TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

Automatic detection of head and neck cancer from PET/MRI imaging using deep learning

Joonas Liedes

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. D OSA – TOM. 1947

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

AUTOMATIC DETECTION OF HEAD AND NECK CANCER FROM PET/MRI IMAGING USING DEEP LEARNING

Joonas Liedes

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. D OSA – TOM. 1947 | MEDICA – ODONTOLOGICA | TURKU 2026

University of Turku

Faculty of Medicine
Clinical Physiology and Nuclear Medicine
Doctoral programme in Clinical Research
Turku PET Centre
Turku University Hospital

Supervised by

Professor Jukka Kemppainen, MD, PhD
Department of Clinical Physiology and Nuclear Medicine
University of Turku and Turku PET Centre
Turku, Finland

Associate Professor Riku Klén, PhD
Imaging instrumentation and detection technologies
University of Turku and Turku PET Centre
Turku, Finland

Reviewed by

Associate Professor Antti Loimaala, MD, PhD
Nuclear Medicine Unit
Helsinki University Hospital
Helsinki, Finland

Professor Lalith Kumar Shiyam Sundar, PhD
Digital Transformation in Radiology
Ludwig Maximilian University of Munich
Munich, Germany

Opponent

Professor Elin Trägårdh, MD, PhD
Clinical Physiology and Nuclear Medicine
Department of Translational Medicine, Lund University
Skåne University Hospital
Malmö, Sweden

ISBN 978-952-02-0534-8 (PRINT)
ISBN 978-952-02-0535-5 (PDF)
ISSN 0355-9483 (Print)
ISSN 2343-3213 (Online)
Painosalama, Turku, Finland 2026

To my family

UNIVERSITY OF TURKU
Faculty of Medicine
Department of Clinical Medicine
Clinical Physiology and Nuclear Medicine
JOONAS LIEDES: Automatic detection of head and neck cancer from PET/MRI images using deep learning
Doctoral Dissertation, 130 pp.
Doctoral Programme in Clinical Research
February 2026

ABSTRACT

Detecting head and neck cancer (HNC) from medical images is a challenging problem due to the complex anatomy of the region and the heterogeneity of the disease. Traditional imaging techniques struggle to differentiate inflammation and scar tissue from tumours, especially with recurrent disease. Hybrid imaging (PET/CT and PET/MRI) utilises metabolic information to distinguish these and is routinely used. PET/MRI has gained popularity due to its improved soft-tissue contrast compared to PET/CT. Moreover, human interpretation of the images is complicated by inter- and intra-observer variability. Recently, deep learning (DL) has been proposed to mitigate these issues. DL has been shown to identify malignancies accurately in various medical imaging tasks using features learned from large annotated datasets. However, its usability in HNC diagnostics from ^{18}F-FDG PET/MRI images has not been thoroughly investigated.

This thesis investigated the applicability of DL in HNC ^{18}F-FDG PET/MRI diagnostics. First, a 2D (two-dimensional) segmentation model was evaluated in a small cohort of 44 patients containing positive and negative samples, yielding a Dice score of 0.84 ± 0.14 per slice for correctly detected lesions. However, the overall accuracy in detecting such lesions was 71%. Next, a 2D PET-only binary classifier was assessed with a cohort of 200 patients (50:50 positive/negative), achieving 78.6% accuracy and an AUC (area under the curve) of 85.1%. The study also indicated that certain subgroups of HNC were more likely to be classified correctly, depending on how frequently they appear in the training data. In addition, models trained with squamous cell carcinoma data only, were able to classify other HNCs accurately as well. A 3D (three-dimensional) classifier trained on the same cohort achieved an accuracy of 90% on a separate test set of 20 patients, compared with a radiologist who reached 100%. The interpretability of the model was examined using gradient-weighted class activation mapping. This method was found to provide useful insights into model decisions. Overall, DL shows promise in HNC PET/MRI analysis, though larger datasets and refined models are required for clinical use.

KEYWORDS: Deep Learning, PET/MRI, Head and Neck Cancer, Interpretable AI

TURUN YLIOPISTO
Lääketieteellinen tiedekunta
Kliininen laitos
Kliininen fysiologia ja isotooppilääketiede
JOONAS LIEDES: Pään ja kaulan alueen syöpien automaattinen tunnistaminen PET/MRI kuvista syväoppimista hyödyntäen
Väitöskirja, 130 s.
Kliininen tohtoriohjelma
Helmikuu 2026

TIIVISTELMÄ

Pään ja kaulan alueen syövän (HNC) tunnistaminen kuvantamisella on haastavaa alueen monimutkaisen anatomian ja vaihtelevan taudinkuvan vuoksi. Tulehduksen tai arpikudoksen erottaminen pahanlaatuisesta muutoksesta on toisinaan vaikeaa tavanomaisilla kuvantamismenetelmillä. Fuusiokuvantaminen (PET/TT ja PET/MRI) hyödyntää metabolista informaatiota ja helpottaa näiden hyvänlaatuisten muutosten erottelua pahanlaatuisista. PET/MRI:n suosio on ollut kasvussa viime aikoina verrattuna PET-TT:hen johtuen sen paremmasta pehmytkudosten erottelukyvystä. Kuvantamismenetelmien tulkintaa vaikeuttavat tulkitsijoiden väliset erot ja heikko toistettavuus. Syväoppimisella (DL) on saavutettu hyviä tuloksia syöpäkuvantamisen eri osa-alueilla hyödyntämällä ennalta opittua hahmontunnistusta laajoista koulutusaineistoista. Tutkimustieto sen käytöstä pään ja kaulan alueen syöpien ^{18}F-FDG PET/MRI kuviin on kuitenkin puutteellista.

Tämän väitöskirjan tavoitteena oli tutkia DL:n käyttöä HNC:n ^{18}F-FDG PET/MRI diagnostiikassa. Segmentoiva 2D-malli koulutettiin ja arvioitiin hyödyntäen 44 potilaan aineistoa, joka sisälsi sekä positiivisia, että negatiivisia kuvauslöydöksiä. Malli ylsi $0{,}84 \pm 0{,}14$ Dice-pisteisiin oikein tunnistettujen tuumoreiden osalta. Kokonaistarkkuus tuumoreiden tunnistuksessa oli 71 %. Luokitteleva 2D-malli koulutettiin käyttäen 200 potilaan aineistoa (50:50 positiivisia/negatiivisia). Paras malli saavutti 78,6 % tarkkuuden ja 85,1 % AUC:n. Tämä tutkimus osoitti myös, että tietyt HNC:n alatyypit luokitellaan sitä todennäköisimmin oikein, mitä useammin ne esiintyvät koulutusaineistossa. Toisaalta vain levyepiteelisyövillä koulutettu malli kykeni tunnistamaan tarkasti myös muita syöpiä pään ja kaulan alueelta. Samalla aineistolla koulutettu 3D-luokittelija saavutti 90 % tarkkuuden erillisellä 20 potilaan testijoukolla, jolla radiologi saavutti 100 %. Tämän mallin läpinäkyvyyttä pyrittiin tutkimaan hyödyntämällä Grad-CAM-tekniikkaa, joka antoi hyödyllistä tietoa mallin päätöksenteon taustoista. DL vaikuttaa lupaavalta menetelmältä HNC:n PET/MRI tulkinnan työkaluna, mutta suurempia koulutusaineistoja ja kehittyneempiä malleja vaaditaan ennen rutiininomaista kliinistä käyttöä.

AVAINSANAT: Syväoppiminen, PET/MRI, Pään ja kaulan alueen syöpä, Tulkittava AI

Table of Contents

Abbreviations

ADC	Apparent diffusion coefficient
AUC	Area under receiver operator characteristic curve
CNN	Convolutional neural network
CRT	Chemoradiation therapy
CT	Computerised tomography
DWI	Diffusion weighted imaging
DL	Deep learning
DSC	Dice score
EBV	Epstein-Barr virus
FDG	Fluorodeoxyglucose
FNA	Fine needle aspiration
FN	False negative
FOV	Field of view
FP	False positive
GAN	Generative adversarial network
GPU	Graphics processing unit
Grad-CAM	Gradient-weighted class activation mapping
GTV	Gross tumour volume
HNC	Head and neck cancer
HNSCC	Head and neck squamous cell carcinoma
HPV	Human papilloma virus
IMRT	Intensity-modulated radiation therapy
LR	Locoregional recurrence
MMFE	Multi-modality and multi-view feature extension method
ML	Machine learning
MRI	Magnetic resonance imaging
MTV	Metabolic tumour volume
NN	Neural network
PET	Positron emission tomography
RT	Radiation therapy
ROC	Receiver operating characteristic

ROI	Region of interest
SUV	Standardised uptake value
T1WI	T1-weighted imaging
T2WI	T2-weighted imaging
TN	True negative
TP	True positive

List of Original Publications

This dissertation is based on the following original publications, which are referred to in the text by their Roman numerals:

I Liedes, J., Hellström, H., Rainio, O., Murtojärvi, S., Malaspina, S., Hirvonen, J., Klén, R., Kemppainen, J. Automatic Segmentation of Head and Neck Cancer from PET/MRI Data Using Deep Learning. *Journal of Medical and Biological Engineering*, 2023; 43: 532–540.

II Hellström, H., Liedes, J., Rainio, O., Malaspina, S., Kemppainen, J., Klén, R. Classification of head and neck cancer from PET images using convolutional neural networks. *Scientific Reports*, 2023; 13, 10528.

III Liedes J., Hirvonen J., Rainio O., Murtojärvi S., Malaspina S., Klén R., Kemppainen J. Deep Learning-Based 3D Classification of Head and Neck Cancer PET/MRI: Radiologist Comparison and Grad-CAM Interpretability. *Clinical Physiology and Functional Imaging*, 2025; 5: e70030.

1 Introduction

Globally, head and neck cancer (HNC) is common with nearly 900 000 new cases in 2020 (Cancer (IARC), n.d.). These cancers can occur in various locations within the otorhinolaryngeal region, including the nasal cavity, sinuses, oral cavity, oropharynx, larynx, and salivary glands. Of these cancers, head and neck squamous cell carcinoma (HNSCC) constitutes the vast majority (Gormley et al. 2022). In addition to squamous cell carcinoma, other cancer types such as adenocarcinomas, lymphomas, blastomas, sarcomas and neuroendocrine tumours also occur in the head and neck region.

Due to the complex anatomy of the head and neck region, diagnosis is often difficult and time consuming. As Early-stage diseases are associated with higher survival rates, accurate and fast diagnosis is essential (Chow 2020). Yet it is not uncommon, that upon diagnosis the disease is already at an advanced stage and the treatment options are limited which impacts on poorer prognosis (Chow 2020). Similarly, any recurrences after treatment are challenging to diagnose due to possible fibrosis and inflammation (Specenier and Vermorken 2008). Traditional imaging methods like CT and MRI often struggle in differentiating between fibrosis, inflammation, and recurrences. However, ^{18}F- fluorodeoxyglucose (FDG) positron emission tomography (PET) combined with CT (PET/CT) has been found to be a reliable method for detecting recurrent disease, with improved fidelity in distinguishing fibrosis, inflammation, and cancer compared to traditional imaging methods (Kao et al. 2009). PET/MRI has also been suggested as an alternative to PET-CT due to its superior ability to visualise soft tissue and the absence of artifacts caused by metallic dental implants (Loeffelbein et al. 2012).

The anatomical information of CT or MRI is combined with the metabolic information of PET. The PET data is analysed visually and semi-quantitatively assessing the standardised uptake value (SUV) of specific regions or lesions. The SUV indicates the tissues propensity to intake the radiotracer used for the imaging, most commonly FDG. An elevated SUV is commonly seen with malignant processes. However, FDG-uptake is regularly observed also in the case of inflammation and benign tumours. There is no clear cut-off point to differentiate malignant processes from the benign. Therefore, the manual analysis of the images

is prone to inter-observer variability and low reproducibility which have negative implications on patient safety when planning treatment (Riegel et al. 2006). Thus, the need for an accurate and reproducible automatic method of assisting in this analysis is evident.

With increased computational capacity and data availability, deep learning (DL) has become a key approach in medical image analysis. Convolutional neural networks (CNNs) are a class of DL algorithms that learn hierarchical image representations and have played a major role in automated biomedical image segmentation (Krizhevsky et al. 2012; Ronneberger et al. 2015). CNN-based methods have also demonstrated expert-level performance in several medical imaging recognition tasks (Esteva et al. 2017; Kermany et al. 2018; Hwang et al. 2019). This study investigates CNN-based DL methods for automatic cancer detection and segmentation from PET/MRI images of the head and neck area.

2 Review of the Literature

2.1 Characteristics of head and neck cancer (HNC)

This chapter gives an overview of HNC: how HNC is defined, diagnosed, and treated, to contextualise the imaging pathways and clinical decision points that motivate the PET/MRI-based detection and segmentation tasks addressed in this thesis.

2.1.1 Definition, aetiology, incidence and risk factors

Head and neck cancer (HNC) encompasses a broad range of malignant tumours in a variety of areas within the head and neck region. These areas include the paranasal sinuses, nasal cavity, nasopharynx, oral cavity, oropharynx, hypopharynx, larynx, salivary glands, thyroid and the parathyroids. However, not all clinical definitions, such as the one by International Agency for Research on Cancer, include thyroid cancers (Gormley et al. 2022). In addition to the affected area, the histology varies greatly. While approximately 90% of HNCs are squamous cell carcinomas (HNSCC) arising from the mucosal epithelium of the oral cavity, pharynx and larynx, other types of histology are observed as well. These include adenocarcinoma, mucoepidermoid carcinoma, and adenoid cystic carcinoma which often stem from the salivary glands. In addition, lymphoma can originate from the lymph nodes of the head and neck region. While thyroid cancers are not always grouped with HNCs, they are common and include several subtypes such as papillary, follicular, medullary, and anaplastic thyroid cancer. In this thesis, thyroid malignancies were included in the study dataset and were analysed alongside other head and neck cancers. Due to the diverse nature of HNCs, they are typically categorised based on their anatomical location using the International Classification of Diseases (ICD-10) by the World Health Organisation (WHO) (Gormley et al. 2022).

In 2020, there were more than 930 000 new HNCs excluding thyroid cancers. With over 1.5 million cases combined, HNC ranks as the seventh or third most common cancer globally, depending on the definition (Sung et al. 2021). The incidence of HNC has been on the rise and is expected to increase by 30% by the end of the decade (Johnson et al. 2020), thus making it a significant concern for public health.

As HNSCC constitutes the vast majority of the cases, understanding risk factors associated with it is key. Alcohol and tobacco consumption are considered as the primary factors responsible for the onset of HNSCC, with an added multiplicative risk when used in combination. Recently, infections with oncogenic Human Papilloma Virus (HPV) strains, HPV-16 in particular, have been recognised as an additional major risk factors for HNSCC of the oropharynx. Epstein-Barr Virus (EBV) infection is also recognised as a specific risk factor for nasopharyngeal carcinoma. Certain occupations, such as cooks, cleaners, and painters, carry a slightly elevated risk of HNSCC in the nasal cavity and paranasal sinuses due to exposure to harmful chemicals like paint fumes, asbestos, and nickel. Furthermore, gastroesophageal reflux disease is associated with a moderate increase in risk for squamous cell carcinoma of the larynx. It is noteworthy that alcohol and tobacco remain by far the most important risk factors, with HPV-16 being an additional major risk factor for oropharyngeal HNSCC (Gormley et al. 2022; Johnson et al. 2020; Anis et al. 2018). The key risk factors for the cancers of the salivary glands are alcohol and tobacco consumption, exposure to radiation and nickel alloys and employment in the rubber industry (Horn-Ross et al. 1997). For thyroid cancers exposure to radiation is the most important risk factor (Pellegriti et al. 2013). Major risk factors for parathyroid cancer are familial hyperparathyroidism, multiple endocrine neoplasia type 1 and irradiation to the head and neck area (Koea and Shaw 1999). The different HNC types are summarised in Table 1.

Table 1. Summary of different HNCs. Abbreviations: EBV (Epstein-Barr Virus); GERD (gastroesophageal reflux disease); LPR (laryngopharyngeal reflux disease)

Location	Histology	Risk factors
Nasal cavity and the paranasal sinuses	Squamous cell carcinoma	Tobacco, alcohol, occupational chemical exposure
Nasopharynx	Squamous cell carcinoma, non-keratinising carcinoma (differentiated or undifferentiated) and basaloid squamous cell carcinoma	EBV infection, alcohol, tobacco
Oral cavity and the oropharynx	Squamous cell carcinoma	Tobacco, alcohol, HPV infection
Hypopharynx and larynx	Squamous cell carcinoma	Tobacco and alcohol use, GERD and LPR
Salivary glands	Adenoid Cystic Carcinoma, Mucoepidermoid Carcinoma, Acinic cell carcinoma Adenocarcinoma	Alcohol, tobacco, occupational radiation exposure
Thyroid and the parathyroids	Papillary, Follicular, Medullary, Anaplastic Carcinoma	Radiation exposure

2.1.2 Clinical Presentation

HNSCC is by far the most common subtype of HNC, and it can be further divided into subgroups based on viral aetiology. These groups are HPV-negative and HPV-positive HNSCC. Median age at diagnosis for HPV-positive patients is approximately 53 years and 66 years for HPV-negative patients, with men having higher risk for both subgroups. The presenting symptoms vary greatly and are dependent on the anatomical location and aetiology of the tumour (Johnson et al. 2020).

HPV-negative HNSCC of the hypopharynx and oropharynx usually lead to symptoms at a later stage due to their obscure location. The patients might exhibit dysphagia, odynophagia or otalgia. Patients with HPV-negative pharyngeal tumours frequently have a history of tobacco and alcohol use. While the decrease in smoking in has had a positive effect on the incidence of HPV-negative HNSCC, it remains a substantial problem worldwide. Moreover, the incidence of HPV-positive HNSCC of the oropharynx has surged in recent years accounting for many new HNSCC cases, especially in North America and Western Europe (Chaturvedi et al. 2011). Compared to the HPV-negative disease, the HPV-positive patients have generally milder symptoms or are even asymptomatic. HPV-positive HNSCC is associated with a better prognosis and the primary risk factors are male sex and multiple sexual partners (Marur et al. 2010).

Patients with tumours of the larynx often present with voice changes or hoarseness, which might lead to an earlier diagnosis. Symptoms in later stages include dyspnoea and airway obstructions. Laryngeal HNSCC is heavily associated with tobacco and alcohol consumption. Patients with HNSCC of the oral cavity often present with a persistent mouth sore or an ulcer that causes pain while speaking and eating. These tumours are commonly associated with smoking, alcohol consumption and poor dental care. Nasopharyngeal HNSCC is often associated with EBV infection (Tsao et al. 2017). These patients typically present with frequent epistaxis and unilateral nasal obstruction.

2.1.3 Diagnosis

A thorough patient interview and a physical examination conducted by a qualified physician are the cornerstones of HNC diagnosis. Patients deemed at risk for malignancy, should undergo contrast-enhanced computed tomography (CT) or magnetic resonance imaging (MRI) to differentiate benign masses from malign. In addition, imaging aids in the planning of fine needle aspiration (FNA) or biopsy procedures, facilitates staging, uncovers hidden disease, and informs treatment decision-making. Upon suspicion of HNC, the diagnosis must be confirmed with a biopsy of the primary tumour or neck mass. If a primary tumour is present, the biopsy

is typically obtained with cup forceps, incisional or excisional biopsy. However, FNA is recommended for patients with suspected malignant neck mass, as open biopsy risks tumour seeding and locoregional recurrence (LR) (Pynnonen et al. 2017).

After obtaining the diagnosis, the disease is staged using the tumour–node–metastasis (TNM) system (Amin et al. 2017). The system was revised in 2017 to include depth of invasion to tumour staging for cancers of the oral cavity, extracapsular nodal extension to nodal staging in non-viral HNSCC and codification of a novel staging system for HPV-positive HNSCC. Nasopharyngolaryngoscopy, CT and MRI are typically used to determine the extent of the disease and stage. In addition, positron emission tomography (PET) is used for staging due to its increased sensitivity in detecting disease that is not palpable or detectable through direct visualisation (Laubenbacher et al. 1995). Moreover, PET is valuable in the management of unknown primary carcinoma of the head and neck region (Miller et al. 2005). In contemporary medical practice, it is common to combine anatomical imaging modalities, such as CT or MRI, with PET scanning to enhance the accuracy of disease staging (Szyszko and Cook 2018). With the recently revised staging system, HPV testing is an essential part of staging for HNSCC tumours, as it affects prognosis significantly (Chaturvedi et al. 2011; Marur et al. 2010).

2.1.4 Treatment

Choice of treatment depends on anatomical location, stage, disease characteristics and functional considerations. Surgery, radiation and chemotherapy are the main curative options for locoregionally confined HNSCC. Single modality treatment is typically sufficient for early-stage diseases and leads to high cure rates. Tumours of the oral cavity are often treated with surgery, whereas pharyngeal and laryngeal disease are usually treated with radiation. HNSCC patients with more advanced tumour or nodal stage often benefit from postoperative radiation or chemoradiation, especially if any pathological risk factors are present. These high-risk features include extra-nodal extension, close or involved resection margins, or perineural invasion indicates an increased risk of recurrence. When these factors are present, the patient will benefit from chemotherapy administered concurrently with radiation (Cooper et al. 2004). Definitive chemoradiation therapy is commonly recommended as a non-surgical alternative for patients with advanced tumour stages (≥T3) or multiple involved nodes (>N1) and cases where function preservation is essential. Some patients with recurrent or metastatic HNSCC can be treated with salvage surgery, re-irradiation and metastasectomy. Immunotherapy, such as pembrolizumab, is considered for the remaining patients.

2.2 PET/MRI in HNC

This section discusses the operating principles of PET/MRI and how it is currently utilised in HNC diagnostics.

2.2.1 Principles of PET/MRI

2.2.1.1 PET

PET scans rely on the use of radioactive tracers to provide functional information about the imaging subject. These tracers may include metabolic molecules such as glucose or amino acids, or cell surface receptor-targeted ligands such as the prostate specific membrane antigen and somatostatin receptors (Dota -peptides), all labelled with a positron-emitting isotope (Rong et al. 2023). The ^{18}F labelled glucose analogue FDG is the most used tracer. The tracer is injected into a peripheral vein and as the tracer decays, it emits positrons that annihilate with electrons producing photons travelling to opposite directions. These photons are then detected by the scanner to produce a 3D image of the tracer distribution in the subject.

The tracer molecule is chosen based on the function wanted to examine. The molecule should exhibit elevated uptake levels of the tracer compared to the surrounding tissues, thereby providing adequate contrast for the analysis. FDG is commonly used as the tracer to differentiate between benign and malignant tissues in cancer patients. Malignant tumours metabolise glucose at a faster rate than benign tumours, thus having increased uptake levels of FDG, which are then reflected as areas of high contrast in the PET scan. These uptake levels are often correlated with the aggressiveness of the tumour that can be quantified with the Ki67-index, which is based on the nuclear protein Ki-67 commonly used as a marker for cell proliferation (Scholzen and Gerdes 2000). To achieve sufficient contrasts PET scans leverage the process of metabolic trapping, where the composition of the FDG molecule prevents it from undergoing normal glucose metabolism and causes it to get trapped within the cell (Gallagher et al. 1978). This process is compounded by the Warburg effect present in most malignant tumours, where the cells have high glucose uptake and prefer to metabolise it through aerobic glycolysis instead of the citric acid cycle (Vander Heiden et al. 2009).

Due to its ability to accurately depict metabolic activity and differentiate benign and malign tissues, PET is commonly used in oncology to identify and characterise tumours, disease staging and monitoring treatment response. PET scans provide semi-quantitative and quantitative measurement of the metabolic activity. The SUV is typically used as a semi-quantitative measure of the tissue's propensity to the radiotracer. It represents the proportion between the radioactivity concentration in

the tissue at a given moment and the amount of radioactivity injected, adjusted for the patient's body weight (Thie 2004). For quantitative measurements, dynamic imaging and tissue specific time-activity curves are required. In practice, diagnostic images are usually static images obtained at a predetermined time post-radiotracer injection. The time varies based on the tracer used. For instance, FDG images are taken approximately 50 minutes after injection. The basic operating principle of PET imaging is shown in Figure 1 (Rong et al. 2023).

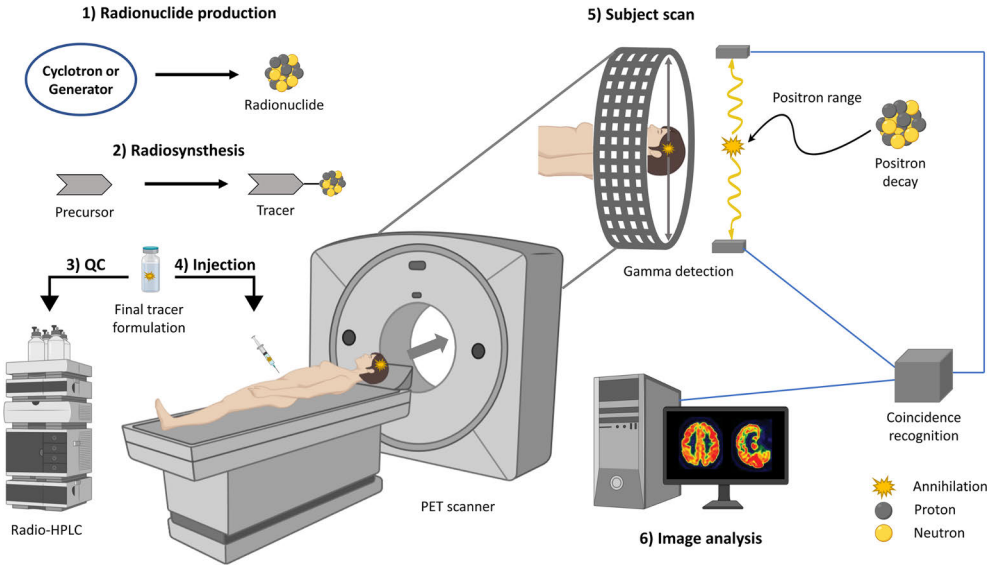

Figure 1. Principle of positron emission tomography (PET) imaging. Reproduced from Rong et al. 2023, under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/) (Rong et al. 2023). No changes were made. Abbreviations: High-Pressure Liquid Chromatography (HPLC), Quality control (QC).

2.2.1.2 MRI

MRI utilises the principle of nuclear magnetic resonance to obtain high resolution anatomical images. A strong magnetic field is used to align hydrogen nuclei (protons) with that field after which, a radiofrequency current is directed through the patient causing the protons to deviate from the equilibrium. The energy released by the protons realigning with the magnetic field can then be detected, after the pulse of radiofrequency has been turned off. Different tissues have different molecular compositions, hence varying in the time and energy it takes to realign with the magnetic field. These differences are then used to discern between different tissues (Vlaardingerbroek and Boer 2013).

MRI offers some distinct advantages over traditional X-rays and CT. First, its working principle does not rely on ionising radiation. Secondly, as hydrogen atoms are abundant in the body in the form of water, the soft tissue contrast of MRI is superior. In recent years, these factors have contributed to its popularity in HNC diagnostics and monitoring treatment response. Its use in hybrid imaging in conjunction with PET has also been proven beneficial compared to PET-CT imaging (Szyszko and Cook 2018; Murtojärvi et al. 2022). However, this does come with the price of being significantly slower than CT.

Two primary MRI sequences are used in HNC imaging, The T1-weighted imaging (T1WI) and the T2-weighted imaging (T2WI). T1WI is based on the T1 relaxation time, which is the time it takes for the protons to realign with the magnetic field after the perturbation caused by the radiofrequency current. In T1WI, tissues with short T1 times appear brighter and those with longer T1 times appear darker. For instance, fat appears bright in T1WI, and water appears dark. T1WI offers excellent anatomical detail and is useful for delineating tumours when used in conjunction with a contrast agent such as gadolinium. Contrarily, T2WI is based on the T2 relaxation time, which is the time it takes for the protons to lose their coherent phase, or the transverse component of the net magnetisation vector, induced by the radiofrequency pulse. Tissues with longer T2 times appear brighter in T2WI than those with shorter T2 times. For example, water has a long T2 time and appears bright in T2WI, thereby making T2WI especially suitable for evaluating conditions with increased water presence, a common feature in inflammation and malignant tumours. Thus, T2WI is often useful in visualising HNC and evaluating peritumoral oedema and inflammation.

An additional MRI technique, called diffusion weighted imaging (DWI), is often used to complement the forementioned sequences. DWI is based on measuring the random motion of water protons to produce images. Water molecule diffusion is affected by its surroundings such as cellular membranes, macromolecules and cytoplasm. The differences in the way these factors influence the diffusion can then be quantified with apparent diffusion coefficient (ADC) maps (Bammer 2003). In HNC, DWI can be used to differentiate benign from malign tissue as cancerous tissue typically shows high signal intensity and associated low ADC values because of their hypercellular nature. Similarly, it can also be used to differentiate necrotic tumours from abscesses and inflammatory nodes from metastatic ones. In addition, DWI is utilised for post-treatment response assessment by differentiating treatment related changes such as fibrosis from recurrent or residual disease. Moreover, pretreatment ADC values can predict the likelihood of treatment response (Connolly and Srinivasan 2018).

2.2.1.3 PET/MRI

PET/MRI can provide complementary metabolic and high–soft-tissue-contrast anatomical information by combining PET with MRI, and integrated systems enable simultaneous acquisition of both datasets (Judenhofer et al. 2008; Musafargani et al. 2018). Simultaneous acquisition supports accurate spatial alignment between modalities and, in integrated systems, has demonstrated low registration variance (Judenhofer et al. 2008). This combination can improve lesion characterisation and may increase diagnostic confidence in selected applications (Musafargani et al. 2018). Compared with PET/CT, PET/MRI avoids CT-related ionising radiation, which can reduce overall examination radiation burden.

The acquisition of PET/MRI can be achieved with either sequential systems or integrated systems (Musafargani et al. 2018). Sequential systems utilise separate scanners located in proximity, typically in the same room. Because of the sequential design, there is no need to address the issue of mutual interference between the modalities, therefore making these systems far less complex than the integrated systems and thus more economical (Musafargani et al. 2018). However, the temporal and spatial correlation between the PET and MRI data suffers when acquired sequentially. The integrated PET/MRI systems are capable of simultaneous acquisition of PET/MRI data, thus reducing this issue (Judenhofer et al. 2008; Musafargani et al. 2018). The drawback to these systems is the need to address the problem of mutual interference between the modalities, making them technically significantly more complex (Judenhofer et al. 2008; Musafargani et al. 2018).

2.2.2 Applications of PET/MRI in HNC

Due to its excellent soft tissue contrast and valuable molecular level metabolic information, PET/MRI has a multitude of applications in HNC diagnostics. Its value in initial diagnosis and staging was shown by Samolyk-Kogaczewska et al. who compared FDG-PET/MRI with CT on 21 HNC patients that received surgical treatment. The authors declared that the outcomes obtained using FDG-PET/MRI where more consistent with the histopathology results than those of the CT images (Samolyk-Kogaczewska et al. 2020). Similar promise in initial staging of HNC with PET/MRI was found by Lee et al. in a cohort of 10 histologically proven HNCs (Lee et al. 2014). FDG-PET/MRI was found to be on par with FDG-PET/CT in more heterogeneous cohorts involving both initial staging and re-staging as well (Szyszko and Cook 2018; Felix P. Kuhn et al. 2014; Kubiessa et al. 2014). S. Partorvi et. al compared FDG-PET/MRI with FDG-PET/CT in terms of lymph node and distant metastasis detection in a similar patient cohort with staging and re-staging scans. They found that FDG-PET/MRI and FDG-PET/CT provide comparable results (Partovi et al. 2014). A retrospective PET/MRI image fusion in a group of HNC

patients, who underwent preoperative MRI and PET/CT for staging, was compared to single modality PET and MRI by Loeffelbein et al. who found that a retrospective fusion was only beneficial in individual recurrent cases (Loeffelbein et al. 2014).

The value of FDG-PET/MRI has been studied for re-staging purposes only in a handful of articles. In a study conducted by Queiroz and colleagues, PET/MRI and contrast-enhanced PET/CT were compared in 87 patients with suspected recurrent head and neck cancer. Although PET/MRI didn't demonstrate superior accuracy, it aided in identifying tumour recurrence, specifying unclear FDG uptake related to possible tumour recurrence (Queiroz et al. 2014). Varoquaux et al. compared the re-staging performance of a sequential PET/MRI with a PET/CT scan in a group of 32 HNC patients (Varoquaux et al. 2014). PET/MRI achieved equivalent performance with PET/CT in terms of qualitative results. However, the authors state that PET/MRI tended to underestimate SUVs in comparison with PET/CT. SUV discrepancy was associated with tumour size, with larger, higher-uptake tumours showing greater PET/MR underestimation. This was attributed in part to technical differences in attenuation correction. Schaarschmidt et al. investigated the accuracy of an integrated PET/MRI for locoregional HNC evaluation compared to PET/CT and MRI. The cohort consisted of 25 HNSCC patients, of which 13 had suspected recurrence and underwent re-staging. No significant differences were found between the modalities (Schaarschmidt et al. 2016). Becker et al. assessed the diagnostic value of PET/MRI with diffusion weighted imaging (PET-DWIMRI) in a prospective group of 74 HNSCC patients with suspected recurrence after (chemo)radiotherapy. The diagnosis was confirmed with either histology in 46 patients, or a mean follow-up of 34 ± 8 months for 28 patients. The authors conclude that PET-DWIMRI yields excellent results in detecting HNSCC recurrence (Becker et al. 2018). Murtojärvi et al. investigated the potential benefits of PET/MRI vs. PET/CT in chemoradiotherapy follow-up imaging. The study cohort consisted of 104 HNSCC patients all of whom had received chemoradiotherapy and underwent either PET/MRI (n=52) or PET/CT (n=52). The results indicated that PET/MRI demonstrated higher sensitivity and superior negative predictive value in detecting LR. PET/CT produced both false-negative and false-positive findings, particularly in advanced disease stages, where PET/MRI showed significantly better performance. Furthermore, PET/CT had false negative findings in the oropharyngeal, laryngeal and nasopharyngeal regions, whereas PET/MRI made no false negative interpretations in these areas. Thus, the authors summarise that PET/MRI might be considered as the modality of choice in HNSCC re-staging, particularly in more advanced disease stages of the oral cavity, larynx and nasopharynx (Murtojärvi et al. 2022).

2.3 Artificial intelligence (AI)

This chapter briefly outlines the basics of AI and deep learning with convolutional neural networks.

2.3.1 AI and machine learning (ML)

Artificial intelligence (AI) can be broadly defined as any machine process, especially computer systems, capable of performing tasks that would normally require human intelligence. AI is often categorised into two subtypes based on its range of capabilities. These subtypes are narrow and general AI. Narrow AI is a system designed to do a specific task, such as image segmentation, and is considered "intelligent" only in this limited context. Contrarily, general AI, or artificial general intelligence (AGI), is a system capable of performing any task a human being would. These systems can learn, adapt and apply knowledge in a wide variety of tasks. Currently, AGI systems do not exist.

Machine learning (ML) is a subset of narrow AI that involves finding patterns from data and making predictions based on these patterns. ML algorithms leverage data to automate analytical model building. These algorithms are fed training data from which they extract patterns that can then be used to give predictions on unseen data sets. In contrast to traditional computer programming, the key characteristic of this approach is the algorithm's ability to make predictions without being explicitly programmed to give them. With ML algorithms, data is used to minimise the amount of human intervention required in the process (Bi et al. 2019). ML can be further subcategorised based on the learning method used. These methods include supervised, unsupervised and reinforcement learning.

In supervised learning the algorithm is fed labelled training data that includes both the data and their labels, which depict the desired output pattern(s) the algorithm is intended to learn. The labels are usually given by human experts and the goal of the algorithm training process is to map the inputs to the outputs (Bi et al. 2019). Supervised learning is further divided into regression and classification algorithms, the difference being the output distribution, where regression is continuous and classification is discrete (Bi et al. 2019). Common supervised learning algorithms include Linear Regression, Decision Trees, Support Vector Machines, and Neural Networks.

Unsupervised learning differs from supervised learning in that no labels are included with the training data. Rather, the algorithm tries to find underlying patterns and structures in the data (Bi et al. 2019). Examples of unsupervised learning algorithms include K-Means Clustering, Hierarchical Clustering, and Principal Component Analysis.

In reinforcement learning the algorithm includes an agent interacting with its environment. The agent is the rewarded for correct decision making and penalised for incorrect decisions. The rewards and penalties then guide the learning process of the algorithm (Bi et al. 2019).

2.3.2 Deep learning (DL)

Deep learning (DL) is a subset of ML that centres around the utilisation of deep neural networks. They exhibit an expanded scale characterised by the presence of multiple hidden layers and associated optimisable parameters. This increased structural complexity allows them to capture increasingly abstract representations of the input data. Consequently, DL facilitates representational learning, wherein the algorithms autonomously extract meaningful patterns without the need for manual feature extraction as was customary in conventional ML approaches (LeCun et al. 2015).

2.3.2.1 Neural networks (NNs)

Neural networks (NNs) are a type of machine learning algorithm inspired by the structure and function of the human brain. These networks consist of interconnected neurons or nodes. A neuron takes in an arbitrary number of inputs and produces an output. The connected neurons form structures called layers. NNs include an input layer, an output layer and between them an arbitrary number of hidden layers. The term hidden stems from the fact that the values and computation performed in the neurons of these layers are not directly observable in the output, nor are they directly affected by the input. These layers handle the bulk of the computation contributing to the final output (Bi et al. 2019). A schematic representation of a NN is depicted in Figure 2.

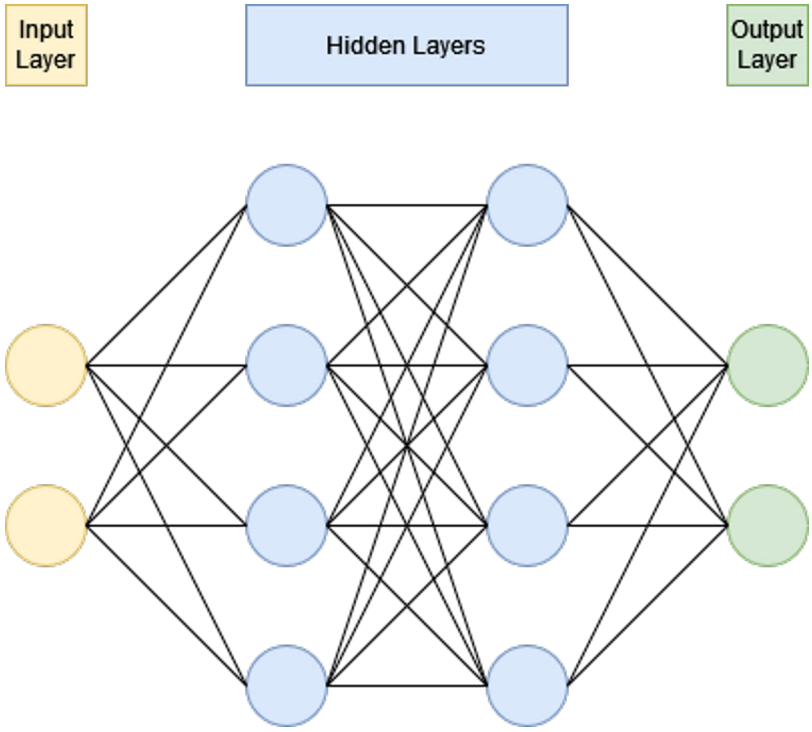

Figure 2. Schematic representation of a neural network.

The connections between the neurons are represented with the terms weight and bias, which are adjusted during training to fit the data. Each connection between two neurons has its own weight and bias. For an arbitrary neuron these terms can be represented as a mathematical formula in the following way:

$$y = \sum_{i=1}^{n} (W_i X_i) + b$$

Where W depicts the weights from the connecting neurons of the previous layer and X the inputs. The bias term is represented as b. The output y is then processed by an activation function that determines whether the neuron is activated or not (Krogh 2008). Typical activation functions used in NNs are for instance sigmoid, ReLU and softmax functions. The computation of a single neuron is depicted in Figure 3.

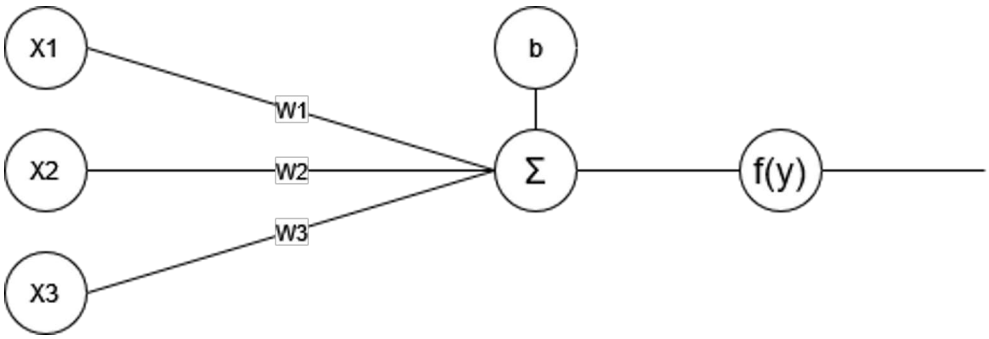

Figure 3. The computation of a single neuron in a neural network.

The training process of a NN involves updating the weight and bias parameters of the neurons iteratively by exposing the NN to the training data. This data consists of the inputs and corresponding desired outputs often called labels, or masks in the case of image segmentation. The goal of this process is to minimise the difference between the desired output and the predicted output. The difference is quantified with a loss function, which acts as measure of the NNs performance. The process involves four steps that can be defined as follows (Krogh 2008; LeCun et al. 2015; Goodfellow et al. 2016):

1. **Forward Propagation**: The NN is given an input, which is processed by each layer of neurons to produce an output. Each neuron takes a weighted sum of its inputs, adds a bias, and applies an activation function to produce its output. Each layer's output functions then as the subsequent layer's input until the output layer is reached.
2. **Loss Calculation**: The difference between the NNs output and the desired output is then measured using the loss function. The result quantifies how well the network's predictions match the desired outputs.
3. **Backpropagation**: The loss is then propagated back through the network. This involves calculating the gradient of the loss function with respect to each weight and bias in the network. This is achieved by calculating the derivative of the loss function in terms of the layer's outputs and the weights and biases, which is then propagated to the previous layers using the chain rule of calculus.
4. **Weight and Bias Update**: The weights and biases of the neurons are adjusted to minimise the loss. This is commonly done using an optimisation algorithm such as gradient descent. The extent by which the weights and biases are adjusted each time is controlled with the learning rate, a parameter that determines how quickly the network learns.

Upon completion of the training phase, the performance of the neural network is typically assessed with cross-validation utilising a distinct dataset, commonly referred to as the validation set, which was not employed during training. It is critical during this evaluation phase to identify potential instances of overfitting or underfitting. Overfitting occurs when the network models the training data excessively well, to the point that it fails to generalise to new, unseen datasets. Conversely, underfitting is a condition in which the network demonstrates an insufficient capacity to learn from the training data.

In cross-validation, the dataset is separated randomly into training and validation sets with the aim of using the majority of the data for the training while still yielding a validation set that adequately represents the whole dataset. With small sample sizes this becomes an issue, as a single random holdout might lead to a distribution of data that causes the model either to overfit or underfit. To prevent this, k-fold cross-validation is often used to account for the randomness. In k-fold cross-validation the holdout is repeated k times ensuring each subset of the data serves as the test set once yielding a more robust estimate of the performance (Goodfellow et al. 2016).

2.3.2.1.1 Convolutional neural networks (CNNs)

Convolutional neural networks (CNNs) are a class of DL algorithms characterised by their ability to efficiently process data in array form, which makes them highly suitable for image classification, object detection and image segmentation, since these are represented as 2D (or 3D for volumetric images) arrays of pixel intensities (Krizhevsky et al. 2012; Simonyan and Zisserman 2014; LeCun et al. 2015; Long et al. 2015). CNNs typically consist of convolutional, pooling, dropout and fully connected layers.

CNNs are motivated by two key properties of visual data: local correlation, meaning neighbouring pixels tend to have related values, and translation invariance, which means a feature can appear anywhere in an image (LeCun et al. 2015). Convolutional layers, as the name suggests, apply convolutional operations to the input. This operation is the fundamental building block of CNNs that enables them to efficiently extract image features. The operation involves moving a kernel, which is matrix of weights, over a portion of the original input matrix while producing a dot product between these matrices. The product is then placed in a new matrix often called the feature or activation map that forms the response of the kernel at each spatial position of the original input matrix. The process is repeated for every position of the input matrix resulting in a full feature map that forms the convolution's output. The number of pixels the kernel moves at a time is determined by the parameter called stride, which affects the size of the output feature map. After a convolutional operation, non-linearity is introduced in the form of an activation

function. This function is commonly the Rectified Linear Unit (ReLU) (LeCun et al. 2015; Goodfellow et al. 2016).

Pooling layers are implemented to merge semantically similar features, thus reducing the spatial size of the feature maps and reducing the computational overhead. In addition, they help reduce overfitting. The pooling kernel is moved across the feature map to produce a summary of the portion of feature map being inspected. This procedure extracts the dominant features within each kernel region, contributing to the spatial invariance of the network. Max pooling, which simply chooses the maximum value of the feature map within the pooling kernel region, is a commonly used type of pooling (LeCun et al. 2015; Goodfellow et al. 2016).

Fully connected layers are typically utilised at the end of a CNN, after the feature extraction and summarisation has been done by the convolutional and pooling layers respectively (LeCun et al. 2015). Fully connected layers function in a similar manor as the traditional NNs described before. Dropout layers are used to prevent overfitting. They operate by setting a fraction of input units to zero randomly (LeCun et al. 2015; Goodfellow et al. 2016). The training process of a CNN is similar to a NNs, where backpropagation and a gradient descent optimisation algorithm are utilised (LeCun et al. 2015). The final output of a CNN classifier is a one-dimensional probability vector describing the input arrays probability of belonging to each category. For instance, Krizhevsky et al. described an architecture with a 1000-way softmax activation since the data set they used included images from a 1000 different categories (Krizhevsky et al. 2012). In image segmentation, the output is a pixel-wise probability map with the same spatial dimensions as the input, as discussed in the next chapter.

2.3.2.1.2 CNNs in image segmentation

Several CNN architectures have been proposed to tackle the difficult problem of medical image segmentation. As described before, typical CNNs are built in a way that outputs a one-dimensional probability array after the fully connected layers and the activation function, giving the probability of an image belonging to a given category. Clearly then, this is a problem for the task of image segmentation, where an output of 2D or 3D arrays is desired.

Long et al. introduced the fully convolutional network to solve this problem (Long et al. 2015). Their key innovation was to replace the fully connected layers with convolutional layers, which allows arbitrary size inputs and pixelwise predictions. In addition, they introduced upsampling layers using transposed convolution that are used to restore the spatial dimensions of the feature maps after downsampling. This allows obtaining outputs in the same dimensions as the input.

Inspired by the fully convolutional network, Ronneberger et al. proposed an architecture called U-Net intended for biomedical image segmentation (Ronneberger et al. 2015). It achieved excellent segmentation performance on relatively small data sets with the help of data augmentation and has since been widely used in image segmentation tasks, especially in the medical field. As the name implies, the architecture has a distinctive U-shape consisting of the downsampling and upsampling paths, often also called the encoder and the decoder respectively. The encoder follows the typical structure of a CNN as described before, while the decoder includes upsampling layers followed by convolutional layers. The U-Net also includes skip connections between the corresponding layers of the encoder and decoder paths. This provides the upsampling process important information that helps localise the output. In addition, they utilise a weighted loss function to handle the class imbalance between the foreground and background, often present in medical images. The key to U-Net's success lies in its ability to simultaneously extract high-level features and combine them with low-level spatial information.

Due to its success, several variations of the U-Net have been proposed. Çiçek et al. suggested a 3D U-Net with the aim of producing dense segmentations of sparsely annotated volumetric data (Çiçek et al. 2016). Milletari et al. introduced an alternative U-Net inspired approach, specifically with medical volumetric segmentation in mind, called the V-Net (Milletari et al. 2016). Their approach differs from the original U-Net in a few important ways. First, V-Net is designed specifically for volumetric data, whereas U-Net was designed for two-dimensional data. Second, a novel objective function based on the maximisation of the Dice coefficient was introduced. This is especially suitable for situations where there is a strong imbalance between the background and foreground voxels, which is often the case with medical imaging. Third, V-Net uses convolutional operations for downsampling instead of traditional pooling. The authors state that this leads to a smaller memory footprint and enhances the interpretability of the network.

Due to the high computational demands of 3D convolutions and the inability to use spatial information of 2D U-Net's, Li et al. proposed a densely connected architecture combining a 2D U-Net for extracting intra-slice features with a 3D U-Net for aggregating volumetric context (Li et al. 2018). With this architecture, the authors achieved excellent results in liver tumour segmentation. Oktay et al. introduced a U-Net with an attention mechanism guiding the algorithm to focus on target features, while implicitly learning to disregard irrelevant areas (Oktay et al. 2018). The authors found consistent improvements in U-Net segmentation performance across datasets. To further improve on the concept of attention, Wang et al. suggest a U-Net model that leverages self-attention, referred to as non-local U-Nets (Wang et al. 2020). The authors describe a novel method for up- and downsampling, called the global aggregation block, which incorporates self-

attention with the up- and downsampling processes. This method enables each position of the output feature maps to depend on all positions of the input feature maps. This approach is opposite to local operations like convolutions and deconvolutions, where each output location has a local receptive field on the input. The authors state that this method generates a more precise segmentation image while simultaneously reducing the number of parameters.

Finally, it is important to note that architectural modifications should be interpreted relative to strong, well-configured baselines. In particular, nnU-Net ("no-new-Net") demonstrated that state-of-the-art biomedical segmentation performance can be achieved across diverse datasets using standard U-Net variants when the overall workflow is systematically adapted to the task, without introducing novel architectural components (Isensee et al. 2021).

2.4 DL in HNC fusion imaging

This section provides an overview of the DL applications in fusion imaging from recent literature. A tabular summary of the most relevant parts of this literature for this thesis is provided in Table 2.

2.4.1 Radiation therapy (RT)

DL techniques, deep CNNs in particular, exhibit great potential in augmenting and facilitating more efficient radiation therapy (RT) protocols in terms of fusion imaging. Their ability to assimilate imaging modalities can for instance be used to generate synthetic CT images for RT planning, as Olin et al. suggested. They investigated the feasibility of using DL generated synthetic CT for PET/MRI-based RT-planning of HNC patients. For this purpose, a 3D U-Net inspired CNN architecture was developed and trained with PET/MRI data from 8 HNC patients referred to RT. Data augmentation and transfer learning were used to compensate for the small sample size. In addition, 3 pilot patients were used to develop a suitable scanning protocol. The resulting synthetic CTs were then compared with routine planning PET/CT's the patients underwent prior to the PET/MRI scans. Synthetic CT-based dose plans deviated by no more than ±1% from CT-based dose plans. For PET, the mean difference was $0.4 \pm 1.2\%$ for SUVmean and $-0.5 \pm 1.0\%$ for SUVmax. The authors concluded that the synthetic CT's generated with the DL-based approach were suitable for clinical use (Olin et al. 2020).

In a subsequent study, the authors proceeded to evaluate the performance of this approach on two separate datasets, one external and one internal with 6 and 17 HNC patients referred for RT respectively. In addition, each patient underwent either a planning CT or a planning PET/CT, which then served as reference to the synthetic

CTs. The mean absolute errors were 78 ± 13 and 76 ± 12 Hounsfield Units for the external and local cohorts respectively. In the external cohort, absorbed dose differences in target volumes were within ± 2.3% and within ± 1% in 95% of cases, with organ-at-risk differences remaining below 2%. Similar outcomes were observed for the local cohort. The DL-based method was found to be robust, producing similar dose calculations for both synthetic CT's and traditional CT's (Olin et al. 2021).

Moreover, DL models have major implications for patient safety when used to aid in RT planning. They can be used in predicting radiosensitivity, facilitating dose escalation or de-escalation strategies. This enhances treatment efficacy while minimising radiation-induced toxicity. For instance, Olin et al. studied the possibility of using a DL-based approach to attenuation correction for PET/MRI. The authors compared the performance of a 3D U-Net inspired deep NN to an atlas-based method provided by the PET/MRI vendor using a CT-based attenuation map as a reference. The average PET voxel error was 0.0 ± 11.4% for the NN and −1.3 ± 21.8% for the atlas-method. The NN exhibited lower mean uptake errors in bone/air (−4%/12%) compared to the atlas-method (−15%/84%) and showed a faster error reduction with distance, limited to areas within 1 cm of bone/air. They concluded that the DL-based attenuation maps were superior to the atlas-based maps (Olin et al. 2022).

2.4.2 Classification

Deep CNN-based approaches have also been adopted for classification tasks in HNC fusion imaging. The inherent advantages of using automated methods for image classification include the lack of inter-observer variability and faster analysis.

Chen et al. proposed an DL-based approach for detecting and classifying lymph node metastasis of HNC patients in PET/CT images. They describe a method that combines traditional handcrafted radiomics with a modern 3D CNN that classifies lymph nodes of the head and neck region into three categories: normal, suspicious and involved. The study included PET/CT images of 59 patients referred for RT due HNSCC. Their method achieved an accuracy of 0.88 and an multiclass area under the curve (AUC) of 0.95 in lymph node classification outperforming radiomics or DL approaches alone (Chen et al. 2019).

Dohopolski et al. Suggested a DL-based binary classification method for predicting lymph node metastasis in PET/CT images of patients with oropharyngeal squamous cell carcinoma. The authors built a 3D CNN inspired by the AlexNet architecture and achieved high classification performance using a dataset consisting of 129 patients who underwent pre-operative PET/CT. The lymph nodes were labelled either malignant or benign based on the pathology reports from the neck dissections and subsequently contoured under PET guidance. These contours were then used to extract image patches of the respective lymph nodes and their

surrounding voxels that were used as inputs for the DL model. In addition, the authors used epistemic and aleatoric uncertainty to quantify the reliability of the predictions. The model achieved an AUC of 0.99 on the test dataset, with sensitivity and specificity of 0.94 and 0.90, respectively. Epistemic and aleatoric uncertainty were significantly higher for false negatives and false positives compared to true negatives and true positives ($p < 0.001$) (Dohopolski et al. 2020). The authors found that measures of uncertainty are useful in quantifying the reliability of the classifications.

The previously described dataset is also used in a study conducted by Chen et al. investigating the feasibility of an attention guided CNN in binary lymph node classification of HNC patients. The proposed model takes a region of interest (ROI) containing the lymph node and surrounding tissues as input. The CNN identifies the discriminative region within the ROI, generating an activation map where voxel values represent their relevance for malignancy prediction. This activation map is multiplied with the ROI to produce the input for the classifier, which outputs the node's malignancy probability. The model achieved sensitivity, specificity, accuracy, and AUC values of 0.91, 0.93, 0.92, and 0.98, respectively. The authors found that using an attention guided training process increased the models classification accuracy when compared to traditional radiomics and CNNs (Chen et al. 2021).

2.4.3 Outcome prediction

Lately, the feasibility of DL algorithms in outcome prediction of HNC patients has been investigated. Deep CNNs can capture intricate imaging biomarkers that might carry prognostic value, especially when combined with other relevant patient characteristics.

Han et al. investigated the possibility of using DL in the binary prediction of locoregional recurrence (LR) of HNSCC. However, no timeframe is explicitly defined for LR. The authors trained 3D CNNs utilising PET/CT, RT planning CT data and baseline clinical characteristics (such as sex, age, tumour stages, and primary disease site) of 157 (out of 298 in total) HNSCC patients available in the public TCIA dataset (Vallières et al. 2017). Several different input combinations were used for training, with the model incorporating CT, PET, dose distributions and clinical characteristics achieving the highest AUC of 0.89 ± 0.07. The authors conclude that the addition of dose distribution and clinical information improves the predictive capabilities of the DL model (Han et al. 2022).

Wang et al. describe the use of multi-modality and multi-view feature extension method (MMFE) with a 2D CNN for LR prediction utilising the PET/CT images and clinical characteristics of 206 HNSCC patients (49 with LR and 157 with no LR).

The median follow-up was 37 months and patients with less than 12 months of follow-up were excluded. The goal of MMFE is to preserve the volumetric information that traditional 3D CNNs benefit from, while reducing model complexity. Similar to Han et al., the highest AUC of 0.81 was obtained using all modalities, thus indicating the added benefit of incorporating clinical characteristics in outcome prediction with DL models. Furthermore, the MMFE outperformed traditional 2D and 3D DL-based methods (Wang et al. 2022).

Le et al. proposed a pseudo-volumetric CNN with a deep preprocessor module and self-attention for predicting distant metastasis, LR and overall survival occurrence probabilities within the 10-year follow-up time. The previously described TCIA dataset was utilised in this study. In addition, an internal dataset of 371 HNSCC patients undergoing RT was used for validation. Ablation experiments were conducted to evaluate the model's key features, yielding AUCs of 0.80, 0.80, and 0.82 for distant metastasis, LR, and overall survival, respectively, on the TCIA dataset. The additional validation achieved comparable AUCs of 0.69 across all outcomes. The authors state that the proposed model is effective in HNSCC outcome prediction when combining multi-modal volumetric data and clinical patient information (Le et al. 2022).

2.4.4 Segmentation

Tumour and lymph node segmentation has been the prevalent research topic for HNC fusion imaging in recent years. The proposed solutions centre around the utilisation of U-Net-like deep CNN architectures both in 2D and 3D variations. Several mechanisms have been proposed to amend the standard U-Net architecture to address the dilemma of capturing the spatial context with 3D models, while maintaining the computational efficiency of the 2D approaches. With the launch of the head and neck tumour (HECKTOR) challenge in 2020, the number of potential solutions for this problem has increased substantially (Oreiller et al. 2022).

Huang et al. studied the possibility of using a deep CNN for gross tumour volume (GTV) segmentation in HNC patients undergoing RT. They describe a 2D U-Net inspired architecture trained on PET/CT images of 22 newly diagnosed HNC patients from 2 separate centres. Data augmentation was used to increase the sample size. Manual delineations by an oncologist and a radiologist served as the gold standard GTV. Leave-one-out cross-validation was used to evaluate the results. The authors report an average Dice score (DSC) of 0.74 ± 0.11 between the gold standard and the automatic delineations and conclude that their method reached high accuracy and efficiency and might be of help to clinicians in HNC management (Huang et al. 2018).

Zhao et al. introduced a similar 2D U-Net based DL model with auxiliary paths for the segmentation of nasopharyngeal carcinoma from PET/CT images. The study included 30 patients with nasopharyngeal carcinoma and utilised data augmentation to increase the size of the dataset. The ground truth was delineated manually by two radiologists on the PET/CT images. The authors reported a mean DSC of 0.88 from 3-fold cross-validation, when compared to the ground truth, and conclude that their proposed method is robust and efficient (Zhao et al. 2019).

Guo et al. proposed a 3D Dense-net for the GTV segmentation of HNSCC patients. PET/CT images of 250 HNSCC patients from the TCIA dataset were utilised in this study. The data was split into training (n=140), validation (n=35), and test sets (n=75). However, no cross-validation procedure is described. The Dense-net approach outperformed the standard 3D U-Net and Dense-nets trained solely on PET or CT images, achieving a mean DSC of 0.71 ± 0.10 when compared to the gold standard set by manual delineations by radiation oncologists. The authors state that their method yields superior results in automatic HNC GTV segmentation while having less trainable parameters than the conventional 3D U-Net, thus indicating it could be successfully applied to RT planning (Guo et al. 2019).

Andrearczyk et al. compared the segmentation performance of a 2D and 3D V-Nets on PET/CT images of 202 HNSCC patients derived from the TCIA dataset. A 4-fold cross-validation was used to evaluate the segmentation results, as the data consisted of scans from four separate institutions. The reported DSCs are an average over 10 cross-validation runs. The 2D V-Net slightly outperformed the 3D-counterpart with DSC being 0.61 ± 0.02 and 0.60 ± 0.02 respectively. The authors also found that bimodal models performed better compared to the single modality models (Andrearczyk et al. 2020).

Groendahl et al. investigated the segmentation capabilities of traditional PET thresholding methods, six classical ML algorithms and a 2D U-Net with RT planning PET/CT images of 197 HNSCC patients. The manual delineations that served as the gold standard were defined by a nuclear medicine physician based on the PET images and further refined by oncology residents based on the CT images and finally reviewed by a senior oncologist for quality control. A 5-fold cross-validation procedure was utilised on the training set (n=157) and the best model for further evaluation on the test set (n=40) was selected based on the DSC. The PET/CT-based 2D U-Net outperformed the other alternatives with a DSC of 0.75 ± 0.09 on the test set (Groendahl et al. 2021).

Moe et al. studied the applicability of 2D U-Net inspired CNNs in automatic primary tumour volume and malignant nodal volume segmentation utilising the same dataset as described by Groendahl et al. (Groendahl et al. 2021). The data was split into training (n=142), validation (n=15), and test sets (n=40). The best performing model was chosen based on the validation performance and further evaluated on the

test set. The authors compared the performance of models trained with single-modalities (CT or PET) as well as multimodality images (PET/CT). In addition, they introduced a framework for structure-wise performance evaluation of multi-structure auto-delineations with the intention of providing additional information to quantify the similarity between the ground truth and the network predictions when more than one contoured structure is present in the ground truth. Qualitative assessment of 15 randomly chosen auto-delineations was also conducted by an oncologist. The multimodality-based model performed best achieving a DSC of 0.71 ± 0.16. Out of the 15 total auto-delineations assessed in terms of their quality, 14 were determined to be of high quality. The study concluded that the multimodality-based CNN provides accurate auto-delineations for HNSCC patients (Moe et al. 2021).

The only study that has included MRI as an anatomical imaging variant in HNC segmentation from fusion imaging was conducted by Ren et al. who described the use of a residual 3D U-Net utilising a dataset of 153 HNSCC patients. The data included RT planning CT, PET and MRI images with clinical delineations, by a team of oncologist, radiologist and a nuclear medicine physician, serving as ground truth. The data was split into training (n=92), validation (n=31), and test sets (n=30). Data augmentation was used to increase the sample size. The authors compared the segmentation performance of all possible bi-modality combinations along with one model trained with all three modalities (CT-PET/MRI). All combinations including PET images achieved similar mean DSC: PET/MRI 0.72 ± 0.15, PET/CT 0.73 ± 0.13, CT/PET/MRI 0.74 ± 0.13. The CT/MRI combination provided inferior segmentation with a mean DSC of 0.58 ± 0.18. The CT/PET/MRI model showed improvements in other key metrics (Hausdorff Distance 95 percentile and Mean Surface Distance) compared to the bi-modality models alone. Ren et al. conclude that the inclusion of PET seems essential for the segmentation task, whereas adding a second anatomical imaging variant to either PET/CT or PET/MRI is of limited benefit (Ren et al. 2021).

De Biase et al. introduced a 3D CNN for oropharyngeal tumour segmentation using an architecture that utilises attention mechanisms and bi-directional long short-term memory to capture inter and intra-slice context. The dataset used in their study consisted of RT planning PET/CT images of 138 oropharyngeal cancer patients with manual primary tumour GTV delineations that were reviewed by a nuclear medicine physician and a radiologist in cases where MRI scan was available as well. The delineations were then approved with a consensus reading by radiation oncologists. Prior to training the model, a region of interest, namely the oral cavity, was automatically extracted from each scan. The described DL model accepts sequences of 3 consecutive 2D PET/CT slices as input. The authors utilise a probability map-based method for depicting the uncertainty of a given segmentation. Moreover, they use 3-fold cross-validation and multi-view (over the coronal, sagittal, and axial

planes) aggregates to produce ensemble models for prediction. They report the highest DSC of 0.81 ± 0.32 with a probability threshold of 0.90 for the multi-view model and conclude that this approach captures the model's spatial uncertainty in auto-delineation and can be used in slice-by-slice adaptive GTV segmentation (De Biase et al. 2023).

Nikulin et al. described a 3D U-Net-based approach modified with a multi-head self-attention block for primary tumour and lymph node classification and metabolic tumour volume (MTV) segmentation from PET/CT images of 698 HNSCC patients. An additional dataset of 181 PET/CT scans was used to assess the generalisability. The annotation used as reference was conducted by two physicians based on the metabolically active areas in the PET that were identified with a semi-automatic adaptive thresholding algorithm. These algorithm-suggested areas were then reviewed by one of the physicians and corrected if necessary. The authors report an DSC of 0.89, 0.81, and 0.87 for primary tumour, lymph node metastases, and the union of both, respectively for the 5-fold cross-validation experiment in the main data set. For the additional dataset, the DSC was 0.85, 0.72, and 0.82 for primary tumour, lymph node metastases, and the union of both, respectively for the ensemble model acquired through the cross-validation procedure. The model was able to distinguish between the primary tumour and lymph nodes with an accuracy of 98.0% for the main dataset and 97.9% for the additional dataset. Nikulin et al. conclude that their approach achieves satisfactory delineation and classification results, thereby having clear potential for supervised clinical use (Nikulin et al. 2023).

Salahuddin et al. proposed a modified 3D nnU-Net with residual skip connections, squeeze-and-excitation channel-wise attention mechanisms at each layer, and grid attention gates at each skip connection for the GTV segmentation of HNSCC from PET/CT images. They utilised the HECKTOR 2022 dataset containing 883 HNSCC patients (Andrearczyk et al. 2023). The authors described a posterior weight space sampling method for estimating model uncertainty and used it to reduce false positive predictions. They achieved an aggregate (accumulated across all images) DSC of 0.77 and 0.76 on the test set with per image false positive reduction of 19.5% and 7.14% using the uncertainty threshold for primary tumour GTV and lymph node GTV, respectively. In addition, the authors report that the radiomics features extracted from nodal GTV in PET and from both tumour GTV and nodal GTV in CT are the most prognostic in terms of recurrence free survival prediction, with the model achieving a C-index of 0.67 on the test set. Salahuddin et al. state that these results demonstrate their models accurate detection capabilities and the possibility for risk stratification (Salahuddin et al. 2023).

Table 2. Summary of DL applications in fusion imaging for outcome prediction and segmentation from recent literature. The studies are listed here in the same order as in chapter 2.4, based on their intended use cases (outcome prediction and segmentation). Abbreviations: CNN (Convolutional Neural Network), DL (Deep learning), DM (Distant Metastasis), DSC (Dice Score), GTV (Gross Tumour Volume), HNC (Head and Neck Cancer), HNSCC (Head and Neck Squamous Cell Carcinoma), LR (Locoregional Recurrence), ML (Machine Learning), MMFE, NPC (Nasopharyngeal Cancer), OS (Overall Survival).

Authors	Objective	Methods	Data	Results	Conclusion
(Han et al. 2022)	Binary LR prediction.	DL, 3D CNN.	157 HNSCC patients. PET/CT.	AUC 0.89 ± 0.07.	Dose and clinical information improve performance.
(Wang et al. 2022)	Binary LR prediction.	DL, 2D CNN with MMFE.	206 HNSCC patients. PET/CT.	AUC 0.81.	Clinical information improves LR prediction performance, MMFE superior to 3D CNN.
(Le et al. 2022)	LR, metastasis, 10yr. survival prediction.	DL, pseudovolumetric 2D CNN.	298 (external data) + 371 HNSCC patients (internal data). PET/CT.	AUCs of 0.80, 0.80, and 0.82 for DM, LR, and OS, respectively (external data). AUC 0.69 for DM, LR, and OS (internal data) .	Proposed model effective in outcome prediction when combining multi-modal volumetric data and clinical information.
(Huang et al. 2018)	HNC GTV segmentation.	DL, 2D CNN.	17 + 5 HNC patients. PET/CT.	Average DSC 0.74 ± 0.11.	DL efficient and accurate compared to gold standard.
(Zhao et al. 2019)	NPC segmentation.	DL, 2D CNN.	30 NPC patients. PET/CT.	Mean DSC 0.88.	DL efficient and accurate compared to gold standard.
(Guo et al. 2019)	HNSCC GTV segmentation	DL, 3D dense-net	250 HNSCC patients. PET/CT.	Mean DSC 0.71 ± 0.10.	Dense net superior to 3D-Unet.
(Andrearczyk et al. 2020)	HNSCC segmentation.	2D and 3D V-Nets with single and bimodal configurations.	202 HNSCC patients. PET/CT.	Mean DSC 0.61 ± 2.1 and 0.60 ± 2.2 for 2D and 3D bimodal V-Nets respectively.	2D bimodal V-Net superior to 3D bimodal V-Net. Bimodal superior to single modality.
(Groendahl et al. 2021)	HNSCC GTV segmentation	2D U-Net, classical ML, PET thresholding	197 HNSCC patients. PET/CT.	Mean DSC 0.75 ± 0.09.	2D bimodal U-Net superior to other methods.
(Moe et al. 2021)	HNSCC GTV segmentation.	2D U-Net.	197 HNSCC patients. PET/CT.	Mean DSC 0.71 ± 0.16.	2D-Unet yields good segmentation results.

Authors	Objective	Methods	Data	Results	Conclusion
(Ren et al. 2021)	HNSCC GTV segmentation.	3D U-Net.	153 HNSCC patients. CT, MRI, PET.	DSC: CT/MRI 0.58 ± 0.18, PET/MRI 0.72 ± 0.15, PET/CT 0.73 ± 0.13, CT/PET/MRI 0.74 ± 0.13.	PET essential for accurate segmentation.
(De Biase et al. 2023)	HNSCC GTV segmentation	3D CNN, single and multi-view ensemble models.	138 oropharyngeal cancer patients. PET/CT.	DSC 0.81 ± 0.32.	Probability output maps and multiview ensembling aid in adaptive segmentation.
(Nikulin et al. 2023)	HNSCC MTV segmentation.	3D U-Net with self-attention.	698 (primary data) + 181(additional data) HNSCC patients. PET/CT.	DSC 0.87 (primary data), DSC 0.82 (additional data).	Accurate segmentation suitable for supervised clinical use.
(Salahuddin et al. 2023)	HNSCC GTV segmentation.	3D U-nnNet.	883 HNSCC patients.	GTVp DSC 0.77, GTVn 0.76.	3D U-nnNet. demonstrates accurate detection capabilities and enables risk stratification.

3 Aims

The aim of this doctoral thesis was to investigate the feasibility of using DL in the automated interpretation of PET/MRI images of head and neck cancer patients.

The specific aims of the thesis were as follows:

1. to investigate the use of a 2D CNN model in the segmentation of head and neck cancer from ^{18}F-FDG PET/MRI images.
2. to explore the possibility of using a 2D CNN model in the binary classification of head and neck cancer from ^{18}F-FDG PET images.
3. to assess the performance and clinical applicability of a 3D CNN model in the binary classification of head and neck cancer from ^{18}F-FDG PET/MRI images.

4 Materials and Methods

4.1 Study population

The study population for the original articles in this doctoral thesis was retrospectively collected from the imaging archives of Turku University Hospital. The first study included 52 HNSCC patients, of which 44 met the inclusion criteria (Table 3), who underwent re-staging FDG-PET/MRI scans 12 weeks after chemoradiation therapy (CRT) from February of 2014 to May of 2017. The cohort included 15 patients with locoregional recurrences (positive subgroup) and 29 patients with no locoregional recurrences (negative subgroup). Of the positive subgroup, 3 patients had follow-up scans, which were included in the study as well. In the negative subgroup 13 patients had follow-up scans, where 2 patients had a follow-up twice, which were also included. These patients are a subgroup of patients initially described by Murtojärvi et al. (Murtojärvi et al. 2022).

In addition to the cohort of the first study, the second and third study included HNC patients that had FDG-PET/MRI scans done in Turku University Hospital between the years 2014–2022. This cohort consisted mostly of HNSCC patients; however, a small number of patients with other types of HNC were also included in these studies to investigate the generalisation properties of the models to other types of cancer. A total of 200 consecutive patients were included in Study II with a 50:50 negative/positive ratio. Study III included 2 additional subjects in the training data compared to Study II: one in the positive sub-group and one in the negative sub-group. These were excluded from the second study to achieve round numbers for the training and test data. The inclusion criteria and the number of subjects are described in Table 3. In addition, a separate test set with 20 subjects (10 with cancer and 10 without cancer) was analysed in Study III. A total of 222 consecutive patients were included in Study III with a 50:50 negative to positive ratio. The follow-up scans included in Study I were excluded from studies II and III, so that only one scan per patient was utilised. Table 4 describes the different histologies of Studies II and III grouped by location. The positive patient that was left out of Study II had a SCC of the hypopharynx and the negative patient that was left out had been treated for a SCC of the right tonsil. Locations and histologies of HNC included in Study III test set are described in Table 5. The mean age of the patients was 63.0 years with a standard

deviation of 11.7 years and their male–female sex ratio was 2.1. The 50:50-ratio of negative and positive patients does not reflect real world circumstances but was chosen to facilitate a more efficient training process. The basis of the ground truth for these cases are clinical readings combined with histological confirmation of disease (when available) or documented absence of disease with at least 1 year of follow-up. Equivocal cases were assigned to positive or negative subgroup based on histological confirmation or clinical follow-up.

Table 3. Inclusion criteria and the number of included subjects for each study in this thesis. Abbreviations: CRT (chemoradiation cherapy), HNC (head and neck cancer).

Study	Subjects	Inclusion criteria
I	44	Histologically confirmed HNSCC, treated with CRT, locoregional recurrence (positive subgroup), no locoregional recurrence during a minimum of 1-year follow-up (negative subgroup), FDG-PET/MRI scan
II	200	Histologically confirmed HNC, FDG-PET/MRI scan, locoregional recurrence or new tumour (positive subgroup), no locoregional recurrence during a minimum of 1-year follow-up (negative subgroup)
III	222	Histologically confirmed HNC, FDG-PET/MRI scan, locoregional recurrence or new tumour (positive subgroup), no locoregional recurrence during a minimum of 1-year follow-up (negative subgroup)

Table 4. Different types of HNC included in Studies II and III, grouped by location. Abbreviations: SCC (Squamous Cell Carcinoma).

Location	Histology	N
Pharynx	SCC	82
	Clear cell carcinoma	1
Oral Cavity	SCC	47
	Adenocystic carcinoma	4
Larynx	SCC	20
Unknown primary	SCC	18
Nasal Cavity	SCC	7
	Epithelial carcinoma	1
Salivary Glands	SCC	2
	Asinus cell carcinoma	1
	Adenocarcinoma	1
	Mucoepidermoid carcinoma	1
	Epithelial carcinoma	1
	Salivary duct carcinoma	1
Oesophagus	SCC	2
	Adenocarcinoma	2
Sinuses	SCC	2
Skin	SCC	2
	Merkel cell carcinoma	1
	Adnexal carcinoma	1
	Angiosarcoma	1
Thyroid	Papillary carcinoma	2
	Follicular carcinoma	1
Skull base	Chondrosarcoma	1

Table 5. Different types of HNC included in the test set of Study III, grouped by location. Abbreviations: SCC (Squamous Cell Carcinoma).

Location	Histology	N
Pharynx	SCC	9
	Clear cell carcinoma	1
	Adenocarcinoma	1
Larynx	SCC	4
Oral cavity	SCC	3
Sinuses	Myoepithelial carcinoma	1
Unknown primary	SCC	1

4.2 Imaging Protocols

4.2.1 Study I

The data for the first study consisted of PET/MRIs that were conducted using a sequential Ingenuity 3T TF PET/MRI system (Philips Healthcare), with a SENSE neurovascular coil. For MRI acquisition, transaxial sequences included T1 TSE, T1 SPIR and T2 TSE. The T1-weighted sequences focused on the region of the primary tumour. Contrast agent was administered with the T1 SPIR scans. T2-weighted sequences were utilised to provide precise anatomical delineation of both the tumour and adjacent lymph node regions. For this study, T1 SPIR sequences were obtained in 56 cases. In instances where T1 SPIR sequences were unavailable (n = 5), T1 TSE was used as an alternative. Similarly, in one instance where both T1 SPIR and T1 TSE sequences were unavailable, a T2 TSE sequence was utilised.

Attenuation correction was achieved using Dixon MRI-based sequences, which spanned from the level of the forehead to the groin. The correction employed a 3-segment model, which distinguished between air, lung tissue, and soft tissue.

PET imaging was initiated immediately following the MRI acquisition. The PET scan employed a transaxial field of view (FOV) of 576 mm. Image reconstruction was performed using the system's default "Blob-OS-TF" algorithm, a 3D ordered subset iterative time-of-flight reconstruction method. This process utilized 3 iterations with 33 subsets. The final PET images were reconstructed on a 144 × 144 matrix, yielding a voxel size of $4 \times 4 \times 4$ mm³. Standard corrections required for quantitative PET imaging, including attenuation, randoms, scatter, dead-time, decay, and detector normalization, were incorporated.

4.2.2 Study II

In the second study, only the PET-images were utilised for model training and evaluation. The scans were performed with a 3 T Philips Ingenuity TF PET/MR scanner (Philips Health Care) until 3/2020, after which a SIGNA™ PET/MR QuantWorks (GE Healthcare) was used. Out of the total 200 PET/MRIs 72 were conducted using the sequential Ingenuity system, of which 29 were positive and 43 negative. The imaging protocol for this machine was the same as is described in the previous chapter for Study I. After 3/2020, the 3 T SIGNA™ PET/MR replaced the Philips Ingenuity scanner. Out of the 200 scans, the remaining 128 were done with this machine, of which 71 were positive and 57 negative. The transaxial FOV this scanner utilised was 600 mm. The system's Q.Clear algorithm was used for image reconstruction. PET images were reconstructed on a 256 × 256 matrix, yielding a voxel size of 2.3 x 2.3 x 2.8mm³. Standard corrections required for quantitative PET

imaging, including attenuation, randoms, scatter, dead-time, decay, and detector normalization, were included.

4.2.3 Study III

In the third study, the model training and evaluation was carried out using PET, MRI, and PET/MRI data. In addition, one positive and one negative patient that were scanned with the Philips Ingenuity were included in this study. The imaging protocols for PET in this study are identical to studies I and II. MRI acquisition included transaxial sequences T1, T1 SPIR and T2. The T1-weighted sequences focused on the region of the primary tumour. T2-weighted sequences provided precise anatomical delineation of both the tumour and adjacent lymph node regions. Contrast agent was administered with the T1 SPIR scans. Of the 30 positive cases from the Ingenuity scanner, 23 were T1 SPIR, 4 were T1, and 3 were T2. Of the 44 negative cases from the same scanner, 28 were T1 SPIR, 15 were T2, and 1 was T1. MRI sequences utilised from the SIGNA scanner were T1 and T2-weighted sequences. Of the 71 positive cases, 64 were T1 and 7 were T2. All 57 negative cases were T2. Only one sequence was included in the analysis per patient. The choice was done based on availability and image quality.

4.3 Image Pre-processing

4.3.1 Study I

The PET/MRI images were resliced into common dimensions, and the PET images were cropped to the level of the corresponding MRI sequence on the transaxial plane. For the positive subgroup, the recurrent tumour was manually delineated using the metabolic information from the PET images and the anatomical information from the MRIs. In addition, local metastases were annotated similarly, if present. Image masks were then created based on these delineations and the resulting tumour volumes were considered as the ground truth for model training and evaluation. Based on the manual delineations 178 2D slices were identified as having cancer. The same number of slices were then randomly chosen from the negative subgroup. The data was split into training and test sets patient-wise with a 80:20 ratio yielding a training set of 290 images and a test set of 66 images. In addition, a validation set of 15% was split from the training data during the training process.

The 2D slices and the corresponding masks were resized from 512 × 512 pixels to 128 × 128 pixels to facilitate a more efficient training process with larger batch sizes and to reduce the compute requirements. PET and MRI pixel intensity ranges differ, therefore a linear scaling between 0 and 1 was applied to both. The global

minimum and maximum from the training data were used for the scaling process. Given a situation where the test set would have had higher pixel values than those of the training set, the normalisation values were set to 1 and 0 for those pixel values that exceeded or fell below the given interval respectively.

To evaluate the model using additional artificial data, an image augmentation system was built. Augmentations used in this study included flipping from left to right and upside down, randomly translating from -10 to 10% and rotating between -15 and 15 degrees. Augmentation was conducted for the training sub-set and the corresponding masks to produce 5 augmented images from a single original image. This resulted in 1450 training images and masks.

4.3.2 Study II

The image reslicing, cropping, delineation and mask creation were carried out similarly to Study I. This process resulted in 995 positive 2D PET slices from 89 HNSCC patients and 119 positive slices from the 11 non-SCC patients. A corresponding number of negative slices were chosen by randomly selecting slices from the negative subgroup. The primary data set consisted of 995 positive HNSCC slices with equally many negative slices, and the additional test set consisted of the 119 positive non-HNSCC slices and the same number of negative slices. The primary data set was used to train and test the models with five-fold cross-validation. The primary training data was divided patient-wise into five possible test sets of 199 negative and 199 positive slices (20% of the primary training data). The non-augmented models were then trained with a data set of 1592 slices and then tested with both a primary test set of 398 slices the additional test set of 238 slices.

The original 2D slice dimensions were reduced to 128 × 128 from 512 × 512 as described in Study I. The images were scaled individually between 0 and 255 and rounded to integers to make them more memory efficient. An augmentation pipeline was utilised in this study as well, quadrupling the size of the training data for half of the performed experiments. The augmentations included clockwise and anticlockwise rotations of 15 degrees and a reflection against the sagittal plane so that the left and the right sides of a transaxial image were switched. The augmented training data set consisted of 6368 images.

4.3.3 Study III

The image reslicing, cropping, delineation and mask creation were carried out similarly to Study I. The images were down-sampled in-plane from 512 × 512 to 64 × 64 to allow bigger batch sizes and more efficient computation. Voxel size of the original images ranged from (0.4, 0.4, 3.4) to (0.7, 0.7, 4.5) in mm. The difference

along the x- and y-axes is approximately two-fold, which could significantly affect accuracy. To address this, voxel sizes were interpolated to match the highest resolution within the dataset. Additionally, the image stacks were cut along the z-axis to a uniform size of 32 slices. For the cancer subgroup, slices were selected based on tumour location, ensuring inclusion of slices from both sides of the central mask slice. For the negative images, 32 consecutive slices were randomly selected. This process yielded 202 PET/MRI images with dimensions of 64 × 64 × 32 pixels. Voxel values were linearly scaled between 0 and 1 to account for the varying pixel value ranges between PET and MRI. Scaling was performed individually for each 3D image using the minimum and maximum values derived from the respective image. To enhance the dataset, a basic augmentation pipeline, including flipping and random 90-degree rotations, was implemented, quadrupling the training set size during the 5-fold cross-validation. This approach yielded 805 training samples for the first two cross-validation folds and 810 samples for the last three. The independent test set underwent similar pre-processing, with the exception that no cropping was performed along the z-axis, as a sliding window approach was employed for inference. This technique allows models to manage inputs of varying sizes by sequentially processing specific regions within the input volume using a fixed-size window that moves systematically across the data.

4.4 Deep Learning Model Implementation

4.4.1 Study I

A 2D U-Net model was implemented for this study. The model was trained to perform segmentation of the primary tumour and potential metastases within the PET/MRI slices. In addition to the combined PET/MRI approach, separate models were also developed using PET slices alone, as well as PET/MRI slices augmented with additional features.

The loss function used during training was binary cross-entropy and the Jaccard similarity coefficient was used as the accuracy metric. Gradient descent optimization was carried out using the Adam optimizer, with a learning rate set to 0.001. Early stopping was implemented to prevent overfitting, with the maximum number of epochs set to 200 and a patience threshold of 50 epochs. During training, the Jaccard similarity coefficient was evaluated on the validation dataset after each epoch, and the highest observed value was stored in memory. The validation data was split from the training data during training and is separate from the test set. After training, the model configuration corresponding to this highest validation Jaccard score was restored. This best-performing model was subsequently used for further performance evaluation on the test set.

4.4.2 Study II

For this study, two CNN architectures were built and compared. Both models utilise convolutional and pooling operations to automatically capture features from the images. The first model was a deep U-Net inspired CNN, where only the contracting path of the U-Net was used, as opposed to the typical contracting and expanding paths of U-Net in image segmentation. The second model implemented was a shallow CNN that differs from the deep model by implementing only two pairs of convolutions. The model architectures are presented in figures 4 and 5.

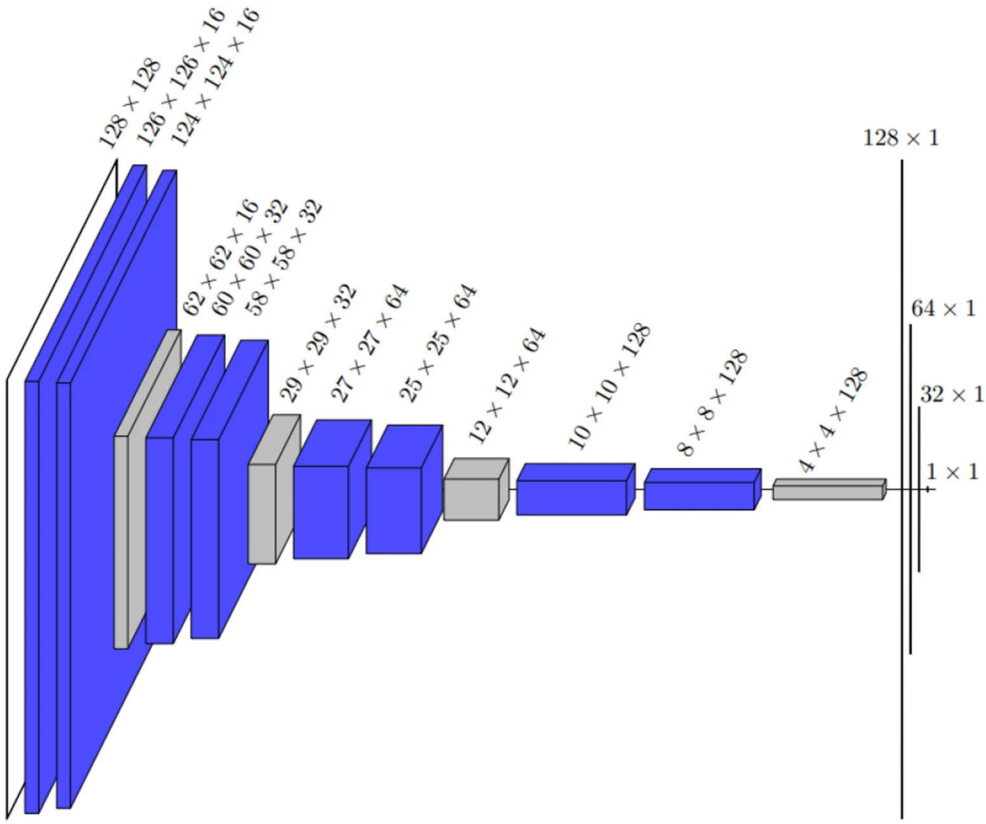

Figure 4. The architecture of the deep CNN used in Study II. Blue layers in the figure represent convolutional layers, and the gray layers represent maximum pooling layers. The numbers denote the dimensions of each layer. The first transparent layer is the input layer.

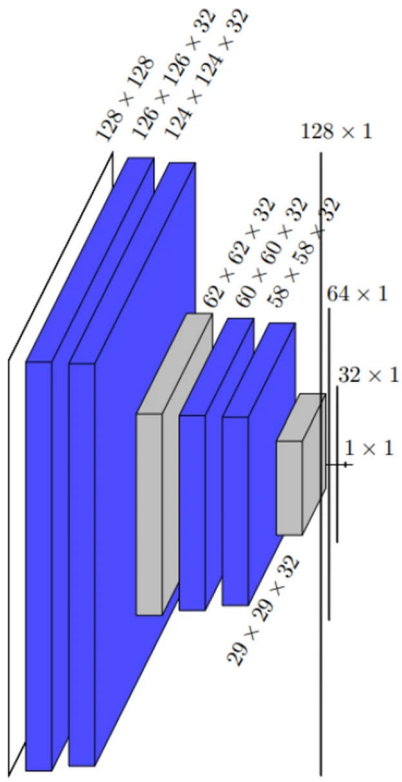

Figure 5. The architecture of the shallow model used in Study II. Figure notation follows the same convention as in figure 4.

During training, binary-cross entropy was used as the loss function and stochastic gradient descent as the optimisation algorithm with a constant learning rate of 0.001 throughout the training process. Early stopping was used to prevent overfitting, with the maximum number of epochs set to 25 and a patience threshold of 15 epochs. This was achieved by comparing consecutive losses in the validation set split during training that contained 30% of the training data. As sigmoid is used as the last activation function in these models, the output is a single real-valued estimate of cancer probability within a given 2D slice. Thus, a threshold for determining the final binary classification result needed to be set. For this purpose, Youden's J-statistic was used. It is defined as follows:

$$J = \frac{TP \cdot TN - FN \cdot FP}{(TP + FN)(FP + TN)}$$

This value can be interpreted as the distance between the receiving operating characteristic (ROC) curve of a randomly predicting model (diagonal line) and the

ROC curve of the model of interest. Therefore, it is logical that maximising this distance yields the best classification results.

4.4.3 Study III

A 3D CNN model incorporating residual connections and an attention mechanism was implemented for this study. Separate models were trained and evaluated for PET, MRI, and PET/MRI data. The architecture begins with an entry block consisting of a 3D convolutional layer combined with batch normalization and ReLU activation. Following the entry block, the data passes through a residual block and an attention block, each with 64 filters. This is succeeded by a second set of residual and attention blocks, each with 128 filters. The attention block refines feature representations by processing them through two parallel 1×1×1 convolutional layers, each learning distinct channel-wise transformations. The resulting feature maps are subsequently combined and reactivated. A global average pooling (GAP) layer is then applied, followed by a dense layer with 512 nodes and a sigmoid activation function to produce the final binary classification. GAP was employed to generate fixed-length feature vectors independent of input size, enabling the model to handle volumes of varying depths while reducing parameters and overfitting. Despite this size invariance, inference on the test set was performed with a sliding window due to practical constraints: limited GPU memory, the localized nature of pathological findings, and the improved robustness achieved by aggregating predictions from overlapping sub-volumes. The model's architecture is illustrated in Figure 6.

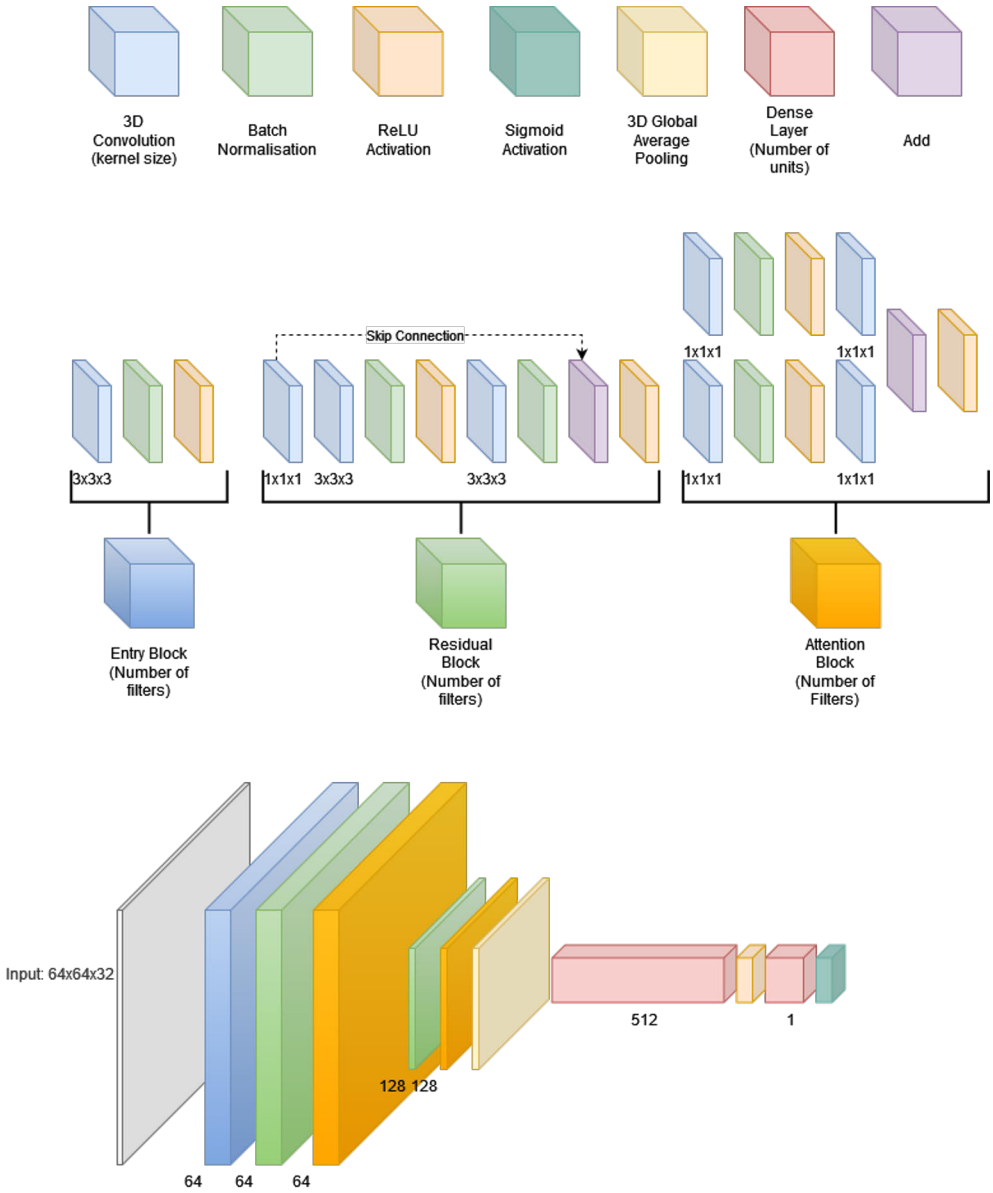

Figure 6. Architecture of the 3D CNN used in Study III.

The Adam optimizer was utilized with an initial learning rate of 0.0001, coupled with an exponential decay rate of 0.96 over 100,000 steps. Focal binary cross-entropy served as the loss function, while AUC was used as the performance metric. Early stopping was used to prevent overfitting, with the maximum number of epochs set to 100 and a patience threshold of 30 epochs. This was achieved by comparing consecutive losses in the validation set split during training that contained 20% of

the training data. As sigmoid is used as the last activation function in these models, the output is a single real-valued estimate of cancer probability within a given 3D image. Thus, a threshold for determining the final binary classification result needed to be set. This was done using the equal sensitivity and specificity-method, which is found by minimizing the absolute value of their difference (Rainio et al. 2024).

4.5 Model evaluation

4.5.1 Study I

Segmentation performance was evaluated using the Dice score (DSC) and Jaccard similarity coefficient. The model was trained 50 times independently and the median of these models was then chosen based on the Dice score to capture the model that would generalise best for other cohorts.

Let A be the predicted binary mask and B the ground truth binary mask. The Jaccard similarity coefficient is calculated as follows:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Where |A∩B| depicts the intersection of A and B. Similarly, |A∪B| represents the union.

The DSC is calculated as follows:

$$D(A,B) = \frac{(2 * |A \cap B|)}{|A| + |B|}$$

Where |A∩B| depicts the intersection of the A and B, and |A| and |B| are the number of pixels with value 1 in A and B respectively.

In addition, the classification capability of the model was assessed with accuracy, sensitivity and specificity given a pair of predicted mask and the corresponding ground truth, where true positive (TP) segmentations are overlapping areas of predicted pixels and the ground truth, true negative (TN) segmentations are image pairs with no predictions nor ground truth, and false positive (FP) and false negative (FN) segmentations are image pairs with predicted pixels but no corresponding ground truth pixels and vice versa. A cut off value of 9 segmented pixels per image was chosen to reflect the real-world interpretation fidelity of radiologists, which is approximately 5 mm for PET/MRI images.

Sensitivity for the classification was calculated as follows:

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity for the classification was calculated as follows:

$$Specificity = \frac{TN}{TN + FP}$$

Accuracy for the classification was calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

4.5.2 Study II

The classification performance of the models (non-augmented shallow, augmented shallow, non-augmented deep, and augmented deep) was evaluated using sensitivity, specificity and accuracy, as defined in section 4.5.1. In addition, F1-scores were determined as follows:

$$F1 = \frac{2 \cdot TPR \cdot PPV}{TPR + PPV}$$

Moreover, ROC curves were drawn for each model to determine the AUC. Each model was trained 30 times using the primary datasets, giving six iterations for each test set in five-fold cross-validation. The models were re-initialised between iterations to maintain independence. After each iteration, Youden's threshold was calculated from the training data predictions, and model performance was assessed on both the cross-validation sets and an additional test set comprising different cancer types.

4.5.3 Study III

The classification performance of the models (PET, MRI, and PET/MRI-based) was evaluated using sensitivity, specificity and accuracy, as defined in section 4.5.1. In addition, ROC curves were drawn for each model to determine the AUC. To ensure a robust evaluation, the models were trained and validated using stratified 5-fold cross-validation. Ensembles of the cross-validation models were then evaluated on a separate test set using the same metrics applied during validation. Test set inference utilised a sliding window approach with a window size of 32 and a stride of 16. Overlapping windows were used to minimise the risk of missing a tumour visible only in a single window. Four aggregation methods were assessed to derive a final

classification from the sliding window predictions: average, majority vote, median, and weighted average. Majority voting classified the test set as positive if most predicted sliding windows were positive; otherwise, it was classified as negative. The weighted average method was designed to linearly decrease the influence of slices closer to the top of the head, proportionate to the number of evaluated sliding windows. This was done to mimic the clinical knowledge that HNC is typically not present above the nasopharynx level, but the images include slices from the brains that have high metabolic activity. This approach to inference was chosen to preserve sensitivity with overlapping windows and to avoid single-window false positives.

To mitigate the inherent "black box" nature of deep CNNs, a Grad-CAM (Gradient-weighted Class Activation Mapping) was implemented to visualize intermediate network gradients. This method provides insight into the model's decision-making process by highlighting the regions in the input image that the model emphasizes when making predictions. The test set was also reviewed by an experienced head and neck radiologist who was blinded to the patients' medical histories (incl. histopathological proof), except for the knowledge that they had HNC and the location of the primary tumour, excluding the one unknown primary case in the test set. The radiologist assessed the PET/MRI images and provided a binary verdict indicating the presence or absence of residual disease activity at the primary site or new disease activity at secondary sites.

4.6 Ethics

Institutional Review Board approval was obtained from the Hospital District of Southwest Finland for all studies. Written informed consent was waived due to the retrospective nature of these studies.

5 Results

5.1 Study I

After training, the models demonstrated effective segmentation of malignant tissue in the test set, when the segmentation performance was evaluated across the entire test set (Table 6). I.e. the individual 2D slices were considered as a single 3D image. The median model trained on un-augmented PET/MRI images achieved a DSC of 0.81 and a Jaccard similarity coefficient of 0.68. The model trained exclusively on PET data yielded a median DSC of 0.68 and a Jaccard similarity coefficient of 0.52. The model trained on augmented PET/MRI data produced a DSC of 0.71 and a Jaccard similarity coefficient of 0.56 across the entire test set.

Table 6. Segmentation performance in Study I across the entire test set.

Model	Median Dice score	Median Jaccard similarity coefficient
PET	0.68	0.52
PET/MRI	0.81	0.68
Augmented PET/MRI	0.71	0.56

The segmentation performance was then evaluated further by considering the mean of the image and mask pairs where segmentation took place. False positive, false negative and true negative segmentations were excluded from this analysis. The PET- and PET/MRI-based models achieved DSCs of 0.79 ± 0.16 and 0.84 ± 0.14 respectively. The augmented PET/MRI model scored 0.87 ± 0.09. Results of this comparison are presented in Table 7. The models were compared and assessed for statistical differences using the Wilcoxon signed-rank test. A statistically significant difference was observed between the PET-based model and the PET/MRI-based model. However, no significant differences were detected between the PET/MRI and augmented PET/MRI models, nor between the PET and augmented PET/MRI models.

Table 7. Mean segmentation performance of the image and mask pairs where segmentation took place. *: Compared to PET/MRI, §: compared to augmented PET/MRI, †: compared to PET.

Model	Mean Dice score ± SD	P-value	Mean Jaccard similarity coefficient ± SD	P-value
PET	0.79±0.16	0.008*	0.68±0.20	0.008*
PET/MRI	0.84±0.14	0.247§	0.75±0.18	0.273§
Augmented PET/MRI	0.87±0.09	0.156†	0.78±0.13	0.145†

Figure 7 demonstrates a successful segmentation of a HNSCC of the nasopharynx where the DSC between the ground truth (a) and prediction (b) is 0.95. The correlation between the number of predicted pixels and ground truth pixels per 2D image slice was also assessed using linear regression. The coefficient of determination (R^2) for the un-augmented PET/MRI model was 0.90.

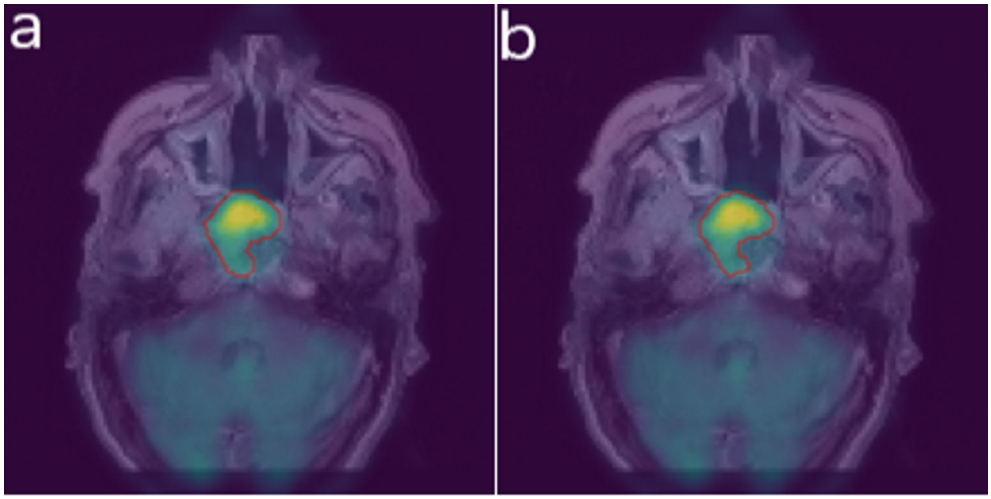

Figure 7. Transaxial FDG-PET/MRI images with a successful segmentation of a HNSCC of the nasopharynx where the Dice score between the ground truth (a) and prediction (b) is 0.95.

The classification performance of the models was moderate. The PET/MRI-based model achieved an accuracy of 0.71, with a specificity of 0.68 and a sensitivity of 0.77. The PET-only model had a higher tendency for false-positive predictions but had a sensitivity comparable to the PET/MRI model. The model trained on augmented PET/MRI data achieved a sensitivity of 0.53, specificity of 0.77, and an overall accuracy of 0.65. Table 8 details the segmentation performances of the median models when true negatives, false positives and false negatives were considered in addition to the true positives. Classification results are presented in Table 9.

Table 8. Segmentation performances of the median models calculated for each individual image slice of the test set. True negatives, false positives and false negatives were considered in these calculations in addition to true positives.

Model	Mean Dice score ± SD	Mean Jaccard similarity coefficient ± SD
PET	0.57±0.46	0.53±0.45
PET/MRI	0.67±0.43	0.64±0.43
Augmented PET/MRI	0.63±0.45	0.60±0.45

Table 9. Classification results of the models on the test set. Abbreviations: TP (true positive), TN (true negative), FP (false positive), FN (false negative), Sens. (sensitivity), Spec. (specificity), Acc. (accuracy).

Data	TP	TN	FP	FN	Sens.	Spec.	Acc.
PET	18	23	19	6	0.75	0.55	0.62
PET/MRI	20	27	13	6	0.77	0.68	0.71
Augmented PET/MRI	17	26	8	15	0.53	0.77	0.65

5.2 Study II

The median values of the evaluation metrics for the predictions on the test set during the 30 iterations for each four models were calculated and are shown in Table 10. The augmented deep CNN scored the highest, with the non-augmented deep CNN and the augmented shallow CNN following and the non-augmented shallow CNN showing the poorest performance. The specificity and AUC values for these models were excellent, while the sensitivity was quite low. This is also reflected in the accuracy and F1-scores.

Table 10. Median values of the evaluation metrics were calculated from the predictions generated across 30 iteration rounds on the primary test set. Abbreviations: F1 (F1-score), AUC (area under the curve).

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1 (%)	AUC (%)
Non-augmented shallow	69.2	65.1	73.6	68.2	74.8
Augmented shallow	74.6	68.6	80.9	72.9	80.7
Non-augmented deep	74.7	70.4	77.6	73.6	82.2
Augmented deep	78.6	72.4	84.9	77.4	85.1

Results of the Wilcoxon tests conducted on six model pairs, comparing evaluation metric values from the primary test sets are presented in Table 11. While Table 10 indicates that all evaluation metrics, except specificity, have higher medians for the non-augmented deep model compared to the augmented shallow model, Table 11 shows that these differences are not statistically significant. However, specificity is significantly higher in the augmented shallow model compared to the non-augmented deep model. Statistically significant differences in accuracy, specificity, F1 score, and AUC were observed across the other five model pairs, with the augmented deep model demonstrating superior performance in both accuracy and AUC compared to the other models.

Table 11. P-values from the Wilcoxon signed-rank test for evaluation metrics derived from predictions on the primary test sets across 30 iteration rounds. Statistically significant values at a 5% significance level are highlighted in bold.

Test	Accuracy	Sensitivity	Specificity	F1	AUC
Non-augmented vs augmented shallow	**< 0.001**	0.497	**< 0.001**	**< 0.001**	**< 0.001**
Non-augmented vs augmented deep	**< 0.001**	0.293	**< 0.001**	**< 0.001**	**0.00159**
Non-augmented shallow vs deep	**< 0.001**	0.251	**< 0.001**	**0.00225**	**< 0.001**
Augmented shallow vs deep	**< 0.001**	**0.00842**	**0.0104**	**< 0.001**	**< 0.001**
Augmented shallow vs non-augmented deep	0.0918	0.141	**< 0.001**	0.718	0.147
Non-augmented shallow vs augmented deep	**< 0.001**	**0.0367**	**< 0.001**	**< 0.001**	**< 0.001**

Table 12 examines the relationship between the median sensitivity values and locations of HNSCC of the primary test set. Higher sensitivity values were seen for all four models for tumours on the root of tongue, in the fossa piriformis, in the oral cavity, and in the larynx. The median sensitivity was lower; however, when the tumour was in the oropharynx (excluding root of tongue), or the nasopharynx. Unknown primaries had higher sensitivity as well, whereas hypopharynx (excluding the piriform fossa), the upper oesophagus, nasal cavity, salivary glands, sinuses, and skin had lower sensitivity. This is most likely due to their very low frequency in the data.

Table 12. Subgroups of head and neck squamous cell carcinoma (HNSCC) patients categorized by tumor location, number of patients per subgroup, total number of positive slices within each subgroup, and the median sensitivity (%) calculated from predictions on the primary test sets over 30 iteration rounds. Values exceeding the overall sensitivity of the primary test sets for the respective model (refer to Table 10) are highlighted in bold.

HNSCC tumour location	Patients	Slices in total	Non-augmented shallow	Augmented shallow	Non-augmented deep	Augmented deep
Oral cavity	18	203	**81.7**	**70.4**	**78.3**	**75.0**
Root of tongue	11	132	**96.8**	**86.7**	**97.7**	**83.3**
Oropharynx (exc. root of tongue)	13	141	50.8	33.3	51.1	35.4
Nasopharynx	8	83	33.3	26.1	39.4	45.2
Larynx	7	54	**65.8**	**73.7**	**73.7**	**77.0**
Fossa piriformis	5	90	**86.7**	**80.5**	**93.3**	**80.2**
Unknown primary	13	147	**90.6**	**88.5**	**90.6**	**88.5**
Other	14	145	50.0	55.8	55.2	67.7

Table 13 presents the evaluation metrics values computed from the additional test set comprising of patients with head and neck cancer diagnoses other than HNSCC. Compared to Table 10, slightly lower performance was seen on this additional test set relative to the primary test sets containing only HNSCC cases.

Table 13. Median evaluation metric values calculated from predictions on the additional test set over 30 iteration rounds.

CNN	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1 (%)	AUC (%)
Non-augmented shallow	63.0	72.3	52.5	65.9	70.1
Augmented shallow	67.6	61.3	73.9	65.6	71.9
Non-augmented deep	70.4	73.5	67.6	70.7	78.1
Augmented deep	75.0	69.7	80.3	73.4	79.4

Table 14 shows that all models accurately detect follicular and papillary carcinoma of the thyroid gland, as well as mucoepidermoid carcinoma of the parotid gland, but fail to identify chondrosarcoma of the neck. Notably, the augmented deep model demonstrates markedly lower sensitivity for adenocystic carcinoma of the oral cavity, adenocarcinoma of the oesophagus, and adnexal carcinoma of the skin compared to the other three models.

Table 14. Subgroups of positive head and neck cancer patients categorized by diagnosis, number of patients per subgroup, total number of positive slices within each subgroup, and median sensitivity (%) calculated from predictions on the additional test set over 30 iteration rounds. Values exceeding the overall sensitivity of the additional test set for the corresponding model (refer to Table 13) are highlighted in bold.

Diagnosis	Patients	Slices in total	Non-augmented shallow	Augmented shallow	Non-augmented deep	Augmented deep
Adenocystic carcinoma of oral cavity	3	24	70.8	**70.8**	70.8	50.0
Adenocarcinoma of oesophagus	2	29	**86.2**	**82.8**	**82.8**	69.0
Chondrosarcoma of neck	1	9	0.00	0.00	0.00	0.00
Papillary carcinoma of thyroid gland	2	33	**100**	**100**	**100**	**100**
Follicular carcinoma of thyroid gland	1	12	**100**	**100**	**100**	**91.7**
Mucoepidermoid carcinoma of parotid gland	1	5	**100**	**100**	**100**	**100**
Adnexal carcinoma of skin on neck	1	7	**100**	**100**	**100**	64.3

5.3 Study III

During cross-validation, some overfitting and volatility between epochs were observed across all imaging modalities, which is expected given the sample size and heterogeneity of the dataset. Among the imaging modalities, PET and PET/MRI produced comparable results, while MRI performed lower overall (Table 15). The PET-based model achieved mean accuracy, sensitivity, specificity, and AUC values of 0.75, 0.72, 0.77, and 0.82, respectively, across the 5 validation folds. For the MRI-based models, these metrics were 0.66, 0.63, 0.68, and 0.73, while the PET/MRI-based model achieved 0.72, 0.69, 0.74, and 0.80. Figure 8 illustrates the AUCs obtained by the PET-based models during the 5-fold cross-validation. Details of true negatives, false positives, false negatives, and true positives for all models and validation folds are provided in Table 16.

Table 15. Average and median performance across the training folds during the 5-fold cross-validation process.

Data	Method	Accuracy	Sensitivity	Specificity	AUC
PET					
	Mean	**0.75**	**0.72**	**0.77**	**0.82**
	Median	**0.78**	**0.75**	**0.80**	**0.86**
MRI					
	Mean	0.66	0.63	0.68	0.73
	Median	0.68	0.65	0.70	0.74
PET/MRI					
	Mean	0.72	0.69	0.74	0.80
	Median	0.73	0.70	0.75	0.81

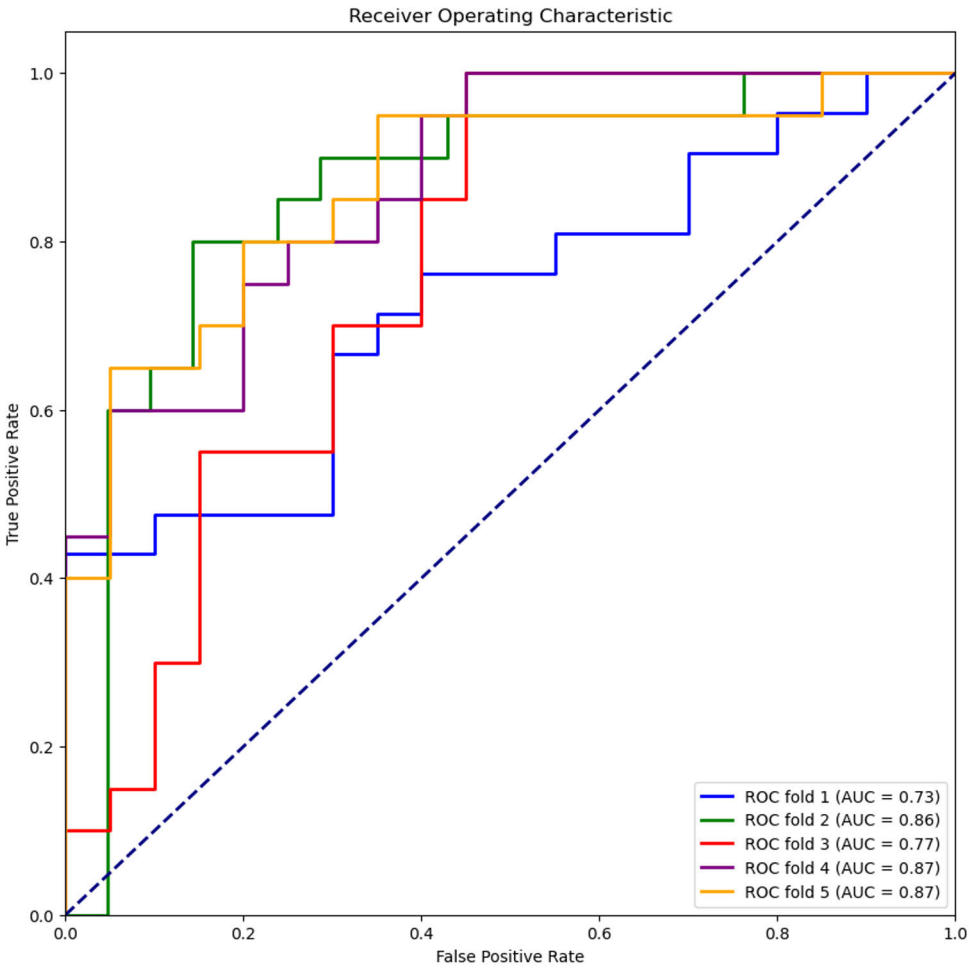

Figure 8. AUC values and the corresponding ROC curves for the cross-validation of the PET-based model.

Table 16. Classification results for each validation fold of each model.

Modality	Validation Fold	True Negatives	False Positives	False Negatives	True Positives
PET	1	14	6	7	14
	2	18	3	4	16
	3	14	6	7	13
	4	16	4	5	15
	5	16	4	5	15
MRI	1	15	5	6	15
	2	13	8	9	11
	3	15	5	6	14
	4	12	8	9	11
	5	14	6	7	13
PET/MRI	1	15	5	6	15
	2	15	6	7	13
	3	13	7	8	12
	4	17	3	4	16
	5	15	5	6	14

Following the 5-fold cross-validation procedure, an ensemble of models for each imaging modality was employed to classify the independent test set. In terms of accuracy, averaging or taking the median of the sliding window predictions yielded the best results for PET, MRI, and PET/MRI-based models, achieving accuracies of 0.90, 0.75, and 0.60, respectively. For PET and PET/MRI models, the highest sensitivities of 1.0 and 0.70 were obtained using averaging, median, or weighted average. The MRI-based model achieved its highest sensitivity of 0.70 with the weighted average method. For PET and MRI models, the highest specificities of 0.80 and 0.90, respectively, were observed using averaging, majority voting, or median. For the PET/MRI model, the highest specificity of 0.50 was consistent across all methods. Regarding AUC, the weighted average performed best for the PET model, scoring 0.94. For the PET/MRI model, the highest AUC of 0.74 was achieved using the median of the sliding windows, while for the MRI model, the best AUC of 0.79 was obtained with averaging. The radiologist identified all cases correctly. These results are shown in Table 17.

Table 17. Results of the sliding window inference on the test set compared with radiologist performance.

Data	Method	Accuracy	Sensitivity	Specificity	AUC
PET/MRI					
	Average	0.60	0.70	0.50	0.72
	Majority Vote	0.55	0.60	0.50	NA
	Median	0.60	0.70	0.50	0.74
	Weighted Average	0.60	0.70	0.50	0.73
PET					
	Average	**0.90**	**1.00**	**0.80**	0.92
	Majority Vote	0.70	0.60	0.80	NA
	Median	**0.90**	**1.00**	**0.80**	0.92
	Weighted Average	0.75	1.00	0.50	**0.94**
MRI					
	Average	0.75	0.60	0.90	0.79
	Majority Vote	0.60	0.30	0.90	NA
	Median	0.75	0.60	0.90	0.77
	Weighted Average	0.70	0.70	0.70	0.78
Radiologist		**1.00**	**1.00**	**1.00**	NA

The highest-performing PET-based model was further evaluated for interpretability using Grad-CAM. The Grad-CAM results indicate that positive predictions are correctly based on regions exhibiting high malignant metabolic activity. This is exemplified in Figure 9, which shows a patient with significant residual disease (SCC of the base of the tongue) correctly classified as positive. A similar example is presented in Figure 10, where a patient with SCC of the oropharynx was also correctly identified as positive.

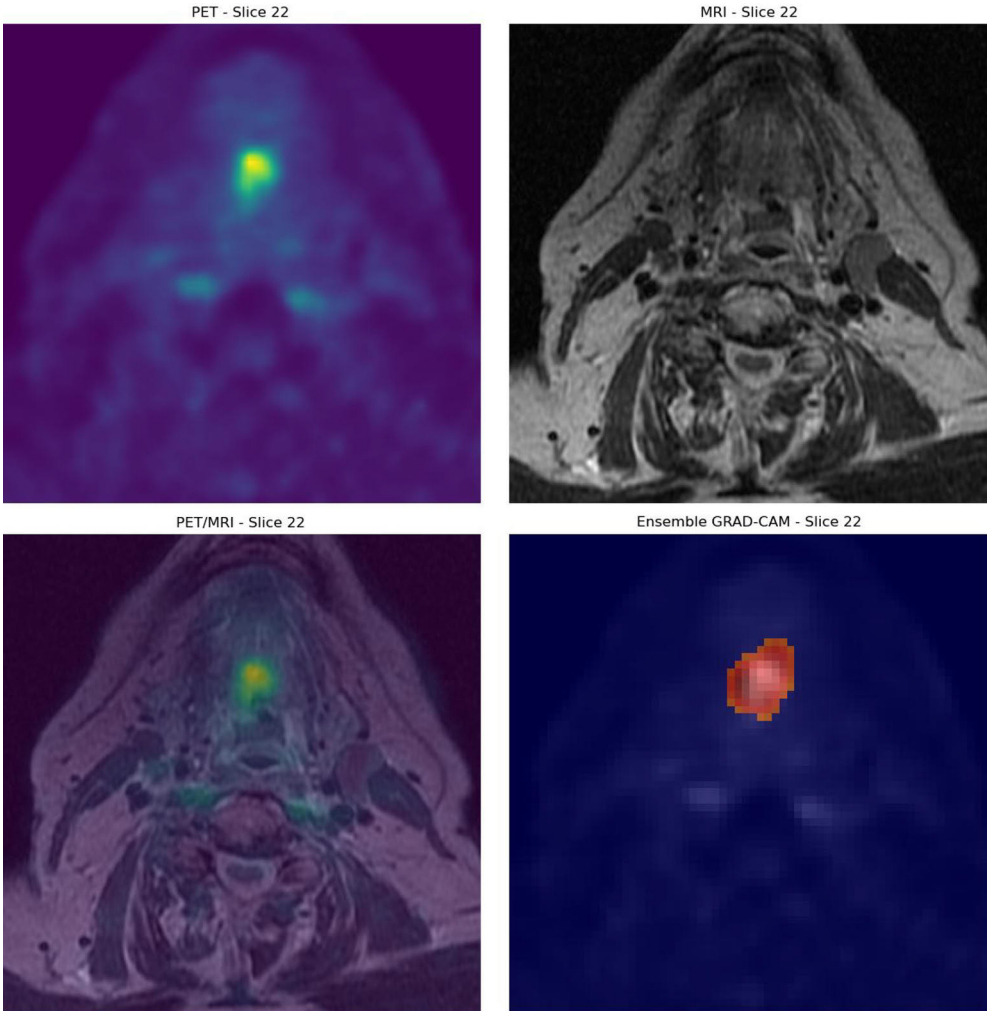

Figure 9. Grad-CAM of patient with a residual SCC of the base of the tongue who was correctly classified as positive.

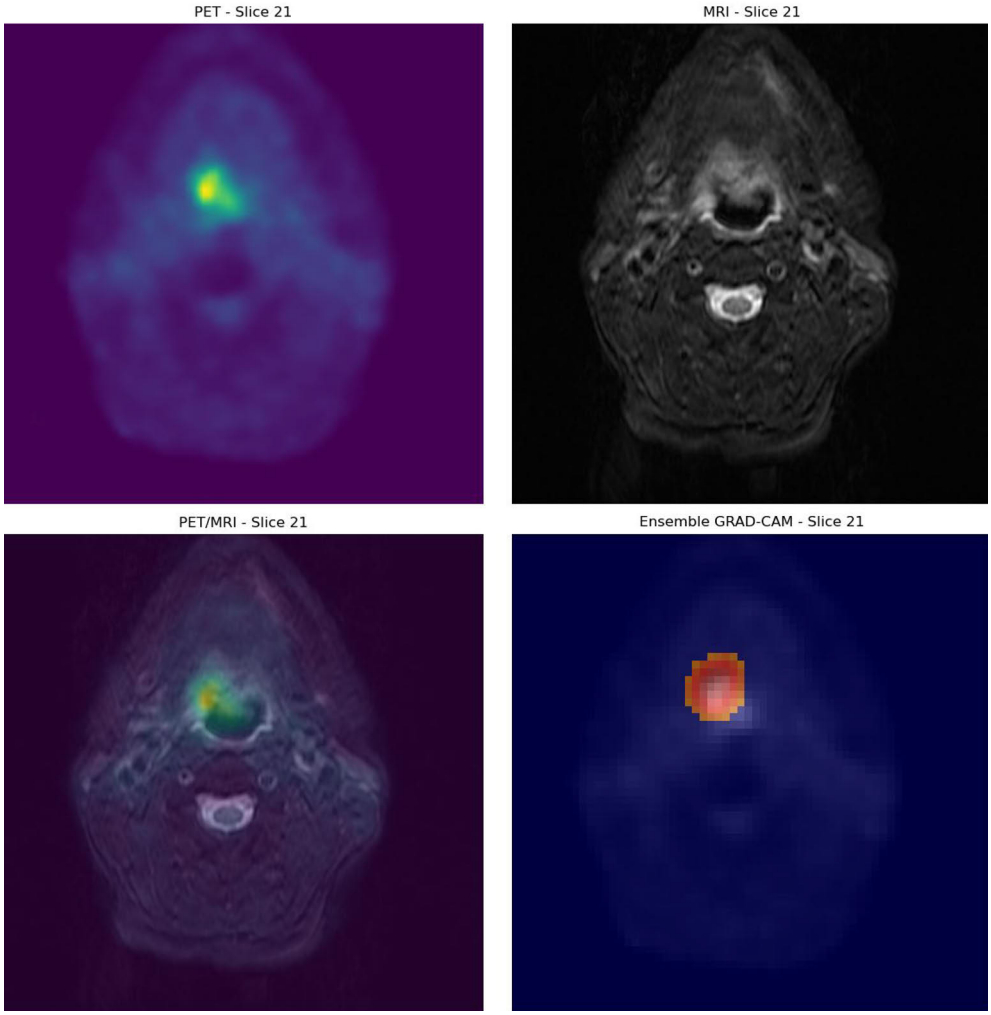

Figure 10. Grad-CAM of patient with a residual SCC of the oropharynx who was correctly classified as positive.

To illustrate the relationship between the PET-based model's observations and its final classification decisions, the correlation between the maximum Grad-CAM pixel intensity per image slice and slices containing cancer was analysed. The results of this analysis for all patients classified as positive (10 true positives and 2 false positives) are shown in Figure 11. In this figure, each box in a column represents a single transaxial PET slice. The left column displays the binary ground truth for each slice, with black indicating a negative slice and white indicating a positive slice, i.e., the presence of cancer in the original images. The right column shows the maximum pixel intensity from the Grad-CAM heatmap for each slice, scaled between 0 and 1. For patients correctly classified as positive (1–10), an overlap was observed between

Grad-CAM activity and slices labeled as positive. As expected, no such overlap was seen for the false positives, as all slices in these cases were negative.

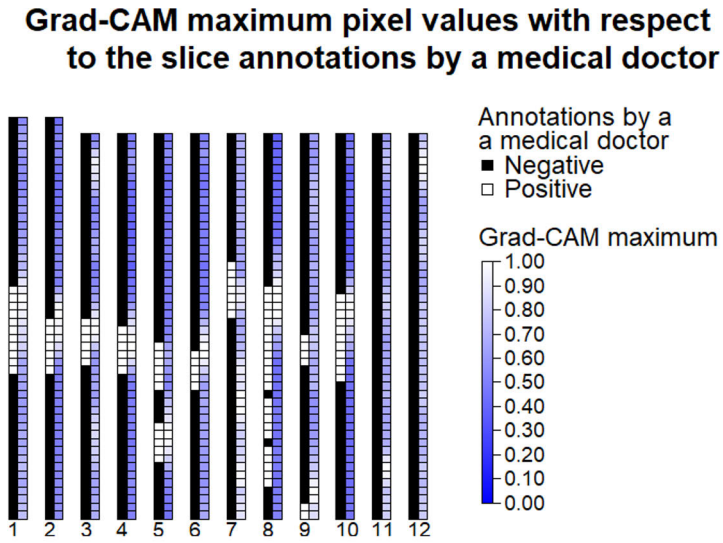

Figure 11. Correlation between the maximum pixel intensity per slice in the Grad-CAM heatmap in patients classified as positive by the PET-based model and the binary ground truth values for the image slices. The x-axis indices 1–10 represent the 10 test set patients correctly classified as positive, while indices 11–12 correspond to negative patients incorrectly classified as positive.

The best-performing model was the PET-based model using averaging (and median) for sliding window aggregation. This model achieved a sensitivity of 100% and a specificity of 80% on the test set, indicating two false-positive predictions. The raw prediction scores for these false positives were 0.92 and 0.88, while the decision threshold for the ensemble model was set at 0.87 (the mean of the five thresholds obtained during cross-validation). The first case involved a patient who had been successfully treated for SCC of the left tonsil. The PET/MRI revealed abnormal metabolic activity in the region, which was determined to be non-malignant. The radiologist also identified non-specific activity in this area. However, the Grad-CAM analysis did not highlight any specific regions of interest in this area. Instead, the model appeared to focus primarily on the shoulder muscles, as illustrated in Figure 12.

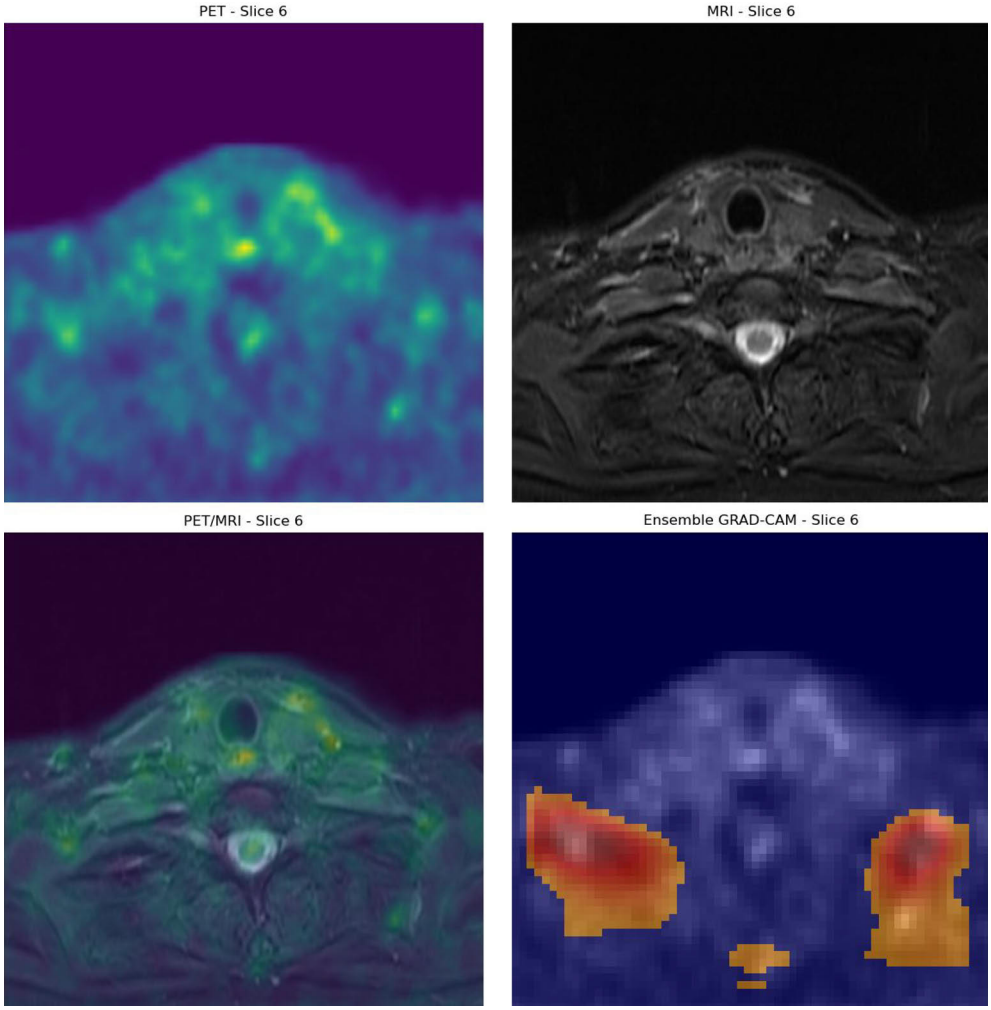

Figure 12. Grad-CAM of the first false positive classification where the PET-based model focused most intensely on the patients shoulder muscles without any apparent reason and ultimately classified the patient incorrectly as positive.

The second false positive involved a patient treated for epithelial cancer of the right sinus. The PET/MRI revealed prominent but diffuse metabolic activity in the sublingual salivary gland, which was determined to be benign. However, Grad-CAM analysis indicated that the model's attention was primarily focused on a region of the skull, as shown in Figure 13.

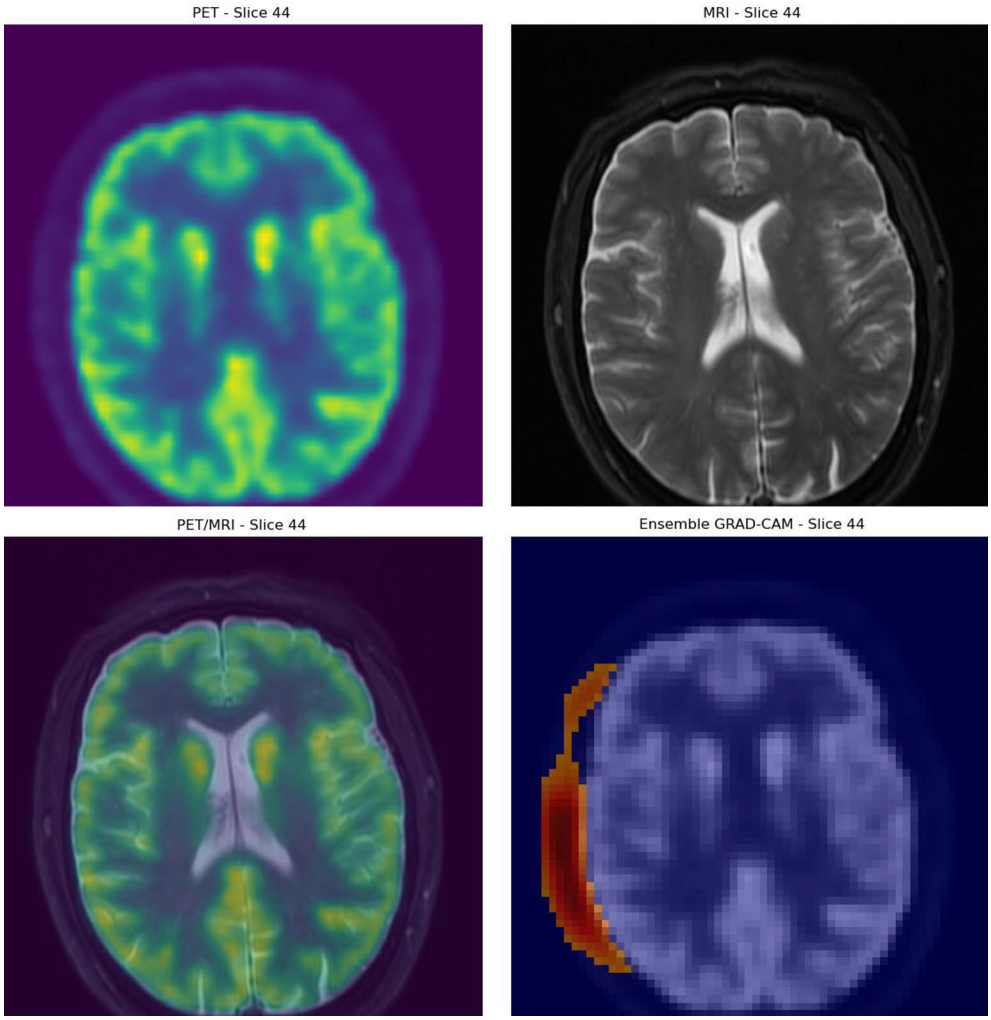

Figure 13. The second false positive case of the test set, where the model's primary focus is located around a section of the skull bone, indicating a clear false positive finding.

6 Discussion

6.1 Performance and Clinical Utility

This section evaluates the performance and clinical utility of the developed models across three key aspects. First, the classification capabilities of the models from studies I, II, and III. Second, the segmentation performance of the models developed in Study I. Third, the interpretability of these models, as analysed in Study III.

6.1.1 Classification

6.1.1.1 Study I

While Study I mainly examined the 2D segmentation performance of the model, classification metrics were evaluated as well, since the training and test data included negative controls. The non-augmented PET/MRI model achieved the highest overall accuracy of 71%, with PET and augmented PET/MRI -models scoring 62% and 65% respectively. The sensitivity and specificity for the PET/MRI model were 77% and 68% respectively. Although augmentation is often expected to improve generalisation, in this work the "classification" outcome is implicitly derived from the segmentation output; consequently, any reduction in lesion detectability or tendency to produce smaller predicted regions can disproportionately increase false negatives and reduce sensitivity, thereby lowering overall accuracy. Given the limited number of unique slices (n = 356) and the resulting redundancy introduced by geometric augmentation, it is plausible that augmentation increased apparent variability without adding new biological variability and may therefore have provided limited benefit for generalisation in this setting. This is consistent with prior observations that augmentation is not universally beneficial; for example, Elgendi et al. reported that several commonly used augmentation strategies reduced validation accuracy compared with training without augmentation in a chest X-ray deep learning classification task (Elgendi et al. 2021). Overall, moderate classification performance is not unexpected because the network was optimised for segmentation rather than dedicated classification.

6.1.1.2 Study II

In Study II CNNs were designed to classify 2D slices of 3D PET volumes. The rationale for this decision was to enable efficient training and inference with the available data and computing resources, as training 3D models requires significantly more computing power for the same data. Moreover, using only PET images ensures compatibility with both PET/CT and PET/MRI scanners. All models showed high specificity and identified negative slices correctly, with the best performing augmented deep model achieving median specificities of 85% and 80% on the primary and additional test sets respectively during the 30 iteration rounds. The sensitivity results were notably lower however, with the augmented deep model achieving median sensitivities of 72% and 70% respectively on the primary and additional test sets. Furthermore, AUC values were relatively high in comparison to the other metrics, which indicates that Youden's threshold may not be optimal for this type of binary classification application. In subsequent work, a sensitivity-weighed threshold was found beneficial in these types of settings, where a high sensitivity is crucial (Rainio et al. 2024). Using all available negative image slices in the dataset might have further increased overall model accuracy, but likely at the cost of sensitivity. In addition, data augmentation improved model performance in the shallow and deep models, with the deep models outperforming the shallow ones, as expected.

The effect of tumour location on sensitivity was notable, with tumours at the base of the tongue and in fossa piriformis performing better than other subgroups. This is likely due to the specific anatomical characteristics of these regions and how frequently they appear in the data. This is further supported by the relatively low sensitivities of more uncommon tumour locations such as the nasopharynx, as these are so scarcely represented in the data. A difference in performance was also seen with the additional test set, where the models performed slightly worse than with the primary test set. This difference was expected, as the models were only trained with HNSCC data. However, the difference between these groups was only a couple of percentage points, suggesting that models trained on HNSCC data can be used to detect other types of cancer and are good candidates for transfer learning. For instance, these models could provide starting points for developing models for detecting tumours of the thyroid or parotid glands. It was also noted that the augmented deep model had lower sensitivity in detecting adenocystic carcinoma of the oral cavity, adenocarcinoma of the oesophagus, and adnexal carcinoma of the skin. This is plausibly related to limited representation of these entities in the training data. As discussed in chapter 6.1.1.1, standard augmentation primarily increases geometric variability and therefore may provide limited additional information for rare subgroups because it does not introduce new biological/pathological variability or substantially different tumour–background contexts. In this setting, augmentation

may therefore have limited benefit—and can even be detrimental—for infrequent histologies. More robust improvement is likely to require additional representative cases and/or imbalance-aware training strategies, while keeping augmentations anatomically plausible.

6.1.1.3 Study III

The classification results obtained in the third study with the 3D based models are promising, especially the PET-based model with a mean accuracy of 75% during cross-validation and 90% for the separate test set. The notably better performance in the test set could be explained by the same issue as described in the previous chapter, where the presence of outliers affects the model performance significantly when the available data is limited. The PET-based model outperformed the PET/MRI and MRI-based models on the test set and in cross-validation, indicating the importance of the metabolic signal in detecting HNC. During cross-validation, the PET/MRI-based model underperformed slightly compared to the PET-based model with a mean accuracy of 72%, whereas the MRI-based model only achieved a mean accuracy of 66%. The poorer performance of the MRI-model was to be expected based on our previous experience in Study I, where the MRI-based model failed to produce any meaningful results in 2D segmentation. Interestingly, the PET/MRI-based model resulted in worse classification results than either modality alone on the test set. The model's inability to benefit from dual-modality images might be related to the small size and heterogeneity of the HNC dataset, where the anatomical signal interferes with the metabolic, instead of providing supporting information. This effect may be particularly prevalent in datasets like ours, where the majority of tumours exhibit high metabolic activity, thus reducing MRI's complementary value. In addition, given the overfitting observed during training, the PET/MRI model could be expecting consistent complementary signals learned from the training data but then struggles when faced with situations where the two modalities conflict.

The trained models were evaluated against the gold standard, defined as histopathological confirmation for positive cases and absence of disease on follow-up for negative cases. For reference, an experienced head and neck radiologist independently reviewed the images. In this dataset, the radiologist's assessments were fully in line with the gold standard, achieving perfect accuracy and surpassing the model performance. However, clinical practice shows that radiologists' interpretations are not always perfect, a larger test set could have resulted in more human errors. Moreover, instead of only a binary classification by the radiologist, a more fine-grained score, such as five-point scale, could have simulated the degree of uncertainty in the interpretation and thus been easier to compare with the models and their decision threshold. Also, unlike the radiologist, the PET/MRI model was not

able to benefit from the joint information. Addressing this issue with more robust data will likely improve model performance.

This study demonstrates the feasibility of analysing entire 3D PET/MRI volumes of variable depths using sliding window inference. This method also provides benefits in terms of efficiency when compared to 2D methods that require aggregating the 2D predictions to produce a patient-wise classification, which is what provides the clinical value. In addition, when compared with human observer's efficiency, the gains are drastic, as these models perform the classification of a whole PET/MRI volume in the fraction of a second. The PET-based model achieved 100% sensitivity and 80% specificity on the test set, suggesting it could be used as an efficient pre-screening tool in clinical practice. Currently, there are no directly comparable classification studies done on HNC PET/MRI images. However, Chen et al. and Dohopolski et al. investigated lymph node classification from PET/CT images using 3D CNNs (Chen et al. 2019; Dohopolski et al. 2020; Chen et al. 2021). Chen et al. were able to achieve an accuracy of 88% in multiclass (normal, suspicious, and involved) classification with a dataset of 59 patients. In a subsequent study they achieved 98% accuracy on binary classification using a dataset of 129 patients. Using the same data, Dohopolski et al. achieved 99% AUC in binary classification. The crucial difference between these studies and studies II and III of this thesis is that the models in them were trained only on cropped images of lymph nodes, whereas in studies II and III whole 2D and 3D images were utilised in training and classification was done for the whole image.

Overall, even though the models showed relatively high classification performance, further improvements and testing are required before clinical adoption.

6.1.2 Segmentation

The 2D U-Net developed with PET/MRI data in Study I was able to produce accurate segmentations on images where cancer was detected, achieving a mean DSC of 84%. The PET-based and the augmented PET/MRI models had DSCs of 79% and 87% respectively. However, as described earlier, the non-augmented PET/MRI model was the most accurate in identifying ROI's for segmentation among the positive and negative test images, thus explaining why the PET/MRI model had a DSC of 81% and the PET and augmented PET/MRI models had 68% and 71% respectively, when the test cases were considered as a single 3D image. Moreover, when including true negative, false positive, and false negative image pairs, the DSCs for PET/MRI, PET, and augmented PET/MRI were 67%, 57%, and 63%, respectively. These results highlight the 2D model's capability of making accurate segmentations once a true positive case is identified, and on the other hand demonstrate the impact of

classification accuracy on the overall segmentation performance, when positive and negative cases are processed with segmentation models.

These results are roughly in line with those published by Ren et al on PET/MRI segmentation and others on PET/CT segmentation (Ren et al. 2021; Moe et al. 2021; Andrearczyk et al. 2020; Huang et al. 2018). Others have also reported higher segmentation accuracies, as shown in Table 2. However, direct comparison between studies is difficult, due to differences in methodology and data. For instance, pre-processing protocols, algorithms used, inclusion of negative controls, and validation methods vary greatly. In terms of this study, the use of MRI as the anatomical counterpart to PET could explain the relatively high DSC for the true positive cases on such limited data. Similarly, the limited data is likely a key contributor to the modest classification accuracy of this model. Moreover, the PET images seem crucial for the segmentation task, at least within this data, since an MRI-based model was also trained, but did not provide any meaningful results. Notably, the augmented PET/MRI model achieved high mean DSC on true positive cases, yet its overall performance decreased once negative controls and misclassifications (false positives/false negatives) were included. This is consistent with the fact that aggregate DSC in a mixed positive–negative evaluation is dominated by ROI identification: missed lesions and spurious predictions substantially lower the overall score even if overlap is strong when a true tumour is detected. As discussed in 6.1.1.1, in a limited dataset, geometric augmentation may add correlated variants without increasing biological/pathological variability, which may reduce generalisation and degrade ROI identification—providing a potential explanation for the lower overall performance of the augmented model.

Despite showing initial promise in selected segmentation cases, increased datasets, refined pre-processing protocols and model architectures are required for creating more robust segmentation models ready for clinical practice. Furthermore, utilising 3D segmentation models to capture spatial information will likely provide the most clinical benefit, as they enable direct MTV and GTV calculations without the need for aggregating 2D segmentations. It is also feasible that combining a classification model with a segmentation model could provide a more efficient interpretation pipeline, where only the cases classified as positive continue to the segmentation model, thus eliminating the need to train the segmentation models with negative controls.

6.1.3 Interpretability

The ‘black box’ nature of DL is a key issue in developing reliable and robust AI applications. This is especially true in safety critical domains, such as medical AI. For use cases like segmentation, where the output of the model is easily illustrated

and verified by an expert, this might not be as big of a problem. In contrast, use cases such as binary classification might only provide a single number as a result that is then thresholded into a classification. Without any additional reasoning or justification, even if correct, this result does not evoke trust or confidence in the end user that is required for routine clinical use. Moreover, interpretability is important to reduce the risk of "Clever Hans" behaviour, where a model appears to perform well by relying on spurious correlations or confounding cues rather than clinically meaningful features (Lapuschkin et al. 2019). In addition to building trust via transparency, interpretability measures can be used to identify imaging biomarkers from the data that might otherwise be missed by the human observer.

The interpretability of the binary classifications produced by the highest performing PET-based 3D CNN model was explored in Study III using Grad-CAM. To visualize the distribution of Grad-CAM heatmap activity for each positively classified patient along the depth axis, the maximum pixel value of each Grad-CAM heatmap per 2D slice of the PET volume was compared with the binary ground truth of the corresponding image slice (Figure 11). The maximum pixel intensity of Grad-CAM was found to correlate with the depthwise locations of the HNC in the positively classified patients, indicating that the true positive classifications on the test set are based on real areas of interest. However, tumours with little or no metabolic activity, such as necrotic lymph nodes, might receive less attention, because the Grad-CAM was based on the PET-model, which explains the transaxial slices in Figure 11, where there is cancer, yet only low or moderate Grad-CAM intensity. Moreover, comparing the maximum heatmap value of a single pixel in a transaxial slice with the ground truth value is a simple proxy to help visualise the depthwise attention distribution of the model. For example, this approach does not take the number of high-intensity pixels (i.e. the size of the ROI) into consideration at all. Some interpretations might also be made based on the presence of normal anatomical structures, which however is not reflected in this analysis either. Furthermore, the Grad-CAM heatmaps for the two false positive classification cases provide no clear indication why the model's peak interest would be on the sections showed in Figures 12 and 13, as these ROIs overlap the patient's shoulder muscles in Figure 12 and a section of the skull for the patient in Figure 13. From a clinical point of view these types of false positive errors are easy to identify and correct. Similarly, it is easy to agree with the heatmaps on true positive findings, as they seem to capture the areas of malignant metabolic activity quite well. Despite providing valuable supportive information, Grad-CAM only gives insight into which areas are relevant for the prediction, but not why they are relevant. This is an important distinction from a clinical point of view. Overall, Grad-CAM provides useful information on model decisions and can help facilitate the transition into clinical practice, provided its limitations are kept in mind.

6.2 Limitations

This chapter discusses the limitations of this thesis and more broadly the limitations encountered in the field of Medical ML. The attempts to mitigate these issues through evaluation guidelines and best practices are also discussed.

6.2.1 Data Quantity and Quality

The number of training examples needed for producing an accurate DL model has been contemplated in the field of computer vision, but no clear values can be set for determining the minimum threshold for a given computer vision problem. For instance, Krizhevsky et al. used a dataset of over 1 million images to train their famous AlexNet in the 2012 Imagenet challenge (Krizhevsky et al. 2012). Many others also report data set sizes with image samples ranging from 10 000 upwards in similar 2D image classification tasks. However, with medical imaging the tasks are often extremely specific, and data is scarce. Cho et al. investigated determining the optimal training data size in a 2D classification problem where they sought to classify whole body CT scans into 6 different anatomical classes (Cho et al. 2015). They found that after approximately 100 training samples per class the model begins to plateau, when at 100 samples per class the model reached an accuracy of 90% and at 200 samples the model scored 96%. The authors also state that to reach a desired accuracy of 99.5%, given the demands of the medical domain, they would need 4092 samples per class based on the predicted learning curve.

In studies I and II the data consisted of 290 and 1990 training samples respectively, with a 50:50 ratio of positive and negative samples. Considering the results obtained by Cho et al. both studies should have enough data to achieve excellent results. However, there are several factors that influence the data requirements of a given problem. First, these studies were binary classification (segmentation is effectively classification on pixel level), whereas Cho et al. conducted multiclass classification. Second, it is unclear how well their findings apply to segmentation problems, even though segmentation can be seen as pixel level classification. Third, the heterogeneity that exists within images of different types of HNC is much greater than in images of normal human anatomy. Finally, Study III involved 3D binary classification of HNC, where 202 patients (50:50 negative to positive ratio) with 32 image slices per each patient were included in the training data, totalling 6464 image slices. According to Cho et al. these 3232 image slices per class should be enough to provide very high accuracies, yet the median accuracy during the with 5-fold cross-validation was 78%, even with prior augmentation to quadruple the data size. It should be noted that not all of the 32 slices for each positive patient contained cancer, but they were included to achieve uniform depth

dimensions for the 3D images. Thus, it is likely that the dimension of the problem (2D vs. 3D) influences its data requirements.

The binary classification of HNC and the multiclass problem of different anatomical sites by Cho et al. are of course not directly comparable, since same anatomical sites tend to resemble each other quite closely forming a distinct class pattern, whereas in HNC the location and the anatomical characteristics of a tumour can differ significantly within the data and yet belong to the same class. Thus, it is clear that the frequency at which a given type of tumour appears in the data of studies II, III, and especially in Study I can be very low. The heterogeneity of the data and its relative scarcity are the two most obvious limiting factors with the largest impact on model performance in Studies I, II, and III. To mitigate this issue, data augmentation was used in these studies to improve performance. However, the drawback of this approach can be overfitting and reduced ability to generalise, as was discussed in 6.1.1.1.

In addition to quantity, data quality is a key concern in developing robust machine learning models. It must be labelled in a coherent and precise manor to avoid giving the model mixed signals during training. In the three studies of this thesis, key factors affecting data quality were image acquisition on two separate PET scanners due to our centre introducing a new machine in March of 2020. This affected studies II and III where roughly 36% of PET data came from the older sequential machine and 64% from the newer machine. Moreover, the MRI sequences used in studies I and III varied based on availability and quality. For instance, in some patients all neck specific sequences were not available due to disease related discomfort laying down in the MRI-machine, or claustrophobia, and movement related artifacts. The sequential scanning method used until 3/2020 also differs from the newer hybrid method. The inclusion of scans from different machines and varying MRI sequences increases the heterogeneity of the data even more. It was in part a necessity due to circumstances related to our centre and individual patients. However, it was also an active decision to include these diverse data to better simulate real world situations. This decision had likely some amount of negative impact on the model performance but can ultimately help generalise these models better in the long run, as more data is gathered. Furthermore, all images were pre-processed before training, including downsampling of the image dimensions to allow a more efficient training procedure on limited computing resources. With this process, some information is inevitably lost. The ratio of positive and negative samples was also set to approximately 50:50 for all studies, which does not reflect the clinical reality, where the majority of cases are negative. Lastly, the data was annotated by a single observer, which inherently comes with a bias. Aggregating an annotation consensus by several experts could provide more accurate annotations in

the future. Overall, HNCs are a diverse group of diseases and going forward there should be even more emphasis on data quantity and quality.

6.2.2 Medical ML Evaluation Guidelines

Some frameworks and best practices for medical ML have been proposed to provide a more robust way of conducting these studies and improving their comparability (Mongan et al. 2020; Müller et al. 2022; Hicks et al. 2022). These guidelines support the use of structured checklists when conducting a study and reporting its findings, such as the CLAIM checklist proposed by Mongan et al. An emphasis should be placed on clearly defining data partitions to avoid leakage. Appropriate metrics to suit the evaluated task should be chosen. For instance, DSC and/or Jaccard similarity coefficient for segmentation, especially when class imbalance is in favour of the background. Class imbalances should be analysed and reported among the training and test data to identify differences in performance across varying prevalence scenarios. In addition, a thorough error analysis should be conducted. Visualisation and expanding on summary statistics (accuracy etc.) with confusion matrices and score distributions is encouraged to reveal any biases or outliers and give insight to the mode of failure. Moreover, explainability techniques are recommended for tying the performance metrics back to clinical relevance and help clinicians to interpret AI decisions. External validation and sharing the code and data to reproduce the results is suggested, when possible. Lastly, constant maintenance and updates to the reporting guidelines through community consensus are crucial to ensure guideline alignment with the rapid advancements in the field.

Despite these proposed frameworks, adherence to them is inconsistent and their formal endorsement by journals is low (Zhong et al. 2023; Koçak et al. 2025). Koçak et al. identified considerable reporting gaps in terms of CLAIM checklist items, where on average a third were missing. Majority of the analysed studies underreported 11 items consistently (missing in ≥50%): De-identification methods, Data imputation, intended sample size and its determination, Flow of participants using a diagram to indicate inclusion and exclusion, Demographic and clinical characteristics for each data partition, statistical measures of significance and uncertainty, explainability and interpretability methods, estimates of diagnostic accuracy and their precision, failure analysis, registration number and name of registry, and access to full study protocol. This was seen as a broader reflection of the challenges in the field of Medical AI, where obtaining sufficient sample sizes, addressing uncertainty and interpretability are some of the core problems.

6.3 Future Prospects

As discussed in the previous section, the key challenge for difficult computer vision tasks such as this, lies in collecting enough quality data that accurately reflects the characteristics of the disease in a balanced way. Addressing this issue will remain the top priority going forward. Achieving clinical levels of accuracy in HNC segmentation and classification will require more robust PET/MRI data sets. However, simply recruiting radiologists or nuclear medicine physicians to produce more research data is problematic in several ways. First, the availability of these professionals for such work is limited due to the already heavy workload. For this reason, a majority consensus annotation, which would be preferred to mitigate some of the inter- and intra-observer variability, is very hard to organise with traditional means. Moreover, to enable smooth annotation the clinicians should have easy and uniform tools with minimal learning curves to enable fast and accurate annotation without needing to consider the details of pre-processing requirements for the deep learning models. With these goals in mind, the development of a cloud-based system with a graphical user interface that allows distributed annotation and incorporates the already built models for DL assisted annotation is on the way. This system will facilitate a more efficient data collection for a variety of medical computer vision models and is designed to be used Finland-wide. Among other clear benefits of such system, is a uniform annotation process that is not dependent on any local university hospital patient record systems nor protocols, supporting the gold standard of majority consensus annotation by several experts.

In addition to streamlining the expert annotation process, automated means of training data generation will play an increasingly important role in the future. Generative adversarial networks were introduced in 2014 by Goodfellow et al. for this purpose and have recently been used successfully in medical image generation as well (Goodfellow et al. 2014; Singh and Raza 2021). GAN's provide several benefits to the annotation process. First, they are able to simulate large quantities of images and corresponding labels extremely fast, which reduces the manual annotation burden greatly. Second, they can be used to address class imbalances by generating a proportionally higher number of the rare cases that exist within the original data. However, GANs are fundamentally constrained by the distribution of their training data and generally do not generate clinically valid phenomena that are not represented in that data. On the other hand, within the range of variability learned from the training set, they can introduce diversity such as imaging conditions and plausible variation in tumour appearance to support better generalisation. Lastly, they may also be used in cross-modality image synthesis. For instance, PET/CT is the prevalent imaging modality for HNC recurrence in many settings due to availability, and GAN-based translation could be explored to synthesise PET/MRI-like representations from PET/CT to support transfer learning, noting that such

synthetic data would require careful validation to avoid unrealistic anatomy or "hallucinated" pathology.

An interesting recent advancement in medical computer vision is the vision transformer. Transformers are perhaps better originally known for their language processing capabilities, which have now been popularised by ChatGPT and others, but they have also been shown to process medical images effectively (Shamshad et al. 2023). By processing images as sequences of patches, they can model longer-range context, which may be useful in anatomically complex head and neck PET/MRI or PET/CT. However, for volumetric (3D) imaging their training is typically computationally demanding and often requires substantially larger, well-curated datasets.

Transformers also support multimodal inputs, so combining imaging with structured clinical variables (e.g., smoking history, HPV status) is a potential direction, but adequate cohort size and rigorous external validation is required. More ambitious text–image models for report drafting remain a longer-term prospect because they require large, paired image–text datasets and reliable localisation of clinically relevant findings within large image volumes.

7 Conclusions

This thesis investigated the feasibility of using DL as a diagnostic aid in PET/MRI imaging of HNC. DL models were evaluated in classification and segmentation of PET/MRI images. Grad-CAM was used to increase classification transparency and its usability as an interpretability measure in this context was assessed.

The key conclusions are as follows:

1. Study I indicated that 2D segmentation of HNC from PET/MRI is feasible; however, more data is required to improve detection of segmentable lesions.
2. Study II showed that 2D binary classification of HNC with only PET images is accurate and models trained exclusively with squamous cell carcinoma can also accurately classify other HNC subtypes. In addition, the study confirmed that along with metabolic activity and location, the prevalence of certain HNC subgroups in the training data is associated with the same group being classified correctly in the test set.
3. Study III demonstrated that 3D classification of HNC PET/MRI provides accurate results but is not on par with radiologist performance. Moreover, Grad-CAM gave valuable insight into model decisions and could help identify false positive findings. However, Grad-CAM merely identifies regions that are relevant for the model – not why they are relevant.
4. All three studies indicate that using DL as a diagnostic aid in HNC PET/MRI analysis is feasible, yet an exceedingly challenging problem. To achieve clinician level performance, high emphasis should be set on data quantity and quality.

Acknowledgements

The work for this thesis was conducted at the Turku PET Centre, University of Turku and Turku University Hospital during the years 2021–2025. The work was financially supported by the University of Turku, Turku University Hospital, Finnish Cancer Foundation, and the Finnish Medical Foundation. The language revision of this thesis was done with the assistance of artificial intelligence (ChatGPT).

I would like to begin by thanking my supervisors, Professor Jukka Kemppainen and Associate Professor Riku Klén. Jukka, your consistent guidance and long experience were of great help throughout this work. Riku, your support and insight, especially regarding some of the mathematical and technical parts of the project, were greatly appreciated. I am grateful for the support both of you have provided. I would also like to thank my follow-up committee member Tommi Noponen for helping me stay on track during the process.

My thanks go to Associate Professor Antti Loimaala and Professor Lalith Kumar Shiyam Sundar for reviewing the dissertation manuscript. Your comments and suggestions improved the clarity and quality of this work tremendously. I am also grateful to Professor Elin Trägårdh for agreeing to act as my opponent in my thesis defence and I look forward to our discussion.

I wish to thank all co-authors and collaborators involved in the studies included in this thesis, particularly Henri Hellström, Oona Rainio and Jussi Hirvonen. Working with you has been both valuable and enjoyable. I also wish to acknowledge Simona Malaspina and Sarita Murtojärvi for their expertise and assistance.

Balancing research, professional commitments, software engineering studies and personal life has been challenging at times, and I would not have managed it without the people closest to me. Above all, I want to thank Tiina**.** Your loving, steady, and wonderfully fun presence has carried me through every stage of this journey. In the final months of this project, as we welcomed our baby boy Väinö into our lives, your warmth, patience, and unfailing support meant more to me than I can express. I am deeply grateful for everything you have made possible and for the joy you bring into our lives.

I am also fortunate to have a supportive family. I want to thank my sister Anna for your help and encouragement over the years. To my parents, Leif and Päivi, thank

you for the long-term support that has shaped both my work and who I am. I would also like to thank my friends, whose presence, encouragement, and conversations have brought much-needed balance throughout this period. I am grateful to all of you for the support you have offered in different ways along the journey.

December 2025
Joonas Liedes

References

Amin, Mahul B., Frederick L. Greene, Stephen B. Edge, et al. 2017. 'The Eighth Edition AJCC Cancer Staging Manual: Continuing to Build a Bridge from a Population-Based to a More "Personalized" Approach to Cancer Staging.' *CA: A Cancer Journal for Clinicians* (United States) 67 (2): 93–99. https://doi.org/10.3322/caac.21388.

Andrearczyk, Vincent, Valentin Oreiller, Moamen Abobakr, et al. 2023. 'Overview of the HECKTOR Challenge at MICCAI 2022: Automatic Head and Neck Tumor Segmentation and Outcome Prediction in PET/CT.' *Head and Neck Tumor Segmentation and Outcome Prediction : Third Challenge, HECKTOR 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings. Head and Neck Tumor Segmentation Challenge (3rd : 2022 : Singapor...* (Switzerland) 13626: 1–30. https://doi.org/10.1007/978-3-031-27420-6_1.

Andrearczyk, Vincent, Valentin Oreiller, Martin Vallières, et al. 2020. 'Automatic Segmentation of Head and Neck Tumors and Nodal Metastases in PET-CT Scans'. In *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, edited by Tal Arbel, Ismail Ben Ayed, Marleen de Bruijne, Maxime Descoteaux, Herve Lombaert, and Christopher Pal, vol. 121. Proceedings of Machine Learning Research. PMLR. https://proceedings.mlr.press/v121/andrearczyk20a.html.

Anis, Mursalin M., Mir-Muhammad Razavi, Xiao Xiao, and Ahmed M. S. Soliman. 2018. 'Association of Gastroesophageal Reflux Disease and Laryngeal Cancer.' *World Journal of Otorhinolaryngology - Head and Neck Surgery* (United States) 4 (4): 278–81. https://doi.org/10.1016/j.wjorl.2017.12.011.

Bammer, Roland. 2003. 'Basic Principles of Diffusion-Weighted Imaging'. *European Journal of Radiology* 45 (3): 169–84. https://doi.org/10.1016/S0720-048X(02)00303-0.

Becker, Minerva, Arthur D. Varoquaux, Christophe Combescure, et al. 2018. 'Local Recurrence of Squamous Cell Carcinoma of the Head and Neck after Radio(Chemo)Therapy: Diagnostic Performance of FDG-PET/MRI with Diffusion-Weighted Sequences.' *European Radiology* (Germany) 28 (2): 651–63. https://doi.org/10.1007/s00330-017-4999-1.

Bi, Qifang, Katherine E Goodman, Joshua Kaminsky, and Justin Lessler. 2019. 'What Is Machine Learning? A Primer for the Epidemiologist'. *American Journal of Epidemiology* 188 (12): 2222–39. https://doi.org/10.1093/aje/kwz189.

Cancer (IARC), The International Agency for Research on. n.d. 'Global Cancer Observatory'. Accessed 26 September 2022. https://gco.iarc.fr/.

Chaturvedi, Anil K., Eric A. Engels, Ruth M. Pfeiffer, et al. 2011. 'Human Papillomavirus and Rising Oropharyngeal Cancer Incidence in the United States'. *Journal of Clinical Oncology* 29 (32): 4294–301. https://doi.org/10.1200/JCO.2011.36.4596.

Chen, Liyuan, Michael Dohopolski, Zhiguo Zhou, et al. 2021. 'Attention Guided Lymph Node Malignancy Prediction in Head and Neck Cancer.' *International Journal of Radiation Oncology, Biology, Physics* (United States) 110 (4): 1171–79. https://doi.org/10.1016/j.ijrobp.2021.02.004.

Chen, Liyuan, Zhiguo Zhou, David Sher, et al. 2019. 'Combining Many-Objective Radiomics and 3D Convolutional Neural Network through Evidential Reasoning to Predict Lymph Node Metastasis

in Head and Neck Cancer.' *Physics in Medicine and Biology* (England) 64 (7): 075011. https://doi.org/10.1088/1361-6560/ab083a.

Cho, Junghwan, Kyewook Lee, Ellie Shin, Garry Choy, and Synho Do. 2015. 'How Much Data Is Needed to Train a Medical Image Deep Learning System to Achieve Necessary High Accuracy?' *arXiv Preprint arXiv:1511.06348*.

Chow, Laura Q. M. 2020. 'Head and Neck Cancer'. *New England Journal of Medicine* 382 (1): 60–72. https://doi.org/10.1056/NEJMra1715715.

Çiçek, Özgün, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 2016. '3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation'. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, edited by Sebastien Ourselin, Leo Joskowicz, Mert R. Sabuncu, Gozde Unal, and William Wells. Springer International Publishing.

Connolly, Michael, and Ashok Srinivasan. 2018. 'Diffusion-Weighted Imaging in Head and Neck Cancer: Technique, Limitations, and Applications.' *Magnetic Resonance Imaging Clinics of North America* (United States) 26 (1): 121–33. https://doi.org/10.1016/j.mric.2017.08.011.

Cooper, Jay S., Thomas F. Pajak, Arlene A. Forastiere, et al. 2004. 'Postoperative Concurrent Radiotherapy and Chemotherapy for High-Risk Squamous-Cell Carcinoma of the Head and Neck'. *New England Journal of Medicine* 350 (19): 1937–44. https://doi.org/10.1056/NEJMoa032646.

De Biase, Alessia, Nanna M. Sijtsema, Lisanne V. van Dijk, Johannes A. Langendijk, and Peter M. A. van Ooijen. 2023. 'Deep Learning Aided Oropharyngeal Cancer Segmentation with Adaptive Thresholding for Predicted Tumor Probability in FDG PET and CT Images.' *Physics in Medicine and Biology* (England) 68 (5). https://doi.org/10.1088/1361-6560/acb9cf.

Dohopolski, Michael, Liyuan Chen, David Sher, and Jing Wang. 2020. 'Predicting Lymph Node Metastasis in Patients with Oropharyngeal Cancer by Using a Convolutional Neural Network with Associated Epistemic and Aleatoric Uncertainty.' *Physics in Medicine and Biology* (England) 65 (22): 225002. https://doi.org/10.1088/1361-6560/abb71c.

Elgendi, Mohamed, Muhammad Umer Nasir, Qunfeng Tang, et al. 2021. 'The Effectiveness of Image Augmentation in Deep Learning Networks for Detecting COVID-19: A Geometric Transformation Perspective'. *Frontiers in Medicine* Volume 8-2021. https://doi.org/10.3389/fmed.2021.629134.

Esteva, Andre, Brett Kuprel, Roberto A. Novoa, et al. 2017. 'Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks'. *Nature* 542 (7639): 115–18. https://doi.org/10.1038/nature21056.

Felix P. Kuhn, Martin Hüllner, Caecilia E. Mader, et al. 2014. 'Contrast-Enhanced PET/MR Imaging Versus Contrast-Enhanced PET/CT in Head and Neck Cancer: How Much MR Information Is Needed?' *Journal of Nuclear Medicine* 55 (4): 551. https://doi.org/10.2967/jnumed.113.125443.

Gallagher, B. M., J. S. Fowler, N. I. Gutterson, R. R. MacGregor, C. N. Wan, and A. P. Wolf. 1978. 'Metabolic Trapping as a Principle of Oradiopharmaceutical Design: Some Factors Resposible for the Biodistribution of [18F] 2-Deoxy-2-Fluoro-D-Glucose.' *Journal of Nuclear Medicine : Official Publication, Society of Nuclear Medicine* (United States) 19 (10): 1154–61.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.

Goodfellow, Ian J, Jean Pouget-Abadie, Mehdi Mirza, et al. 2014. 'Generative Adversarial Nets'. *Advances in Neural Information Processing Systems* 27.

Gormley, Mark, Grant Creaney, Andrew Schache, Kate Ingarfield, and David I. Conway. 2022. 'Reviewing the Epidemiology of Head and Neck Cancer: Definitions, Trends and Risk Factors'. *British Dental Journal* 233 (9): 780–86. https://doi.org/10.1038/s41415-022-5166-x.

Groendahl, Aurora Rosvoll, Ingerid Skjei Knudtsen, Bao Ngoc Huynh, et al. 2021. 'A Comparison of Fully Automatic Segmentation of Tumors and Involved Nodes in PET/CT of Head and Neck Cancers.' *Physics in Medicine and Biology* (England), ahead of print, February 11. https://doi.org/10.1088/1361-6560/abe553.

Guo, Zhe, Ning Guo, Kuang Gong, Shun'an Zhong, and Quanzheng Li. 2019. 'Gross Tumor Volume Segmentation for Head and Neck Cancer Radiotherapy Using Deep Dense Multi-Modality

Network'. *Physics in Medicine and Biology* 64 (20): 205015–205015. PubMed (31514173). https://doi.org/10.1088/1361-6560/ab440d.

Han, Kyumin, Joonyoung Francis Joung, Minhi Han, Wonmo Sung, and Young-Nam Kang. 2022. 'Locoregional Recurrence Prediction Using a Deep Neural Network of Radiological and Radiotherapy Images.' *Journal of Personalized Medicine* (Switzerland) 12 (2). https://doi.org/10.3390/jpm12020143.

Hicks, Steven A., Inga Strümke, Vajira Thambawita, et al. 2022. 'On Evaluation Metrics for Medical Applications of Artificial Intelligence'. *Scientific Reports* 12 (1): 5979. https://doi.org/10.1038/s41598-022-09954-8.

Horn-Ross, P. L., B. M. Ljung, and M. Morrow. 1997. 'Environmental Factors and the Risk of Salivary Gland Cancer.' *Epidemiology (Cambridge, Mass.)* (United States) 8 (4): 414–19. https://doi.org/10.1097/00001648-199707000-00011.

Huang, Bin, Zhewei Chen, Po-Man Wu, et al. 2018. 'Fully Automated Delineation of Gross Tumor Volume for Head and Neck Cancer on PET-CT Using Deep Learning: A Dual-Center Study'. *Contrast Media & Molecular Imaging* 2018 (October): 8923028. https://doi.org/10.1155/2018/8923028.

Hwang, Eui Jin, Sunggyun Park, Kwang-Nam Jin, et al. 2019. 'Development and Validation of a Deep Learning–Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs'. *JAMA Network Open* 2 (3): e191095–e191095. https://doi.org/10.1001/jamanetworkopen.2019.1095.

Isensee, Fabian, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. 2021. 'nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation'. *Nature Methods* 18 (2): 203–11. https://doi.org/10.1038/s41592-020-01008-z.

Johnson, Daniel E., Barbara Burtness, C. René Leemans, Vivian Wai Yan Lui, Julie E. Bauman, and Jennifer R. Grandis. 2020. 'Head and Neck Squamous Cell Carcinoma'. *Nature Reviews Disease Primers* 6 (1): 92. https://doi.org/10.1038/s41572-020-00224-3.

Judenhofer, Martin S, Hans F Wehrl, Danny F Newport, et al. 2008. 'Simultaneous PET-MRI: A New Approach for Functional and Morphological Imaging'. *Nature Medicine* 14 (4): 459–65. https://doi.org/10.1038/nm1700.

Kao, Johnny, Ha Linh Vu, Eric M. Genden, et al. 2009. 'The Diagnostic and Prognostic Utility of Positron Emission Tomography/Computed Tomography-Based Follow-up after Radiotherapy for Head and Neck Cancer'. *Cancer* 115 (19): 4586–94. https://doi.org/10.1002/cncr.24493.

Kermany, Daniel S., Michael Goldbaum, Wenjia Cai, et al. 2018. 'Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning'. *Cell* 172 (5): 1122-1131.e9. https://doi.org/10.1016/j.cell.2018.02.010.

Koçak, Burak, Fadime Köse, Ali Keleş, Abdurrezzak Şendur, İsmail Meşe, and Mehmet Karagülle. 2025. 'Adherence to the Checklist for Artificial Intelligence in Medical Imaging (CLAIM): An Umbrella Review with a Comprehensive Two-Level Analysis'. *Diagn Interv Radiol*.

Koea, Jonathan B., and James H. F. Shaw. 1999. 'Parathyroid Cancer: Biology and Management'. *Surgical Oncology* 8 (3): 155–65. https://doi.org/10.1016/S0960-7404(99)00037-7.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. 2012. 'ImageNet Classification with Deep Convolutional Neural Networks'. In *Advances in Neural Information Processing Systems*, edited by F. Pereira, C. J. Burges, L. Bottou, and K. Q. Weinberger, vol. 25. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Krogh, Anders. 2008. 'What Are Artificial Neural Networks?' *Nature Biotechnology* 26 (2): 195–97. https://doi.org/10.1038/nbt1386.

Kubiessa, K., S. Purz, M. Gawlitza, et al. 2014. 'Initial Clinical Results of Simultaneous 18F-FDG PET/MRI in Comparison to 18F-FDG PET/CT in Patients with Head and Neck Cancer.' *European Journal of Nuclear Medicine and Molecular Imaging* (Germany) 41 (4): 639–48. https://doi.org/10.1007/s00259-013-2633-2.

Lapuschkin, Sebastian, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2019. 'Unmasking Clever Hans Predictors and Assessing What Machines Really Learn'. *Nature Communications* 10 (1): 1096. https://doi.org/10.1038/s41467-019-08987-4.

Laubenbacher, C., D. Saumweber, C. Wagner-Manslau, et al. 1995. 'Comparison of Fluorine-18-Fluorodeoxyglucose PET, MRI and Endoscopy for Staging Head and Neck Squamous-Cell Carcinomas.' *Journal of Nuclear Medicine : Official Publication, Society of Nuclear Medicine* (United States) 36 (10): 1747–57.

Le, William Trung, Eugene Vorontsov, Francisco Perdigón Romero, et al. 2022. 'Cross-Institutional Outcome Prediction for Head and Neck Cancer Patients Using Self-Attention Neural Networks.' *Scientific Reports* (England) 12 (1): 3183. https://doi.org/10.1038/s41598-022-07034-5.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. 'Deep Learning'. *Nature* 521 (7553): 436–44. https://doi.org/10.1038/nature14539.

Lee, Soo Jin, Hyo Jung Seo, Gi Jeong Cheon, et al. 2014. 'Usefulness of Integrated PET/MRI in Head and Neck Cancer: A Preliminary Study.' *Nuclear Medicine and Molecular Imaging* (Germany) 48 (2): 98–105. https://doi.org/10.1007/s13139-013-0252-2.

Li, Xiaomeng, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. 2018. 'H-DenseUNet: Hybrid Densely Connected UNet for Liver and Tumor Segmentation From CT Volumes'. *IEEE Transactions on Medical Imaging* 37 (12): 2663–74. https://doi.org/10.1109/TMI.2018.2845918.

Loeffelbein, Denys J., Michael Souvatzoglou, Veronika Wankerl, et al. 2012. 'PET-MRI Fusion in Head-and-Neck Oncology: Current Status and Implications for Hybrid PET/MRI'. *Journal of Oral and Maxillofacial Surgery* 70 (2): 473–83. https://doi.org/10.1016/j.joms.2011.02.120.

Loeffelbein, Denys J., Michael Souvatzoglou, Veronika Wankerl, et al. 2014. 'Diagnostic Value of Retrospective PET-MRI Fusion in Head-and-Neck Cancer.' *BMC Cancer* (England) 14 (November): 846. https://doi.org/10.1186/1471-2407-14-846.

Long, Jonathan, Evan Shelhamer, and Trevor Darrell. 2015. 'Fully Convolutional Networks for Semantic Segmentation'. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431–40.

Marur, Shanthi, Gypsyamber D'Souza, William H. Westra, and Arlene A. Forastiere. 2010. 'HPV-Associated Head and Neck Cancer: A Virus-Related Cancer Epidemic.' *The Lancet. Oncology* (England) 11 (8): 781–89. https://doi.org/10.1016/S1470-2045(10)70017-6.

Miller, Frank R., David Hussey, Mural Beeram, Tony Eng, H. Stan McGuff, and Randal A. Otto. 2005. 'Positron Emission Tomography in the Management of Unknown Primary Head and Neck Carcinoma'. *Archives of Otolaryngology–Head & Neck Surgery* 131 (7): 626–29. https://doi.org/10.1001/archotol.131.7.626.

Milletari, Fausto, Nassir Navab, and Seyed-Ahmad Ahmadi. 2016. 'V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation'. *2016 Fourth International Conference on 3D Vision (3DV)*, 565–71.

Moe, Yngve Mardal, Aurora Rosvoll Groendahl, Oliver Tomic, Einar Dale, Eirik Malinen, and Cecilia Marie Futsaether. 2021. 'Deep Learning-Based Auto-Delineation of Gross Tumour Volumes and Involved Nodes in PET/CT Images of Head and Neck Cancer Patients'. *European Journal of Nuclear Medicine and Molecular Imaging* 48 (9): 2782–92. PubMed (33559711). https://doi.org/10.1007/s00259-020-05125-x.

Mongan, John, Linda Moy, and Charles E. Kahn. 2020. 'Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers'. *Radiology: Artificial Intelligence* 2 (2): e200029. https://doi.org/10.1148/ryai.2020200029.

Müller, Dominik, Iñaki Soto-Rey, and Frank Kramer. 2022. 'Towards a Guideline for Evaluation Metrics in Medical Image Segmentation'. *BMC Research Notes* 15 (1): 210. https://doi.org/10.1186/s13104-022-06096-y.

Murtojärvi, Sarita, Simona Malaspina, Ilpo Kinnunen, et al. 2022. 'Diagnostic Accuracy of 18F-FDG-PET/CT and 18F-FDG-PET/MRI in Detecting Locoregional Recurrence of HNSCC 12 Weeks after the End of Chemoradiotherapy: Single-Center Experience with PET/MRI'. *Contrast Media & Molecular Imaging* 2022 (August): 8676787. https://doi.org/10.1155/2022/8676787.

Musafargani, Sikkandhar, Krishna Kanta Ghosh, Sachin Mishra, Pachaiyappan Mahalakshmi, Parasuraman Padmanabhan, and Balázs Gulyás. 2018. 'PET/MRI: A Frontier in Era of Complementary Hybrid Imaging'. *European Journal of Hybrid Imaging* 2 (1): 12. https://doi.org/10.1186/s41824-018-0030-6.

Nikulin, Pavel, Sebastian Zschaeck, Jens Maus, et al. 2023. 'A Convolutional Neural Network with Self-Attention for Fully Automated Metabolic Tumor Volume Delineation of Head and Neck Cancer in [Formula: See Text]F]FDG PET/CT.' *European Journal of Nuclear Medicine and Molecular Imaging* (Germany), ahead of print, April 20. https://doi.org/10.1007/s00259-023-06197-1.

Oktay, Ozan, Jo Schlemper, Loic Le Folgoc, et al. 2018. 'Attention U-Net: Learning Where to Look for the Pancreas'.

Olin, Anders B., Adam E. Hansen, Jacob H. Rasmussen, et al. 2020. 'Feasibility of Multiparametric Positron Emission Tomography/Magnetic Resonance Imaging as a One-Stop Shop for Radiation Therapy Planning for Patients with Head and Neck Cancer'. *International Journal of Radiation Oncology, Biology, Physics* 108 (5): 1329–38. https://doi.org/10.1016/j.ijrobp.2020.07.024.

Olin, Anders B., Adam E. Hansen, Jacob H. Rasmussen, et al. 2022. 'Deep Learning for Dixon MRI-Based Attenuation Correction in PET/MRI of Head and Neck Cancer Patients.' *EJNMMI Physics* (Germany) 9 (1): 20. https://doi.org/10.1186/s40658-022-00449-z.

Olin, Anders B., Christopher Thomas, Adam E. Hansen, et al. 2021. 'Robustness and Generalizability of Deep Learning Synthetic Computed Tomography for Positron Emission Tomography/Magnetic Resonance Imaging-Based Radiation Therapy Planning of Patients With Head and Neck Cancer.' *Advances in Radiation Oncology* (United States) 6 (6): 100762. https://doi.org/10.1016/j.adro.2021.100762.

Oreiller, Valentin, Vincent Andrearczyk, Mario Jreige, et al. 2022. 'Head and Neck Tumor Segmentation in PET/CT: The HECKTOR Challenge'. *Medical Image Analysis* 77: 102336. https://doi.org/10.1016/j.media.2021.102336.

Partovi, S., A. Kohan, J. L. Vercher-Conejero, et al. 2014. 'Qualitative and Quantitative Performance of ^{18}F-FDG-PET/MRI versus ^{18}F-FDG-PET/CT in Patients with Head and Neck Cancer.' *AJNR. American Journal of Neuroradiology* (United States) 35 (10): 1970–75. https://doi.org/10.3174/ajnr.A3993.

Pellegriti, Gabriella, Francesco Frasca, Concetto Regalbuto, Sebastiano Squatrito, and Riccardo Vigneri. 2013. 'Worldwide Increasing Incidence of Thyroid Cancer: Update on Epidemiology and Risk Factors'. *Journal of Cancer Epidemiology* 2013 (May): 965212. https://doi.org/10.1155/2013/965212.

Pynnonen, Melissa A., M. Boyd Gillespie, Benjamin Roman, et al. 2017. 'Clinical Practice Guideline: Evaluation of the Neck Mass in Adults'. *Otolaryngology–Head and Neck Surgery* 157 (2_suppl): S1–30. https://doi.org/10.1177/0194599817722550.

Queiroz, Marcelo A., Martin Hüllner, Felix Kuhn, et al. 2014. 'PET/MRI and PET/CT in Follow-up of Head and Neck Cancer Patients.' *European Journal of Nuclear Medicine and Molecular Imaging* (Germany) 41 (6): 1066–75. https://doi.org/10.1007/s00259-014-2707-9.

Rainio, Oona, Jonne Tamminen, Mikko S. Venäläinen, et al. 2024. 'Comparison of Thresholds for a Convolutional Neural Network Classifying Medical Images'. *International Journal of Data Science and Analytics*, ahead of print, June 18. https://doi.org/10.1007/s41060-024-00584-z.

Ren, Jintao, Jesper Grau Eriksen, Jasper Nijkamp, and Stine Sofia Korreman. 2021. 'Comparing Different CT, PET and MRI Multi-Modality Image Combinations for Deep Learning-Based Head and Neck Tumor Segmentation'. *Acta Oncologica* 60 (11): 1399–406. https://doi.org/10.1080/0284186X.2021.1949034.

Riegel, Adam C., Anthony M. Berson, Sylvie Destian, et al. 2006. 'Variability of Gross Tumor Volume Delineation in Head-and-Neck Cancer Using CT and PET/CT Fusion'. *International Journal of Radiation Oncology, Biology, Physics* 65 (3): 726–32. https://doi.org/10.1016/j.ijrobp.2006.01.014.

Rong, Jian, Ahmed Haider, Troels E. Jeppesen, Lee Josephson, and Steven H. Liang. 2023. 'Radiochemistry for Positron Emission Tomography'. *Nature Communications* 14 (1): 3257. https://doi.org/10.1038/s41467-023-36377-4.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. 2015. 'U-Net: Convolutional Networks for Biomedical Image Segmentation'. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, edited by Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi. Springer International Publishing.

Salahuddin, Zohaib, Yi Chen, Xian Zhong, et al. 2023. 'From Head and Neck Tumour and Lymph Node Segmentation to Survival Prediction on PET/CT: An End-to-End Framework Featuring Uncertainty, Fairness, and Multi-Region Multi-Modal Radiomics.' *Cancers* (Switzerland) 15 (7). https://doi.org/10.3390/cancers15071932.

Samolyk-Kogaczewska, Natalia, Ewa Sierko, Dorota Dziemianczyk-Pakiela, Klaudia Beata Nowaszewska, Malgorzata Lukasik, and Joanna Reszec. 2020. 'Usefulness of Hybrid PET/MRI in Clinical Evaluation of Head and Neck Cancer Patients.' *Cancers* (Switzerland) 12 (2). https://doi.org/10.3390/cancers12020511.

Schaarschmidt, Benedikt Michael, Philipp Heusch, Christian Buchbender, et al. 2016. 'Locoregional Tumour Evaluation of Squamous Cell Carcinoma in the Head and Neck Area: A Comparison between MRI, PET/CT and Integrated PET/MRI.' *European Journal of Nuclear Medicine and Molecular Imaging* (Germany) 43 (1): 92–102. https://doi.org/10.1007/s00259-015-3145-z.

Scholzen, T., and J. Gerdes. 2000. 'The Ki-67 Protein: From the Known and the Unknown.' *Journal of Cellular Physiology* (United States) 182 (3): 311–22. https://doi.org/10.1002/(SICI)1097-4652(200003)182:3<311::AID-JCP1>3.0.CO;2-9.

Shamshad, Fahad, Salman Khan, Syed Waqas Zamir, et al. 2023. 'Transformers in Medical Imaging: A Survey'. *Medical Image Analysis* 88: 102802. https://doi.org/10.1016/j.media.2023.102802.

Simonyan, Karen, and Andrew Zisserman. 2014. 'Very Deep Convolutional Networks for Large-Scale Image Recognition'. *arXiv Preprint arXiv:1409.1556*.

Singh, Nripendra Kumar, and Khalid Raza. 2021. 'Medical Image Generation Using Generative Adversarial Networks: A Review'. In *Health Informatics: A Computational Perspective in Healthcare*, edited by Ripon Patgiri, Anupam Biswas, and Pinki Roy. Springer Singapore. https://doi.org/10.1007/978-981-15-9735-0_5.

Specenier, Pol M, and Jan B Vermorken. 2008. 'Recurrent Head and Neck Cancer: Current Treatment and Future Prospects'. *Expert Review of Anticancer Therapy* 8 (3): 375–91. https://doi.org/10.1586/14737140.8.3.375.

Sung, Hyuna, Jacques Ferlay, Rebecca L. Siegel, et al. 2021. '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): 209–49. https://doi.org/10.3322/caac.21660.

Szyszko, T. A., and G. J. R. Cook. 2018. 'PET/CT and PET/MRI in Head and Neck Malignancy.' *Clinical Radiology* (England) 73 (1): 60–69. https://doi.org/10.1016/j.crad.2017.09.001.

Thie, Joseph A. 2004. 'Understanding the Standardized Uptake Value, Its Methods, and Implications for Usage.' *Journal of Nuclear Medicine : Official Publication, Society of Nuclear Medicine* (United States) 45 (9): 1431–34.

Tsao, Sai Wah, Chi Man Tsang, and Kwok Wai Lo. 2017. 'Epstein–Barr Virus Infection and Nasopharyngeal Carcinoma'. *Philosophical Transactions of the Royal Society B: Biological Sciences* 372 (1732): 20160270. https://doi.org/10.1098/rstb.2016.0270.

Vallières, Martin, Emily Kay-Rivest, Léo Jean Perrin, et al. 2017. 'Radiomics Strategies for Risk Assessment of Tumour Failure in Head-and-Neck Cancer'. *Scientific Reports* 7 (1): 10117. https://doi.org/10.1038/s41598-017-10371-5.

Vander Heiden, Matthew G., Lewis C. Cantley, and Craig B. Thompson. 2009. 'Understanding the Warburg Effect: The Metabolic Requirements of Cell Proliferation.' *Science (New York, N.Y.)* (United States) 324 (5930): 1029–33. https://doi.org/10.1126/science.1160809.

Varoquaux, Arthur, Olivier Rager, Antoine Poncet, et al. 2014. 'Detection and Quantification of Focal Uptake in Head and Neck Tumours: (18)F-FDG PET/MR versus PET/CT.' *European Journal of Nuclear Medicine and Molecular Imaging* (Germany) 41 (3): 462–75. https://doi.org/10.1007/s00259-013-2580-y.

Vlaardingerbroek, Marinus T, and Jacques A Boer. 2013. *Magnetic Resonance Imaging: Theory and Practice*. Springer Science & Business Media.

Wang, Rongfang, Jinkun Guo, Zhiguo Zhou, et al. 2022. 'Locoregional Recurrence Prediction in Head and Neck Cancer Based on Multi-Modality and Multi-View Feature Expansion.' *Physics in Medicine and Biology* (England) 67 (12). https://doi.org/10.1088/1361-6560/ac72f0.

Wang, Zhengyang, Na Zou, Dinggang Shen, and Shuiwang Ji. 2020. 'Non-Local u-Nets for Biomedical Image Segmentation'. *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (04): 6315–22.

Zhao, Lijun, Zixiao Lu, Jun Jiang, Yujia Zhou, Yi Wu, and Qianjin Feng. 2019. 'Automatic Nasopharyngeal Carcinoma Segmentation Using Fully Convolutional Networks with Auxiliary Paths on Dual-Modality PET-CT Images.' *Journal of Digital Imaging* (United States) 32 (3): 462–70. https://doi.org/10.1007/s10278-018-00173-0.

Zhong, Jingyu, Yue Xing, Junjie Lu, et al. 2023. 'The Endorsement of General and Artificial Intelligence Reporting Guidelines in Radiological Journals: A Meta-Research Study'. *BMC Medical Research Methodology* 23 (1): 292. https://doi.org/10.1186/s12874-023-02117-x.

List of Figures and Tables

Figures

Tables

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

Painosalama, Turku, Finland 2026

ISBN 978-952-02-0534-8 (PRINT)
ISBN 978-952-02-0535-5 (PDF)
ISSN 0355-9483 (Print)
ISSN 2343-3213 (Online)