dominic, Nandita Bhaskhar, Arjun D. Desai, Andrew Schmidt, Elka Rubin, Beliz Gunel, Garry E. Gold, Brian A. Hargreaves, Leon Lenchik, Robert Boutin, and Akshay S. Chaudhari. Improving Data-Efficiency and Robustness of Medical Imaging Segmentation Using Inpainting-Based Self-Supervised Learning. *Bioengineering*, 10(2), 2 2023.

[26] Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B. Gotway, and Jianming Liang. Models Genesis. 4 2020.

[27] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, Vivek Natarajan, and Mohammad Norouzi. Big Self-Supervised Models Advance Medical Image Classification. 1 2021.

[28] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. 6 2020.

[29] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. 7 2018.

[30] Michael U Gutmann. Noise-Contrastive Estimation of Unnormalized Statistical Models, with Applications to Natural Image Statistics Aapo Hyvärinen. Technical report, 2012.

[31] Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron van den Oord. Data-Efficient Image Recognition with Contrastive Predictive Coding. 5 2019.

[32] Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert. 3D Self-Supervised Methods for Medical Imaging. 6 2020.

[33] Evgin Goceri. Medical image data augmentation: techniques, comparisons and interpretations. *Artificial Intelligence Review*, 56(11):12561–12605, 11 2023.

[34] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning Deep Features for Discriminative Localization. Technical report.

[35] Harsh Panwar, P. K. Gupta, Mohammad Khubeb Siddiqui, Ruben Morales-Menendez, Prakhar Bhardwaj, and Vaishnavi Singh. A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-Scan images. *Chaos, Solitons & Fractals*, 140:110190, 11 2020.

[36] Nicholas Heller et al. The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct, 2023.

[37] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F. Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, and Klaus H. Maier-Hein. nnu-net: Self-adapting framework for u-net-based medical image segmentation, 2018.

[38] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[39] lightly-ai. Lightly: A python library for self-supervised learning. `https://github.com/lightly-ai/lightly`, 2023.

[40] SPEECHCOG. Contrastive predictive coding implementation in pytorch. `https://github.com/SPEECHCOG/cpc_pytorch`, 2023.

[41] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation. Technical report.

[42] Python Core Developers. Python documentation. Python Software Foundation, 2020. `https://www.python.org/doc/`.

[43] Pytorch: An imperative style, high-performance deep learning library. `https://pytorch.org`.

[44] NVIDIA. Cuda toolkit documentation. NVIDIA Corporation. `https://developer.nvidia.com/cuda-toolkit`.

[45] Overleaf: Collaborative writing and publishing system. Overleaf. `https://www.overleaf.com`.

[46] OpenAI. Chatgpt: Optimizing language models for dialogue. `https://openai.com/chatgpt/`.

[47] Grammarly: Free online writing assistant. Grammarly Inc., 2023. `https://www.grammarly.com`.

Appendix

Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Glib Manaiev**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Medical Image Classification with Limited Data,

   supervised by Dmytro Fishman and Joonas Ariva.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Glib Manaiev
15/05/2024