SCIENTIFIC Publishing Limited

ENT Updates

https://ojs.ukscip.com/index.php/ENTU

RESEARCH ARTICLE

Advances in Deep Learning for Head and Neck Cancer: Datasets and Applied Methods

Tabasum Majeed and **Assif Assad** *

Department of Computer Science and Engineering, Islamic University of Science and Technology, Awantipora 192122, Jammu & Kashmir, India

* Correspondence: assif.assad@islamicuniversity.edu.in

Received: 19 November 2024; **Revised:** 24 December 2024; **Accepted:** 27 December 2024; **Published:** 10 January 2025

Abstract: Head and neck cancers (HNCs) include malignancies of the oral cavity, salivary glands, thyroid, oropharynx, and nasopharynx, with risk factors such as tobacco use, alcohol consumption, viral infections, and environmental exposures contributing to over half a million global cases annually. Despite treatment advances, poor prognosis underscores the need for accurate diagnosis and continuous monitoring. Medical imaging plays a critical role in HNC evaluation but is often limited by the complexity of anatomy and tumor biology. Recent advances in artificial intelligence (AI), particularly deep learning, offer opportunities to enhance diagnostic accuracy and optimize treatment strategies. This study reviews the application of deep learning in HNC imaging, evaluating different architectures and addressing challenges like limited annotated datasets, high computational demands, and ethical concerns. Overcoming these challenges will revolutionize HNC diagnostics, redefine precision oncology, and improve patient care. The future integration of explainable AI models and multimodal data will be crucial in advancing diagnostic precision, ensuring clinical applicability, and addressing ethical and resource challenges. As AI progresses, its effective integration into clinical workflows will not only enhance healthcare delivery but also reduce inequalities, accelerating significant advancements in HNC management and transforming patient outcomes.

Keywords: Deep Learning; Head and Neck Cancer; Histopathology Images; Attention Mechanisms; Imaging Modalities; Survival Prediction; Healthcare Decision-Making

1. Introduction

Head and neck cancers (HNCs) originate from a variety of anatomical sites, including the craniofacial bones, soft tissues, and mucosal linings of the oral cavity, salivary glands, thyroid, oropharynx, and nasopharynx. Key risk factors include tobacco use, heavy alcohol consumption, areca (betel) nut, paan masala (Gutkha), exposure to gamma and ultraviolet radiation, prolonged sunlight exposure, a family history of cancer, and increasing age. Additionally, human papillomavirus (HPV) and Epstein–Barr virus (EBV) are strongly associated with the development of squamous cell carcinoma (SCC) in the oropharynx and nasopharynx. Globally, the incidence of HNCs is rising, with more than 500,000 cases reported annually and around 12,000 new cases each year in the UK alone, reflecting a 20% increase over the past decade. The prognosis remains poor, with 5-year survival rates ranging from 28% to 67%, depending on the stage of diagnosis [1]. In contrast, cancers such as breast and prostate cancer demonstrate significantly better outcomes, with five-year survival rates of approximately 91.2% [2] and 98% [3], respectively. Similarly, thyroid cancer has a five-year survival rate of around 98.4% [4], while testicular cancer boasts a rate

of about 95% [5]. Even melanoma, a form of skin cancer, shows an overall survival rate of approximately 92%, particularly high for localized cases [6]. These comparisons underscore the critical need for early detection and effective treatment strategies to improve survival rates for HNC patients. Medical imaging (MI) modalities, including radiological imaging, endoscopic and clinical imaging, hyperspectral imaging, multimodal optical imaging, and histopathology whole-slide imaging, are crucial for evaluating HNCs. However, interpreting these images is challenging due to the intricate anatomy, varied tumor biology, and similarities between malignant and benign lesions. In addition, traditional imaging methods often rely on human expertise, which can be subjective and prone to inconsistency. As a result, there is a growing need for advanced technologies and tools to improve the diagnosis and management of HNCs.

Advancements in artificial intelligence (AI) have revolutionized various fields, including computer vision [7], natural language processing [8], genomics [9] and robotics [10]. AI has shown promising results in medical domain such as lung disease detection [11], breast cancer detection [12], brain stroke prediction [13], brain tumor detection [14] and identifying surgical actions [15] by utilizing various image modalities. This research explores the application of deep learning (DL) techniques for detecting and analyzing head and neck cancer (HNC) using diverse imaging modalities. The study critically examines various deep learning architectures, assessing their effectiveness across different imaging contexts. A comprehensive literature review captures the current state of the field and includes a detailed analysis of the datasets commonly used in this area. The findings notably reveal a significant research gap in fully exploring the potential of histopathology imaging techniques for HNC. Specifically, there is a distinct lack of studies investigating the use of histopathology imaging for survival prediction in Head and Neck Squamous Cell Carcinoma (HNSCC), highlighting an untapped opportunity for further research and advancements in this field. Consequently, this research investigates the use of histopathology images to predict patient survival outcomes, underscoring how these advanced methods can enhance personalized treatment strategies. Predicting survival after hospitalization is crucial for both healthcare providers and patients. For doctors, it helps assess the severity of the condition and plan appropriate treatments, allowing them to prioritize patients based on the seriousness of their condition. For patients and their families, it offers critical time to make necessary arrangements, facilitates timely prevention and treatment, and helps avoid poor decisions, such as overtreatment or delayed supportive care. The structure of this paper is as follows: Section 1 presents the research background, the motivation behind it, and its objectives. Section 2 explores the deep learning architectures employed for diagnosing HNC. Section 3 provides an extensive literature review on the use of various imaging modalities in HNC, along with their corresponding datasets. Section 4 discusses the challenges encountered in implementing deep learning models. In Section 5, we identify critical gaps in the field, ongoing debates, and potential future innovations. Section 6 presents the experimental results related to survival prediction using histopathology images. Finally, Sections 7 and 8 examine and interpret the results, highlighting the key insights and the significance of the research findings.

2. Deep Learning Architectures for Head and Neck Cancer Diagnosis

2.1. Convolutional Autoencoders for Image Enhancement and Feature Learning

Convolutional autoencoders (CAEs) are advanced deep-learning models designed for tasks such as image enhancement, feature extraction, and data compression. By combining an encoder, which compresses high-dimensional input images into compact latent representations, and a decoder, which reconstructs the images, CAEs have become a valuable tool in medical imaging. Their architecture effectively addresses challenges such as noise, low contrast, and artifacts in imaging modalities like computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET). These challenges are particularly pronounced in the diagnosis and staging of Head and Neck Squamous Cell Carcinoma (HNSCC), where precise imaging is critical for detecting tumor margins, assessing lymph node involvement, and planning radiation therapy. CAEs have been successfully used to enhance image quality, reduce noise, and extract features crucial for tasks such as tumor segmentation, low-dose imaging, and artifact removal, thereby contributing to more accurate and personalized treatment planning [16, 17].

2.1.1. Encoder Architecture

The encoder is responsible for transforming input images into a low-dimensional latent representation while preserving critical diagnostic features. Its components and associated applications are as follows:

- Convolutional Layers: These layers detect spatial features such as edges, textures, and anatomical structures. In HNC imaging, they capture tumor boundaries, tissue heterogeneity, and structural abnormalities, supporting tasks like tumor segmentation and classification.
- Pooling Layers: By reducing the spatial resolution of feature maps, pooling layers improve computational efficiency and invariance to small transformations. This ensures robust feature extraction, even from noisy images, aiding in noise reduction and artifact removal.
- Latent Representation: The compressed representation encodes essential features of the input image in a low-dimensional format. This step is crucial for applications such as low-dose imaging, where CAEs reconstruct high-quality images from reduced data, minimizing radiation exposure without sacrificing diagnostic detail.

2.1.2. Decoder Architecture

The decoder complements the encoder by reconstructing the original image from the latent representation, refining its quality and ensuring the preservation of critical features. Its components and their impact include:

- Upsampling Layers: These layers restore the spatial dimensions of the latent representation. Techniques like transpose convolution are particularly effective for learning the reconstruction process, enabling artifact removal and resolution enhancement. This is essential in tasks requiring high-fidelity imaging, such as the visualization of intricate anatomical structures.
- Convolutional Layers: These layers refine the upsampled feature maps, ensuring the reconstructed images retain fine details and closely match the original input. This capability is critical for applications like tumor margin delineation, image deblurring, and enhancing low resolution images, improving the interpretability of medical scans.

CAEs excel at addressing the challenges inherent in HNC imaging by enhancing image clarity, removing artifacts, and optimizing features for diagnostic and therapeutic applications. Their ability to preprocess data, extract critical features, and reconstruct high-quality images has established them as a cornerstone in oncological imaging. As such, CAEs contribute significantly to improving diagnostic accuracy, treatment planning, and patient outcomes.

2.2. Adversarial Networks for High-Resolution Image Generation

Generative Adversarial Networks (GANs) are an advanced deep learning framework composed of two interdependent models: a generator and a discriminator. The generator creates synthetic images from latent vectors, while the discriminator evaluates these images to determine whether they are real or artificially generated. Through this adversarial training process, GANs iteratively refine their outputs, making them highly effective for generating high-resolution medical images and addressing challenges in imaging modalities such as computed tomography (CT), magnetic resonance imaging (MRI), and histopathology slides [18].

In the context of Head and Neck Cancer (HNC), GANs have demonstrated significant potential by enhancing the resolution of CT and MRI scans, enabling more accurate visualization of tumor margins and critical anatomical features. This improved resolution supports better tumor detection and segmentation, leading to enhanced diagnostic accuracy [19]. Furthermore, GANs effectively remove imaging artifacts caused by noise, motion, or hardware limitations, resulting in cleaner and more reliable diagnostic images. These capabilities are particularly beneficial when artifacts obscure subtle pathological features essential for treatment planning [20]. Beyond resolution enhancement, GANs have been employed to generate high-resolution synthetic histopathology slides, aiding in the training of diagnostic models and providing pathologists with detailed visualizations of malignant features [21].

How GANs Work

Generative Adversarial Networks (GANs) are a type of deep learning framework that harnesses the interplay between two neural networks: the generator and the discriminator as shown in **Figure 1**. These networks are trained simultaneously in an adversarial setup, where each tries to outperform the other in their respective tasks.

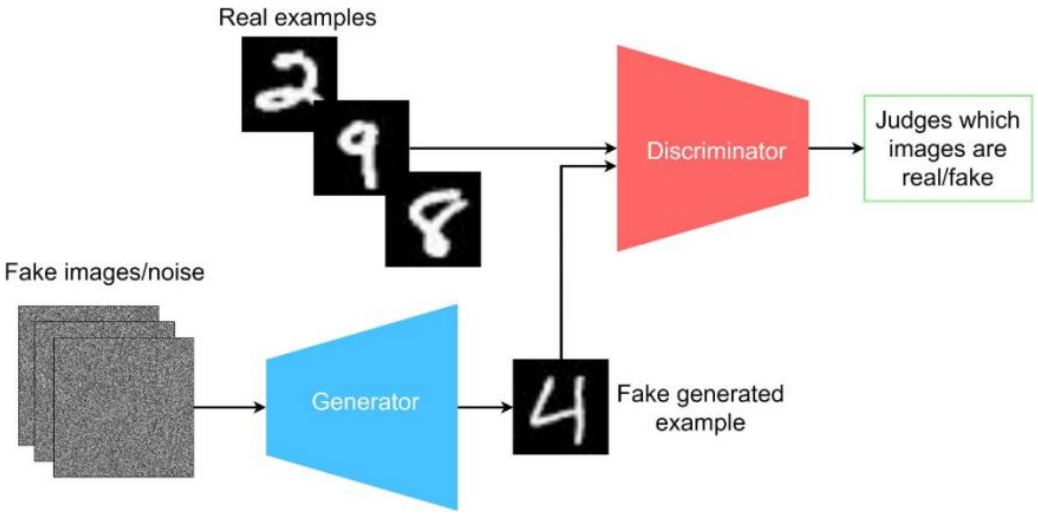

Figure 1. Structure of a Generative Adversarial Network (GAN), illustrating the interaction between the generator and the discriminator (adapted from [22]).

The generator starts by creating synthetic data, such as images, based on a random input vector sampled from a latent space. This vector, often drawn from a Gaussian distribution, serves as the seed for generating data. The generator processes this input through its layers, which typically include dense and convolutional layers, transforming the vector into structured outputs. The generator's goal is to produce synthetic samples that mimic the real data in the original dataset, aiming to "fool" the discriminator into believing that the generated samples are real. On the other side, the discriminator acts as a binary classifier, evaluating whether the input data is real (from the actual dataset) or fake (produced by the generator). It processes both real and synthetic samples through its network layers, extracting features and assigning a probability score indicating the authenticity of each input. The discriminator's objective is to correctly classify real and fake samples, becoming more effective at identifying synthetic data as training progresses.

The adversarial training process involves a continuous feedback loop. The discriminator provides feedback to the generator, highlighting areas where the synthetic data deviates from the real data. In response, the generator adjusts its parameters to create more convincing outputs. Meanwhile, the discriminator improves its ability to distinguish between real and fake data. This competitive interaction continues iteratively, with the generator and discriminator driving each other to perform better. The training reaches equilibrium when the generator produces synthetic data that the discriminator can no longer reliably distinguish from real data. At this point, the GAN achieves its goal of generating highly realistic synthetic data, making it a powerful tool for applications such as image generation, super-resolution, and data augmentation.

2.3. Vision Transformers

Vision Transformers (ViTs) represent a paradigm shift in deep learning models for image analysis, offering an alternative to traditional convolutional neural networks (CNNs). Unlike CNNs, which rely on convolutional layers for localized feature extraction, ViTs treat images as sequences of patches, enabling the model to capture both local and global spatial relationships. By leveraging transformer architectures originally designed for natural language processing, ViTs process visual data with remarkable effectiveness, making them particularly suited for complex medical imaging tasks as shown in **Figure 2**. In the domain of Head and Neck Cancer (HNC), ViTs have shown significant potential by addressing challenges inherent in imaging modalities such as computed tomography (CT), magnetic resonance imaging (MRI), and histopathology slides. These modalities demand precise analysis of intricate textures, high-resolution requirements, and global spatial dependencies, all of which ViTs handle effectively through their global self-attention mechanisms. This capability allows ViTs to excel in tasks like tumor detection, segmentation, classification, and multi-modal image analysis, revolutionizing the approach to medical image processing [23, 24].

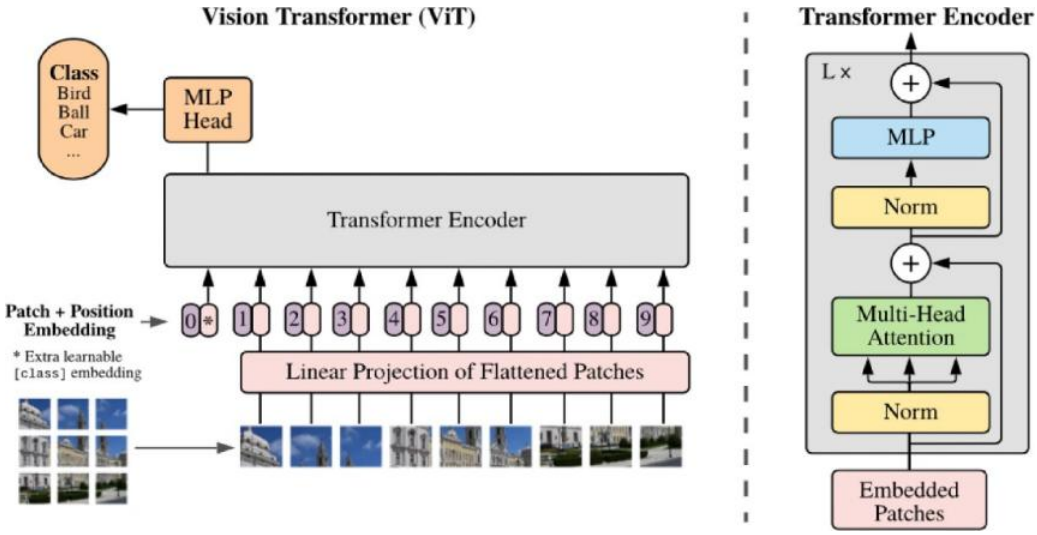

Figure 2. Architecture of the Vision Transformer (ViT), illustrating the patch embedding and transformer encoding process (adopted from [25]).

ViTs enable highly accurate tumor detection and localization by analyzing spatial relationships across patches, identifying subtle abnormalities in CT, MRI, and histopathology slides. Their ability to capture both local and global contextual information makes them exceptional for delineating tumor boundaries, a critical step in treatment planning such as radiation therapy. In histopathology, ViTs process high-resolution slides to detect cancerous regions, classify tissue types, and identify cellular structures indicative of malignancy. Furthermore, ViTs integrate data from multiple imaging modalities, such as CT, MRI, and PET, providing a comprehensive view of tumor characteristics and improving diagnostic accuracy. Another significant advantage of ViTs lies in their ability to perform risk stratification and staging. By analyzing features associated with tumor growth, spread, and lymph node involvement, they support accurate cancer staging and facilitate risk assessment. Additionally, ViTs monitor treatment response by comparing pre- and post-treatment images, tracking changes in tumor size and morphology, and offering insights into treatment efficacy, thereby guiding adaptive therapy. Their global self-attention mechanisms enable awareness of spatial relationships across entire images, which is particularly useful for analyzing large, high-resolution medical datasets. Unlike CNNs, which rely on hierarchical feature extraction, ViTs preserve fine details necessary for diagnostics, processing high-resolution images effectively. This flexibility makes them suitable for multi-modal analysis, enabling seamless adaptation to diverse imaging modalities and cross-domain applications. These unique capabilities, combined with their adaptability and precision, position ViTs as transformative tools in the diagnosis and management of HNC. By improving diagnostic precision, enhancing treatment planning, and facilitating patient outcome monitoring, Vision Transformers are poised to revolutionize medical image analysis.

ViT Architecture

The architecture of Vision Transformers processes images differently from traditional CNNs. The key components include:

- Patch Embedding Layer: Input images are divided into fixed-size, non-overlapping patches, akin to words in a sentence for natural language processing. Each patch is flattened and passed through a linear projection layer to generate a patch embedding, which represents the feature vector of that patch.
- Positional Encodings: To retain spatial information, positional encodings are added to the patch embeddings. These encodings allow the model to understand the relative positions of patches within the image.
- Transformer Encoder: The sequence of patch embeddings, augmented with positional encodings, is fed into a transformer encoder. The encoder uses multi-head self-attention mechanisms to capture relationships between patches and identify patterns across the entire image. By leveraging global self-attention, ViTs excel at capturing long-range dependencies and spatial relationships, making them particularly effective for high-resolution medical imaging tasks.

- Classification or Feature Extraction: The output embeddings are processed by a classification head or used for downstream tasks like segmentation or feature extraction, depending on the application.

2.4. Pre-Trained Models

Pre-trained models are a foundational component of modern deep learning, offering a highly efficient approach to tackling complex tasks in computer vision. These neural networks, pre-trained on large benchmark datasets such as ImageNet, are fine-tuned for specialized applications, significantly reducing the need for extensive data and computational resources. By leveraging the rich feature representations learned during pre-training, researchers can adapt these models to new domains with remarkable efficiency and accuracy. In medical imaging, particularly for Head and Neck Cancer (HNC), pre-trained models have demonstrated exceptional promise. Their ability to generalize from pre-learned features enables them to excel in critical tasks such as tumor detection, segmentation, and classification. Fine-tuning pre-trained models on HNC-specific datasets allows researchers to extract meaningful features from imaging modalities like CT, MRI, and histopathology slides with high precision, even in low-data scenarios [26]. For instance, pre-trained models like VGG19 have achieved notable success in HNC diagnostics, with studies reporting up to 76% accuracy in distinguishing between normal tissue, precancerous lesions, and malignancies [27].

Pre-trained models excel at identifying tumors in medical images by leveraging hierarchical feature extraction to detect patterns in complex textures and structures. This capability supports precise tumor localization and segmentation, which are critical for radiation therapy and treatment planning. Furthermore, these models process high-resolution histopathology slides to identify cancerous regions and classify tissue types, aiding pathologists in diagnosing malignancies. Their versatility extends to multi-modality analysis, where they integrate and analyze data from multiple imaging sources, such as CT and MRI, providing comprehensive assessments that enhance diagnostic accuracy. One of the most significant advantages of pre-trained models lies in their effectiveness in scenarios with limited labeled data. By generalizing from pre- learned features, these models maintain high performance even in low-data conditions, a common challenge in medical imaging. Additionally, pre-trained models accelerate development workflows by reducing the time and computational resources required for training. This efficiency allows researchers to focus on domain-specific fine-tuning, enabling rapid advancements in HNC diagnostics and management. The adaptability and performance of pre-trained models have made them transformative tools in medical imaging. In HNC diagnosis and management, they have proven invaluable for improving diagnostic precision, optimizing treatment planning, and enhancing patient outcomes. By reducing computational costs and addressing the challenges of limited data, pre-trained models are driving innovation in medical image analysis and paving the way for future advancements.

Pre-Trained Model Architecture

Pre-trained models utilize a structured approach to extract and process features from images, enabling them to generalize effectively across domains. Key architectural components include:

- Convolutional Layers: These layers detect low-level features such as edges and textures, progressively building toward higher-level abstractions like shapes and patterns. This hierarchical feature extraction is foundational to pre-trained models' success.
- Transfer Learning Capabilities: Pre-trained models can be fine-tuned by freezing some layers while updating others to adapt to specific tasks. This process allows the model to leverage existing knowledge while specializing in new domains.
- Classification Head: The final layers of a pre-trained model are tailored to the specific task at hand, such as tumor classification or segmentation. This modularity makes pre-trained models highly adaptable.

2.5. Siamese Neural Networks

Siamese neural networks (SNNs) are a unique deep learning architecture designed to compare pairs of inputs and determine their similarity or dissimilarity. This architecture consists of two identical subnetworks that share the same structure and weights, ensuring consistent feature extraction from the input pairs. Each subnetwork processes one input, and their outputs are combined using a similarity function, such as Euclidean distance or

cosine similarity, to generate a similarity score.

In the context of medical imaging, particularly for Head and Neck Cancer (HNC), Siamese neural networks have proven to be highly effective for tasks involving small datasets and image comparison. These networks are particularly useful in scenarios where obtaining large amounts of labeled data is challenging, as they learn to identify relationships between image pairs rather than relying solely on absolute classification. For example, in vocal cord leukoplakia classification, a study demonstrated the ability of Siamese networks to handle limited sample sizes while achieving accurate classification results by comparing vocal cord images [28].

Architecture and Working of Siamese Neural Networks

The architecture of Siamese neural networks is structured to compare and analyze the relationships between pairs of inputs. The key components and their functionalities are as follows:

- Input Pairs and Processing: In an SNN, the input consists of pairs of data points. Each pair is processed independently by two identical subnetworks, which are designed to extract meaningful features from the inputs. The inputs can be images, text, or other types of data, depending on the application.
- Feature Extraction: The identical subnetworks, often referred to as twin networks, are responsible for feature extraction. These subnetworks typically consist of convolutional layers (for images) or recurrent layers (for sequential data), followed by fully connected layers. The extracted features are represented as high-dimensional vectors, or embeddings, that capture the essential characteristics of the inputs.
- Comparison Using Similarity Functions: After feature extraction, the SNN compares the embeddings using a similarity function. This function quantifies how similar or dissimilar the inputs are, based on their feature representations.

2.6. Graph Neural Networks

Graph Neural Networks (GNNs) are specialized deep learning models designed to process and analyze graph-structured data. Unlike traditional neural networks, which operate on Euclidean data such as grids or sequences, GNNs excel at handling non-Euclidean data, where relationships between entities are explicitly represented as edges connecting nodes. This flexibility allows GNNs to model complex relationships and dependencies, making them particularly well-suited for advanced tasks in medical imaging and oncology, including Head and Neck Cancer (HNC). In the context of HNC, GNNs have demonstrated substantial potential by integrating diverse data types and capturing intricate relationships critical for advanced diagnostic and prognostic tasks. For instance, multi-level data fusion is one of the most impactful applications of GNNs. By integrating multi-modal imaging data, such as PET and CT scans, with clinical features, GNNs provide a comprehensive analysis of patient health. A notable example is the Multi-Level Fusion Graph Neural Network (MLF-GNN), which combines PET and CT imaging data for risk stratification in HNC patients. This approach enhances prognosis prediction by capturing interactions between imaging features and clinical attributes, thereby facilitating improved decision-making in cancer treatment [29].

GNNs also play a pivotal role in risk stratification and prognosis prediction by analyzing relationships between imaging biomarkers, clinical data, and patient outcomes. This capability enables the stratification of patients into risk groups, aiding clinicians in predicting disease progression and optimizing treatment strategies. Additionally, GNNs can process genomic and molecular interaction networks, uncovering biomarkers associated with HNC and supporting targeted therapies by identifying key molecular pathways. Beyond individual-level analysis, GNNs offer insights into tumor-environment interactions by modeling the tumor microenvironment as a graph. This allows the analysis of relationships between cancer cells, immune cells, and surrounding tissues, providing valuable information on tumor progression and potential therapeutic targets.

The core mechanism of GNNs is message passing, where nodes iteratively update their representations by aggregating information from their neighbors. This process enables GNNs to capture both local and global structures within a graph, allowing effective inference on complex relationships between entities. Common tasks addressed by GNNs include node classification, link prediction, and graph classification. Popular GNN architectures include:

- Graph Convolutional Networks (GCNs): Extend the principles of convolutional neural networks to graph-structured data, enabling the extraction of localized features.

- Graph Attention Networks (GATs): Introduce attention mechanisms to prioritize information from the most relevant neighbors during message passing.

2.7. Attention Mechanism

The attention mechanism is a transformative concept in deep learning, enabling models to dynamically focus on specific regions of input data while disregarding irrelevant parts. Originally developed for sequence-to-sequence (Seq2Seq) tasks in natural language processing, such as machine translation, attention has since been adapted to diverse domains, including image captioning, object detection, and medical image analysis. By assigning varying levels of importance (weights) to different input elements, attention mechanisms enhance feature extraction, contextual understanding, and model interpretability, making them particularly effective for tasks requiring precise localization and integration of complex data. In medical imaging, particularly for Head and Neck Cancer (HNC) diagnostics, attention mechanisms have demonstrated significant utility. For example, attention is widely used to integrate multi-modal imaging data, such as PET and CT scans. In 18FDG PET-CT imaging, attention mechanisms enhance tumor segmentation by focusing on complementary features across modalities, improving diagnostic accuracy. Additionally, attention aids in accurate tumor segmentation by directing the model's focus to key regions within the imaging data, ensuring precise delineation of tumor boundaries. This is critical for effective treatment planning, such as radiation therapy.

Another major advantage of attention mechanisms in HNC diagnostics is their ability to analyze feature importance. By visualizing attention weights, such as heatmaps, clinicians gain insights into which regions of an image contribute most significantly to model predictions, thereby increasing the interpretability and trustworthiness of AI-driven results. In cases where large and complex medical datasets are involved, attention mechanisms also reduce computational complexity by prioritizing significant features, allowing models to process data efficiently without sacrificing performance. Attention mechanisms also excel in dynamically adapting their focus during various stages of processing, enhancing contextual understanding and accuracy. This dynamic focus is particularly beneficial for the fusion of multi-modal data, where complementary information from modalities like PET and CT can be seamlessly integrated to provide a holistic view of tumor characteristics. The ability of attention mechanisms to process longer sequences by considering all encoder states dynamically further enhances their application in longitudinal imaging studies or time-series medical data. The adoption of attention mechanisms in medical imaging has revolutionized how models process, interpret, and extract valuable insights from complex datasets. In HNC diagnostics, these mechanisms empower models to deliver precise and interpretable predictions, improve tumor segmentation accuracy, and facilitate multi-modal data integration. As attention-based models continue to evolve, their potential to advance diagnostic precision and treatment planning in oncology is becoming increasingly evident.

How Attention Works

The central idea behind attention is to dynamically compute context vectors by considering the relevance of each input element at different stages of processing. This mechanism enhances the limitations of traditional fixed-length representations, especially for longer or more complex inputs. The key steps include:

- Score Computation: A scoring function evaluates the relevance of each input element to the current processing state. For example, in an encoder-decoder model, scores are calculated for all encoder states based on the current decoder state.
- Attention Weights: Scores are normalized using a softmax function to produce attention weights, which sum to 1 and represent the relative importance of each input element.
- Context Vector: A weighted sum of all input elements, based on the attention weights, is computed. This context vector dynamically represents the most relevant information for the current task.
- Integration: The context vector is concatenated with the decoder's previous output and used to generate the next prediction in the sequence.

This process is repeated at each time step, allowing the model to focus on different parts of the input as needed.

3. Literature Review

HNC is diagnosed using various imaging modalities, such as radiological imaging, endoscopic/clinical imaging, hyperspectral imaging, multimodal optical imaging and histology whole-slide imaging. **Figure 3** illustrates the investigations conducted in the field of HNC detection, employing diverse imaging modalities. Remarkably, the findings reveal an evident research gap in fully exploring the potential of histopathology imaging techniques for HNC. Notably, there is a distinct absence of studies that specifically investigate the use of histopathology imaging for survival prediction in HNSCC, highlighting an untapped avenue for further research and potential advancements in this field. In the following sections, we will discuss each of these modalities in detail, highlighting their strengths, limitations, and clinical applications in head and neck imaging.

3.1. Radiological Imaging

Radiological imaging is a medical specialty that employs various imaging techniques to visualize and diagnose internal structures of the body. These techniques include computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), and ultrasound. Different research studies used different radiological imaging techniques for the evaluation of HNC.

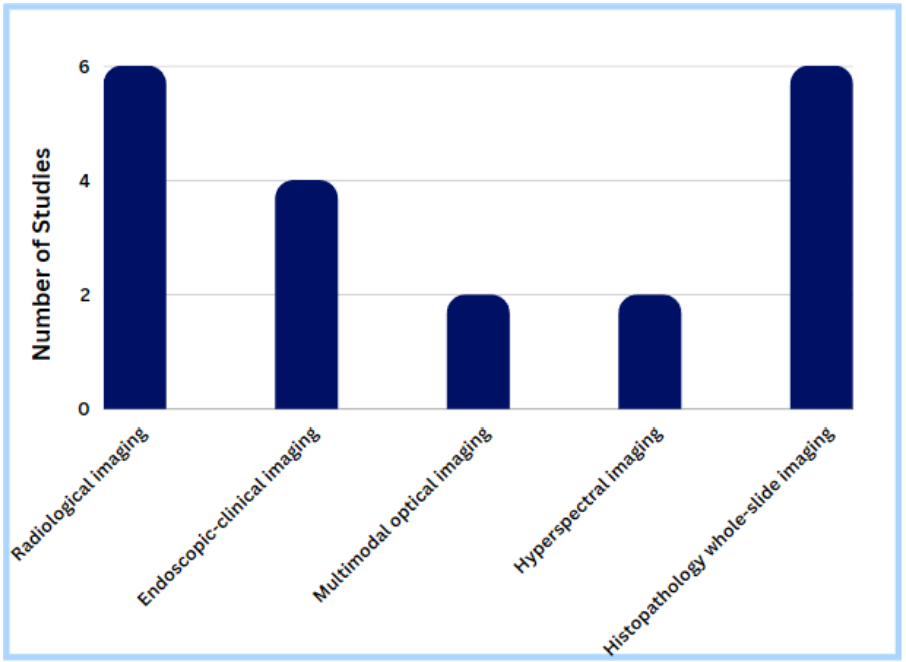

Figure 3. Distribution of studies across imaging modalities for HNC detection.

Al Ajmi et al. [30] created a machine learning method that utilized spectral dual-energy CT (DECT) information from multi-energy virtual monochromatic image datasets to detect the energy-dependent variations in tissue attenuation for categorizing typical benign parotid gland tumors (Warthin tumor and pleomorphic adenoma). Their method achieved an accuracy of 92%. In a study, the authors Ranjbar et al. [31] conducted research to evaluate the potential of computed tomography (CT)-based texture analysis in determining the human papillomavirus (HPV) status of oropharyngeal squamous cell carcinoma (OPSCC) with an accuracy of 75.7%. In another research study [32], both morphologic and functional information has been utilized for automatic detection and classification of nasopharyngeal carcinoma (NPC) on PET/CT using support vector classifier. Their system was validated with 25 PET/CT examinations from 10 NPC patients, and results showed 99.3% sensitivity in identifying hypermetabolic lesions larger than 1 cm in size and excluding normal physiological uptake.

An ultrasound-based CAD system has been proposed by Siebers et al. [33] for the differential diagnosis of parotid gland lesions based on supervised classification using tissue-describing features derived from ultrasound radio-frequency (RF) echo signals and image data. Their proposed system aims at automating the differentiation between malignant and benign cases with an AUC of 0.91.

The researchers Huang et al. [34] aimed to investigate the application of MRI features on a deep learning-based image super-resolution reconstruction algorithm, the optimized convolutional neural network (OPCNN) algorithm, for the diagnosis of nasopharyngeal carcinoma (NPC). Their study included 54 patients with NPC, and their MRI images were processed using the traditional CNN model, U-net network model, and the OPCNN algorithm and concluded that the OPCNN algorithm can improve the quality of MRI images, and its effect is better than the traditional deep learning models. By integrating various MRI modalities such as DCE-MRI, T2WI, and DWI, the diagnostic accuracy can be significantly enhanced, resulting in an impressive accuracy rate of 93.2%. Ramkuma et al. [35] explore the potential of MR imaging–based texture analysis to differentiate between sinonasal inverted papilloma and squamous cell carcinoma, and to compare the classification performance of the machine learning algorithm with the neuroradiologist's review. A total of 46 adult patients who had inverted papilloma or squamous cell carcinoma resected were included in the study. The results showed that the machine learning algorithm achieved similar accuracies of 89.1% for both the training and validation datasets, while the accuracy of the algorithm was better than that of the neuroradiologist's ROI review but not significantly different from the neuroradiologist's review of the tumors or entire images.

Figure 4 provides an overview of how deep learning is utilized in the analysis of radiological images.

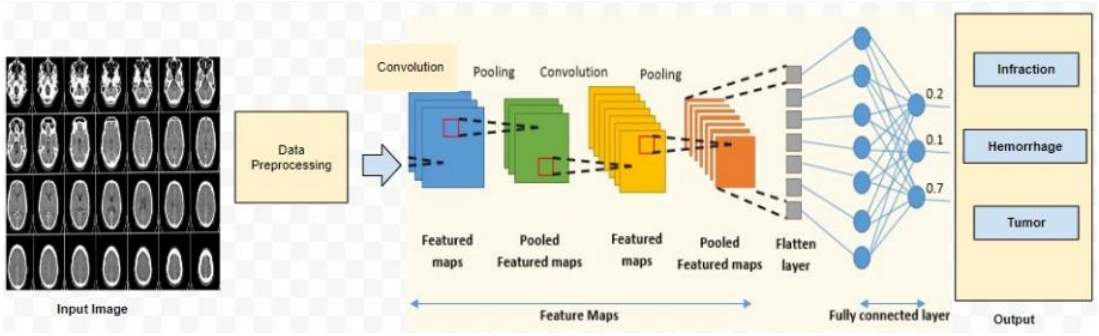

Figure 4. Deep learning framework for radiological image analysis.

Radiological imaging provides detailed, high-resolution images that can be used to diagnose and monitor a wide range of diseases and conditions, making it an essential tool in medical image analysis with numerous applications. However, radiological imaging can be expensive, expose patients to radiation, produce large amounts of data, and be affected by factors such as patient motion and image artifacts, which can affect the accuracy and reliability of deep learning models.

3.1.1. Radiological Datasets

- Head-Neck-PET-CT: This dataset consists of FDG-PET/CT and radiotherapy planning CT imaging data from 298 patients diagnosed with histologically confirmed HNC. These patients, treated at four different institutions in Québec, underwent pre-treatment FDG-PET/CT scans between April 2006 and November 2014, with scans conducted a median of 18 days (range: 6–66 days) before treatment[1].
- Head-Neck Cetuximab: The Head-Neck Cetuximab collection includes CT, PT, RTSTRUCT, RTPLAN, and RTDOSE images from 945 patients, all gathered as part of the RTOG 0522 trial[2].
- HaN-Seg: This publicly available dataset is designed for the segmentation of organs-at-risk (OARs) in the head and neck region using both computed tomography (CT) and magnetic resonance (MR) imaging. It aims to enhance segmentation accuracy for radiotherapy planning by combining CT and MR images to improve the visibil-

[1] https://www.cancerimagingarchive.net/collection/head-neck-pet-ct/
[2] https://www.cancerimagingarchive.net/collection/head-neck-cetuximab/

ity of OARs less discernible in CT scans. The dataset includes images from 56 patients, each with both CT and T1-weighted MR images, as well as manually segmented binary masks for 30 OARs. It is intended for research and development in medical imaging, particularly for improving radiotherapy treatment planning by accurately delineating target volumes and OARs[3].

- HNSCC-3DCT-RT: This dataset is a collection of high-resolution three-dimensional computed tomography (CT) images from 31 patients diagnosed with HNSCC. It includes CT scans taken during pre-treatment, mid-treatment, and post-treatment phases, allowing researchers to visualize changes in tumor characteristics throughout the radiotherapy process. The dataset also provides additional clinical information, such as tumor volume, treatment-related toxicities, and patient demographics, which can aid in the study of treatment outcomes and the development of predictive models[4].
- Hyper-Scale Multimodal Imaging Dataset: This dataset comprises 4.5 million images collected from various medical imaging modalities including CT, MRI, PET, and X-ray. It was created by combining 102 medical imaging datasets and aims to classify these images by their modality type. This dataset is particularly valuable for training deep learning models and improving diagnostic outcomes in clinical settings [5].

3.2. Endoscopic-Clinical Imaging

Endoscopic-clinical imaging is a diagnostic technique that combines endoscopy and imaging technology to examine the internal organs and tissues of the body. Several studies have utilized clinical data from endoscopic imaging to detect cancer in the head and neck. In a research study [36], the authors used CNN (Inception V3) to detect nasopharyngeal malignancies on endoscopic images of the nasopharynx with 28,966 endoscopic images in a training set and 1,430 endoscopic images in a testing set and achieved an accuracy of 88.7% on the test set. The authors of reference [37] proposed a novel approach for the early detection of oral cancer using an optimized computational model. The authors combined Echo State Neural Networks (ESNNs) and Gravitational Search Algorithm (GSA) for oral cancer detection by optimizing the ESNN parameters using GSA. The results of the experiments demonstrate that the proposed model achieves an accuracy of 98.47% on clinical X-rays images and outperformed other state-of-the-art approaches for oral cancer detection. The authors of reference [38] proposed a texture-based machine learning approach using local binary patterns and statistical analysis for the classification of laryngeal tissue (normal vs. malignant) in endoscopic images, which can aid in the early diagnosis of laryngeal cancer. The authors reported an overall classification accuracy of 94.8%. A DL based system has been proposed for auto automatic classification of oral dysplasia and malignancy tissue images captured using a dual-modality smartphone-based imaging system [39]. The system achieved an overall accuracy of 87.6% in classifying oral dysplasia and malignancy images, with high sensitivity and specificity in detecting oral malignancy.

Figure 5 provides an overview of how deep learning is utilized in the detection, segmentation, and classification of endoscopic images.

Endoscopic-clinical imaging in deep learning has several advantages as a diagnostic tool, as it provides high-resolution real-time images of the gastrointestinal tract, allowing for accurate diagnosis and treatment planning. However, this technique requires a high level of technical skill and expensive equipment, which can limit its accessibility in certain areas. Additionally, the risk of infection and potential complications such as bleeding and perforation of the organ being examined may cause concern for some patients. The limited view provided by this technique may also make it difficult to diagnose certain conditions.

Endoscopic Datasets

Narrow Band Imaging (CE-NBI): This dataset includes 11,144 images from 210 adult patients with vocal fold conditions. The CE-NBI technique enhances vascular pattern visualization, aiding in the differentiation between benign and malignant lesions. Various machine learning methods applied to this dataset have achieved high diagnostic accuracy and reliability. The complete dataset, including all images and labels, is available in the Zenodo Repository [6].

[3] https://han-seg2023.grand-challenge.org
[4] https://www.cancerimagingarchive.net/collection/hnscc-3dct-rt/
[5] https://research-portal.st-andrews.ac.uk/en/publications/classification-of-hyper-scale-multimodal-imaging-datasets/datasets
[6] https://zenodo.org/

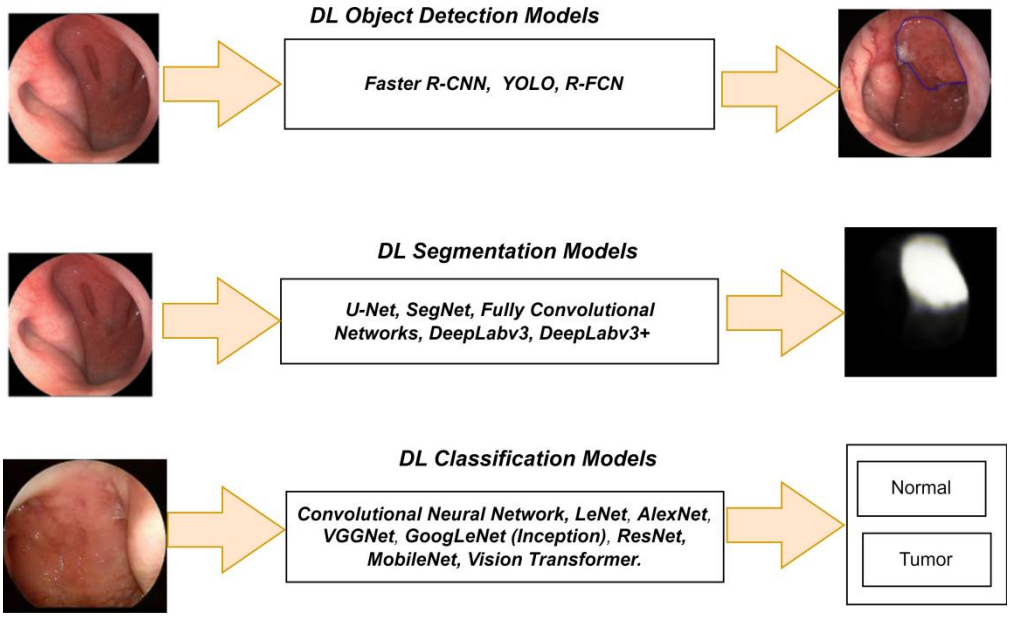

Figure 5. Deep learning framework for endoscopic image analysis.

3.3. Multimodal Optical Imaging

Multimodal optical imaging is an advanced imaging technique that combines different optical imaging modalities to obtain a more comprehensive view of biological structures and processes. It typically involves using a combination of techniques such as fluorescence imaging, confocal microscopy, and optical coherence tomography (OCT). In research [40], the authors utilized multispectral wide-field optical imaging, such as white-light reflectance, autofluorescence, narrow-band reflectance, and cross-polarized imaging, to distinguish between oral cancer and non-cancerous mucosa by evaluating the contrast in the images. The results indicated that autofluorescence imaging using a 405 nm excitation wavelength produced the highest contrast images. The red-to-green fluorescence intensity ratio calculated from these images was the best indicator for identifying cancer versus non-cancerous tissue with a sensitivity of 100% and specificity of 85%. However, the method accurately identified malignant tissue from non-cancerous tissue but had some limitations in identifying precancerous lesions. In another research study [41], the authors detected Oral Neoplasia in "Vivo" aimed to evaluate the performance of a multimodal optical imaging approach for the detection of oral squamous cell carcinoma (OSCC). This approach involved using autofluorescence imaging to identify high-risk regions within the oral cavity, followed by high-resolution microendoscopy to confirm or rule out the presence of neoplasia. This study included data from 92 sites to develop algorithms for the automatic identification of OSCC in vivo, which were then prospectively evaluated using images from 114 sites. Diagnostic accuracy was assessed based on confirmed histological diagnoses from biopsies or surgical specimens. The study found that the multimodal imaging approach was highly accurate in detecting benign lesions, with a 100% classification accuracy, and an 85% accuracy in detecting cancerous lesions. In cases where a surgical specimen was available, the imaging approach correctly classified 100% of benign sites and 61% of neoplastic sites. **Figure 6** presents how deep learning is utilized to diagnose retinal diseases.

The above studies suggest that multimodal optical imaging with automated image analysis could potentially improve the accuracy and efficiency of oral cancer screening. Furthermore, this imaging technique can improve the accuracy and specificity of image analysis tasks, such as identifying and tracking specific cell types or molecules. Multimodal imaging can also help to reduce the impact of imaging artifacts and noise, leading to clearer and more accurate images. However, there are some challenges associated with multimodal imaging. For one, integrating data from multiple modalities can be technically complex, requiring specialized equipment and software. Additionally, multimodal imaging can be more time-consuming and resource-intensive, as data from each modality must be acquired and processed separately before being combined.

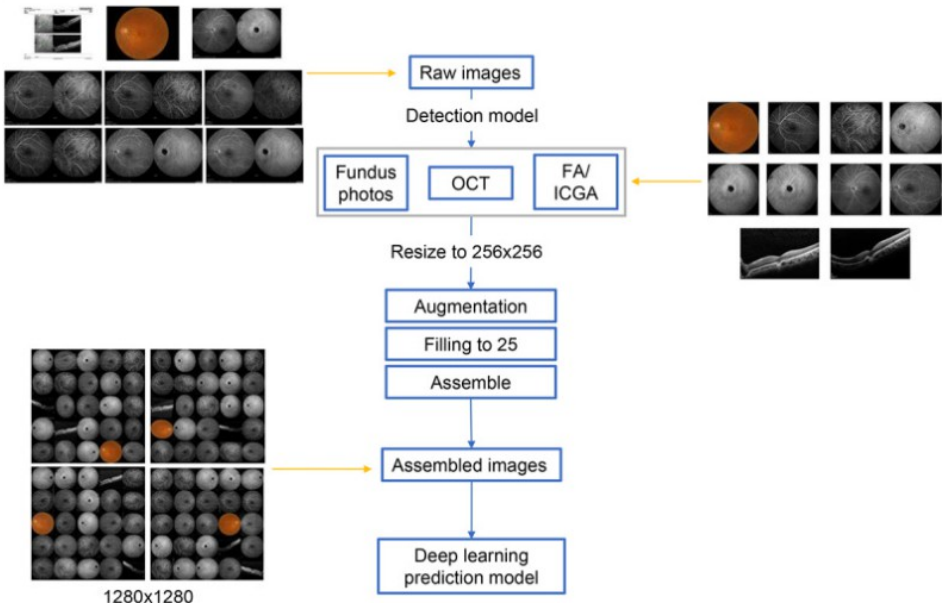

Figure 6. Utilization of deep learning for the diagnosis of retinal diseases (adopted from [42]).

Another potential limitation is that different modalities may have different limitations and biases, which can affect the overall accuracy and reliability of the results.

Multimodal Optical Datasets

- MEMO Dataset: The MEMO dataset includes multimodal retinal images, specifically pairs of Enhanced Depth Imaging (EMA) and OCTA images. While the exact number of images is not specified, it is designed for studying multimodal retinal image registration and includes labeled matched points for research purposes. The MEMO dataset contains 30 pairs of EMA and OCTA images. For each image pair, 6 corresponding point pairs were manually annotated. In addition, each EMA image comes with a carefully annotated vessel segmentation mask[7].
- CF-FA Dataset: The CF-FA dataset comprises 59 pairs of images, including color fundus (720 × 576, RGB) and fluorescein angiography (720 × 576, grayscale) images. Of these, 29 pairs are from healthy individuals, while the remaining 30 pairs are from patients diagnosed with retinopathy[8].

3.4. Hyperspectral Imaging

Hyperspectral Imaging is a powerful technology that combines imaging and spectroscopy to capture and analyze data from across the electromagnetic spectrum. A novel method proposed by Halicek [43] employed Hyperspectral imaging (HSI) to perform real-time optical biopsies of ex-vivo surgical specimens collected from 21 patients undergoing surgical cancer resection. The method shows promising results in distinguishing squamous cell carcinoma (SCC) from normal tissues with an accuracy of 81% using CNN, as well as sub-classifying normal oral tissues into epithelium, muscle, and glandular mucosa using a decision tree method, with an accuracy of 90%. The authors also developed a CNN architecture for differentiating between thyroid carcinoma and normal thyroid and achieved an accuracy of 81% . Partitioned Deep CNNs offer a way to enhance efficiency and scalability, making them useful in resource-constraint environments. A similar architecture has been proposed by Jeyaraj and Samuel Nadar [44] to classify and label regions of interest in multidimensional hyperspectral images. The algorithm achieved a classification accuracy of 91.4% for a 100-image training dataset for the classification of malignant and benign cancer, and a classification accuracy of 94.5% for 500 training patterns for the classification of malignant cancer and normal tissue. **Figure 7** represents the deep learning method for cancer detection with hyperspectral imaging. Although Hyperspectral imaging is a promising approach for image analysis in deep learning, the complexity and high dimen-

[7] https://chiaoyiwang0424.github.io/MEMO/
[8] https://chiaoyiwang0424.github.io/MEMO/

sionality of the hyperspectral data can make it challenging to train deep learning models effectively. Additionally, the large amounts of data generated by hyperspectral imaging require significant computational resources and specialized expertise for processing and analysis. Obtaining high-quality hyperspectral data also presents a great challenge for training deep learning models.

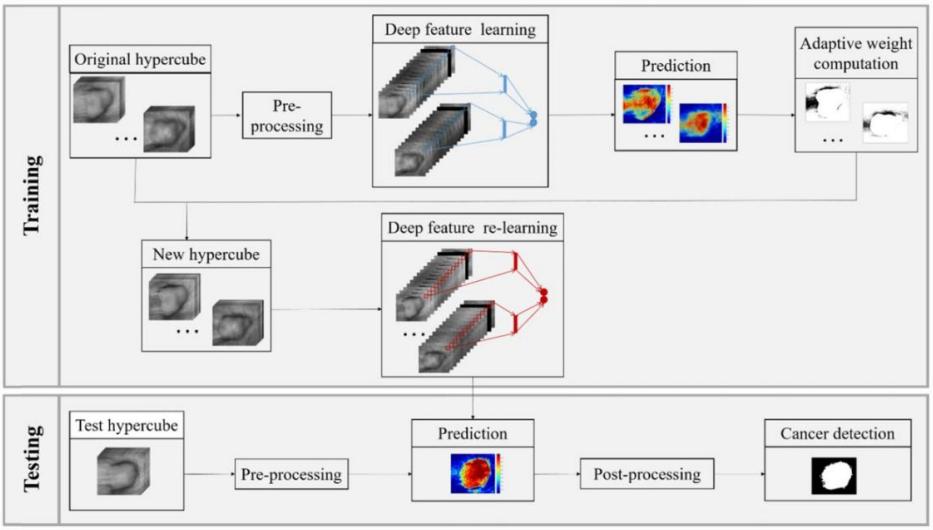

Figure 7. Deep learning method for cancer detection using hyperspectral imaging (adopted from [45]).

3.5. Histopathology Whole-Slide Imaging

In the realm of deep learning, histopathology images stand out as a highly advantageous modality. Their high-resolution and standardized imaging protocols allow for accurate visualization of tissue structures, while their status as the gold standard for the diagnosis of many diseases provides a reliable reference point for training deep learning models. The relative ease of collection makes histopathology images a practical choice for developing accurate and robust deep-learning algorithms. Together, these advantages make histopathology images a valuable tool for advancing our understanding of disease and improving diagnostic accuracy.

In a study conducted by Halicek [46], the authors recognized the potential of histopathology images and employed a convolutional neural network (CNN) to differentiate between normal and abnormal histopathology images of HNSCC. The CNN was trained using a dataset comprising 381 images from 156 patients, achieving an impressive accuracy of 95%. However, during the experiment, the researchers discovered limitations of the algorithm caused by artifacts originating from whole-slide images. These artifacts included out-of-focus regions, tissue folding, and tearing, which were identified after the completion of the study. Unfortunately, these artifacts significantly affected the classification accuracy, particularly in the case of the squamous cell carcinoma (SCC) testing dataset.

Another research study conducted by He et al. [47] utilized a deep CNN (InceptionV3) model for the diagnosis of laryngeal squamous cell carcinoma (LSCC) based on Narrow-Band Imaging (NBI) endoscopy and pathological images. The model was trained, tested and validated using 4,591 patient laryngeal NBI scans and 3,458 pathological images in the ratio of 70:15:15. Remarkably, the deep learning model achieved high accuracy in diagnosing LSCC on the test set, with an area under the curve (AUC) of 0.87 and 0.98 for NBI and pathology groups, respectively.

By employing shape, texture, and color features from whole slide images, the authors [48] successfully distinguished normal and malignant Oral Squamous Cell Carcinoma (OSCC) utilizing 42 original whole slides. Remarkably, their approach achieved an impressive accuracy, specificity, sensitivity, and precision of over 99%.

The researchers in the study conducted by Rahman et al. [49] successfully categorized microscopic images of oral squamous cell carcinoma into benign and malignant categories using histological slides. They employed texture features extracted from the images using GLCM (Gray-Level Co-occurrence Matrix) and histogram techniques. The classification was performed using a linear Support Vector Machine (SVM), and the results showcased the remarkable effectiveness of this approach, achieving a perfect accuracy rate of 100%.

The authors Rodner et al. [50] utilized a CNN to differentiate between cancer, normal epithelium, background stroma, and other tissue types in 114 images from 12 patients, for the diagnosis of HNSCC. The study achieved average and overall recognition rates of 88.9% and 86.7%, respectively, for the four classes.

Researchers Tang et al. [51] used convolution neural network-based algorithms for the detection of lymph node metastasis in HNSCC with an accuracy of 86%. However, there are some limitations in this study. The training sample size consists of only 20 patients (collected from only one institution); further, the test set consists of only a few patients, thus the results are not reliable.

Histological Datasets

Histopathology OSCC Dataset: comprises H&E whole slide images of the normal oral cavity epithelium and images depicting Oral Squamous Cell Carcinoma (OSCC). This dataset comprises a total of 1,224 histopathological images. These images are categorized into two sets, each with different resolutions. The first set contains 89 images showing the normal oral cavity epithelium and 439 images depicting OSCC, all captured at 100x magnification. The second set includes 201 images of the normal oral cavity epithelium and 495 images of OSCC, with each image taken at 400x magnification [8].

OSCC Histopathology Dataset: This dataset comprises H&E whole slide images of the normal oral cavity epithelium and images depicting Oral Squamous Cell Carcinoma (OSCC). These images are available at 100x and 400x magnification levels, providing a comprehensive and detailed view for our analysis and experimentation. This dataset has three subsets: training, validation, and test sets. Each of these subsets serves a crucial role in our research study, contributing to the training and evaluation of our models for detecting Oral Squamous Cell Carcinoma (OSCC). **Figure 2** depicts the distribution of the dataset, where the training set comprises 2,435 samples of normal cases and 2,511 samples of oral squamous cell carcinoma. The validation set has 28 normal cases and 92 squamous cell carcinoma cases, while the test set includes 31 normal cases and 95 cases of oral squamous carcinoma [9].

In this study, **Table 1** provides a detailed summary of the studies focused on medical imaging (MI) techniques for head and neck cancer (HNC).

Table 1. Studies of Medical Imaging Techniques for Head and Neck Cancer.

Related Work	Imaging Technique	Work Accomplished	Methods	Output	Performance
[30]	CT	Classification of benign parotid gland tumors into Pleomorphic adenoma and Warthin tumor	Random Forest	Accuracy	92%
[31]	CT	Detection of human papillomavirus of oropharyngeal squamous cell carcinoma	Quadratic discriminant analysis	Accuracy	75.7%
[32]	PET/CT	Classification of nasopharyngeal squamous cell carcinoma	Support Vector Machine	Sensitivity	99.3%
[33]	Ultrasound radio frequency echo data	Differentiating between malignant and benign parotid gland lesions	Texture feature-based maximum likelihood classifier	AUC	91%
[34]	MRI	Diagnosis of Nasopharyngeal Carcinoma	Optimized convolutional neural network	Accuracy	93.2%
[35]	Endoscopic images	Differentiate between sinonasal inverted papilloma and squamous cell carcinoma	Support Vector Machines	Accuracy	89.1%
[36]	MRI	Detect nasopharyngeal malignancies	Convolutional neural network (inception)	Accuracy	88.7%
[37]	X-Rays	Detect Oral Cancer	Gravitational search optimized echo state neural networks	Accuracy	98.47%
[38]	Endoscopic images	Classification of laryngeal tissue (normal vs. malignant)	Support Vector Machine	Accuracy	94.8%
[39]	Endoscopic images	Oral cancer classification of dysplasia and malignancy tissue from normal ones	Convolutional Neural Network Architecture	Accuracy	87.6%
[40]	Narrowband reflectance, autofluorescence, and polarized reflectance images	Identification of malignant tissue from non-cancerous tissue	Linear Classifier, Decision Tree	Sensitivity	100%
[41]	Multimodal optical images	Detection of oral squamous cell carcinoma	Linear Discriminate Analysis	Accuracy	85%
[43]	Hyperspectral images	Distinguishing squamous cell carcinoma from normal tissues	Convolutional Neural Network	Accuracy	81%

[9] https://www.kaggle.com/datasets/ashenafifasilkebede/dataset

In this research study, **Table 2** provides a comprehensive summary of studies conducted on medical imaging (MI) techniques for head and neck cancer (HNC).

Table 2. Studies of medical imaging techniques for head and neck cancer.

Related Work	Imaging Technique	Work Accomplished	Methods	Output Performance
[44]	Multi-dimensional hyperspectral images	Classification of malignant and benign cancer	Convolutional Neural	Accuracy: 91.4%
[46]	Histopathology images	Distinguish between normal and abnormal histopathology images	Convolutional Neural Network	Accuracy: 95%
[47]	Histopathology images	Diagnosis of laryngeal squamous cell carcinoma	Deep convolutional neural network (inceptionV3	AUC: 87%
[48]	Histopathology images	Distinguish normal and malignant Oral Squamous Cell Carcinoma	Decision Tree Classifier, SVM, and Logistic regression	Accuracy: 100%
[49]	Histopathology images	Diagnosisoflaryngeal squamous cell carcinoma	Deep convolutional neural network (inceptionV3)	AUC: 87%
[50]	Histopathology images	Classification of oral squamous cell carcinoma into benign and malignant	Linear SVM	Accuracy: 100%
[51]	Histopathology images	Classification between cancer, normal epithelium, background stroma in head and neck carcinoma	Fully Convolutional Neural Network	Accuracy: 88.9%

4. Challenges in Implementing Deep Learning Models for Head and Neck Cancer Imaging

4.1. Annotation Practices in Medical Imaging

The annotation of HNC images obtained from primary or secondary sources presents a substantial challenge for researchers in the field of medical imaging. The involvement of medical professionals in both the collection and annotation processes is crucial for establishing a well-organized dataset. Nevertheless, inconsistencies frequently arise, as medical experts may annotate images based on their individual experiences and varying levels of expertise. In instances where multiple regions exhibiting structural abnormalities are present within a single medical image, experts may prioritize labeling the most conspicuous abnormality, potentially overlooking others. This variability in annotation practices can lead to mislabeling, which adversely affects classification outcomes and results in increased rates of false positives and false negatives.

4.2. Role of Dataset Size in Imaging Studies

Due to the lack of publicly available benchmark imaging datasets for head and neck cancer, many researchers have resorted to compiling their own datasets, which are often small or incomplete. Several studies [52, 53] have reported that the limited amount of training data has hindered the effective training of their models, resulting in unreliable outcomes when tested with real-world data. Furthermore, researchers in studies countered significant obstacles during data collection, including issues related to missing data, patient non-consent due to privacy concerns, and the denial of private hospitals to share their data, citing confidentiality regulations.

4.3. Data Quality Challenges

The effectiveness of deep learning models in HNC imaging is significantly influenced by the quality of the training data. Ensuring high data quality presents considerable challenges across various imaging modalities, including CT, MRI, PET, ultrasound, and histopathology. Each modality has its own unique issues. For instance, factors such as patient movement, machine calibration problems, and environmental conditions can introduce noise and artifacts in images from CT, MRI, PET, and ultrasound. In histopathology, additional challenges such as inconsistent staining, variations in lighting during slide examination, and differences in operator expertise further compromise data quality.

4.4. Requirements for Computational Power

Deep learning models, particularly those used in complex fields such as medical imaging, require extensive computational power to process large datasets and perform intricate calculations. Graphics processing units (GPUs) are essential for accelerating the training of these models, as they can handle parallel processing tasks much more efficiently than traditional central processing units (CPUs). Specialized hardware, such as tensor processing units (TPUs) or field-programmable gate arrays (FPGAs), can further enhance performance by optimizing specific tasks related to deep learning. However, the substantial computational requirements associated with training and deploying these models can present significant challenges for smaller clinical centers and individual researchers. Many of these institutions may lack the financial resources to invest in high-performance computing infrastructure, which can limit their ability to develop and implement advanced deep learning solutions. As a result, they may miss out on the potential benefits of these technologies, which could improve diagnostic accuracy, treatment planning, and patient outcomes.

4.5. Accessing Unpublished Datasets

A key issue identified in this review is that researchers tend to achieve better results when using publicly available datasets, as compared to when they utilize their own private or unpublished datasets. One reason for this disparity is that public datasets are often well-preprocessed, refined, and balanced, leading to more robust outcomes. In contrast, unpublished or private datasets often lack sufficient preprocessing and data balance, which can negatively impact results. Several studies [54, 55] have relied on unpublished datasets. While authors have compared their results with other similar methods to validate their findings, the authenticity of private datasets remains uncertain until they are tested by multiple researchers using different classification techniques.

4.6. Class Imbalance

Head and neck cancer (HNC) imaging datasets often suffer from class imbalance, where certain cancer types or stages are underrepresented compared to others. This lack of balance in the distribution of classes can significantly impact the performance of deep learning models trained on these datasets. The problem of class imbalance manifests differently across various HNC imaging modalities. In CT and MRI scans, early-stage cancers may be less prevalent in the dataset, leading to an underrepresentation of these classes. PET imaging can exhibit an imbalance due to varying metabolic activity patterns among HNC subtypes. Ultrasound datasets may be skewed by anatomical variations and tumor accessibility, while histopathology samples can be biased towards specific subtypes or stages based on biopsy practices or referral patterns. This class imbalance can have serious consequences on the predictive capabilities of deep learning models. Models tend to learn more from the majority classes, which are well-represented in the dataset. As a result, they may develop a bias towards predicting the majority classes, even when minority classes are present. This leads to reduced sensitivity for detecting rare HNC subtypes or early-stage cancers, resulting in higher false-negative rates. In extreme cases, models may completely ignore the minority classes and focus solely on predicting the majority classes, leading to overfitting and poor generalization. Moreover, imbalanced datasets can skew performance metrics like accuracy, making the model appear to perform well overall while masking poor performance in the minority classes. This can give a false sense of confidence in the model's capabilities and lead to misleading conclusions about its effectiveness in clinical settings.

4.7. Ethical Considerations

The use of deep learning models in medical imaging, particularly in HNC imaging, raises critical ethical concerns that must be addressed. Data privacy is a primary issue, as these models rely heavily on sensitive patient information, necessitating robust security measures to prevent unauthorized access and breaches that could lead to identity theft and privacy violations. Additionally, algorithmic fairness is a significant concern. Biases in training datasets can result in disparities in diagnostic accuracy and treatment outcomes, particularly for marginalized populations. Therefore, it is essential to use diverse datasets and implement continuous monitoring to identify and mitigate these biases. Lastly, accountability poses challenges due to the opaque nature of many deep learning algorithms, making it difficult to determine responsibility when errors occur.

5. Critical Gaps, Ongoing Debates, and Future Innovations

5.1. Standardization and Benchmarking: The Need for a Common Framework Across Imaging Modalities

A significant challenge in applying deep learning to HNC, particularly across imaging modalities such as histopathology, radiology (CT, MRI, PET), and molecular imaging, is the lack of standardization and benchmarking. Researchers employ a wide variety of datasets, image preprocessing methods, and training strategies, making direct comparisons between models difficult. The lack of uniformity in data acquisition, model architectures, and performance metrics has led to fragmented progress, with inconsistent results across studies and institutions. This variation prevents a comprehensive assessment of models' clinical effectiveness, ultimately slowing their integration into routine practice. In histopathology, differences in slide preparation, staining protocols, and digitization techniques often result in images of varying quality. Similarly, in radiology and molecular imaging, disparities in scanner settings, image resolutions, and noise reduction techniques complicate model development and evaluation. These variations in imaging practices make it challenging to produce models that are both generalizable and clinically viable. Moreover, inconsistent reporting of critical performance metrics such as accuracy, sensitivity, specificity, and survival predictions, further hampers the ability to compare results across studies and determine the best-performing models. To address these issues, future research must prioritize the establishment of a common framework that spans all imaging modalities. Publicly available, well-annotated, and standardized datasets for each modality are essential. In histopathology, this involves standardizing staining protocols, slide digitization resolutions, and preprocessing techniques, such as patch extraction, stain normalization and image augmentation. In radiology, the focus should be on harmonizing scanning protocols and ensuring uniform image resolutions to minimize the variability introduced by different scanners. These datasets should be representative of diverse populations and clinical settings to ensure that models trained on them are broadly applicable.

5.2. Comprehensive Data Fusion

Most deep learning studies in HNC imaging have traditionally focused on single-modality data, such as MRI, CT, or PET scans. While these modalities offer valuable insights into tumor characteristics, relying solely on one type of imaging can limit the scope and accuracy of model predictions. Integrating multimodal data such as combining anatomical (CT or MRI) with functional (PET) imaging has the potential to enhance model performance by providing a more holistic understanding of tumor biology. This multimodal approach allows models to leverage complementary information, improving the precision of tumor detection, characterization, and treatment response prediction. Histopathology images, a crucial modality in cancer diagnosis and prognosis, offer detailed insights into the microscopic structure of tumor tissues. Incorporating histopathological data alongside radiological images can provide a deeper understanding of tumor aggressiveness, heterogeneity, and response to treatment. For example, combining high-resolution histopathology images with radiological scans can improve survival prediction and guide personalized treatment plans by offering insights into both the macroscopic and microscopic features of the tumor. In addition to multi-modal imaging, integrating temporal information from longitudinal imaging datasets can further improve model performance. Analyzing changes in tumor characteristics over time—such as those captured in sequential MRI or PET scans—enables models to predict tumor progression, recurrence, or response to therapy more accurately. Temporal data can highlight subtle changes in tumor size, shape, or metabolic activity that may not be apparent in single-time point imaging, offering more nuanced predictions of patient outcomes. Furthermore, the inclusion of clinical information, such as patient demographics, tumor histology, treatment protocols, and comorbidities, can significantly enhance the predictive power of deep learning models. For instance, integrating patient-specific factors like age, gender, tumor stage, and histopathological grade with imaging data allows models to deliver more personalized prognostications. This holistic approach aligns with the growing trend toward precision oncology, where treatment strategies are tailored to the unique characteristics of each patient's disease. Future research should focus on the development of deep learning models that can seamlessly integrate multimodal imaging (including histopathology), temporal data, and clinical information. Such models would offer a comprehensive framework for personalised treatment planning, prognostication, and follow-up care in HNC management. By combining diverse data sources, these models have the potential to improve diagnostic accuracy, refine treatment response predictions, and ultimately enhance patient outcomes.

5.3. Explainability, Interpretability, and Authenticity

Deep learning models, particularly CNNs, have demonstrated remarkable performance in various medical imaging tasks, including HNC diagnosis and prognosis. However, these models are frequently regarded as "black boxes" due to their intricate architectures and the lack of transparency in their decision-making processes. This lack of clarity can hinder the adoption of deep learning methods in clinical settings, as clinicians may be cautious about trusting a model's predictions without understanding the reasoning behind them.

Explainability: To address this challenge, researchers have been focusing on developing explainable deep learning models. Explainability refers to the ability to provide clear and understandable explanations for a model's predictions. By incorporating explainability into the design of DL models for HNC imaging, researchers can help clinicians understand the reasoning behind a model's decisions. This can be achieved through techniques such as attention mechanisms, which highlight the regions of an image that are most important for a particular prediction, and layer-wise relevance propagation, which traces the contribution of each feature to the final prediction.

Interpretability: Bridging the Gap between AI and Clinical Practice In addition to explainability, interpretability is another crucial aspect of building trust in deep learning models. Interpretability refers to the ability to understand the internal workings of a model and the relationships between its inputs and outputs. By developing interpretable deep learning models for HNC imaging, researchers can provide clinicians with a clear understanding of how the model arrives at its predictions. This can be achieved through techniques such as visualization of feature maps, which allow clinicians to see the features that the model has learned to recognize, and sensitivity analysis, which helps identify the most important features for a particular prediction.

Authenticity: Ensuring Trustworthiness and Robustness While explainability and interpretability are important for building trust in deep learning models, it is also crucial to ensure that these models are authentic and reliable. Authenticity refers to the ability of a model to perform consistently and accurately in real-world clinical settings. To ensure the authenticity of deep learning models for HNC imaging, researchers should focus on developing methods to assess the trustworthiness and robustness of these models in the face of noisy, incomplete, or adversarial data. This can be achieved through techniques such as uncertainty quantification, which helps identify the regions of an image where the model is most uncertain about its predictions, and adversarial training, which helps improve the model's robustness to adversarial attacks.

6. Experimentation Results

In this section, we present the results obtained from the proposed survival prediction system. The experimentation process is divided into two key stages: (1) Dataset Description and Data Preprocessing, where we describe the characteristics of the dataset and the steps undertaken to prepare the data for model training, and (2) Model Training and Evaluation, where we provide detailed performance metrics and an analysis of the model's outcomes.

6.1. Dataset Description and Data Preprocessing

The Cancer Genome Atlas Head and Neck Squamous Cell Carcinoma (TCGA-HNSC) cohort includes diagnostic slides from 200 subjects, which are used in the experimental setup. These diagnostic slides, offering detailed insights into tissue phenotypic heterogeneity [56], are specifically chosen for their critical role in histologic analysis. Each whole slide image (WSI) is labeled as either Survival or Not Survival, with a label of 0 assigned to Survival and 1 to Not Survival. To ensure a robust and effective analysis, the dataset is divided into two parts: 80% of the data is used for training, while the remaining 20% is reserved for testing the models. Gigapixel histopathology images, which are too large for direct deep learning (DL) model training, are effectively processed using the OpenSlide library [57]. This specialized tool reads and divides these large images into smaller, manageable tiles at 20X magnification, making them suitable for DL training while preserving the necessary detail for accurate analysis. During preprocessing, between 5,000 and 30,000 tiles are typically extracted from each gigapixel image. To avoid introducing noise, tiles that are blank, black, or blurry are rigorously filtered out based on average pixel values. Tiles with average pixel values above 220 (overly white) or below 60 (mostly black) are discarded, and a manual inspection further removes tiles with visual artifacts or other anomalies, ensuring high-quality data for DL training.

6.2. Model Training and Evaluation

In this study, we extracted features from all tiles within each whole slide image (WSI) using a pre-trained Inception v3 model, generating a 2048-dimensional feature vector for each tile. This process resulted in a feature matrix of size N × 2048 for each subject, where N represents the number of tumor tiles in the WSI, varying significantly between samples. The feature vectors from all tiles of the same patient were then averaged into a single 2048-dimensional vector, which served as the input for an attention-based CNN aimed at predicting survival outcomes.

The architecture of the attention-based CNN model consists of a series of convolutional blocks designed to extract increasingly complex and abstract features from the input data. The initial block applies a convolutional layer with 32 filters and a 3 × 3 kernel size, followed by batch normalization to stabilize and accelerate the training process. A LeakyReLU activation function is employed to introduce non-linearity, and the block concludes with a MaxPooling layer that downsamples the spatial dimensions, thereby reducing the computational load for subsequent layers. This pattern is repeated across two additional convolutional blocks, with the number of filters increasing to 64 and 128, respectively. These blocks follow the same structure: convolution, batch normalization, LeakyReLU activation, and MaxPooling. The increasing filter sizes enable the network to capture more sophisticated features as the data progresses through the layers. Following the convolutional blocks, the output is flattened into a 1D vector, which is then passed through a dropout layer with a 50% dropout rate. This dropout layer helps prevent overfitting by randomly deactivating neurons during training, ensuring that the network does not become overly reliant on specific features. The flattened vector is then fed into a custom attention layer, which enhances the model's focus on the most relevant features by assigning different weights to various parts of the input data. This prioritization allows the model to effectively concentrate on critical information for the task at hand. The attention-enhanced features are subsequently passed through a fully connected layer with 128 units, followed by batch normalization and another LeakyReLU activation. A second dropout layer is applied at this stage to further mitigate the risk of overfitting. The model also includes an additional dense layer with 64 units, batch normalization, and LeakyReLU activation, adding another level of feature processing. The architecture concludes with an output layer consisting of a single dense unit with a sigmoid activation function, producing the final prediction as a probability, making the model suitable for binary classification tasks. Overall, this architecture combines the powerful feature extraction capabilities of CNNs with an attention mechanism that refines the focus on the most pertinent data, while also employing regularization techniques to enhance the model's generalization to new data. The attention-based CNN model exhibited balanced performance in distinguishing between "Survival" and "Not Survival" WSIs, as shown in the confusion matrix of **Figure 8**. The precision for the "Survival" class was 0.74, meaning 74% of the predictions were correct. For "Not Survival", the precision was slightly higher at 0.76. The recall was 0.78 for "Survival" and 0.72 for "Not Survival", indicating the model correctly identified 78% of actual survival cases and 72% of actual non-survival cases. The F1-scores were 0.76 for "Survival" and 0.74 for "Not Survival", reflecting strong overall performance. The model's overall accuracy was 0.75, correctly classifying 75% of the WSIs. The confusion matrix further illustrates the model's effectiveness: it correctly identified 14 out of 18 survival cases but misclassified 4 as non-survival, and it correctly predicted 13 out of 18 non-survival cases, with 5 being mislabeled as survival. These results emphasize the model's balanced performance, crucial in clinical settings where both false positives and false negatives can have significant consequences.

The consistency in precision and recall values highlights the model's robustness and reliability as a predictive tool for survival outcomes in HNC patients.

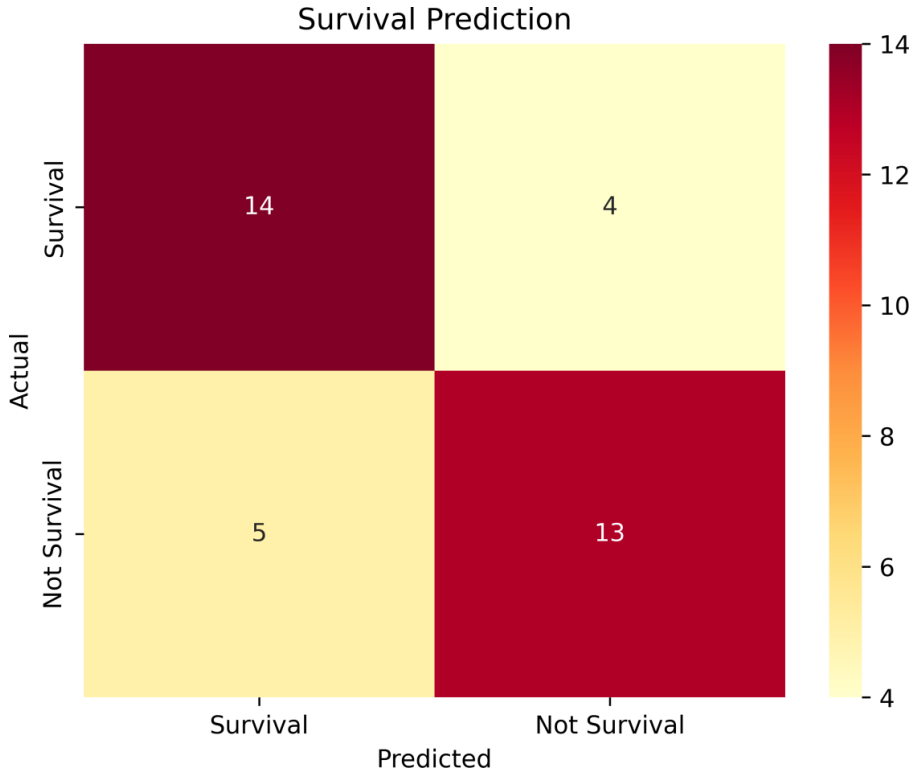

Figure 8. Confusion matrix of survival prediction model.

7. Discussion

The integration of various imaging modalities and advanced artificial intelligence (AI) techniques has shown great promise in enhancing the diagnosis and treatment of HNSCC. This review comprehensively examined the roles of radiological, endoscopic, multimodal optical, hyperspectral, and histopathological imaging techniques, while highlighting the growing impact of deep learning methods across these imaging modalities. Despite these advancements, several critical challenges remain that must be addressed to fully utilize the potential of AI in improving clinical outcomes for HNSC patients. The role of imaging in HNSC diagnosis and monitoring cannot be underestimated, as it offers essential insights into tumor characteristics, disease progression, and patient outcomes. Radiological imaging modalities such as CT, MRI, and PET provide detailed anatomical and functional information that assists in the identification and staging of tumors. The findings of this review underscore that endoscopic imaging, particularly with deep learning models like CNNs, has demonstrated high accuracy in detecting nasopharyngeal malignancies with accuracies approaching 89%. Similarly, multimodal optical imaging has shown promise in early cancer detection by combining different techniques like autofluorescence and narrow-band reflectance imaging. Hyperspectral imaging, although still in its early stages for HNSC, offers potential due to its ability to capture a wide range of spectral data, which helps distinguish cancerous from non-cancerous tissues. Studies included in this review demonstrate accuracies as high as 91% when using convolutional neural networks (CNNs) for tissue classification. Furthermore, histopathology whole-slide imaging (WSI) has been a critical modality for diagnosing HNSC at the cellular level. Employing deep learning methods, such as InceptionV3 models, has enabled effective classification of histopathological images, with accuracy rates as high as 95%.

AI, and specifically deep learning, has revolutionized the analysis of medical images, particularly in fields like oncology where accurate and timely diagnosis can significantly impact treatment outcomes. In HNSC, deep learning models such as convolutional autoencoders (CAEs) and Vision Transformers (ViTs) have proven highly effective for tasks ranging from image enhancement and noise reduction to tumor detection and classification. CAEs, in particular, excel in extracting meaningful features from noisy images, improving downstream tasks like tumor segmentation. One of the most significant findings from the review is the application of Generative Adversarial Networks

(GANs) for high-resolution image generation. GANs ability to convert low-resolution images into high-resolution ones enhances the visibility of tumor features, which is particularly valuable in radiological imaging. This advancement could lead to more precise tumor detection, better segmentation, and improved clinical decision-making. The adaptation of Transformer models, originally developed for natural language processing, to medical image analysis has opened new doors. Vision Transformers (ViTs), by segmenting images into patches, capture long-range dependencies and spatial relationships that traditional CNNs often neglect. This method, which showed competitive results in medical imaging tasks, particularly in HNSC histopathology, positions ViTs as viable alternatives to CNNs.

Despite these advances, several challenges remain in the application of AI to HNSC imaging. A key issue is the lack of standardized datasets and uniform imaging protocols, particularly for histopathology and radiological images. The variety of data sources, differences in image acquisition methods, and the absence of consistent benchmark datasets lead to inconsistent progress. This review highlighted that some datasets are often incomplete, which results in models being trained on non-representative or imbalanced samples, thus limiting their generalizability to real-world clinical settings. Another critical challenge lies in the quality of the data itself. Imaging data can be affected by noise, artifacts, and patient movement, especially in radiological modalities such as CT and MRI. Histopathology images, though highly informative, suffer from artifacts such as tissue folding, tearing, or out-of-focus regions, which can degrade model performance. These limitations call for the development of robust preprocessing techniques to filter out artifacts and ensure that only high-quality data is fed into

deep learning models. Additionally, class imbalance remains a persistent issue across several studies in HNSC imaging. In many datasets, the minority classes, such as early-stage cancers or specific subtypes, are underrepresented, leading to biased model training. Addressing this issue requires class balancing techniques, such as oversampling of minority classes or the development of custom loss functions that penalize misclassifications of rare classes. AI's growing role in healthcare raises important ethical concerns, particularly around data privacy and algorithmic fairness. Patient data used to train AI models must be handled with extreme care, ensuring compliance with privacy regulations such as HIPAA and GDPR. Additionally, there is a need to ensure algorithmic transparency and fairness, as biases in training datasets can lead to unequal diagnostic outcomes, particularly for underrepresented populations. From a technical perspective, the computational power required for training deep learning models on medical images is substantial. Graphics processing units (GPUs) and tensor processing units (TPUs) are often necessary to handle the parallel computations required for large-scale datasets. However, not all institutions, particularly those in low-resource settings, have access to such high-end infrastructure. This raises concerns about the global applicability of AI solutions, as the models may not be feasible for smaller clinical centres without the necessary computational resources.

To fully capitalize on AI's potential in HNSC imaging, several steps must be taken. First, the standardization of imaging datasets is crucial. Future research should focus on creating publicly available, well-annotated, and balanced datasets that cover a wide range of cancer subtypes, stages, and treatment outcomes. These datasets should be representative of diverse patient populations to ensure that AI models generalize well to all clinical environments. In histopathology, standardizing slide preparation, staining protocols, and digitization techniques will help mitigate variability in image quality and improve the performance of deep-learning models. The use of multimodal data fusion is another promising avenue. Integrating data from multiple imaging modalities (e.g., combining radiological with histopathological images) along with clinical data (e.g., demographics, treatment history) could significantly enhance model performance. Such a holistic approach would provide a deeper understanding of tumor biology, improving both diagnostic accuracy and personalized treatment planning. Finally, developing explainable AI models will be critical to bridging the gap between AI and clinical practice. Models that can provide transparent reasoning for their predictions will foster greater trust among clinicians and lead to wider adoption of AI in healthcare. Techniques like attention mechanisms and visualization tools will enable models to highlight key features that influence their decisions, allowing clinicians to make more informed choices.

The experimentation using the TCGA-HNSC cohort involved processing gigapixel histopathology images into smaller tiles, which were then analyzed using a pre-trained Inception v3 model to extract 2048-dimensional feature vectors. These vectors were averaged for each patient and input into an attention-based CNN model for survival prediction. The model achieved a precision of 0.74 for "Survival" and 0.76 for "Not Survival", with recalls of 0.78 and 0.72, respectively. The F1-scores were 0.76 for "Survival" and 0.74 for "Not Survival", and the overall accuracy was 0.75. The importance

of histopathology images lies in their detailed microscopic level of tissue analysis, which provides critical insights into the phenotypic heterogeneity and cellular characteristics of tumors. Unlike other imaging techniques, such as CT or MRI, which offer broader structural views, histopathology images provide the underlying cellular and molecular features essential for accurate diagnosis and prognosis. This detailed information is pivotal for predicting patient survival and tailoring personalized treatment plans, ultimately enhancing clinical decision-making and patient outcomes.

8. Conclusions

This study highlights the transformative potential of deep learning and advanced imaging modalities in the diagnosis, analysis, and treatment of head and neck cancers. The integration of radiological, endoscopic, multimodal optical, hyperspectral, and histopathological imaging techniques, coupled with AI, has shown significant promise in improving diagnostic accuracy and enabling personalized treatment strategies. Deep learning models have demonstrated high precision in predicting patient outcomes, particularly through histopathology images that provide critical microscopic insights into tumor heterogeneity. The experimental findings on survival prediction underscore AI's potential to personalize care by identifying high-risk patients and adapting treatments accordingly. However, challenges remain, such as the need for standardized, well-annotated datasets, computational resources, and addressing ethical concerns around data privacy and algorithmic fairness. The global applicability of AI solutions also poses concerns, particularly in low-resource settings. Moving forward, developing explainable AI models and integrating multimodal data to improve diagnostic accuracy will be crucial for fostering clinical trust and maximizing the potential of AI in healthcare. By overcoming these challenges, AI has the potential to revolutionize HNC treatment, improve survival rates, and enhance patient outcomes.

Author Contributions

T.M. and A.A. contributed to the conceptualization and design of the study. T.M. performed the experiments, wrote the manuscript, and generated the figures. A.A. critically revised and edited the manuscript and provided supervision throughout the study. Both authors read and approved the final manuscript.

Funding

No funding was obtained for this study.

Institutional Review Board Statement

This article contains no studies with human participants or animals performed by any of the authors.

Informed Consent Statement

All participants provided informed consent prior to their inclusion in this study, ensuring their voluntary participation and confidentiality.

Data Availability Statement

The authors confirm that the data associated with the experiments can be made available upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

References

1. Barsouk, A.; Aluru, J.S.; Rawla, P.; et al. Epidemiology, Risk Factors, and Prevention of Head and Neck Squamous Cell Carcinoma. *Med. Sci.* **2023**, *11*, 42.
2. Giaquinto, A.N.; Sung, H.; Miller, K.D.; et al. Breast Cancer Statistics, 2022. *CA: Cancer J. Clin.* **2022**, *72*, 524–541.
3. Dai, X.; Park, J.H.; Yoo, S.; et al. Survival Analysis of Localized Prostate Cancer with Deep Learning. *Sci. Rep.*

2022, *12*, 17821.
4. Boucai, L.; Zafereo, M.; Cabanillas, M.E. Thyroid Cancer: A Review. *JAMA* **2024**, *331*, 425–435.
5. Cheng, L.; Albers, P.; Berney, D.M.; et al. Testicular Cancer. *Nat. Rev. Dis. Primers* **2018**, *4*, 29.
6. Song, X.; Zhao, Z.; Barber, B.; et al. Overall Survival in Patients with Metastatic Melanoma. *Curr. Med. Res. Opin.* **2015**, *31*, 987–991.
7. Tolstikhin, I.O.; Houlsby, N.; Kolesnikov, A.; et al. MLP-Mixer: An All-MLP Architecture for Vision. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 24261–24272.
8. Devlin, J.; Chang, M.-W.; Lee, K.; et al. Pre-training of Deep Bidirectional Transformers for Language Understanding. **2018**, arXiv preprint.
9. Boulesteix, A.-L.; Wright, M. Artificial Intelligence in Genomics. *Hum. Genet.* **2022**, *141*, 1449–1450.
10. Kumar, Y. Artificial Intelligence & Robotics–Synthetic Brain in Action. **2018**, SSRN 3325115.
11. Bhatia, S.; Sinha, Y.; Goel, L. Lung cancer detection: a deep learning approach. In *InSoft Computing for Problem Solving: SocProS 2017*. Springer: Singapore, 2019; Volume 2, pp. 699–705.
12. Guo, R.; Lu, G.; Qin, B.; et al. Ultrasound Imaging Technologies for Breast Cancer Detection and Management: A Review. *Ultrasound Med. Biol.* **2018**, *44*, 37–70.
13. Sailasya, G., Kumari, G.L. Analyzing the performance of stroke prediction using ML classification algorithms. *Int. J. Adv. Comput. Sci. Appl.* **2021**, *12*.
14. Sajid, S.; Hussain, S.; Sarwar, A. Brain Tumor Detection and Segmentation in MR Images Using Deep Learning. *Arabian J. Sci. Eng.* **2019**, *44*, 9249–9261.
15. Khalid, S.; Goldenberg, M.; Grantcharov, T.; et al. Evaluation of Deep Learning Models for Identifying Surgical Actions and Measuring Performance. *JAMA Netw. Open* **2020**, *3*, 201664.
16. Roy, M.; Kong, J.; Kashyap, S.; et al. Convolutional Autoencoder Based Model Histocae for Segmentation of Viable Tumor Regions in Liver Whole-Slide Images. *Sci. Rep.* **2011**, *11*, 139.
17. Chen, H.; Zhang, Y.; Zhang, W.; et al. Low-Dose CT via Convolutional Neural Network. *Biomed. Opt. Express* **2017**, *8*, 679–694.
18. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; et al. Generative Adversarial Networks. *Commun. ACM* **2020**, *63*, 139–144.
19. Michelutti, L.; Tel, A.; Zeppieri, M.; et al. Generative Adversarial Networks (GANs) in the Field of Head and Neck Surgery: Current Evidence and Prospects for the Future—A Systematic Review. *J. Clin. Med.* **2024**, *13*, 3556.
20. Son, J.; Park, S.J.; Jung, K.-H. Towards Accurate Segmentation of Retinal Vessels and the Optic Disc in Fundoscopic Images with Generative Adversarial Networks. *J. Digit. Imaging* **2019**, *32*, 499–512.
21. Wolterink, J.M.; Leiner, T.; Viergever, M.A.; et al. Generative Adversarial Networks for Noise Reduction in Low-Dose CT. *IEEE Trans. Med. Imaging* **2017**, *36*, 2536–2545.
22. Pan, Z.; Yu, W.; Yi, X.; et al. Recent progress on generative adversarial networks (GANs): A survey. *IEEE Access.* **2019**, *7*, 36322–36333.
23. Li, X.; Zhang, L.; Yang, J.; et al. Role of Artificial Intelligence in Medical Image Analysis: A Review of Current Trends and Future Directions. *J. Med. Biol. Eng.* **2024**, 1–3.
24. Atabansi, C.C.; Nie, J.; Liu, H.; et al. A survey of Transformer Applications for Histopathological Image Analysis: New Developments and Future Directions. *BioMed. Eng. OnLine* **2023**, *22*, 96.
25. Han, K.; Wang, Y.; Chen, H.; et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110.
26. Majeed, T.; Masoodi, T.A.; Macha, M.A., et al. Addressing data imbalance challenges in oral cavity histopathological whole slide images with advanced deep learning techniques. *Int. J. Syst. Assur. Eng. Manag.* **2024**, *26*, 1–9.
27. Kavyashree, C.; Vimala, H.S.; Shreyas, J. Improving oral cancer detection using pretrained model. In Proceedings of the 2022 IEEE 6th conference on information and communication technology (CICT), Gwalior, India, 18–20 November 2022; pp. 1–5.
28. You, Z.; Han, B.; Shi, Z.; et al. Vocal Cord Leukoplakia Classification Using Siamese Network Under Small Samples of White Light Endoscopy Images. *Otolaryngol. Head Neck Surg.* **2024**, *170*, 1099–1108.
29. Peng, J.; Zhou, Z.; Han, X.; et al. Multi-Level Fusion Graph Neural Network: Application to PET and CT Imaging for Risk Stratification of Head And Neck Cancer. *Biomed. Signal Process. Control* **2024**, *92*, 106137.
30. Al Ajmi, E.; Forghani, B.; Reinhold, C.; et al. Spectral Multi- Energy CT Texture Analysis with Machine Learning for Tissue Classification: An Investigation Using Classification of Benign Parotid Tumours as a Testing Paradigm. *Eur. Radiol.* **2018**, *28*, 2604–2611.

31. Ranjbar, S.; Ning, S.; Zwart, C.M.; et al. Computed Tomography-Based Texture Analysis to Determine Human Papillomavirus Status of Oropharyngeal Squamous Cell Carcinoma. *J. Comput. Assist. Tomogr.* **2018**, *42*, 299–305.
32. Wu, B.; Khong, P.-L.; Chan, T. Automatic Detection and Classification of Nasopharyngeal Carcinoma on PET/CT with Support Vector Machine. *Int. J. Comput. Assist. Radiol. Surg.* **2012**, *7*, 635–646.
33. Siebers, S.; Zenk, J.; Bozzato, A.; et al. Computer Aided Diagnosis of Parotid Gland Lesions Using Ultrasonic Multi-Feature Tissue Characterization. *Ultrasound Med. Biol.* **2010**, *36*, 1525–1534.
34. Huang, R., Zhou, Z., Wang, X., et al. Magnetic resonance imaging features on deep learning algorithm for the diagnosis of nasopharyngeal carcinoma. *Contrast Media Mol. Imaging* **2022**, 3790269.
35. Ramkumar, S.; Ranjbar, S.; Ning, S.; et al. Mri-Based Texture Analysis to Differentiate Sinonasal Squamous Cell Carcinoma from Inverted Papilloma. *Am. J. Neuroradiol.* **2017**, *38*, 1019–1025.
36. Li, C.; Jing, B.; Ke, L.; et al. Development and Validation of an Endoscopic Images-Based Deep Learning Model for Detection with Nasopharyngeal Malignancies. *Cancer Commun.* **2018**, *38*, 1–11.
37. Al-Ma'aitah, M.; AlZubi, A.A. Enhanced Computational Model for Gravitational Search Optimized Echo State Neural Networks Based Oral Cancer Detection. *J. Med. Syst.* **2018**, *42*, 1–7.
38. Moccia, S.; De Momi, E.; Guarnaschelli, M.; et al. Confident Texture-Based Laryngeal Tissue Classification for Early Stage Diagnosis Support. *J. Med. Imaging* **2017**, *4*, 034502.
39. Song, B.; Sunny, S.; Uthoff, R.D.; et al. Automatic Classification of Dual-Modalilty, Smartphone-Based Oral Dysplasia and Malignancy Images Using Deep Learning. *Biomed. Opt. Express* **2018**, *9*, 5318–5329.
40. Roblyer, D.; Kurachi, C.; Stepanek, V.; et al. Comparison of Multispectral Wide-Field Optical Imaging Modalities to Maximize Image Contrast for Objective Discrimination of Oral Neoplasia. *J. Biomed. Opt.* **2010**, *15*, 066017.
41. Quang, T.; Tran, E.Q.; Schwarz, R.A.; et al. Prospective Evaluation of Multimodal Optical Imaging with Automated Image Analysis to Detect Oral Neoplasia in Vivomultimodal Optical Imaging to Detect Oral Neoplasia In Vivo Cancer Prev. Res. **2017**, *10*, 563–570.
42. Kang, E.Y.-C.; Yeung, L.; Lee, Y.-L.; et al. A Multimodal Imaging–Based Deep Learning Model for Detecting Treatment-Requiring Retinal Vascular Diseases: Model Development and Validation Study. *JMIR Med. Inform.* **2021**, *9*, 28868.
43. Halicek, M.; Little, J.V.; Wang, X.; et al. Optical biopsy of head and neck cancer using hyperspectral imaging and convolutional neural networks. In *Optical Imaging, Therapeutics, and Advanced Technology in Head and Neck Surgery and Otolaryngology*; SPIE: Bellingham, WA, USA. 2018; Volume 10469, pp. 8–16.
44. Jeyaraj, P.R.; Samuel Nadar, E.R. Computer-Assisted Medical Image Classification for Early Diagnosis of Oral Cancer Employing Deep Learning Algorithm. *J. Cancer Res. Clin. Oncol.* **2019**, *145*, 829–837.
45. Ma, L.; Lu, G.; Wang, D.; et al. Adaptive Deep Learning for Head and Neck Cancer Detection Using Hyperspectral Imaging. *Vis. Comput. Ind. Biomed. Art* **2019**, *2*, 1–12.
46. Halicek, M.; Shahedi, M.; Little, J.V.; et al. Head and Neck Cancer Detection in Digitized Whole-Slide Histology Using Convolutional Neural Networks. *Sci. Rep.* **2019**, *9*, 14043.
47. He, Y.; Cheng, Y.; Huang, Z.; et al. A deep convolutional neural network-based method for laryngeal squamous cell carcinoma diagnosis. *Ann. Transl. Med.* **2021**, *9*.
48. Rahman, T.Y.; Mahanta, L.B.; Das, A.K.; et al. Automated Oral Squamous Cell Carcinoma Identification Using Shape, Texture and Color Features of Whole Image Strips. *Tissue Cell* **2020**, *63*, 101322.
49. Rahman, T.; Mahanta, L.; Chakraborty, C.; et al. Textural Pattern Classification for Oral Squamous Cell Carcinoma. *J. Microsc.* **2018**, *269*, 85–93.
50. Rodner, E.; Bocklitz, T.; Eggeling, F.; et al. Fully Convolutional Networks in Multimodal Nonlinear Microscopy Images for Automated Detection of Head and Neck Carcinoma: Pilot Study. *Head Neck* **2019**, *41*, 116–121.
51. Tang, H.; Li, G.; Liu, C.; et al. Diagnosis of Lymph Node Metastasis in Head and Neck Squamous Cell Carcinoma Using Deep Learning. *Laryngoscope Investig. Otolaryngol.* **2022**, *7*, 161–169.
52. Ariji, Y.; Fukuda, M.; Kise, Y.; et al. Contrast-Enhanced Computed Tomography Image Assessment of Cervical Lymph Node Metastasis in Patients with Oral Cancer by Using a Deep Learning System of Artificial Intelligence. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol.* **2019**, *127*, 458–463.
53. Abram, T.J.; Floriano, P.N.; James, R.; et al. Development of a Cytology-Based Multivariate Analytical Risk Index for Oral Cancer. *Oral Oncol.* **2019**, *92*, 6–11.
54. Tapak, L.; Shirmohammadi-Khorram, N.; Amini, P.; et al. Prediction of Survival and Metastasis in Breast Cancer Patients Using Machine Learning Classifiers. *Clin. Epidemiol. Glob. Health* **2019**, *7*, 293–299.
55. Zhang, Q.; Xiao, Y.; Dai, W.; et al. Deep Learning Based Classification of Breast Tumors with Shear-Wave Elastography. *Ultrasonics* **2016**, *72*, 150–157.

56. Chen, R.J.; Lu, M.Y.; Wang, J.; et al. Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis. *IEEE Trans. Med. Imaging* **2020**, *41*, 757–770.
57. Goode, A.; Gilbert, B.; Harkes, J.; et al. Openslide: A Vendor-Neutral Software Foundation for Digital Pathology. *J. Pathol. Inf.* **2013**, *4*, 27.