Deep Learning for Paranasal Anomaly Classification

Dissertation (cumulative) approved by the Doctoral Degree Committee of Hamburg University of Technology

in pursuit of the academic degree of

Doktor-Ingenieur (Dr.-Ing.)

written by

Debayan Bhattacharya

From

Shillong, India

2025

Reviewers:

1. Prof. Dr.-Ing. Alexander Schlaefer
2. Prof. Dr.-Ing. Tobias Knopp
3. Prof. Dr.-Med. Christian Stephan Betz

Date of Oral Examination:

11 February 2025

TABLE OF CONTENTS

ACKNOWLEDGMENT

This dissertation is the result of collective efforts, and I extend my sincere gratitude to those who have contributed to its success.

Foremost, I am deeply grateful to my advisor, Alexander Schlaefer, for his unwavering support, mentorship, and invaluable guidance in both research and academic writing. His encouragement has been instrumental in shaping my development as a researcher. I also sincerely thank my clinical advisor, Christian Betz, for providing essential clinical insights and direction. Special appreciation is due to Anna Sophie Hoffmann, whose clinical problem formed the foundation of this work, as well as to Dennis Eggert for his crucial assistance in refining my research focus and facilitating administrative matters.

I acknowledge my colleagues and collaborators, particularly Finn Behrendt, whose insightful discussions greatly enriched my research. I am also thankful to Sarah Latus, Lennart Maack, Maxmillian Neidhart, Marcel Bengs, Lennart Holstein, Stefan Gerlach, Robin Mieling, Sarah Grube, and Johanna Sprenger for their support, both academic and personal. I extend my gratitude to Michael Freude for IT support, Martin Fischer for prototype development, and Katrin Rausch for administrative assistance.

I express deep appreciation for my high school teachers, whose dedication fostered my passion for learning. I thank Martin Gromniak, who initially served as my project supervisor and later became a friend, for the many enriching discussions on deep learning. I am grateful to my friend Markynsai Lamar for his constant encouragement and to my brother Debanjan Bhattacharya and sister-in-law Soumee Bhattacharya for their steadfast support. Above all, I owe immense gratitude to my parents, Bhupesh Bhattacharya and Jana Bhattacharya, whose unwavering support has been the foundation of my journey. Finally, I am profoundly

thankful to my wife, Jayasmita Bhattacharjee, whose love and encouragement have been my greatest source of strength.

Abstract

The susceptibility of human sinuses, including the frontal, ethmoid, sphenoid, and maxillary sinuses, to both allergic and non-allergic infections poses a significant challenge in identification due to their varied manifestations—ranging from mucosal wall thickening to diverse polypoid masses visible in magnetic resonance imaging. Previous endeavors linking these incidental findings to underlying health and lifestyle factors, such as sex, smoking habits, and allergies, heavily relied on labor-intensive manual diagnoses. This manual approach, especially in prospective studies, leads to heightened resource consumption and contributes to clinician fatigue. The integration of computer-aided diagnosis (CAD) stands poised to alleviate these challenges, promising enhanced patient care while alleviating the workload burden on clinicians.

Deep learning research, burgeoning since 2010, initially showcased success in ImageNet classification before expanding into tasks like detection and segmentation. While medical imaging analysis benefited from this progress, research in classifying paranasal sinus anomalies has been relatively sparse. Early deep learning methods primarily centered on supervised learning, necessitating large annotated datasets—a challenge in medical imaging due to costly and limited access to labelled data requiring expert clinician supervision. Paradigms like unsupervised and self-supervised learning have been explored to augment supervised learning in medical imaging to address these challenges. Additionally, novel architectures such as vision transformers have emerged as alternatives to convolutional neural networks (CNNs). However, this progress has primarily focused on other medical imaging domains like chest

x-rays and brain anomalies, with limited emphasis on diagnosing paranasal anomalies.

This thesis endeavors to bridge this gap by proposing diverse deep learning methodologies to tackle maxillary sinus opacifications. Our approaches encompass unsupervised anomaly detection, utilizing its location information for a tailored self-supervised learning task. We introduce a novel hybrid architecture combining CNNs and transformers, showcasing its efficacy in classifying paranasal anomalies. Furthermore, we present techniques leveraging contrastive loss and multiple instance ensembling to bolster supervised learning performance. The culmination of these methods is employed as a CAD system, unraveling correlations with crucial clinical variables - showing real-world application.

In essence, this thesis aims to spotlight varied learning (supervised, self-supervised, unsupervised) and architectural strategies, along with data processing suggestions, particularly advantageous for paranasal anomaly classification in the maxillary sinus. Additionally, we shed light on challenges and limitations inherent in these proposed methods. Lastly, we emphasize the need for using our methods in other sinuses other than the maxillary sinus, exploring lightweight 2D and 2.5D architectures, and generating synthetic datasets for future research in this domain.

Chapter 1

Introduction

The nasal cavity, a cylindrical midline airway, extends from the anterior nasal ala to the posterior choana [101]. It is divided along the midline by the nasal septum. The cavity's superior boundary comprises the frontal, ethmoid, and sphenoid sinuses, and the maxillary sinuses form its lateral boundaries, arranged in an anterior-to-posterior orientation. This sinonasal structure represents a complex network of segmented airways and drainage pathways, enabling interconnectivity among the sinuses. Collectively, these four sinuses constitute the paranasal sinuses. Their primary functions include air humidification and facilitating the respiratory system's immune response [101]. Furthermore, these sinuses contribute to reducing the skull's weight, augmenting voice resonance, providing facial trauma protection, and insulating sensitive structures within the nose from rapid temperature variations [101]. However, these sinuses are prone to infections which can be fatal if not treated properly. Figure 1.1 shows the anatomy of the paranasal sinuses with elucidations of the four different sinuses.

One of the common ailments pertaining to the paranasal sinuses is inflammation. Commonly termed as sinusitis, it means the inflammation of the sinuses. Since the sinus passages are contiguous with the nasal passages, rhinosinusitis is often a more appropriate term. Its estimated incidence is 12.3% in the USA, 10.9% in Europe and 13% in China [50, 105, 57].

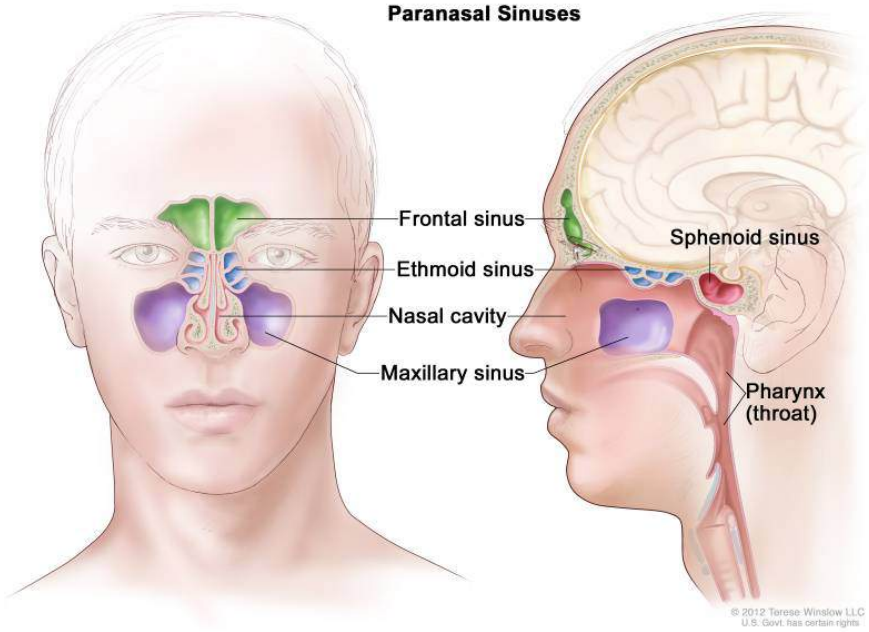

Figure 1.1 Anatomy of paranasal sinuses [16]

Rhinosinusitis can be classified into acute (lasting less than 4 weeks), subacute (lasting between 4 and 12 weeks), and chronic (lasting more than 12 weeks). Allergic rhinosinusitis contributes significantly to have an adverse impact on sleep, memory ability, quality of life, academic performance, and work productivity [117, 35, 123, 135, 17]. In one study it was estimated that acute rhinosinusitis results in 3.4 billion dollars of direct medical costs annually [90]. In another large-scale study, the cost in loss of productivity due to rhinosinusitis was estimated at $2.4 billion to $4.5 billion annually [31].

Rhinosinusitis manifests in the frontal, ethmoid, sphenoid and maxillary sinus in multiple ways. Allergic and non-allergic infections of the frontal sinus lead to inflammations and these manifests externally as facial swelling, erythema, and edema, particularly in the periorbital area and can cause deep tenderness upon percussion of the forehead and decreased visual acuity in cases of frontal sinus mucoceles [126]. Maxillary sinusitis can result in maxillary dental pain, abnormal sinus transillumination, poor response to nasal decongestants or antihistamines, and colored nasal discharge with mucopurulent secretions [39]. The sphenoid sinus drains posterior to the superior turbinate into the sphenoethmoid recess through the

sphenoid ostium. An obstruction in this region gives rise to the sphenoethmoid pattern of sinusitis [40]. Ethmoid sinusitis can cause pain between the eyes and tenderness when touching the bridge of the nose and may also lead to more eye-related symptoms due to its proximity to the eyes [88].

Sinusitis in adult and pediatric population for both acute and chronic rhinosinusitis is made clinically and then followed by imaging [136, 98]. Plain radiography may be used as a screening method for various pathological conditions of sinuses, but computed tomography (CT) remains the study of choice for the imaging evaluation of acute and chronic rhinosinusitis [98]. In acute sinusitis, CT is indicated in patients with symptoms persisting after 10 days of appropriate therapy and in patients with suspected complications (especially in the brain and in the orbit). In addition to CT scanning, magnetic resonance imaging (MRI) of the sinuses, orbits, and brain should be performed whenever extensive or multiple complications of sinusitis are suspected [98]. Imaging findings of acute sinusitis are non-specific and can be seen in a large number of asymptomatic patients. Some of the possible findings in acute sinusitis include mucosal thickening, air-fluid levels, and complete opacification of the involved sinus [102]. This dissertation primarily focuses on the detection of maxillary sinusitis. The maxillary sinusitis is common and numerous disorders can affect this anatomical area. Various categories of abnormalities of the maxillary sinus include non-neoplastic, benign neoplastic, and malignant neoplastic. Non-neoplastic conditions encompass inflammatory processes, infections, cysts, polyps, and mucoceles. Benign neoplasms consist of papillomas, fibro-osseous, and mesenchymal tumors, while malignant tumors affecting the maxillary sinus include Squamous Cell Carcinoma, adenocystic carcinoma, adenocarcinoma, and sarcomas [128]. Several studies seek to uncover links between these unexpected irregularities and the patient's health and lifestyle, aiming for a deeper understanding of their causes.

Over the years, there have been several studies to estimate the prevelance of paranasal incidental findings in large cohort of participants [49, 79, 3, 46, 112, 127, 127, 115, 30, 22]. The

main reasons behind these studies have been mostly that previous works have analysed the prevalence of paranasal incidental findings on those referred for diagnositic imaging however, there was no study on the prevelance of paranasal findings on the general population other than HUNT MRI study [49]. Further, studies have been performed to understand the effect of sex and seasonal variations towards paranasal sinusitis [131, 115]. A common element in all of these studies is that all these studies involved clinicians manually diagnosing large cohort of participant's CT or MRIs after which correlations with important clinical variables were calculated. While this results in accurate diagnosis, it also increases the fatigue of clinicians. Computer aided diagnosis (CAD) systems can be used here to expedite the diagnosis of paranasal sinusitis as well as decrease the workload of clinicians.

The main reasoning behind the rise of CAD systems is to improve the quality of a clinician's workflow. This could be through rapid diagnosis, localisation of regions of interest or provide a secondary opinion to aid in the diagnosis for a particular case. Recently, deep learning has emerged as a fundamental driver supporting the rise of CAD systems in various medical imaging applications [13, 139, 61, 20, 92]. Current deep learning based CAD systems mostly employ the supervised learning mechanism wherein an input and its corresponding label is present and the model is tasked to predict the label. This setup works properly as long as there is access to large labelled dataset. However, large datasets are particularly scarce in medical imaging [5] and labelling them leads to added burden to clinicians which CAD systems want to avoid in the first place. Hence, there have been alternate learning mechanisms which are not majorly dependent on labelled dataset but dependent on access to abundant unlabelled dataset. Unsupervised learning is a learning mechanism where a deep learning model attempts to learn structure of the data through reconstructing the an input image from a latent representation [114]. A key area where unsupervised learning has been particularly useful is anomaly localization with applications found in non-medical [130] and medical applications [9, 10]. A caveat of this setup is that it is dependent on the access to images belonging to one class for example in brain anomaly detection, the deep

learning model is trained only on healthy brain images making the unsupervised learning method a weakly supervised learning scheme. Nonetheless, the depenedence on labelled dataset is reduced because access to brain images with pathologies is not a requirement for this setup to work. Self-supervised learning has also emerged as another technique which attempts to learn the structure from unlabelled dataset with the caveat being that it learns primarily by constructing a pseudo label. For this setup, weakly supervised requirement is circumvented. Existing self-supervised methods rely on non-linear compression [108, 144], denoising [19], and the alignment of features of randomly augmented versions of the same image [48, 26, 59, 25]. While these have shown promising results, they have mostly been developed with the downstream task of ImageNet [34] classification. Self-supervision tasks, designed with the downstream task in mind, can improve the acquisition of more transferable representations [97].

Considering these factors, this dissertation primarily focuses on classifying maxillary sinus anomalies downstream using a deep learning-based CAD system. The necessity originates from the Hamburg City Health Study (HCHS), a comprehensive population study involving non-selected participants [63] where automated diagnosis is required. The anomalies of interest exclusively pertain to non-neoplastic conditions. Classifying healthy and anomalous maxillary sinuses poses a significant challenge for deep learning-based CAD systems for two primary reasons. First, anomalies like mucosal thickening, polyps, and cysts in the maxillary sinus display morphological variations [49, 79], complicating the development of CAD systems susceptible to overfitting specific anomaly types. Second, the maxillary sinus exhibits anatomical variations [1, 124, 94], making it difficult to generalize to unseen maxillary sinuses. Third, confronted with a substantial cohort, our available resources consist of a restricted labeled dataset and an extensive unlabeled dataset. Consequently, our imperative lies in the exploration of optimal strategies to extract superior representations from the labeled dataset. Simultaneously, we aim to probe the effective utilization of the vast unlabeled dataset to enhance performance on the labeled dataset. Addressing these

challenges, this dissertation explores diverse deep learning techniques to aid in diagnosing paranasal anomalies in the maxillary sinus. Moreover, it proposes methodologies to enhance generalization to test sets comprising unknown maxillary sinuses and various anomaly distributions using labelled and unlabelled dataset, aiming to automate the diagnosis process in population studies for expedited health monitoring.

1.0.1 Primary Contributions

In this work we propose several deep learning methods that explore learning mechanisms, architectural changes and data processing techniques that are particularly beneficial for classifying anomalies in the maxillary sinus.

Unsupervised anomaly detection

Unsupervised anomaly detection (UAD) is a technique which uses so-called healthy samples and effectively learns its distribution using an autoencoder[114]. This method proves advantageous in scenarios with limited labeled datasets, where modelling the distribution of healthy/normal samples is comparatively more tractable than modelling the distribution of anomalous samples. The two main benefits of this method are as follows: (1) It is not required to acquire samples from the anomaly or pathology class thereby saving labelling effort and not requiring to learning the diverse anomalous maxillary sinus distribution. (2) A benefit of using this method is that the reconstruction error provides a coarse location of anomalies when inferring. We investigate the feasibility of using UAD as a technique to classify between normal and anomalous maxillary sinuses by training a convolutional autoencoder (CAE) on normal maxillary sinus [13]. We also qualitatively describe how it can be used to effectively localise anomalies using the reconstruction error computed during inference.

Self-supervised learning

Self-supervised learning utilizes unlabeled data to develop effective representation techniques, where 'representations' refer to the features extracted by neural networks. Effectively, we investigate how to properly leverage the unlabelled dataset to improve the classification performance. The efficacy of these representations hinges on both the self-supervised learning task and the downstream task [104]. We propose that anomaly localization can serve as a potent self-supervision task. Given the need for voxel-level annotation of anomalies in the maxillary sinus, we adopt a more feasible approach using approximate localization information derived from training CAE within the UAD framework. This study examines the impact of pretraining a CNN on reconstructing residual maxillary sinus volumes, and its subsequent effect on the performance of downstream classification tasks [14].

Improving supervised learning using deep learning architectural modifications

Vision transformers (ViT) have surpassed Convolutional Neural Networks (CNNs) as the leading models in computer vision tasks, including classification [86] and segmentation [80] in medical imaging, as noted by Dosovitskiy et al. [38]. Despite their advanced modeling capabilities, ViTs are hindered by a significant demand for data, especially when trained from scratch, and require substantial computational resources. These drawbacks are exacerbated in scenarios with limited datasets and computational capacity. To address these challenges, we propose a hybrid architecture that integrates convolutional and transformer blocks. This design capitalizes on the benefits of inductive bias and facilitates the creation of long-range dependencies through self-attention. We introduce a convolutional transformer network tailored for paranasal anomaly classification in the maxillary sinus [**?**]. Our research investigates how this hybrid architecture mitigates the limitations of ViT and potentially exceeds the classification efficacy of standalone CNNs and ViT models.

Improving supervised learning using contrastive loss

In our study, which focuses on classifying normal versus anomalous maxillary sinus volumes, the prevalent approach is to train CNNs using cross-entropy loss. While effective in learning discriminative features, cross-entropy loss can result in narrow margins between different classes in the training set, potentially compromising robustness to unseen test samples [96]. Supervised contrastive loss, by contrast, enhances label differentiation by ensuring that normalized features of the same class are more closely clustered than those of different classes [68]. This approach promotes dense clustering of same-class samples in feature space. Our research evaluates the effectiveness of both cross-entropy loss and supervised contrastive loss in classifying paranasal anomalies [11]. Furthermore, we introduce a novel loss function that merges cross-entropy and supervised contrastive loss, leveraging the distinct advantages of both. This study also demonstrates how this combined loss approach contributes to enhanced label efficiency.

Improving supervised learning through data processing and ensembling

Current methodologies for analyzing paranasal sinuses in existing literature typically adopt a two-stage process involving the localization of the sinus followed by its classification via deep learning models [70, 103, 69]. These methods predominantly rely on deep learning for sinus localization, which inherently depends on the dataset used. This dependence necessitates specialized annotations, increasing the workload for clinicians. To address these limitations, we introduce an alternative approach that does not depend on deep learning. Our method models the centroids of the left and right maxillary sinuses using Gaussian distributions, enabling us to sample centroid points for cropping the sinuses from larger cranial MRIs. Additionally, we propose a technique to extract multiple maxillary sinus volume candidates from each patient's MRI, thereby augmenting our training dataset. During inference, we extract several maxillary sinus instances from each patient's MRI, compute predictions for

each instance, and then apply an ensemble method. This approach, which we term Multiple Instance Ensembling (MIE), demonstrates enhanced classification performance across various deep learning architectures, as evidenced by our rigorous experimental evaluations [15].

Clinical application

Research at both population [49] and smaller scales has been conducted to investigate the influence of sex and seasonal variations on paranasal sinusitis [131, 115]. A recurring aspect of these studies is the reliance on clinicians to manually diagnose large cohorts using CT or MRI scans, followed by correlational analyses with key clinical variables. Although this method ensures accurate diagnoses, it significantly contributes to clinician fatigue. To mitigate this, we propose the implementation of a CAD system utilizing the MIE approach [12]. We enhance this system by ensembling models trained on varied proportions of the dataset, divided via cross-validation. Our approach uncovers statistically significant correlations between participants with maxillary sinus anomalies and relevant clinical variables, such as sex and smoking habits. This study demonstrates how deep learning can streamline population-level studies, greatly enhancing the efficiency of clinician-driven diagnostic processes.

1.0.2 Outline

The dissertation is organized as follows: Chapter 2 presents an overview of related work in Unsupervised Anomaly Detection (UAD), self-supervised learning, supervised learning, and deep learning applications for paranasal anomalies. Chapter 3 formalizes the components of the methodologies investigated for the classification of paranasal anomalies. In Chapter 4, we elucidate the key results obtained from these methods. Chapter 5 offers a summary, highlights crucial considerations for the classification of paranasal anomalies and provides possible avenues for future work. The appendix includes relevant publications that support

the content of this dissertation.

Chapter 2

Background

2.1 Brief history on unsupervised learning and unsupervised anomaly detection

Unsupervised learning is a critical paradigm in machine learning where algorithms discern patterns from unlabelled data. This approach is epitomized by clustering methods, such as the K-means clustering algorithm, which derive labels directly from the data [47]. K-means clustering exemplifies a non-parametric unsupervised learning technique. In contrast, parametric models like neural networks were first applied to unsupervised learning by Schmidhuber [120]. In his seminal work, Schmidhuber introduced the use of a recurrent neural network (RNN) to predict subsequent inputs in a process termed the 'pretext task'. This concept laid the groundwork for self-supervised pretraining, now a cornerstone in the development of large language models [74]. Schmidhuber's theory posited that pretext task pretraining without labels fosters the development of internal representations beneficial for subsequent supervised learning tasks.

With the advent of deep learning, marked by breakthroughs like AlexNet [75], unsupervised learning has found extensive applications particularly through CNNs [27]. A notable

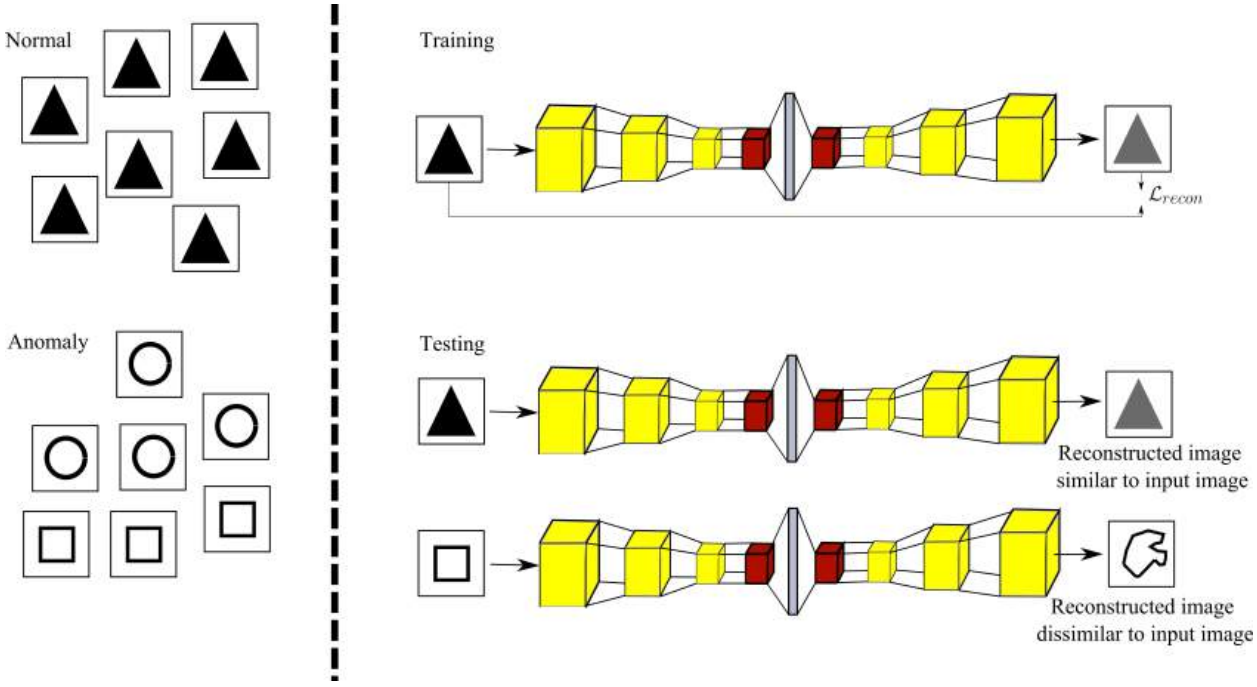

Figure 2.1 (LEFT) Images with triangles represent the normal class whereas images with circles and squares represents the anomaly class. (RIGHT) An autoencoder is trained to reconstruct the image in the training step using $\mathcal{L}_{recon}$. Once trained, the autoencoder is used for testing. It can mostly reconstruct the images with triangles but fails to reconstruct the images with square.

use of unsupervised learning is in unsupervised anomaly detection (UAD). UAD typically assumes a dataset predominantly comprising samples from a 'normal' or 'healthy' class, interspersed with anomalies. If the dataset contains only 'normal' samples, the approach is referred to as 'weakly supervised' or 'semi-supervised' anomaly detection, although in some prior works it is also called *unsupervised* anomaly detection [84]. A fundamental neural network used in UAD is an autoencoder [87, 56]. An autoencoder, comprising an encoder and a decoder, compresses input images and subsequently attempts to reconstruct them. This process enables the autoencoder to learn a dense representation of the image distribution, primarily focusing on the 'normal' class. During inference, it is hypothesized that an autoencoder, when confronted with an anomalous sample, will fail to accurately reconstruct the anomalous aspects of the image.

The process of detecting anomalous samples in UAD hinges on a well-designed scor-

ing function. This function is pivotal in distinguishing between representation-based and reconstruction-based UAD methodologies. In the representation-based approach, deep neural networks are utilized to distill entire images into informative vectors. The anomaly score is then determined by the distance metric between the embedded vectors from test images and the reference vectors, which represent normality, derived from the training dataset. This technique, leveraging embedded vectors for anomaly detection, has been effectively applied in the industrial sector [32].

In contrast, reconstruction-based UAD employs a scoring function that typically calculates the mean reconstruction error. Here, samples with a mean reconstruction error exceeding a specific threshold are identified as anomalies. While this method can differentiate between normal and anomalous samples, its primary utility lies in localizing anomalies. An illustration of UAD is shown in figure 2.1 The underlying assumption is that areas with elevated reconstruction error are indicative of anomalies. Consequently, this approach leverages *normal* data to implicitly generate cost-effective segmentation maps that localize the anomalies, a strategy extensively studied in the context of unsupervised brain anomaly detection [7].

In the domain of brain anomaly detection, the repertoire of architectures extends beyond autoencoders. Innovations include variational autoencoders (VAE) [24], generative adversarial networks (GAN) [64], denoising autoencoders [67], transformers [110], and diffusion models [8, 109]. All these methods share a common objective: to learn representations of healthy brain MRI. The expectation is that upon encountering an unhealthy brain MRI, these networks will notably struggle to reconstruct the anomalous regions. It is important to highlight that a significant advantage of UAD lies in its ability to produce anomaly maps, which effectively localize the anomalies. While the field of UAD has seen extensive research and development in the context of brain anomaly detection, its application in the area of paranasal anomaly detection remains relatively underexplored. This disparity in research focus highlights a significant opportunity for the advancement of UAD techniques in the

diagnosis and analysis of paranasal anomalies.

2.2 Brief history on self-supervised learning

Self-supervised learning (SSL) constitutes a machine learning paradigm where neural networks extract structures and representations from unlabelled datasets. SSL, a subset of unsupervised learning, necessitates the creation of pseudo labels to train neural networks. Notably in deep learning, SSL has excelled in natural language processing by engaging in tasks like guessing masked words in unlabelled text, known as the *pretext* task [18, 111]. In non-medical computer vision, SSL methods have outperformed supervised learning models on competitive datasets like ImageNet [34, 133, 53]. These methods have also found application across diverse modalities such as time series, audio, and video [142, 81, 119].

In medical imaging, SSL has gained widespread adoption owing to the amplified challenge of acquiring labelled medical data compared to general computer vision tasks. Annotating medical data demands expert knowledge and substantial time investment. For instance, in the CheXpert dataset, labelling chest X-ray images consumed an estimated 2-5 minutes per study, whereas ImageNet's image samples were labelled at an average rate of 50 images per minute [62, 118].

SSL strategies in constructing the *pretext* task can be broadly categorized based on their employed methodologies.

2.2.1 Spatial Relationship

The *pretext* task in such scenarios aims to comprehend the intrinsic relationships within the data without requiring additional labels. It encompasses learning through classification or regression loss tasks. Pretraining models on such custom-designed tasks facilitates learning visual features adept at solving the given task, albeit with potential limitations for downstream tasks. Spatial relationship mining tasks, such as predicting rotation angles [43],

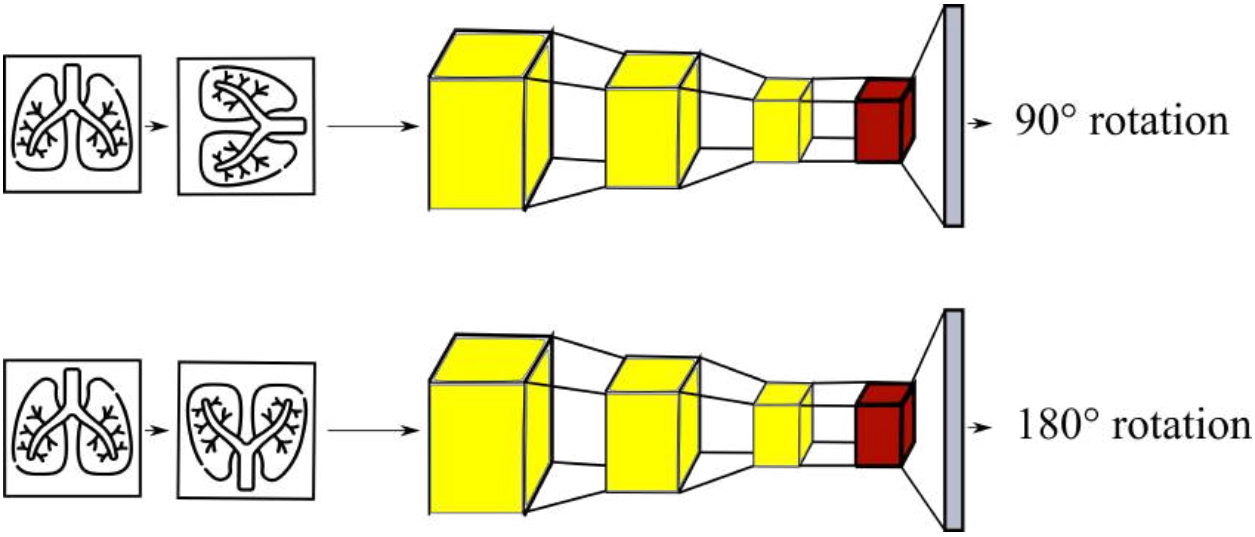

Figure 2.2 Example of predicting the rotation angle as an SSL task

solving image jigsaw puzzles [99], or ascertaining relative positions of image patches [37], exemplify these methodologies. These tasks have also found application in medical imaging [129]. Figure 2.3 shows an illustration of predicting the rotation of an image as a SSL task.

2.2.2 Feature alignment via contrastive learning tasks

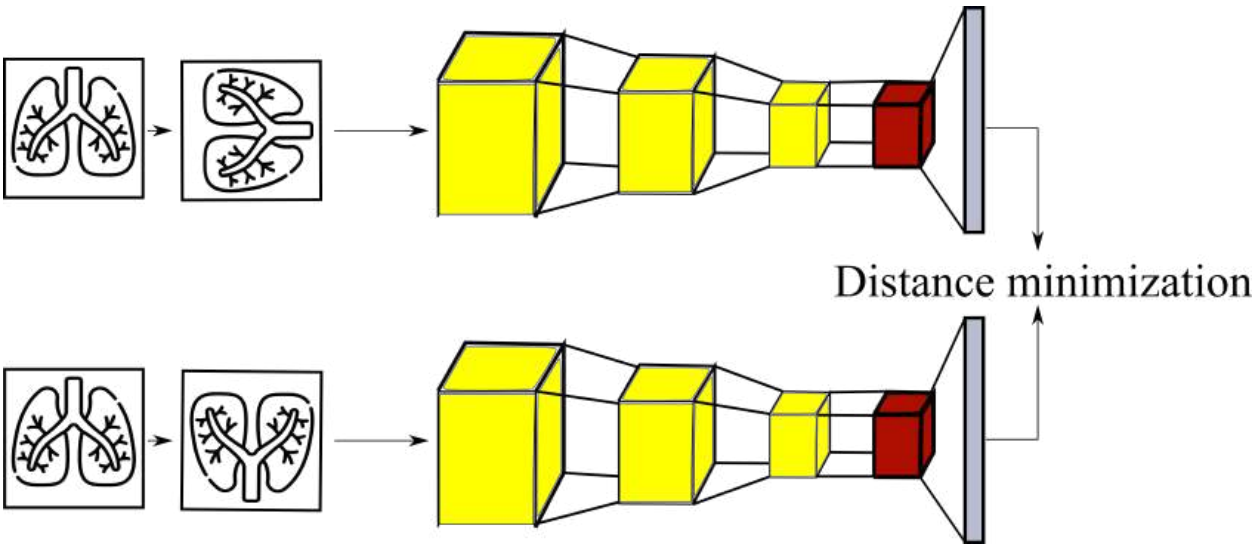

Figure 2.3 Illustration of constrastive pretraining where augmented versions of the image which carry the same semantic meaning extract features which are similar.

Self-supervised methods based on feature alignment via contrastive learning tasks operate under the premise that augmenting an image retains its semantic meaning, ensuring similar neural network-produced features. In contrastive learning, positive and negative pairs are defined: positive pairs consist of an image and its augmented view, while other images or their augmented versions serve as negative pairs in relation to a given image. During training, features from positive pairs converge in the latent space, while those from negative pairs diverge. A chosen distance metric aligns positive pair features and separates negative pair features. SimCLR [25], a popular contrastive learning method, merges features of an image and its augmented version while pushing apart features of negative pairs using cosine similarity. This method utilizes the InfoNCE loss [137]. However, its effectiveness often requires a large batch size. Momentum Contrast (MoCo) [53] mitigates this need by incorporating an encoded queue to manage negative samples. Other contrastive self-supervised methods like DINO [23], BYOL [48], and SimSiam [26] reduce reliance on negative pairs, concentrating solely on pulling features of positive pairs together. Contrastive self-supervised techniques have found extensive use in medical imaging, spanning vision and language tasks [28], disease detection [143, 29], and segmentation [125].

2.2.3 Generative tasks

Generative models, exemplified by autoencoders (AE) [121], VAE [72], GAN [45], and diffusion models [58], grasp the underlying distribution of training data by reconstructing input images or generating synthetic instances. By mastering the reconstruction of unlabelled data, these models acquire effective latent representations, aligning with the goals of SSL tasks. Autoencoders, for instance, excel in learning representations by denoising input images [138, 44]. VAEs enhance latent spaces for subsequent tasks [41], while GANs adeptly learn to fill missing data gaps via self-supervised collaborative learning [21]. Self-supervised diffusion models have found application in anomaly segmentation [76].

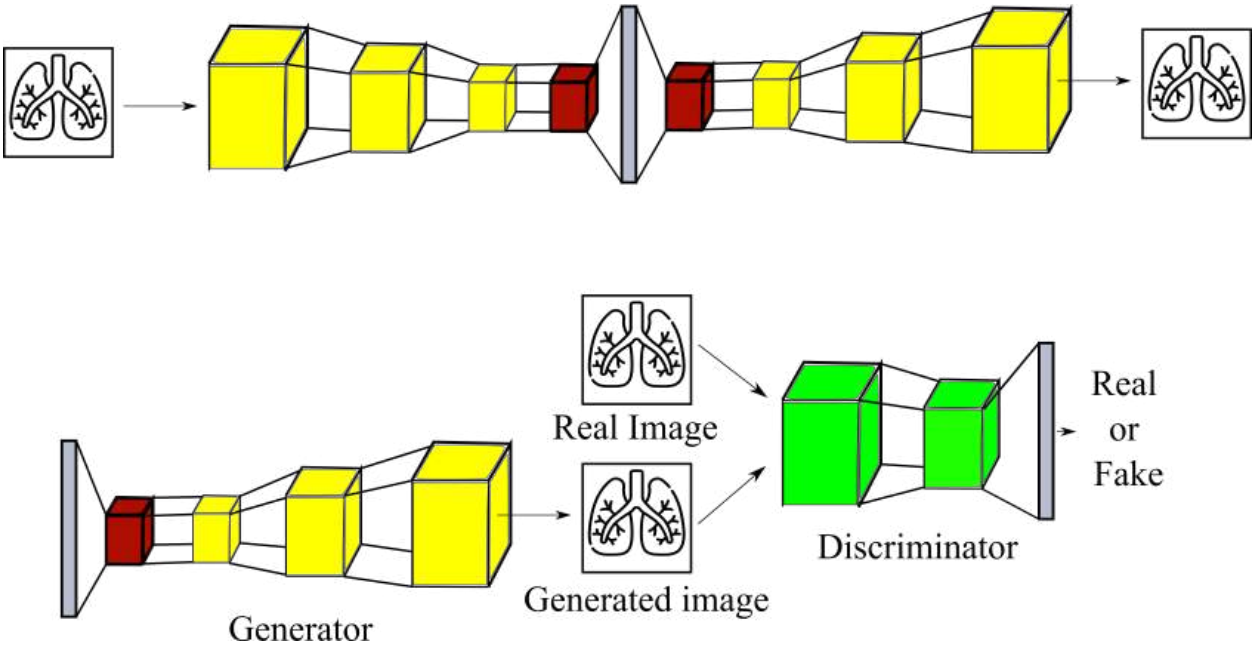

Figure 2.4 (TOP) A generative pretraining using an autoencoder (BOTTOM) Illustration of Generative Adversarial Network

2.2.4 Masking and inpainting

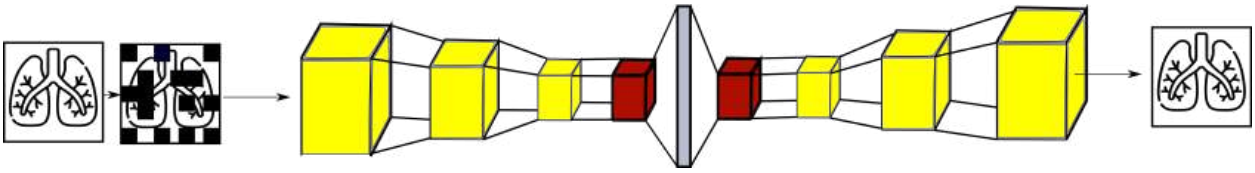

Figure 2.5 Illustration of a masked autoencoding where a neural network is required to inpaint the masked regions shown with black rectangles. Typical neural network used for this task is a Vision Transformer.

The concept of masking and inpainting originated in Natural Language Processing (NLP), where cutting-edge models underwent pre-training via Masked Language Modeling, predicting missing words in sentences [120, 36]. This approach transitioned into training CNNs, where instead of masking words, patches of images were masked, tasking CNNs with inpainting these regions [107]. Vision Transformers (ViT) embraced this technique, employing it as a Masked Autoencoder to reconstruct masked images, often termed as masked image modeling. This strategy proved remarkably successful, achieving state-of-the-art perfor-

mance across various natural image benchmarks [52, 6]. Masked image modeling (MAE) has demonstrated efficacy in diverse medical imaging analyses, including 3D image classification and segmentation [28, 80]. Although MAE pretraining is popular for ViT, recently MAE has been adapted to CNNs using masKed modeling with hierarchy (SparK) [132]. This approach leverages sparse convolutions to enable CNNs to compute solely on unmasked regions, preventing information leakage and seamlessly integrating with any CNN without modifying its backbone. Sparse convolutions efficiently compute only at visible locations, addressing concerns like "pixel distribution shift" and "mask pattern vanishing.".

2.2.5 Training strategy

The aforementioned methods serve as the model's pretraining phase. Once the model is pretrained based on our specified SSL task, we proceed to the fine-tuning stage. Here, supervised training occurs using a labeled dataset, typically involving unfreezing and end-to-end training of the encoder and decoder weights. While akin to transfer learning, a drawback of the latter lies in deriving weights from natural image training, rendering features less suitable for medical imaging tasks. Consequently, self-supervised methods have surfaced, enabling training on unlabeled medical imaging datasets. As a result, weights acquired through the SSL task prove more conducive to supervised fine-tuning, given that both the unlabeled and labeled datasets stem from the same distribution.

2.3 Supervised learning and strategies to boost supervised learning

Despite the proliferation of unsupervised and SSL methodologies, supervised learning remains a foundational approach that consistently yields superior task performance. This methodology involves training a model to output corresponding values when presented with

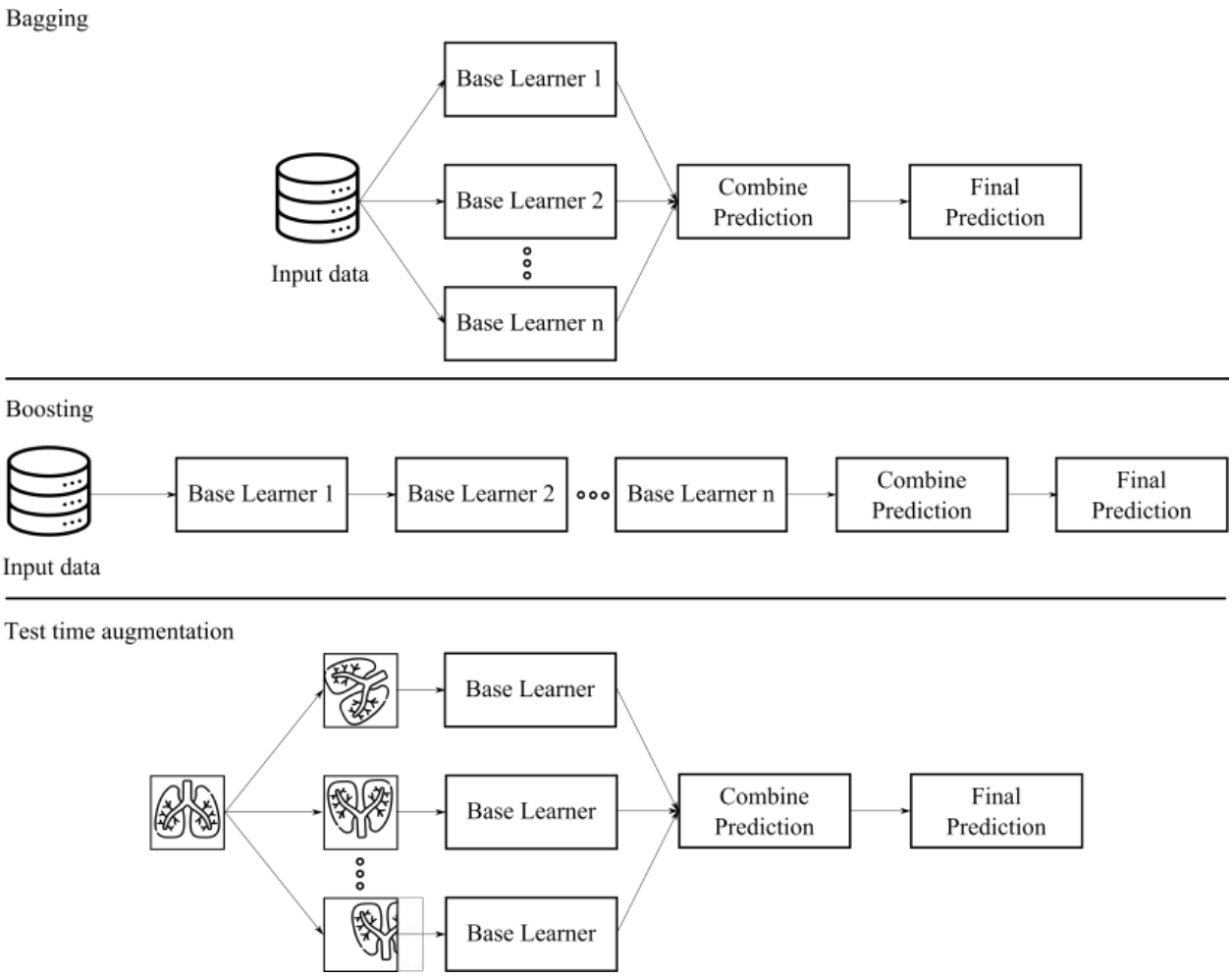

Figure 2.6 (TOP) Illustration of bagging ensemble technique (MIDDLE) Illustration of boosting ensemble technique (BOTTOM) Illustration of test-time augmentation

input images. Effective execution of supervised learning hinges upon the availability of datasets containing paired inputs and outputs. Supervised learning can be categorized primarily into two streams: classification, where the task involves mapping input variables to discrete output variables, and regression, a predictive modeling task mapping input variables to continuous output variables.

The realm of supervised learning has found extensive utility in medical imaging, encompassing applications like brain anomaly detection [83], breast cancer classification [141, 89], cervical and oral cancer classification [145, 4], lung cancer classification [78, 113], and skin cancer classification [2, 77]. However, a well-established concern exists regarding the propen-

sity of deep learning models to exhibit excessive confidence and susceptibility to classification and regression errors, constraining their reliability in safety-critical domains such as medical imaging [93].

To mitigate these challenges, a commonly employed strategy is ensemble learning. This paradigm involves training multiple 'base' learners individually and combining their predictions to enhance performance and generalizability. Ensemble learning addresses the high variance and bias issues inherent in machine learning algorithms, particularly when dealing with imbalanced or small datasets [91]. By combining multiple models, ensemble methods effectively mitigate variance and bias errors associated with individual models.

In deep learning, especially in image classification and regression, two prevalent ensemble techniques are utilized: bagging and boosting. Bagging involves training models independently in parallel and subsequently amalgamating their predictions to reduce variance. Bagging has shown efficacy in improving chest X-ray predictions [100] and skin cancer classification [42]. Conversely, boosting constructs models where each model feeds its output to the next, aiming to reduce bias. Boosting techniques have been successful in breast cancer classification [146]. Beyond boosting and bagging, a recent emerging approach is ensemble via augmentation, also known as test-time augmentation. This technique involves augmenting an image multiple times and feeding these augmented versions into the machine learning model. The predictions from all augmented versions are aggregated to produce a final prediction. This methodology fosters robust predictions by averaging across different augmentations of a single image. Test-time augmentation has shown promise in colorectal polyp segmentation [66]. Illustrations of bagging, boosting and test-time augmentation is shown in figure 2.6.

In the domain of paranasal anomaly classification, ensemble techniques remain unexplored, presenting an opportunity for further investigation and application.

2.4 Deep learning in paranasal sinus disease classification

Deep learning has played a pivotal role in numerous medical imaging tasks, notably in the detection of paranasal sinus anomalies. The comprehensive task of identifying anomalies within the paranasal sinuses is typically tackled in a two-stage process. Initially, localization of the sinuses is performed on the radiological image, followed by classification of the localized area. For instance, Jeon *et al.* [65] conducted sinusitis screening using deep learning on 1535 patients with Waters' and Caldwell view radiography. Their deep learning model was utilized to diagnose frontal, ethmoid, and maxillary sinusitis on both Waters' and Caldwell views. They categorized images where both the left and right sinuses were visible into either sinusitis or no sinusitis. Although this approach reduced the number of images required per patient, it lacked the granularity to classify whether the left, right, or both sinuses were affected by sinusitis. The schematic representation of their pipeline is elucidated in Figure 2.7.

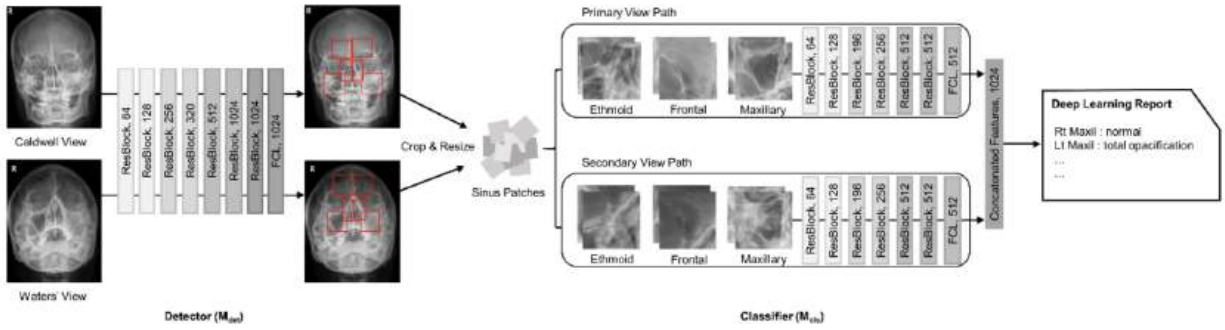

Figure 2.7 Classification of sinusitis of paranasal sinuses from Waters' and Caldwell radiography using a two step process of localisation followed by classification [65]

In a separate study, Murata *et al.* [95] conducted an analysis using a deep learning model on cone beam computed tomography (CBCT) images, extracting regions of interest that displayed both maxillary sinuses from a dataset comprising 490 patients. The initial step involved manual localization of the paranasal sinuses in the CBCT. Subsequently, a deep neural network was trained using images containing both left and right maxillary sinuses to execute binary classification, distinguishing between healthy and inflamed maxillary sinuses.

Furthermore, the application of GANs was explored. Conditional GANs were employed to generate synthetic maxillary sinus views with and without sinusitis [73]. In this study, the authors utilized an internal dataset of 250 patients along with an external dataset of 105 patients. They employed conventional data augmentation techniques on existing data samples and generated new data samples using GANs. This strategy aimed to enhance the training dataset, consequently improving the classification performance. An overview of their proposed methodology is depicted in Figure 2.8.

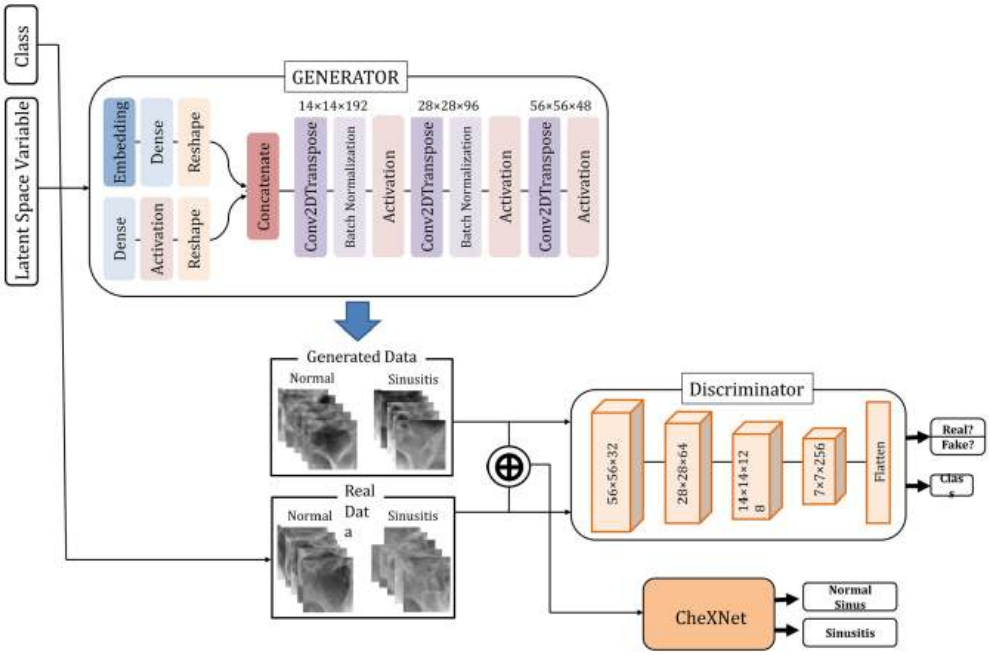

Figure 2.8 GAN is used to generate synthetic images of maxillary sinus to increase the training dataset size [73]

In a different study focused on chronic rhinosinusitis (CRS), a condition known for its poor prognosis and tendency for recurrence, a multi-task deep learning network was proposed. This network addressed sinus segmentation and CRS recurrence prediction simultaneously

[55]. The study involved a dataset comprising 265 paranasal sinus computed tomography (CT) images. The deep neural network exhibited satisfactory performance in both sinus segmentation and recurrence prediction. Moreover, the automation of volumetric analysis for paranasal sinus inflammations has been explored using deep neural networks. Humphries *et al.* [60] employed a CNN to segment paranasal sinuses on CT images, enabling volumetric quantification of sinonasal inflammations in a dataset encompassing 690 patients. Figure 2.9 visually illustrates the CNN's outcomes, showcasing air-filled areas and opacifications segmented in green and red colors, respectively.

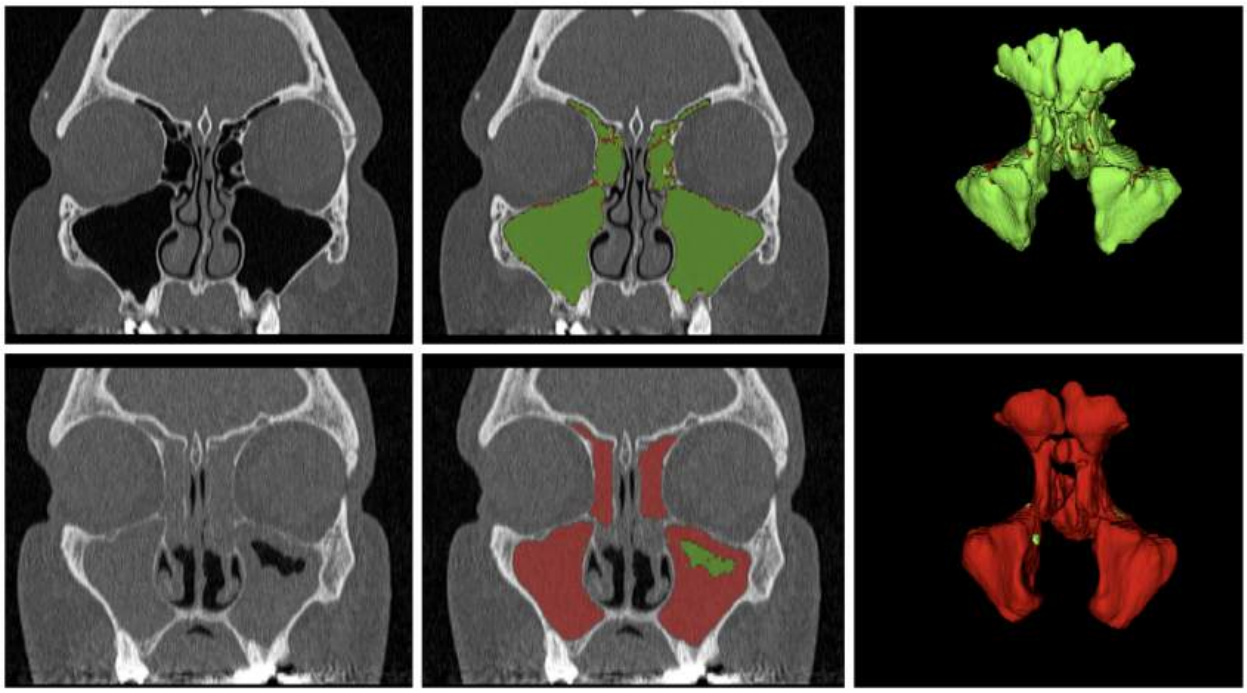

Figure 2.9 Automatic segmentation results generated by the CNN with air (green) and opacifications (red). Left to right: original CT image, image with overal of CNN segmentation: 3 dimensional surface rendering of sinus segmentation [60]

In a separate study utilizing MRI images, a deep learning network was trained to differentiate between inverted papilloma tumors and tumors that had transformed into squamous cell carcinoma [81]. The study employed MRI data from 90 patients and utilized a 3D CNN for classification. Similarly, in another study, a deep neural network was employed to distinguish between maxillary sinus fungal ball, CRS, and healthy controls using CT images. The

study utilized 512 patients in the internal dataset and 64 patients in the external dataset. Employing a two-stage approach, the first stage involved training a 2D CNN to identify key slices containing the maxillary sinus, as not all CT image slices encompassed this area. In the subsequent step, the identified key slices were stacked, cropped, and divided into two parts. The left and right maxillary sinuses were separately processed through a 3D CNN for classification into maxillary sinus fungal ball, CRS, and healthy controls for each sinus in a single patient. Figure 2.10 provides an overview of their data processing pipeline.

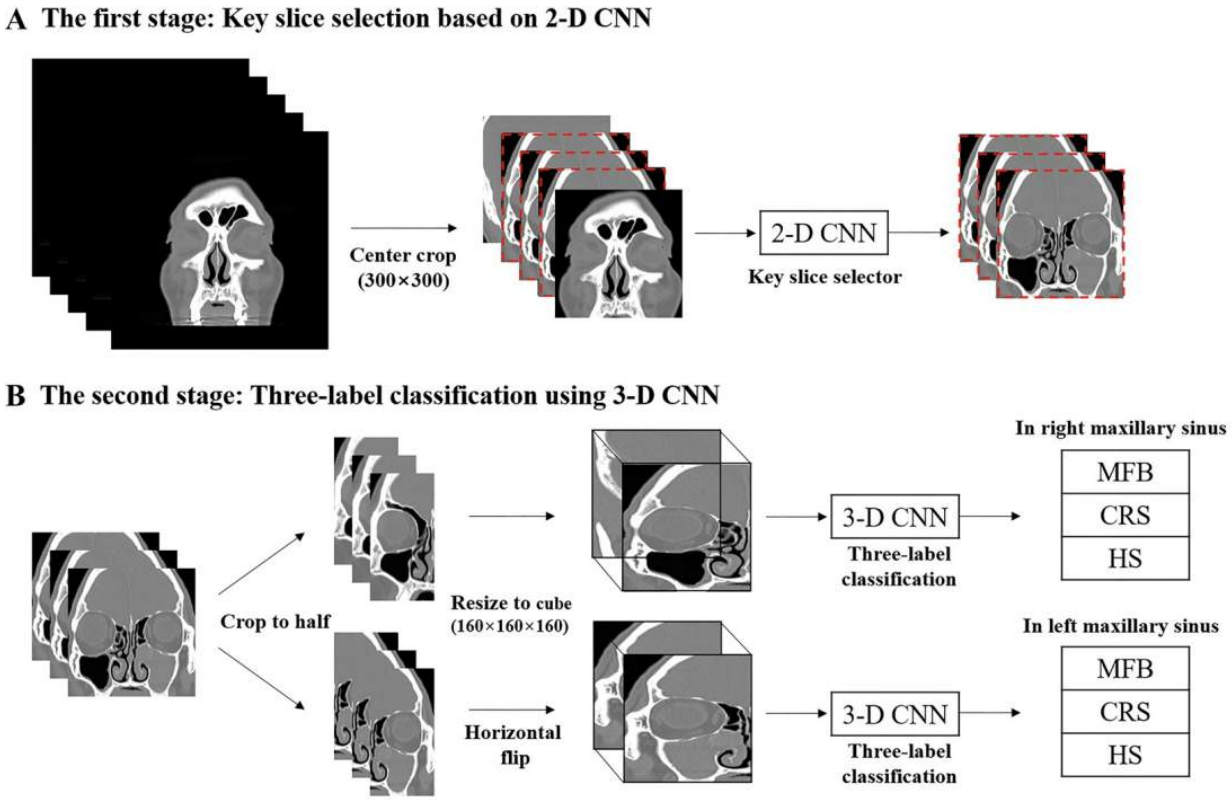

Figure 2.10 Overview of [71]. In the first stage, key slices are extracted using 2D CNN. In the second stage, disease classification was performed by a 3D CNN by processing a 3D stack of only key CT slices as input.

Chapter 3

Methods

3.1 Dataset

The Hamburg City Health Study (HCHS) [63] stands as a single-center, prospective, population-based cohort investigation. Samples images of MRIs from the study are shown in figure 3.1. Out of the intended 45,000 participants, a subset of 2,619 individuals underwent cranial MRI scans. These MRI scans were conducted between February 08, 2016, and November 30, 2018, specifically on individuals aged between 45 and 74 years. The imaging was performed using a 3-T Siemens Skyra MRI scanner (Siemens, Erlangen, Germany). The 3D T2-weighted fluid attenuated inversion recovery (FLAIR) images were captured with the following sequence parameters: TR = 4700 ms, TE = 392 ms, 192 axial slices, slice thickness (ST) = 0.9 mm, and in-plane resolution (IPR) = 0.75 $\times$ 0.75 mm. The resolution of each MRI scan was 173 mm x 319 mm x 319 mm along the sagittal, coronal, and axial directions, respectively, with a voxel size of 0.53 mm x 0.75 mm x 0.75 mm.

The participant data encompassed laboratory measurements of leukocytes/µL (LK) and high-sensitivity CRP (hCRP). Additionally, self-reported questionnaires were used to document alcohol consumption per day, smoking habits, and diagnoses of chronic bronchitis or chronic obstructive pulmonary disease (COPD) and allergic bronchial asthma. Other

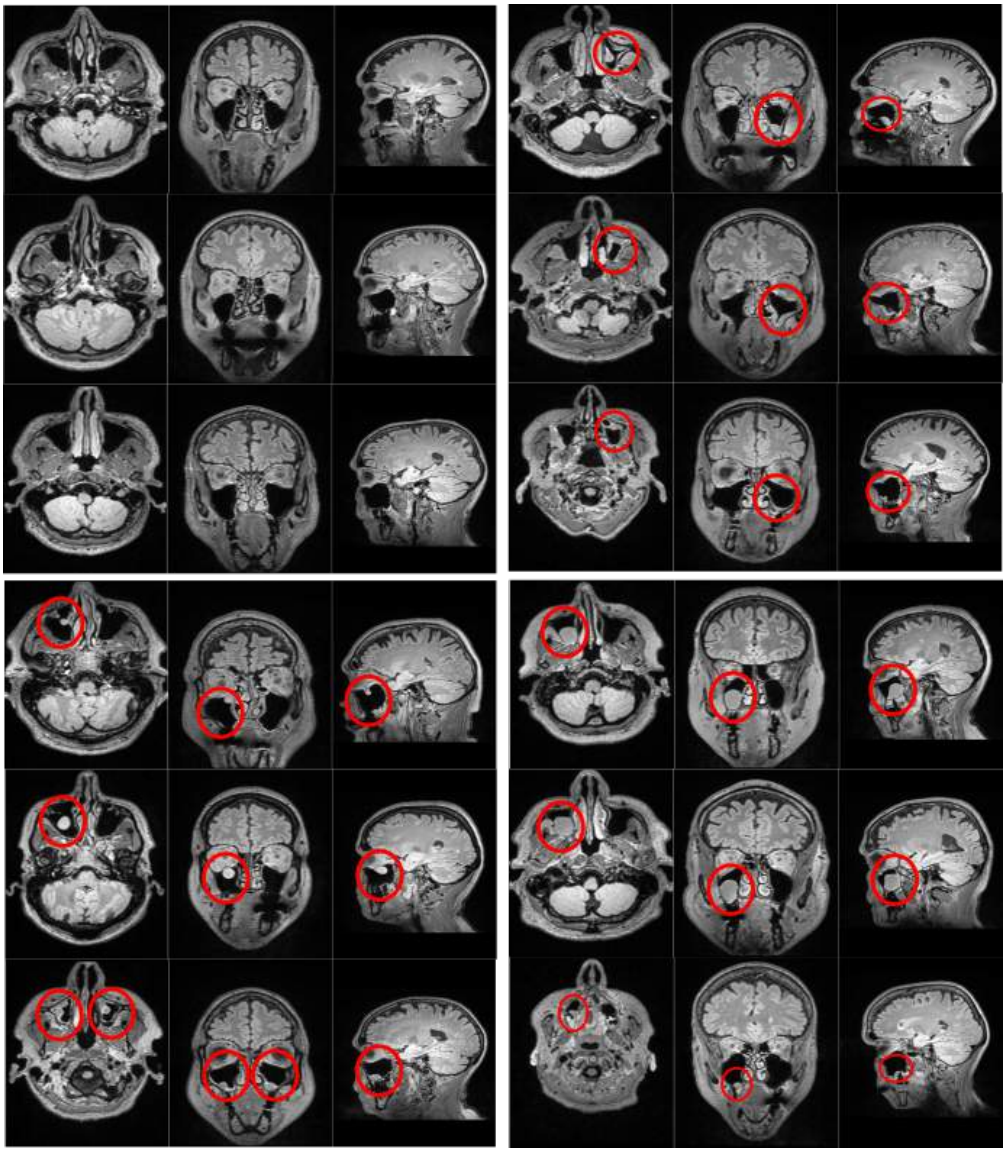

Figure 3.1 Samples of our dataset showing the axial, coronal and sagittal planes of participants. Top left images show 3 participants with no opacifications. Top right images show 3 participants with mucosal thickening pathology. Bottom left shows 3 participants with polyp pathology. Bottom right shows 3 participants with cyst pathology. Note that the pathologies are differently located, have varied intensities and occur in diverse shapes and sizes.

recorded participant details included Body Mass Index (BMI), age, sex, and allergies.

Among the 2,619 participants (56.05% men, 43.95% women) with a mean age of 63.98 (σ=8.32) years, a subset of 1072 individuals (56.6% men, 43.4% women) with a mean age of 63.93 (σ=8.24) years were manually annotated. Here, σ is the standard deviation. The

annotation process involved the collaboration of two Ears, Nose, and Throat surgeons and one Ears, Nose, and Throat specialized radiologist. These clinicians recorded observed opacifications solely within the maxillary sinus, classifying three pathologies: mucosal thickening (mucosal wall $> 2\ mm$), polyps, and cysts. Notably, only one pathology was found within the maxillary sinus. A derived category, termed 'fully occupied,' was considered to denote the presence of polyp or cyst opacification completely subsuming a maxillary sinus. 489 participants had both normal left and right maxillary sinuses whereas 583 had opacifications on atleast one maxillary sinus. A detailed statistics of the different participant groups is shown in table 3.1. The statistics of the opacification type observed in the participants is shown in table 3.1.

Table 3.1 Participant Groups and number of participants in each group of our labelled dataset D_l. LMS - Left maxillary sinus, RMS - Right maxillary sinus

Participant Group	**Number of Participants**
LMS and RMS normal	489
LMS opacification and RMS normal	149
LMS normal and RMS opacification	160
LMS opacification and RMS opacification	274

Table 3.2 Statistics of opacifications within our labelled dataset D_l

Opacification type	**Number of maxillary sinus**
Normal	1287
Mucosal Thickening	321
Polyps	421
Cysts	79
Fully occupied	36

3.2 Preprocessing strategy

The MRI scans of each participant encompass the entire head, although much of this information is irrelevant for the specific task of classifying opacifications in the maxillary sinus. To efficiently extract the 3D volumes of the maxillary sinus, we devised a strategy

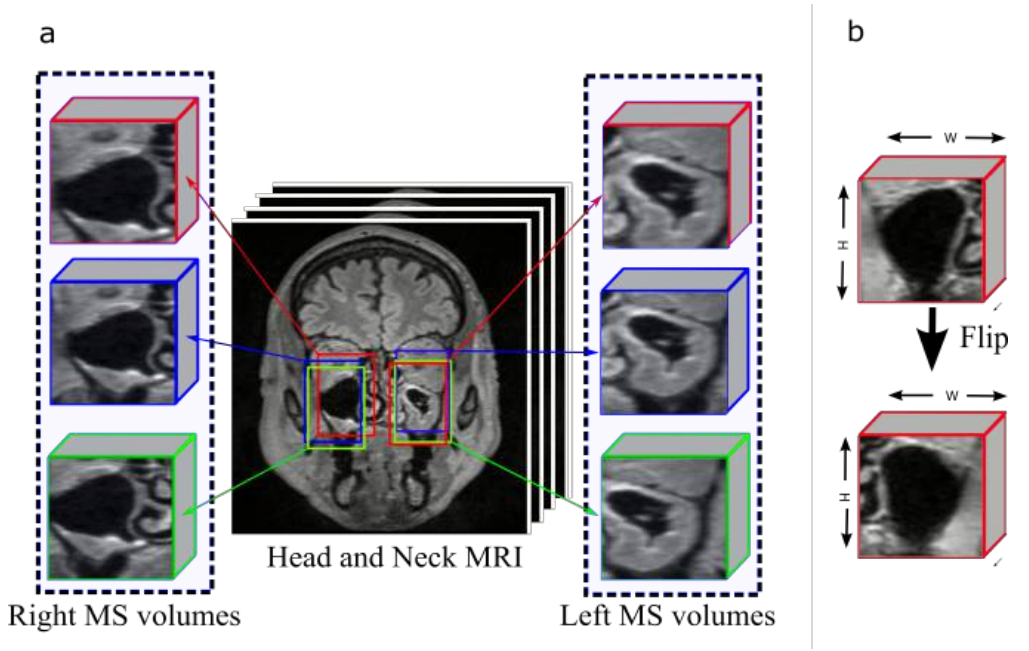

Figure 3.2 Overview of our preprocessing strategy [15]. a) Illustration of extraction N=3 maxillary sinus (MS) volumes from left and right side of head and neck MRI. b) Flipping of the coronal planes of the right maxillary sinus to make it look similar to the left maxillary sinus.

involving manual recording of centroid locations for the left and right maxillary sinuses from 20 patients which we mention in our paper [15]. From these coordinates, we computed the mean $(\mu(x), \mu(y), \mu(z))$ and standard deviation $(\sigma(x), \sigma(y), \sigma(z))$ of the centroid locations. Subsequently, we initialized Gaussian distributions $(\mathcal{N}(\mu(x), \sigma^2(x))$, $\mathcal{N}(\mu(y), \sigma^2(y))$, $\mathcal{N}(\mu(z), \sigma^2(z)))$ based on these values to sample centroid locations within the head and neck MRI, specifically for the maxillary sinus volumes. Notably, due to different mean and standard deviation values for the left and right maxillary sinuses, a total of six Gaussian distributions are utilized in practice. For instance, considering the left maxillary sinus, the mean (μ) and standard deviation (σ) values for x, y, and z coordinates are 75 mm, 231 mm, 121 mm and 1.47 mm, 1.56 mm, 1.76 mm, respectively. Correspondingly, for the right maxillary sinus, these values are 149 mm, 232 mm, 118 mm for μ and 1.90 mm, 1.66 mm, 6.47 mm for σ. We sample N volumes for both the left and right maxillary sinuses from each head and neck MRI (where N represents the sample size). This sampling approach

enables extraction of maxillary sinuses of arbitrary sizes by determining the centroid locations. Furthermore, to enhance symmetry between the left and right maxillary sinuses, we horizontally flip the coronal planes of the right maxillary sinus, creating the appearance of symmetry with the left maxillary sinus volume. Figure 3.4 a shows the extraction of multiple maxillary sinus volumes and figure 3.4 b shows the horizontal flipping of coronal planes of the right maxillary sinus.

3.3 Formalisation of methods in paranasal anomaly detection

Let D_l^N represent the labeled dataset containing maxillary sinus opacifications and no opacifications, and N denote the number of extracted maxillary sinus volumes from the left or right side of each participant's MRI. Within D_l^N, consider $D_{l/n}^N \subset D_l^N$ as a subset solely comprising normal maxillary sinus volumes. Furthermore, let D_u^N denote the unlabeled dataset encompassing 3D maxillary sinus volumes with unknown conditions.

Define $x \in \mathbb{R}^{H \times W \times D}$ as a 3D maxillary sinus volume extracted from the left or right maxillary sinus region within the head and neck MRI. Correspondingly, let $y \in \{0, 1\}$ represent the label, where 1 denotes opacification and 0 denotes no opacification/normal. We also consider $f(x)$ as a deep network model that we will define explicitly in the corresponding sections.

3.3.1 Overview of all methods

The schematic representation in Figure 3.3 elucidates the comprehensive methodology employed in this study. It encompasses the preprocessing procedures applied to generate both labeled and unlabeled datasets. The UAD and SSL methods are depicted, along with the integration of hybrid networks. Additionally, the approach involves enhancing super-

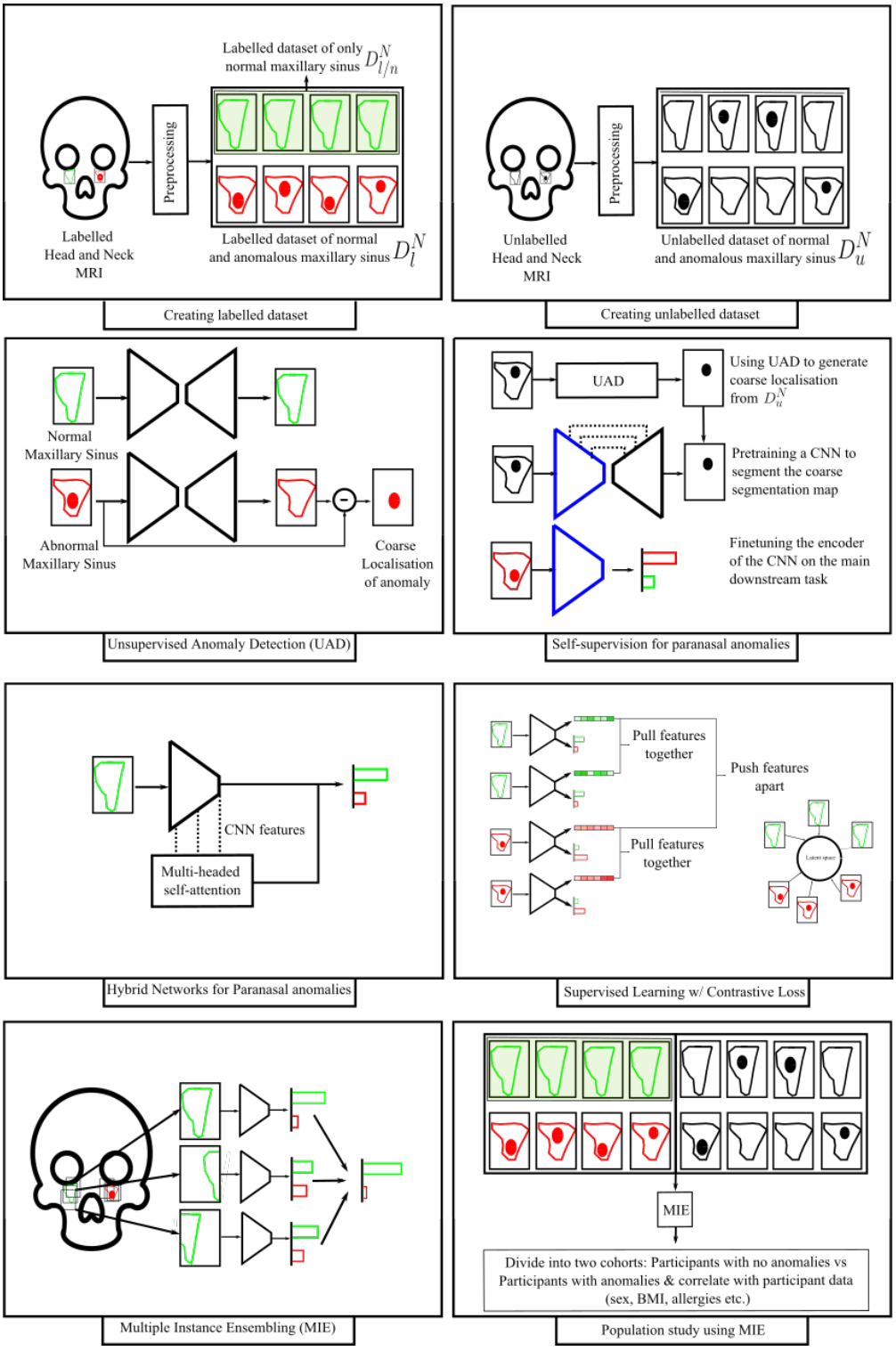

Figure 3.3 Overview of all methods

vised learning through contrastive loss and multiple instance learning. Ultimately, the MIE method, identified as the most effective, is utilized to ascertain correlations between maxillary sinus opacification and participant attributes such as sex, age, BMI, among others. In essence, these methods aim either to employ the labelled dataset for acquiring effective representations that subsequently enhance classification performance or alternatively, utilize the unlabeled dataset to improve performance on the labelled dataset.

3.3.2 Unsupervised Anomaly Detection

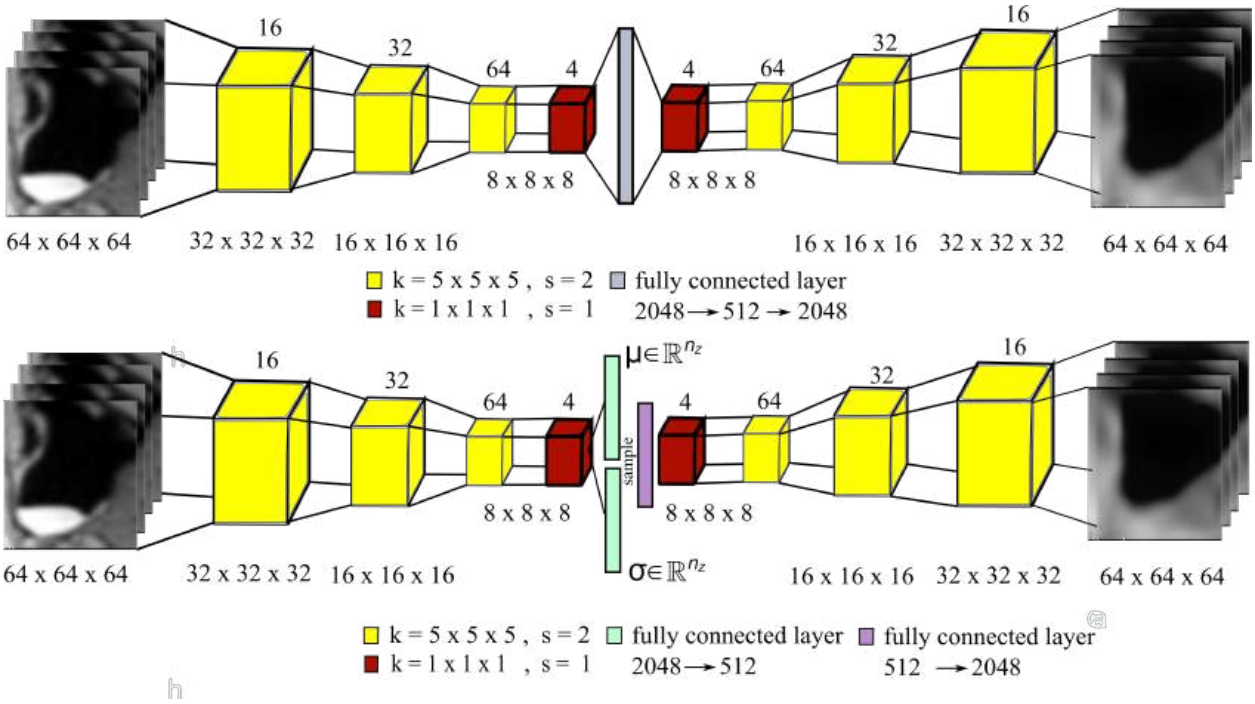

Figure 3.4 UAD of paranasal anomalies [13]: Observation of a maxillary sinuses with a polyp in the input and the CAE and VAE failing to reconstruct the polyp in the output.

The objective of UAD is to enhance performance on the labelled dataset and minimize labelling efforts, relying solely on access to normal or healthy samples. In our experimental setup, we explore the viability of employing UAD for the classification of paranasal anomalies. For our unsupervised anomaly detection task as shown in [13], we sample x from the dataset $D^N_{l/n}$. Note, $N = 1$, $|D^N_l| = 199$, $|D^N_{l/n}| = 106$ for this method. Consider $\hat{x} = f(x)$ as the

reconstructed maxillary sinus volume. In this context, $f(.)$ could denote a CAE or a VAE. Specifically, $f(x) = dec(z)$ where $z = enc(x)$, collectively $f(x) = dec(enc(x))$. While these functions are parameterized, for brevity, we omit their explicit representation. Here, $enc(.)$ and $dec(.)$ refer to the encoder and decoder functions, respectively, with z being the latent vector. We train our CAE/VAE $f(.)$ to learn the distribution $\mathcal{X}_{D^N_{l/n}}$.

For CAE training, we employ the L1 reconstruction loss, defined as:

$$\mathcal{L}_{CAE} = \sum_{i=1}^{n} |x_i - \hat{x}_i| \tag{3.1}$$

where n is the mini-batch size, and x_i and $\hat{x}_i$ index the i-th element from the mini-batch.

In VAE training, we incorporate L1 reconstruction loss and Kullback-Liebler (KL) Divergence. The VAE learns mean μ_z and variance σ_z from which a sample is drawn and reconstructed. The VAE's loss is:

$$\mathcal{L}_{VAE} = \mathcal{L}_{CAE} + \lambda_{KL} D_{KL}(q(z|x)||p(z)) \tag{3.2}$$

Here, $D_{KL}(.||.)$ represents the KL divergence between the parameterized latent distribution $q(z|x) \sim N(\mu_z, \sigma_z)$ and the prior $p(z)$, following a multivariate normal distribution. $z \in \mathbb{R}^{n_z}$ represents the latent vector. Setting λ_{KL} as 1 in our experiments, VAE projects the input maxillary sinus volume to $q(z|x)$, with KL divergence minimizing the distance to the prior $p(z)$ [72].

We utilize the trained $f(.)$ to reconstruct maxillary sinus volumes in the validation set. For each volume, we compute the L1 and L2 reconstruction losses, denoted as t_{L1} and t_{L2} respectively. Optimal thresholds are chosen via precision-recall curve analysis, selecting the threshold with the highest F1 score. During inference, $\hat{x}_i$ with $L1$ loss $> t_{L1}$ or $L2$ loss $> t_{L2}$ is classified as a maxillary sinus with opacification.

For anomalous maxillary sinus volumes, further analysis involves calculating voxel-wise intensity differences $D_k = |x_i - \hat{x}_i|$, followed by a median filter application (kernel size: 5)

to mitigate sporadic reconstruction errors. Regions exhibiting high reconstruction errors aid in coarse anomaly localization.

3.3.3 Self-supervised learning in paranasal anomaly detection

Our investigation in Section 3.3.2 demonstrated the utility of CAE and VAE architectures within the UAD framework for localizing anomalies within the maxillary sinus. We hypothesized that training $f(.)$, our 3D CNN structured as a UNet [116] comprising an encoder $enc(.)$ and decoder $dec(.)$, to reconstruct residual maxillary sinus volumes (i.e., volumes derived from the difference between generated and original maxillary sinus volumes) could implicitly enable anomaly localization and this learning could improve the extraction of better transferrable features for our downstream classification task of differentiating between opacification and no opacification of maxillary sinuses. This method aims to systematically explore the optimal utilization of the unlabeled dataset to enhance the classification performance on the constrained labeled dataset.

Given the ample unlabelled dataset D_u^N, we aimed to train an autoencoder $A(.)$ within the UAD framework to reconstruct residual volumes ($x' = A(x)$). This trained autoencoder $A(.)$ leveraged L1 reconstruction loss ($\|x - x'\|$) on $D_{l/n}^N$. Subsequently, $A(.)$ generated residual volumes on D_u^N. This was followed by using $f(.)$ to reconstruct the residual volumes which constituted our SSL task.

Once $f(.)$ underwent SSL, we excluded $dec(.)$ and retained only $enc(.)$ for fine-tuning on our labeled dataset D_l^N. Fine-tuning involved supervised training, utilizing binary cross-entropy loss over D_l^N. This process aimed to finetune the encoder $enc(.)$ for better performance in classification. Note, $N = 1$, $|D_l^N| = 1067$, $|D_u^N| = 1559$ for this method. Figure 3.5 gives an overview of our SSL method.

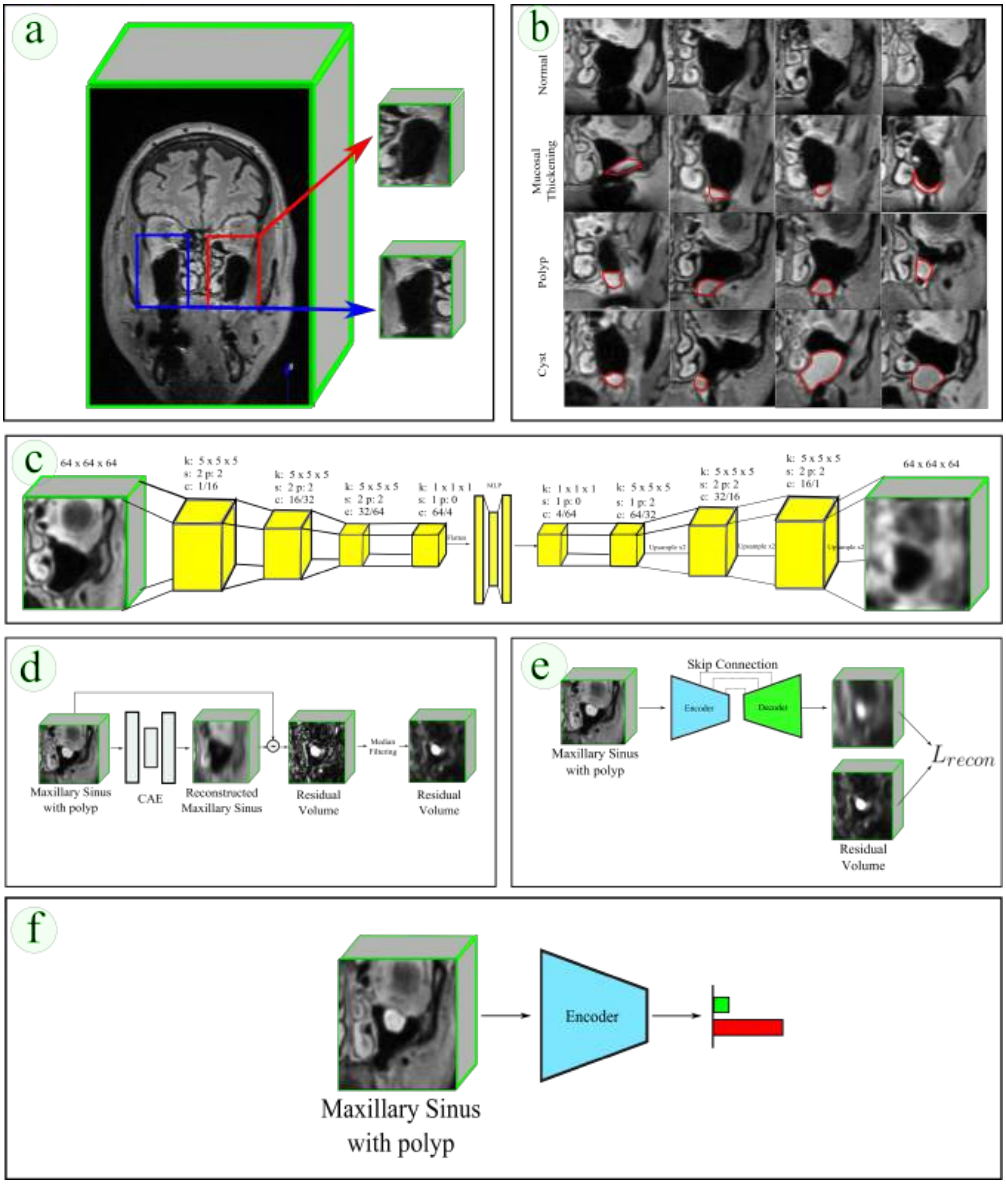

Figure 3.5 Overview of our SSL in paranasal anomalies [**?**]: a) Extraction of maxillary sinus volumes from cranial MRI b) Exemplary coronal images depicting a normal maxillary sinus volume and maxillary sinus instances showcasing mucosal thickening, polyps, and cyst anomalies c) Description of our CAE architecture. Here, k represents kernel size, s denotes stride, p signifies padding, and c indicates channel information. For instance, 1/16 signifies an input channel of 1 and an output channel of 16. Each stage of the encoder and decoder involves 3D convolution followed by batch normalization and leaky ReLU. Upsample denotes trilinear upsampling. d) Generation of residual volumes essential for the self-supervision task using our CAE e) The self-supervision task involving the training of the encoder and decoder to reconstruct the residual volume f) The downstream task wherein the self-supervision trained encoder is further trained to classify between normal and anomalous maxillary sinus.

3.3.4 Improving supervised learning using architectural modifications

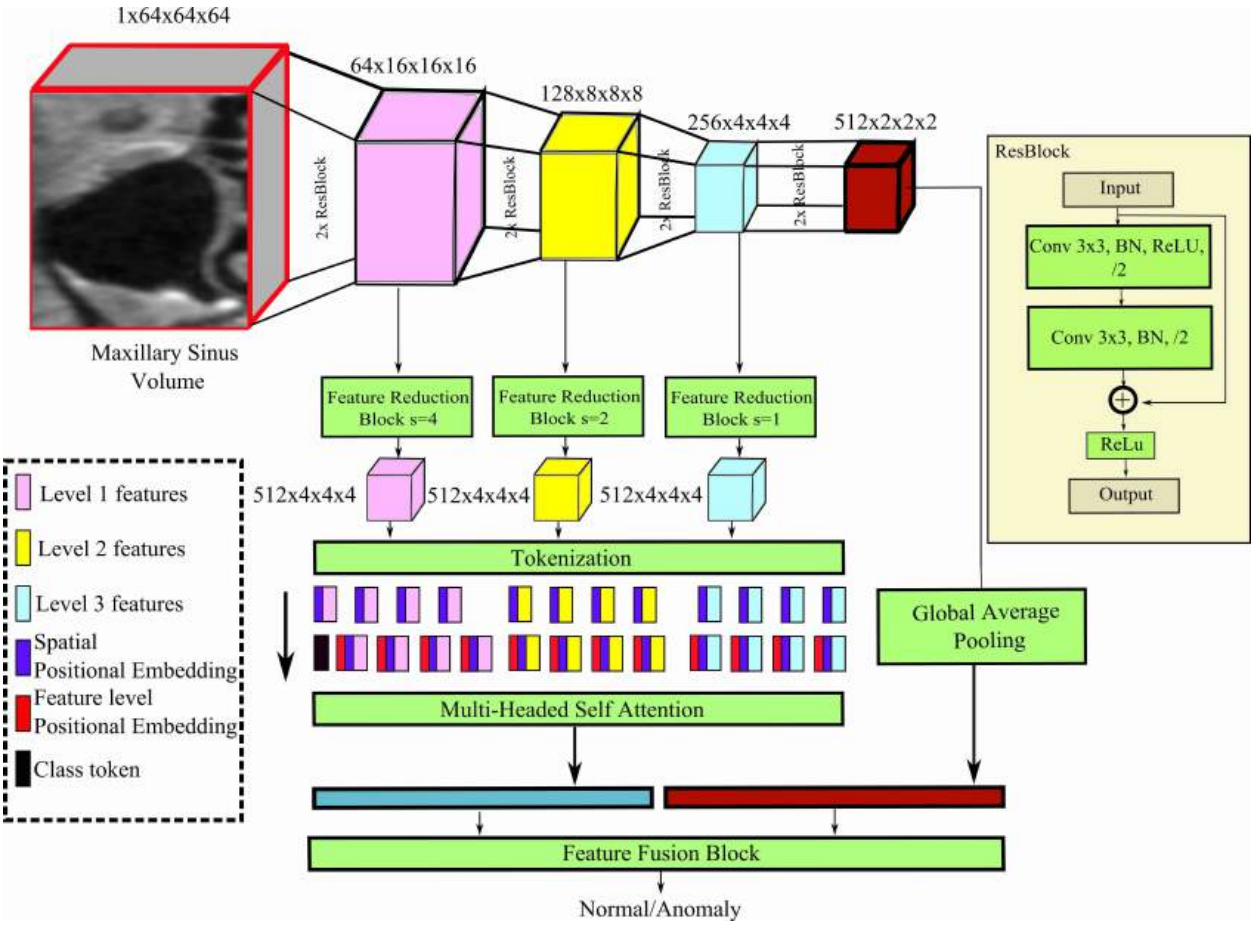

Figure 3.6 Overview of ConTra-Net [11]: The s in the Feature Reduction Block denotes the stride used in the convolution operation. The pink, yellow, and cyan features correspond to the low, mid, and high-level features extracted by the CNN.

In our proposed method, we introduce a hybrid architecture termed Convolutional Transformer Network (ConTra-Net), leveraging both the inherent convolutional bias and the capability of capturing long-range feature dependencies enabled by the multi-headed self-attention block (MHSA) within the transformer. The underlying motivation for this method is to explore deep learning architectures that utilize the labelled dataset effectively for improving classification performance.

The input maxillary sinus volume x undergoes processing through a 3D CNN $f(.)$ com-

posed of L stages of 3D residual blocks. Formally, $f_l = \mathcal{F}_l(x;\theta) \mid x \in \mathbb{R}^{C_{l-1}\times\frac{H}{2^{l-1}}\times\frac{W}{2^{l-1}}\times\frac{D}{2^{l-1}}}$, where $f_l \in \mathbb{R}^{C_l\times\frac{H}{2^l}\times\frac{W}{2^l}\times\frac{D}{2^l}}$ denotes the feature maps at level l of the 3D CNN $f(.)$. Here, C_l signifies the channel dimension at the l-th feature level, and θ represents the CNN parameters.

To enable our ConTra-Net to capture global context while maintaining computational efficiency, we downsample the resolution and increase the channel dimension of f_l. This is achieved through feature transformations employing 3D depthwise separable convolutions. The depthwise convolution operation spatially downsamples f_l, while the pointwise convolution operation increases the channel dimension. Formally, $\hat{f}_l = \Psi_l(\Upsilon(f_l;k,s);c_{in},c_{out})$, where $\Upsilon(.;k,s)$ represents the depthwise convolution with kernel size k and stride s, controlling the downsampling factor, and $\Psi_l(.;c_{in},c_{out})$ is a 3D convolution operation with kernel size 1 and input/output channel dimensions c_{in} and c_{out}, respectively. Notably, $c_{in} = C_l$.

The resulting multi-scale features f_l are transformed into $\hat{f}_l \in \mathbb{R}^{c_{out}\times h\times w\times d}$ via a feature reduction block. These $\hat{f}_l$ serve as tokens for the subsequent MHSA block, being flattened into a sequence $x_l \in \mathbb{R}^{N_{seq}\times c_{out}}$, where x_l represents the sequence of the l-th feature level, and $N_{seq} = h \cdot w \cdot d$ signifies the sequence length. c_{out} denotes the embedding dimension.

To maintain spatial positional information, we incorporate a learnable positional embedding $p_{spatial} \in \mathbb{R}^{N_{seq}\times c_{out}}$. Consequently, $\hat{f}_l^{pos} = \hat{f}_l + p_{spatial}$. We then concatenate these spatially-informed features $\hat{f}_l^{pos}$ from different feature levels l into a consolidated sequence $f_t \in \mathbb{R}^{N_{total}\times c_{out}}$, where $N_{total} = N_{seq} \times (L-1)$ spans multiple feature levels, excluding the L-th feature block. Additionally, for retaining positional information of different level features, we introduce another learnable positional embedding $p_{level} \in \mathbb{R}^{N_{total}\times c_{out}}$, leading to $\hat{f}_t = f_t + p_{level}$. Finally, the class token $cls \in \mathbb{R}^{1\times c_{out}}$ is concatenated with $\hat{f}_t$ ($\hat{f}_t = (\hat{f}_t \oplus cls) \in \mathbb{R}^{N_{total}+1\times c_{out}}$). This resulting matrix $\hat{f}_t$ undergoes multiple layers of MHSA blocks and feedforward layers, culminating in feature $F_t \in \mathbb{R}^{N_{total}+1\times c_{out}}$.

The feature fusion block combines MHSA and CNN features for class predictions. It utilizes the cls vector from F_t as the representative MHSA feature vector encoding global

context. Additionally, it concatenates the CNN feature vector f_L with the MHSA feature vector. The resultant vector is then processed through feedforward layers to yield the final class prediction. ConTra-Net is trained using binary cross-entropy loss. Note, $N = 1$, $|D_l^N| = 299$ and $|D_{l/n}^N| = 174$ for this method.

3.3.5 Improving Supervised learning using contrastive loss

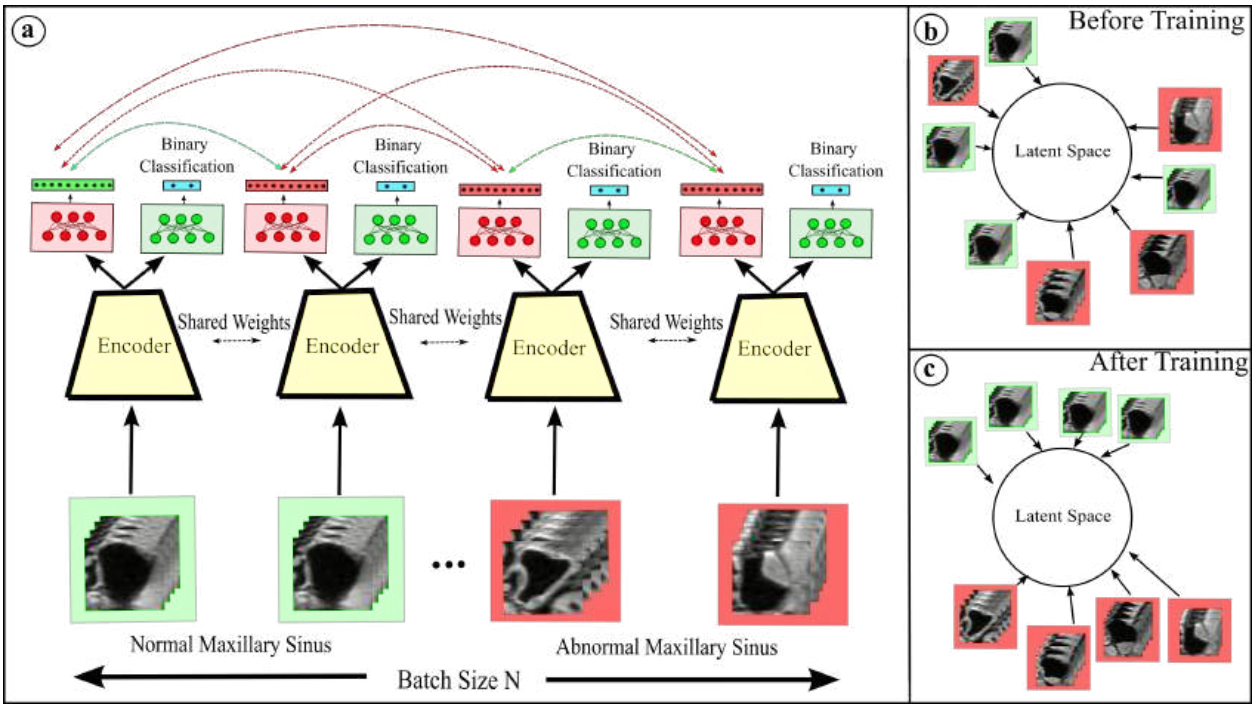

Figure 3.7 Overview of method to improve supervised learning using contrastive loss [11] (a) Our proposed method depicts similar representations as curved green lines and dissimilar representations as curved red lines. $(b) - (c)$ These figures illustrate the latent space embedding of normal and anomalous maxillary sinuses, both pre- and post-training of the encoder, respectively.

As in Section 3.3.4, the primary objective remains the more efficient utilization of the labeled dataset to enhance classification performance. We investigate the usage of contrastive loss along side cross-entropy loss for improved representation learning. In the domain of contrastive learning, such as in SimCLR [25], the fundamental objective is to map samples having the same content but augmented into two different versions have the similar encodings.

However, the assumption that images within a mini-batch are inherently dissimilar does not hold in our context. Our dataset consists of maxillary sinus volumes belonging to one of two classes (opacification or no opacification), where semantic similarities and dissimilarities exist both within and across these classes. Consequently, the contrastive loss mechanism in SimCLR is incapable of forming meaningful clusters as it is dependent on augmentations to learn meaningful features.

To address this limitation and incorporate class-specific information, we adopt a different strategy. Instead of randomly sampling n samples, we explicitly sample $n/2$ volumes without opacification and $n/2$ volumes with opacification from our dataset D_l^N. Each of the mini-batch $B = \{x_1, x_2, ..., x_n\}$ undergoes random transformations using the set T twice. Denoting the collection of all augmented volumes across the C classes as $\mathbb{M} = \bigcup_{c=1}^{C} M_c$ (where $C = 2$ in our case), M_c represents the subset of augmented volumes belonging to a single class, with $|M_c| = n$ denoting its cardinality. $m_i, i \in \mathbb{I} = \{1, 2, ..., 2N\}$ denotes the augmented volumes in set $\mathbb{M}$ and $m_{k(i)}$ is its corresponding volume augmented from the same volume in B.

$$L_{simclr} = -\sum_{c=1}^{C} \frac{1}{|M_c|} \sum_{i \in I_c} log \frac{e^{sim(Z_i, Z_{k(i)})/\tau}}{e^{sim(Z_i, Z_{k(i)})/\tau} + \sum_{j \in \mathbb{I} \setminus I_c} e^{sim(Z_i, Z_j)/\tau}} \tag{3.3}$$

The loss function (3.3) (L_{simclr}) considers the class priors when forming positive and negative pairs. τ is a scalar temperature parameter, $Z_i = f^{con}(m_i)$ denotes the normalized feature vector, $k(i)$ signifies the index of the corresponding volume in $\mathbb{M}$ augmented from the same volume in B, and $sim(.)$ denotes the cosine similarity function. Notably, this loss primarily constructs negative pairs where the volumes belong to different classes, yet it also generates positive pairs by augmenting the same volume using T, relying on transformations to learn meaningful representations.

However, relying solely on these transformations does not guarantee anatomical and anomaly invariance. The encoder is not incentivized to produce similar representations for volumes belonging to the same class in the mini-batch. To overcome this, we introduce

a supervised contrastive loss L_{sc} (equation (3.4)) that constructs multiple positive pairs. This involves matching every volume with every other volume belonging to the same class in the mini-batch. This incentivizes the encoder to generate similar representations for volumes within the same class, facilitating the learning of anatomical and anomaly invariant representations.

$$L_{sc} = -\sum_{c=1}^{C} \frac{1}{|M_c|} \sum_{i \in I_c} log \frac{\sum_{j \in I_c \setminus \{i\}} e^{sim(Z_i, Z_j)/\tau}}{\sum_{j \in I_c \setminus \{i\}} e^{sim(Z_i, Z_j)/\tau} + \sum_{j \in \mathbb{I} \setminus I_c} e^{sim(Z_i, Z_j)/\tau}} \tag{3.4}$$

where I_c represent indices of augmented maxillary sinus volumes belonging to class c such that $\mathbb{I} = \bigcup_{c=1}^{C} I_c$.

Additionally, we incorporate a regular cross-entropy loss L_{ce} (equation (3.5)) to preserve the discriminative ability of our 3D CNN. Finally, our combined loss function $L_{ours} = L_{sc} + \lambda L_{ce}$ (equation (3.6)) integrates the supervised contrastive loss with the cross-entropy loss, where λ is set to 1 in our experiments.

$$L_{ce} = -\frac{1}{N} \sum_{i \in \mathbb{I}} y_i log(f^{class}(m_i)) \tag{3.5}$$

Therefore, our combined loss function is

$$L_{ours} = L_{sc} + \lambda L_{ce} \tag{3.6}$$

f^{class} and f^{con} are parameterised functions (3D CNN) having the same backbone and two different multi-layer perceptrons for transforming the backbone encoded features to features suitable for cross-entropy loss and contrastive loss respectively. Note, $N = 1$, $|D_l^N| = 199$ and $|D_{l/n}^N| = 106$ for this method.

3.3.6 Improving supervised learning using multiple instance ensembling

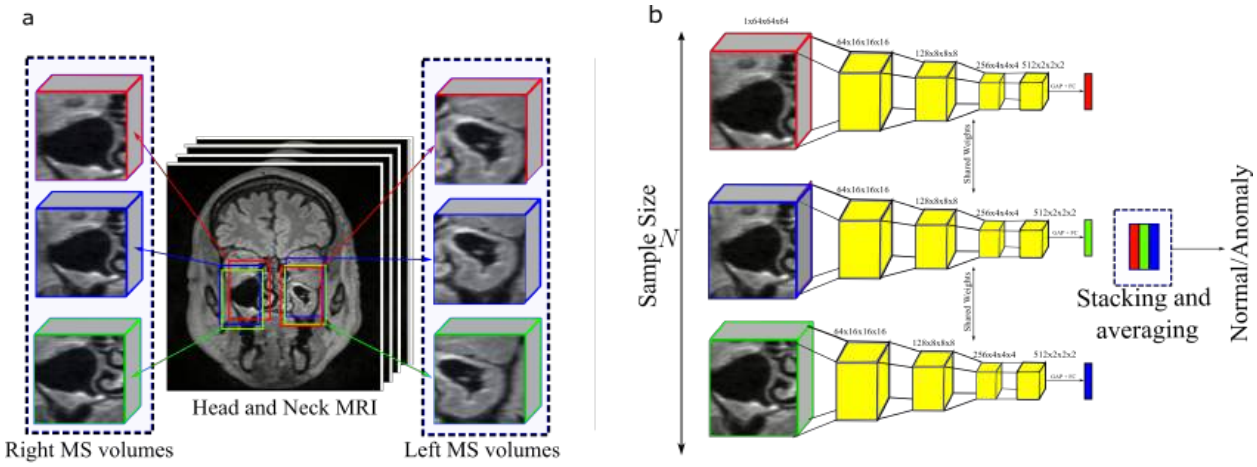

Figure 3.8 Overview of our method to improve supervised learning for paranasal anomaly classification using multiple instance ensembling. [15] (*a*) This figure illustrates our strategy for extracting maxillary sinus (MS) volumes, displaying three MS volumes for both the left and right MS. (*b*) This figure depicts our Multiple Instance Ensembling (MIE) prediction strategy employed during inference, where GAP signifies Global Average Pooling and FC represents the Fully Connected Layer.

The overarching goal of this method is similar to the methods in sections 3.3.2, 3.3.4 and 3.3.5 which is to improve the classification performance by learning better representations strictly from the labelled dataset. Our proposed end-to-end approach for improving supervised learning combines a non-deep learning solution for localizing maxillary sinus volumes with a deep learning method for classifying maxillary sinus opacifications. The unique localization strategy, employing Gaussian sampling of centroid coordinates, substantially expands the dataset by extracting multiple overlapping instances of the maxillary sinus. Section 3.2 details this method. Note, $N = 15$, $|D_l^N| = 299$ and $|D_{l/n}^N| = 174$ for this method. This process results in extracting N maxillary sinus volumes from each participant, effectively amplifying the training dataset by a factor of N. During inference, we extract N maxillary sinus volumes from both the left and right sides of the cranial MRI and perform ensembling. Specifically, for prediction, we average the softmax scores of the 3D CNN classifier

$f(.)$ obtained from the multiple maxillary sinus volumes x_i. Formally,

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} \text{softmax}(f(x_i)) \tag{3.7}$$

3.3.7 Correlation with patient data

Numerous large cohort studies have examined paranasal anomalies' incidence and their correlation with diverse clinical variables [131, 49, 30]. However, these approaches heavily relied on labor-intensive manual diagnoses. To address the limitations of manual diagnosis, we employed the multiple instance ensembling approach [13] and further enhanced its performance through ensemble modelling. Leveraging this ensemble model (EM), trained on a labeled dataset D_l^N, we conducted inferences on our unlabeled dataset D_u^N, delineating two groups: participants exhibiting maxillary sinus opacifications (referred to as 'cases') and those without (referred to as 'control'). We used an exclusion criterion for the unlabelled MRIs where EM confidence $<$0.90 for any MS were excluded. More details can be found in our paper [12].

Subsequently, we investigated statistically significant differences between these groups across various clinical variables (including smoking, alcohol consumption, Body Mass Index, asthma, bronchitis, sex, age, leukocyte count, C-reactive protein, and allergies). We employed a χ^2 test to explore associations among categorical variables and utilized the point-biserial correlation, a special case of the Pearson correlation, to assess relationships involving continuous variables.

Chapter 4

Results

4.1 Unsupervised anomaly detection

Table 4.1 Unsupervised Anomaly Detection Performance using two thresholds t_{L1} and t_{L2}. When the mean reconstruction error of the maxillary sinus (MS) volume is above t_{L1} and t_{L2}, the volume is classified as anomaly. Bold text indicates the highest value in the column.

Method	MS Size	Precision		Recall		F1		AUPRC	
		t_{L1}	t_{L2}	t_{L1}	t_{L2}	t_{L1}	t_{L2}	t_{L1}	t_{L2}
VAE	small	0.69	0.76	0.63	0.62	0.64	0.68	0.76	0.80
VAE	medium	0.63	0.63	**0.84**	**0.91**	0.72	**0.75**	0.70	0.75
VAE	large	0.61	0.64	0.81	0.86	0.70	0.73	0.65	0.69
CAE	small	**0.75**	0.81	0.74	0.66	0.74	0.73	**0.81**	0.83
CAE	medium	0.73	**0.77**	0.62	0.74	0.67	**0.75**	0.80	**0.85**
CAE	large	0.68	0.74	0.82	0.73	**0.74**	0.73	0.73	0.78

For our classification task, initial thresholds t_{L1} and t_{L2} are computed using a dedicated validation set. Notably, our extraction process yields maxillary sinuses of varying sizes categorized as *small*, *medium*, and *large*. Detailed information on threshold computation and maxillary sinus volume sizes is available in our paper [13].

Table 4.1 illustrates that all VAEs exhibit relatively lower Area Under Precision-Recall Curve (AUPRC) scores compared to CAEs. The highest AUPRC of 0.85 is observed for CAE

using the *small* maxillary sinus. Moreover, our findings demonstrate that incorporating L2 loss in computing the anomaly score and utilizing t_{L2} as the threshold consistently enhances performance across all CAEs and VAEs.

Despite its relatively poor classification performance, UAD offers a significant advantage: cost-effective anomaly localization, as depicted in Figure 4.1. Regions exhibiting high reconstruction error serve as indicators of potential anomalies. Exploiting this information forms a crucial aspect of our self-supervised learning, as detailed in Section 3.3.3.

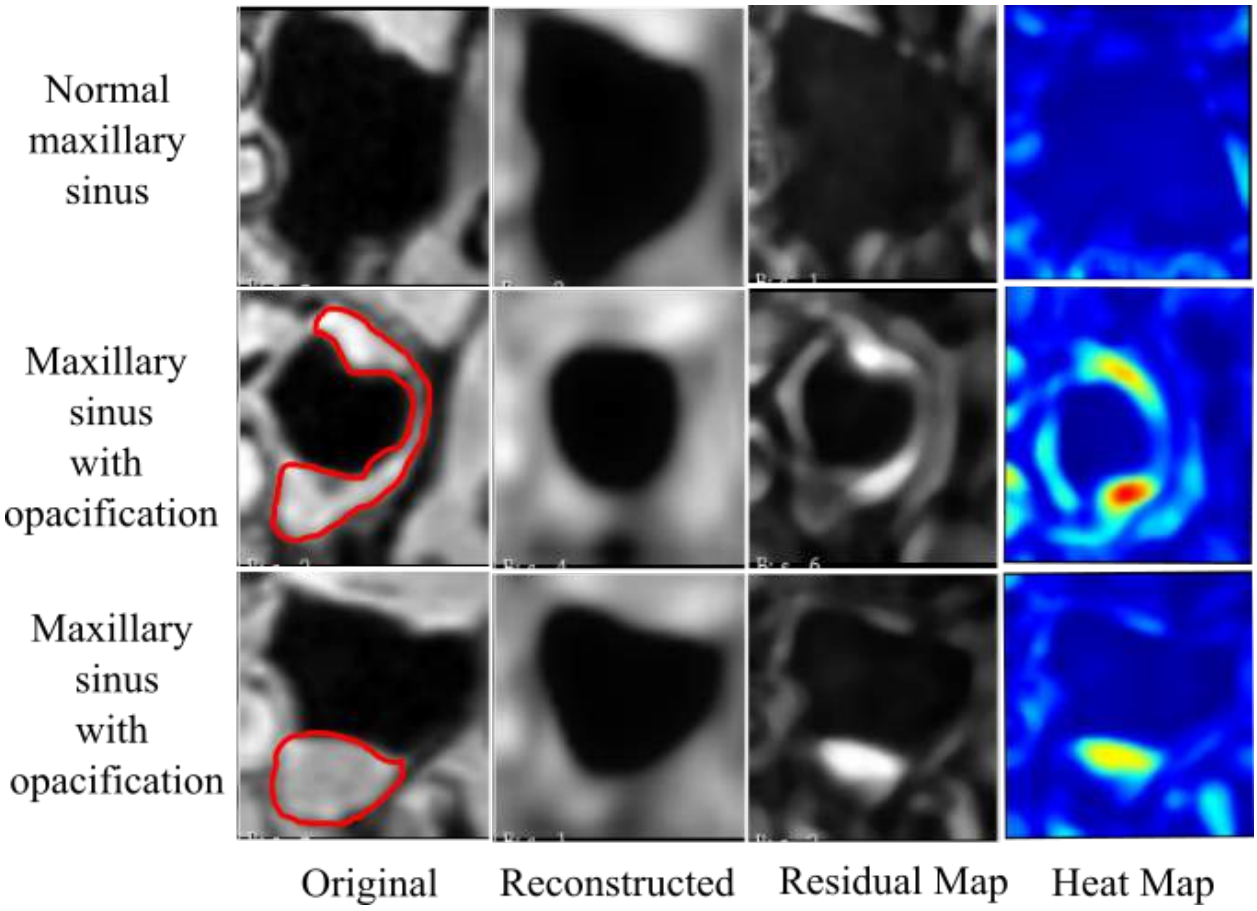

Figure 4.1 Coronal images displaying original, reconstructed, and residual maxillary sinus volumes from one normal and two opacification samples. Additionally, heat maps are presented for visualization purposes, where red pixels indicate regions with high reconstruction errors, while blue pixels signify areas of low reconstruction error. As the CAE or VAE struggles to reconstruct anomalous regions accurately, the reconstruction error offers preliminary coarse localization information of the anomalies.

4.2 Self-supervised learning in paranasal anomaly detection

Table 4.2 Results of our SSL for paranasal anomaly classification [14]: The table presents the mean and 95% confidence intervals for metrics assessing the model's performance in the downstream classification task. The models were initialized using different SSL methods before supervised training and subsequently, trained with different proportions of D_l^N in a supervised fashion. Bold text indicates the highest value in the column.

Method	Training Set Percentage D_l	AUROC	AUPRC	F1
No pretraining	10%	0.74 (0.64-0.84)	0.69 (0.56-0.82)	0.64 (0.59-0.69)
Transfer Learning	10%	0.77 (0.72-0.82)	0.73 (0.66-0.79)	0.63 (0.57-0.69)
AE	10%	0.73 (0.68-0.79)	0.68 (0.62-0.74)	0.55 (0.43-0.67)
DAE	10%	0.74 (0.73-0.76)	0.68 (0.66-0.69)	0.62 (0.60-0.64)
BYOL	10%	0.79 (0.76-0.81)	0.75 (0.70-0.79)	0.63 (0.59-0.69)
SimSiam	10%	0.77 (0.72-0.83)	0.74 (0.68-0.79)	0.62 (0.53-0.72)
SimCLR	10%	0.78 (0.74-0.81)	0.73 (0.68-0.78)	0.63 (0.59-0.68)
SparK MAE	10%	0.78 (0.77-0.80)	0.75 (0.73-0.76)	0.65 (0.63-0.67)
Ours	10%	**0.81 (0.74-0.88)**	**0.79 (0.71-0.87)**	**0.67 (0.58-0.77)**
No pretraining	20%	0.81 (0.79-0.82)	0.78 (0.76-0.79)	0.67 (0.65-0.69)
Transfer Learning	20%	0.84 (0.79-0.88)	0.81 (0.74-0.88)	0.68 (0.62-0.75)
AE	20%	0.81 (0.76-0.86)	0.78 (0.72-0.83)	0.67 (0.60-0.74)
DAE	20%	0.79 (0.77-0.81)	0.74 (0.70-0.79)	0.67 (0.64-0.70)
BYOL	20%	0.82 (0.80-0.84)	0.79 (0.77-0.82)	0.70 (0.68-0.71)
SimSiam	20%	0.84 (0.82-0.86)	0.81 (0.78-0.84)	0.70 (0.67-0.74)
SimCLR	20%	0.81 (0.79-0.83)	0.77 (0.74-0.81)	0.68 (0.67-0.69)
SparK MAE	20%	0.80 (0.78-0.82)	0.76 (0.73-0.79)	0.67 (0.65-0.68)
Ours	20%	**0.85 (0.83-0.87)**	**0.82 (0.81-0.83)**	**0.72 (0.70-0.75)**
No pretraining	100%	0.90 (0.89-0.91)	0.89 (0.88-0.90)	0.80 (0.78-0.82)
Transfer Learning	100%	0.92 (0.91-0.93)	0.91 (0.90-0.93)	0.82 (0.80-0.83)
AE	100%	0.92 (0.91-0.93)	0.91 (0.90-0.93)	0.82 (0.80-0.84)
DAE	100%	0.90 (0.88-0.92)	0.89 (0.88-0.91)	0.79 (0.77-0.82)
BYOL	100%	0.89 (0.89-0.90)	0.88 (0.87-0.89)	0.78 (0.76-0.81)
SimSiam	100%	0.92 (0.91-0.93)	0.91 (0.90-0.92)	0.81 (0.79-0.83)
SimCLR	100%	0.90 (0.88-0.91)	0.89 (0.87-0.91)	0.79 (0.77-0.80)
SparK MAE	100%	0.87 (0.85-0.88)	0.86 (0.84-0.87)	0.75 (0.73-0.76)
Ours	100%	**0.93 (0.91-0.94)**	**0.92 (0.90-0.93)**	**0.83 (0.80-0.86)**

In our UAD approach for paranasal anomaly classification [13], we discovered that coarse anomaly localization could be achieved using a CAE or VAE. Leveraging an unlabelled dataset of maxillary sinus volumes D_u, we speculated that employing this coarse anomaly localization might enhance downstream classification between no opacification and opacification of maxillary sinuses. To explore this, we utilized a trained CAE to generate residual volumes (depicted in Figure 4.1), acting as basic segmentation masks. Subsequently, we

trained a 3D CNN $f(.)$ to reconstruct these residual volumes, forming the basis of our self-supervised learning (SSL) task, detailed in Section 3.3.3.

The results in Table 4.2 illustrate the performance of our SSL method across various labeled dataset scenarios (10%, 20%, 100% of D_l^N), outperforming other techniques in terms of Area Under Receiver Operator Characteristic (AUROC), AUPRC, and F1 scores. Notably, our method demonstrates a substantial improvement in AUROC (3.34% and 4.93% over SimSiam) and AUPRC (5.33% over BYOL and 5.12% over AE) for 10% and 20% dataset scenarios, respectively. Pretraining significantly enhances AUPRC by 14.49% and AUROC by 9.45% compared to no pretraining. Models trained with SparK performed worse than other methods, especially as the amount of training data increased. In contrast, our method achieved an AUPRC 8.21% higher than the transfer learning (TL) when fine-tuned on a 10% training set. Upon finetuning with a 100% dataset, our method achieves the highest scores, exhibiting similar performance to AE and SimSiam. Relative to no pretraining, our method increases AUPRC by 3.33% at 100% training set. We conducted additional experiments to scrutinize the impact of loss function choice in the SSL task and determine the influence of the normal maxillary sinus dataset $D_{l/n}^N$ size on the downstream classification task. Comprehensive details and results of this analysis are provided in our paper [14].

4.3 Improving supervised learning using architectural modifications

Table 4.3 Improving supervised learning using architectural modifications [13]

Method	AUPRC	Precision	Recall	F1
ResNet 50 [54]	$\mu = 0.93, \sigma = 0.05$	$\mu = \mathbf{0.94}, \sigma = \mathbf{0.08}$	$\mu = 0.68, \sigma = 0.19$	$\mu = 0.77, \sigma = 0.13$
3D ViT [38]	$\mu = 0.69, \sigma = 0.08$	$\mu = 0.69, \sigma = 0.14$	$\mu = 0.52, \sigma = 0.20$	$\mu = 0.56, \sigma = 0.10$
ConTra-Net	$\mu = \mathbf{0.95}, \sigma = \mathbf{0.04}$	$\mu = 0.92, \sigma = 0.10$	$\mu = \mathbf{0.83}, \sigma = \mathbf{0.15}$	$\mu = \mathbf{0.86}, \sigma = \mathbf{0.07}$

As indicated in Table 4.3, ResNet50 [54] emerged as the second-best-performing method,

while 3D ViT exhibited the poorest performance. ConTra-Net showcased the highest scores in terms of AUPRC, recall, and F1, surpassing ResNet50 by 2.15%, 22.05%, and 11.68% respectively. We also analysed the influence of low, mid and high-level features extracted by the CNN backbone of ConTra-Net on the classification task. The results of the analysis can be found in our paper [13].

4.4 Improving supervised learning using contrastive loss

Table 4.4 Improving supervised learning using contrastive loss [11]. Bold text indicates the highest value in the column.

Method	L_{ce}	L_{simclr}	L_{sc}	Accuracy	F1(weighted)	AUROC	AUPRC
Randomly initialised	✓			0.68 ± 0.03	0.57 ± 0.04	0.66 ± 0.10	0.53 ± 0.12
Randomly initialised (with augmentation)	✓			0.69 ± 0.03	0.58 ± 0.05	0.68 ± 0.10	0.56 ± 0.10
SimCLR [25]		✓		0.65 ± 0.03	0.58 ± 0.04	0.54 ± 0.09	0.41 ± 0.09
SupCon [68]			✓	0.72 ± 0.05	0.70 ± 0.06	0.73 ± 0.08	0.61 ± 0.08
Ours	✓		✓	$\mathbf{0.80 \pm 0.02}$	$\mathbf{0.78 \pm 0.03}$	$\mathbf{0.85 \pm 0.03}$	$\mathbf{0.78 \pm 0.03}$

We introduced a novel loss function integrating binary cross-entropy loss (L_{ce}) and supervised contrastive loss (L_{sc}) [11]. These combined losses aimed to elevate the quality of learned representations, leading to enhanced classification performance, as illustrated in Table 4.4. Notably, our method outperforms 3D CNN initialized with random weights, both without and with data augmentation. Additionally, we observed that 3D CNN trained using SimCLR and SupCon techniques performed less effectively compared to our proposed approach. Moreover, our examination of performance with limited labeled datasets (60% and 80%) showcased notable improvements in classification accuracy through the use of our proposed loss function. For a comprehensive analysis, please refer to our paper [11].

4.5 Improving supervised learning using multiple instance ensembling

As described in section 3.2, we sample N maxillary sinus volumes from left and right side of each participant's cranial MRI. Analysis from Table 4.5 reveals a consistent improvement in all reported metrics as the sample size N increases, up until $N = 15$, beyond which a decrease in metrics is observed.

In Table 4.5, we display the classification performance of various deep neural network architectures, evaluating the consistency of improvement at $N = 15$ and with the inclusion of MIE. Our findings consistently demonstrate enhanced classification performance with the application of MIE and the extraction of N=15 maxillary sinus volumes. Furthermore, our extensive experimentation across diverse deep neural network, as shown in Table 4.6, architectures confirms that MIE consistently enhances the classification of paranasal anomalies, regardless of the architecture.

In pursuit of explainability, activation maps were generated using GradCAM [122] to assess the alignment between activated regions and clinically relevant areas for diagnosing maxillary sinus opacifications. Figure 4.2 demonstrates a direct overlap of activation maps with numerous clinically relevant areas across different opacification conditions. False predictions were also scrutinized to identify conditions where the MIE model encountered difficulties. Qualitative observations revealed frequent misclassifications of small cysts and polyps. Moreover, mucosal wall thickening around 2mm was occasionally misclassified as no opacification. Notably, cases of hypoplastic maxillary sinuses were misclassified as opacification.

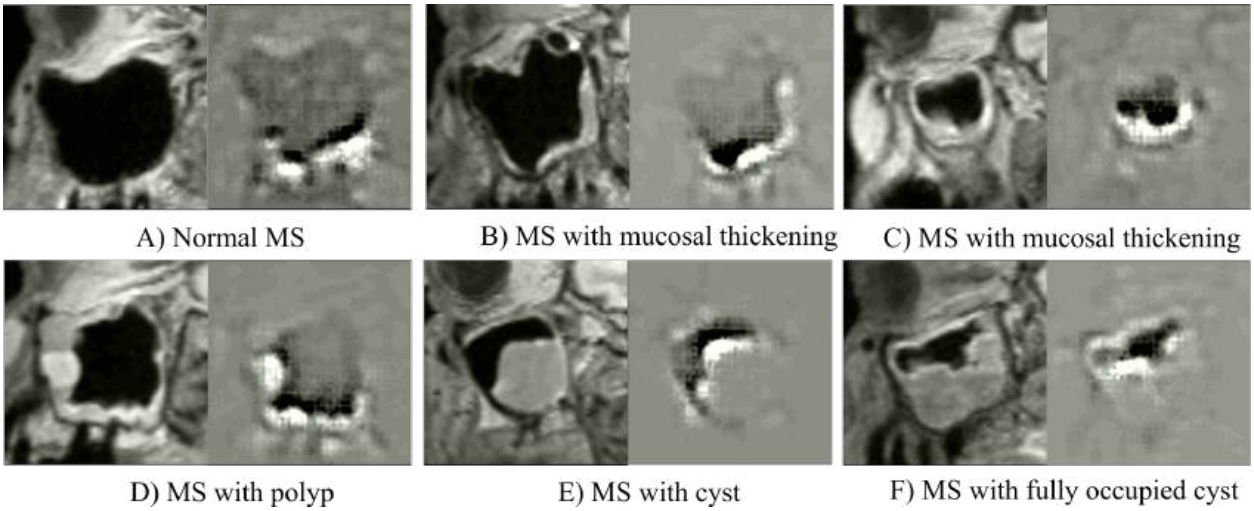

Figure 4.2 GradCAM activation taken from [12] - Sagittal plane images, accompanied by activation maps (white voxels denoting high activation), depict various conditions of the maxillary sinus – namely, no opacification, mucosal thickening, polyp, cyst, and fully occupied cyst. In A), heightened activation is concentrated on the maxillary sinus walls. B) and C) reveal localized high activation on thickened mucosa. D) displays activation within the polyp mass. E) showcases activation inside and on the edges of the cyst mass. F) demonstrates activation within and on the edge of the fully occupied cyst mass.

4.6 Correlation with patient data

Table 4.7 delineates the performance metrics of individual models within their respective cross-validation sets and showcases the EM outcomes. EM boosts the classification of MS anomalies, yielding noteworthy metrics: an AUROC of 0.95, precision and sensitivity at 0.85, and specificity of 0.90. In contrast, the next best-performing model, a 3D-CNN trained on the first fold, achieved an AUROC of 0.93 in the same test set.

From the unlabelled dataset, 1360 MRI scans met the inclusion criteria. These were supplemented with participants from the labeled dataset, culminating in a total of 2429 MRIs for further analysis. This pool of 2429 MRIs was segregated into two cohorts: the first comprising individuals with 'no opacification' in both left and right MS (our *control* group), and the second including those with at least one MS showing opacification (our *cases* group). A subsequent comparative analysis between these cohorts involved self-evaluated questionnaires and blood reports sourced from the HCHS for each participant. Comprehensive results

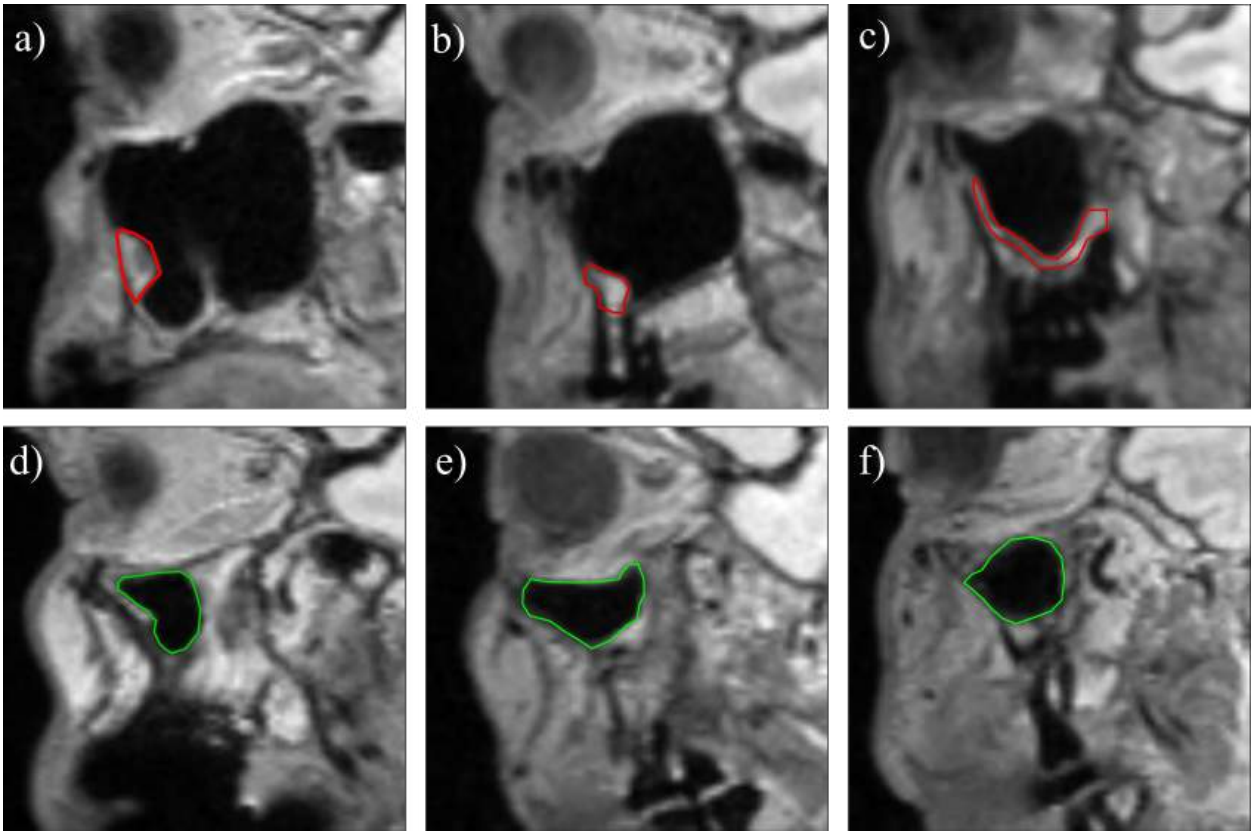

Figure 4.3 Misclassification cases taken from [12] - Sagittal plane images depict instances of misclassification: a) a small cyst in the anterior wall of the maxillary sinus misclassified as no-pathology, b) a small polyp in the bottom wall of maxillary misclassified as no-pathology, c) mucosal thickening slightly larger than 2 mm misclassified as "no pathology," and d,e,f) hypoplastic maxillary sinus misclassified as pathology. The red contours emphasize anomalous regions, while the green contours delineate the boundary walls of hypoplastic maxillary sinus.

Table 4.5 Results of classification performance using MIE [15] and varying N

N	MIE	AUPRC	F1
1		0.80±0.12	0.70±0.13
5		0.85±0.03	0.77±0.10
5	✓	0.87±0.04	0.76±0.10
10		0.85±0.04	0.75±0.08
10	✓	0.89±0.05	0.79±0.10
15		0.88±0.07	0.81±0.12
15	✓	**0.92±0.06**	**0.85±0.09**
20		0.87±0.04	0.77±0.05
20	✓	0.91±0.02	0.78±0.07

of these analyses are detailed in Table 4.8 and Table 4.9. Figure 4.4 shows the maxillary sinus conditions that cause low confidence in our EM. Detailed analysis on the statistical tests performed as well as additional details on the cohort sizes between *control* and *cases* group for different lifestyle and allergy variables can be found in our paper [12].

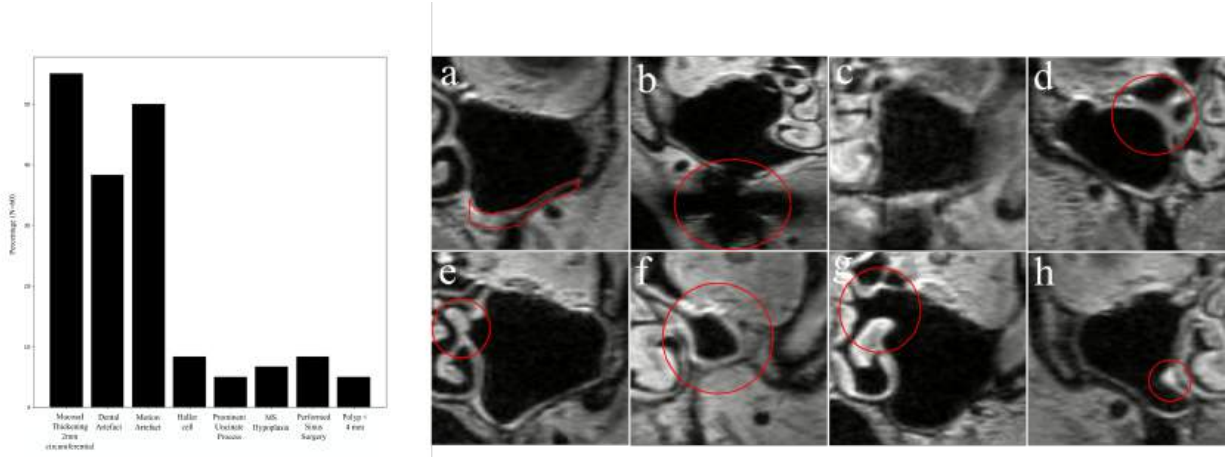

Figure 4.4 Analysis of low confidence scored maxillary sinus [12] (LEFT) Bar chart of special conditions causing low confidence scores of CNN (RIGHT) The 3D CNN predicts confidence scores below 0.90 for the following scenarios: (a) Maxillary sinuses exhibiting mucosal thickening of around 2mm. (b) maxillary sinuses affected by dental artifacts. (c) maxillary sinuses with motion artifacts. (d) maxillary sinuses featuring Haller cells. (e) maxillary sinuses displaying prominent uncinate processes. (f) maxillary sinuses showing signs of hypoplasia. (g) maxillary sinuses undergoing post-surgical changes. (h) maxillary sinuses hosting polyps smaller than 4mm. Regions of interest are demarcated using a red contour.

Table 4.6 Results of classification performance using MIE for different deep neural network architectures [15]

CNN	N	MIE	AUPRC	F1
3D ResNet18	1		0.80±0.12	0.70±0.13
3D ResNet18	15		0.88±0.07	0.81±0.12
3D ResNet18	15	✓	0.92±0.06	0.85±0.09
3D ResNet50	1		0.72±0.13	0.59±0.19
3D ResNet50	15		0.82±0.11	0.71±0.19
3D ResNet50	15	✓	0.85±0.07	0.74±0.13
3D ResNet101	1		0.73±0.10	0.59±0.04
3D ResNet101	15		0.85±0.04	0.69±0.10
3D ResNet101	15	✓	0.90±0.06	0.79±0.14
3D ResNet152	1		0.66±0.06	0.57±0.08
3D ResNet152	15		0.83±0.07	0.76±0.11
3D ResNet152	15	✓	0.89±0.05	0.80±0.08
3D ResNet200	1		0.60±0.21	0.45±0.39
3D ResNet200	15		0.86±0.05	0.79±0.10
3D ResNet200	15	✓	0.90±0.04	0.83±0.07
3D DenseNet121	1		0.86±0.11	0.80±0.07
3D DenseNet121	15		0.86±0.10	0.81±0.06
3D DenseNet121	15	✓	0.92±0.05	0.83±0.12
3D DenseNet169	1		0.81±0.09	0.76±0.11
3D DenseNet169	15		0.91±0.05	0.82±0.04
3D DenseNet169	15	✓	0.94±0.03	0.86±0.09
3D DenseNet201	1		0.88±0.07	0.72±0.07
3D DenseNet201	15		0.88±0.04	0.72±0.08
3D DenseNet201	15	✓	0.93±0.06	0.78±0.07
3D DenseNet264	1		0.84±0.09	0.81±0.07
3D DenseNet264	15		0.88±0.05	0.82±0.12
3D DenseNet264	15	✓	0.93±0.01	0.85±0.09

Table 4.7 Performance Metrics of the ensemble model [12]

Metric	**CV 1**	**CV 2**	**CV 3**	**EM**
Precision	0.82	0.86	0.83	0.85
Sensitivity	0.85	0.77	0.77	0.85
Specificity	0.88	0.91	0.89	0.90
F1	0.84	0.81	0.80	0.85
AUROC	0.93	0.92	0.92	0.95

Table 4.8 Comparison of control and cases in at least one MS with respect to health and lifestyle factors [12]

Variable	Control	Cases	P-value
Smoking habits			
	N=1245	**N=1162**	
Yes	218 (17.51%)	202 (17.38%)	0.97
No	1027 (82.49%)	960 (82.62%)	
Alcohol consumption (g/day)			
	N=1170	**N=1089**	
Mean (95% CI)	15.56 (-30.18 - 61.31)	20.65 (-34.83 - 76.15)	$6.79 \times e^{-8}$
BMI			
	N=1210	**N=1126**	
Mean (95% CI)	26.26 (17.33 - 35.21)	27.11 (18.99 - 35.33)	$3.85 \times e^{-6}$
Intrinsic Asthma			
	N=1156	**N=1072**	
Yes	78 (6.75%)	99 (9.24%)	0.03
No	1078 (93.25%)	973 (90.76%)	
Extrinsic Asthma			
	N=1162	**N=1053**	
Yes	66 (5.68%)	89 (8.45%)	0.01
No	1096 (94.32%)	964 (91.55%)	
Chronic bronchitis or COPD			
	N=1155	**N=1069**	
Yes	67 (5.8%)	74 (6.92 %)	0.31
No	1088 (94.2%)	995 (93.08%)	
Sex			
	N=1249	**N=1165**	
Male	544 (43.55%)	795 (68.24%)	$5.5 \times e^{-34}$
Female	705 (56.45%)	370 (31.76%)	
Age (years)			
	N=1249	**N=1165**	
Mean (95% CI)	63.97 (47.5 - 80.44)	64.01 (47.83 - 80.2)	0.90
LK			
	N=1220	**N=1134**	
Mean (95% CI)	6.19 (2.24 - 10.15)	6.21 (2.77 - 9.66)	0.76
hCRP			
	N=1214	**N=1126**	
Mean (95% CI)	0.22 (-0.64 - 1.1)	0.23 (-0.54 - 1.0)	0.91

Table 4.9 Comparison of control and cases in at least one MS for different allergies [12]

Variable	Control	Cases	P-value
Hay fever			
	N=1173	**N=1085**	
Yes	224 (19.1%)	272 (25.07%)	0.0007
No	949 (80.9%)	813 (74.93%)	
Bee/wasp venom allergy			
	N=1110	**N=1038**	
Yes	63 (5.68%)	37 (3.56%)	0.02
No	1047 (3.56%)	1001 (96.44%)	
Food allergy			
	N=1134	**N=1053**	
Yes	102 (8.99%)	107 (10.16%)	0.39
No	1032 (91.09%)	946 (89.84%)	
House dust allergy			
	N=1145	**N=1049**	
Yes	102 (8.91%)	124 (11.82%)	0.02
No	1043 (91.09%)	925 (88.18%)	
Allergy to animal hair			
	N=1152	**N=1065**	
Yes	85 (7.38%)	96 (9.01%)	0.18
No	1067 (92.62%)	969 (90.99%)	
Contact allergy			
	N=1132	**N=1052**	
Yes	73 (6.45%)	49 (4.66%)	0.08
No	1059 (93.55%)	1003 (95.34%)	
Medication allergy			
	N=1110	**N=1027**	
Yes	154 (13.87%)	132 (12.85%)	0.52
No	956 (86.13%)	895 (87.15%)	
Other allergy			
	N=1094	**N=1006**	
Yes	93 (8.5%)	81 (8.05%)	0.76
No	1001 (91.5%)	925 (91.95%)	

Chapter 5

Discussion and Conclusion

Paranasal anomalies, frequently discovered incidentally during routine radiological screenings, hold significant clinical relevance. Identifying their prevalence not only informs individual treatment plans, often involving surgical interventions, but also serves as a crucial metric for assessing sinonasal health within broader population cohorts [30]. Particularly in the latter context, correlating symptomatic and asymptomatic presentations with these anomalies aims to establish causal relationships [49]. However, the reliance on experienced radiology specialists for diagnosis presents inherent challenges, especially within large cohorts, demanding substantial manpower and resources. This becomes especially burdensome in prospective studies, where repeated manual diagnoses drain labor and resources. Consequently, CAD systems emerge as a promising solution. Automating diagnosis through CAD not only ensures accuracy and reliability but also alleviates clinician fatigue, marking just one of its many potential advantages.

Within the realm of CAD systems for paranasal anomaly detection, supervised learning has been the dominant approach, as evidenced by prior studies [65, 95, 60, 55, 81]. While some endeavors have utilized conditional GANs to augment datasets, these efforts remained within the scope of supervised learning [73]. Despite the prevalence of supervised methods, there is a growing need to explore alternative approaches such as unsupervised learning,

contrastive learning, and self-supervised learning. These avenues hold promise for yielding superior representations and potentially enhancing generalization. Each of these learning paradigms presents distinct advantages. Unsupervised learning reduces dependency on labelled datasets. Contrastive learning, even within supervised settings, facilitates improved feature learning from the same training samples, potentially enhancing generalization to test sets. Self-supervised learning harnesses unlabelled datasets to develop representations beneficial for subsequent supervised training. Moreover, architectural adaptations in deep learning networks could also yield advantages by fostering the acquisition of better features compared to conventional convolutional architectures.

We investigated UAD for the detection of maxillary sinus anomalies. In our approach, we exclusively utilized healthy maxillary sinus data to train CAE and VAE. Our findings revealed relatively poor classification performance, as outlined in Table 4.1. Several factors contribute to this outcome. Primarily, the diverse morphologies of opacifications pose a challenge. For instance, opacifications like mucosal thickening exhibit thickened walls, making them easily mistaken for a healthy maxillary sinus. Polyps and cysts vary in size, location and intensity within sinus walls, rendering judgment based solely on mean reconstruction error challenging. Additionally, our method employed a greedy search to determine the optimal threshold for classification, utilizing a validation set containing a limited number of maxillary sinuses with opacifications. This limited representation of the opacifications in the validation set may have led to sub-optimal thresholds, impacting classification accuracy. However, it is essential to note that the primary advantage of employing UAD was not solely for classification purposes. As depicted in Figure 4.1, the reconstruction errors served as indicators of poorly reconstructed areas, correlating with potential opacification locations. This aspect offers a valuable benefit by generating a coarse segmentation map of maxillary sinus opacifications, obviating the need for expensive ground truth annotations. We used this coarse segmentation map to our advantage in our work on SSL.

For the SSL task, we utilized the CAE trained in UAD framework to generate coarse

segmentation masks from our unlabeled dataset. These 'residual volumes' were used in our SSL task where a UNet style CNN was required to reconstruct the residual volumes given the corresponding maxillary sinus volume. It is to be noted that we enhanced our SSL task by mitigating spurious reconstruction errors through median filtering, resulting in performance improvement. Subsequently, after training the UNet-style encoder-decoder to reconstruct the residual volume, we fine-tuned the encoder specifically for our downstream classification task. Our systematic exploration encompassed scenarios simulating limited labelled datasets, revealing that in extremely constrained scenarios (e.g., 10% and 20% labelled data), our SSL-trained encoder outperformed state-of-the-art SSL methods, as detailed in Table 4.2. Our approach centered on a single hypothesis: utilizing SSL to localize opacifications could offer distinct advantages for subsequent classification tasks. This hypothesis was substantiated, demonstrating superior performance compared to transfer learning and generic SSL methods like autoencoding, denoising autoencoding, SimCLR, SimSiam, BYOL and MAE.

Beyond investigating unsupervised and self-supervised learning approaches, we proposed architectural modifications' potential benefits for paranasal anomaly classification. Introducing a novel deep learning network amalgamating CNNs and transformers, we aimed to leverage their individual advantages. CNNs inherently possess an induction bias suited for image classification, yet their early layer features exert minimal influence on later layers. In contrast, vision transformers, with their multi-headed self-attention mechanism, facilitate long-range dependencies among features. However, their computational demands and data hunger limit their effectiveness without extensive pretraining. Hence, we proposed ConTra-Net, integrating a CNN backbone with multi-headed self-attention blocks to enable interactions between low and high-level features, as depicted in Figure 3.6. This facilitated the formation of long-range dependencies, ultimately fostering superior representation learning, detailed in Table 4.3.

Beyond architectural modifications, our exploration led to enhancing representation learning through contrastive loss within a supervised setting. Conventionally, supervised learning

relies on cross-entropy loss, encouraging the learning of discriminative features crucial for classification. However, extensive research highlights the limitations of cross-entropy loss in fostering intra-class compactness and inter-class separability [82, 85, 33]. Supervised contrastive loss [68] emerged as a remedy, addressing these limitations by aligning features from images within the same class while pushing apart features from different classes, fostering intra-class compactness and inter-class separation. In our study, the aim was twofold: harnessing discriminative features while ensuring intra-class compactness and inter-class separation. To achieve this, we proposed a CNN employing two Multi-Layer Perceptron (MLP) heads from a single backbone. One MLP head facilitated contrastive learning through the supervised contrastive loss, while the other handled cross-entropy learning, outlined in Figure 3.7. This combined approach yielded significant advantages, resulting in the acquisition of enhanced representations, ultimately improving generalization performance on the test set, as detailed in Table 4.4.

The most impactful method we discovered involved refining the data processing pipeline and employing an instance-level ensemble strategy. Similar to previous studies [65, 55, 60], our data processing followed a two-stage approach. However, unlike prior methods reliant on DL for region-of-interest extraction, we introduced a non-DL approach tailored to extract maxillary sinus volumes. Furthermore, our innovation lay in a multiple instance ensembling approach. We extracted multiple candidate maxillary sinuses from each side (left and right) of an individual's MRI, each sinus deliberately offset from the others. This approach conferred dual advantages: augmenting the dataset by introducing diverse candidate sinuses and enabling ensemble learning during test time via a single CNN model. This strategy notably improved classification performance. The enhancement stemmed from two primary factors: the enlarged and diversified training dataset and the test-time ensemble, which effectively is a version of test-time augmentation. Collectively, our distinctive data processing and multiple instance learning approach emerged as the most beneficial strategy for maxillary sinus opacification classification.

Employing the multiple instance learning approach, we incorporated an ensemble model within a 3-fold cross-validation framework to enhance classification robustness, detailed in Table 4.7. This ensemble model facilitated the creation of distinct *control* and *cases* groups integrating both labeled and unlabeled datasets. Our analysis, outlined in Table 4.8, revealed noteworthy associations between health and lifestyle factors. Notably, males within the 'cases' group exhibited a higher incidence of maxillary sinus opacifications, aligning with findings in prior studies [49, 131, 30, 51]. Additionally, a higher prevalence of opacifications correlated with an increased risk of allergic rhinitis [134]. Participants within the *cases* group displayed elevated Body Mass Index and higher alcohol consumption, possibly linked to a higher male percentage. Surprisingly, no significant correlation between the *cases* group and smoking habits emerged, consistent with prior research [49, 131, 30, 51]. Similarly, our results did not indicate substantial associations between age, blood parameters, and maxillary sinus opacifications, with limited existing literature exploring these connections. Further correlation analysis regarding allergy-related variables, detailed in Table 4.9, indicated increased tendencies for hay fever and house dust allergy among participants with maxillary sinus opacifications, findings not extensively explored in prior studies. In summary, our CNN-based study using multiple instance ensembling demonstrates rapid computation of correlations with reduced labelling effort, showcasing the practical utility of the methods outlined in this dissertation.

In our clinical work [12], we analyzed the volumes of maxillary sinuses that did not meet our predefined inclusion criteria. This involved manually scrutinizing the maxillary sinus conditions that the CNN identified with confidence scores below 0.90 for either normal or anomalous classifications. Out of the 190 MRI scans excluded from our study, we randomly selected 60 (31.5%) for a more detailed examination. Our goal was to understand why these scans were misclassified compared to those correctly identified by our CNN. We identified several factors contributing to the lower confidence scores (see Figure 4.4). Notably, 55% (33 out of 60) of the participants exhibited maxillary sinuses with mucosal thickening of at

least 2mm in circumference in at least one sinus. Additionally, 38.3% (23 out of 60) of the MRI scans displayed artefacts such as missing features in the maxillary sinus images due to dental restorations like crowns, fillings, and orthodontic appliances. Motion during imaging sessions also led to blurred features in some scans, further reducing confidence scores, affecting a total of 50% (30 out of 60) of the MRI scans. Furthermore, we observed anatomical variations in the maxillary sinuses contributing to the lower confidence scores. Notable anomalies included Haller cell (8.33% or 5 out of 60), variations in the uncinate process (5% or 3 out of 60), and hypoplasia (6.67% or 4 out of 60). Surgical interventions also introduced deformations in the maxillary sinuses, resulting in reduced confidence in 8.33% (5 out of 60) of the cases. Additionally, 5% (3 out of 60) of the MRIs exhibited pathological conditions such as polyps in the maxillary sinuses under 4 mm. Importantly, we did not identify any new pathologies that were not already present in our labeled dataset. Our analysis highlights that the primary factors contributing to the lower confidence scores were MRI artefacts and instances of maxillary sinus thickening of 2mm circumferential. These cases represent ambiguous scenarios, displaying characteristics of both normal and abnormal in the maxillary sinuses. To address these challenges, we propose several strategies. One potential approach is to redefine mucosal thickening cases as those with swelling significantly greater than 2mm, for example, 5mm. This adjustment aims to accentuate the disparity between normal and anomalous maxillary sinuses within the pixel space, potentially improving diagnostic accuracy and confidence levels. Moreover, addressing artefacts emerged as crucial. Proposed strategies include utilizing data augmentation techniques to simulate motion and dental artefacts and expanding the training dataset to include MRI scans with representative artefacts. However, it is essential to note that dental artefacts may lead to the removal of relevant image information around the maxillary sinuses, making diagnosis challenging for such cases. Additionally, addressing anatomical variations, including rare ones like Haller cells, uncinate process variations, hypoplasia, and surgical interventions, could be achieved by augmenting the training dataset to incorporate these variations in training the CNN.

Finally, addressing small polypoid masses under 4mm, which also exhibited low confidence scores, could involve increasing the inclusion of such cases in the dataset, adjusting the decision threshold, or generating synthetic maxillary sinus images with small polypoid masses using generative deep learning methods. Figure 4.4 shows the bar chart and the sample maxillary sinus exhibiting the special conditions that cause low confidence scores by CNN.

This study is subject to several limitations. Firstly, the methods discussed are not compared using the same dataset due to the labeling process spanning three years. The UAD (Section 3.3.2) and contrastive learning-based method (Section 3.3.5) were initially assessed over a dataset of 199 participants. Subsequently, 100 additional patients were annotated, leading to exploratory analyses through architectural modifications (Section 3.3.4) and MIE (Section 3.3.6). The dataset was ultimately expanded to include 1069 labeled participants, enabling SSL experiments (Section 3.3.3) and a population study (Section 4.6). Secondly, despite methodological advancements, cases of misclassification persist, as outlined in Section 3.3.6, necessitating careful consideration in prospective studies. Lastly, this study is retrospective, underscoring the importance of prospective investigations to gain a deeper understanding of failure and edge cases and address these issues effectively.

This study aimed to identify effective methods for paranasal anomaly classification under conditions of limited labeled data and in scenarios where both labeled and unlabeled datasets are available simultaneously. The dissertation explores various learning mechanisms, including unsupervised, self-supervised, and supervised approaches, introducing a novel deep learning architecture. The dissertation assesses the advantages and limitations of each method in the classification of maxillary sinus paranasal anomalies. Furthermore, we showcase the practical applicability of one of the most effective methods by swiftly identifying correlations with critical clinical variables. Our work stands out by replacing labor-intensive manual diagnoses with rapid automation, granting immediate access to crucial clinical insights. Deep learning-based CAD systems not only improve the diagnosis of maxillary sinus opacifications but also expedite extensive population studies. This rapid insight generation

capability holds promise for real-time monitoring and analysis, especially as cohort sizes expand in prospective studies, enabling effective, long-term health monitoring of sizable populations.

Chapter 6

Future Work

While this dissertation presents various methodologies, several promising research directions warrant further exploration.

6.1 Multi-class classification

Our study simplified the problem into a binary classification, consolidating all opacification types (polyps, cysts, mucosal thickening) into a single class. However, future investigations into classifying specific types of opacifications could yield valuable insights. This approach could offer a more nuanced understanding of observed opacifications and enable correlation analyses with existing variables, potentially providing deeper insights into the phenomenon.

6.2 Exploration into Other Sinuses

Our investigation focused solely on maxillary sinus opacifications, primarily due to the higher prevalence observed within our dataset compared to other sinuses. While our work centered on the maxillary sinus, previous studies considered sinusitis classification, encompassing frontal and ethmoid sinuses [65]. Some works have even explored frontal sinus segmentation,

yet diagnostic attempts for opacifications remain unexplored [147]. Looking ahead, developing deep learning-based CAD systems for other sinuses, such as the frontal, ethmoid, and sphenoid sinuses, holds promise for a comprehensive MRI analysis of an individual. Such systems could potentially yield more profound clinical insights, correlating with relevant clinical variables and offering a more holistic perspective on paranasal anomalies.

6.3 Exploration of 2D and 2.5D Architectures

Throughout this dissertation, all discussed methodologies revolved around 3D CNNs processing 3D data. However, this approach comes with potential downsides, including increased model parameter count and a risk of overfitting. To mitigate these challenges, alternative strategies like 2D and 2.5D architectures warrant exploration. The 2D approach involves processing individual MRI image slices sequentially, offering a lightweight model with the potential for utilizing multiple slices extracted from coronal, axial, and sagittal planes, thereby expanding the training dataset and potentially enhancing paranasal anomaly classification. The 2.5D approach merges information across multiple 2D slices, harnessing the advantages of 3D models while maintaining the efficiency of 2D models. Investigating both these approaches holds promise for improving generalization capabilities in paranasal anomaly classification.

6.4 Synthetic Data Generation Using Generative Models

In the domain of paranasal anomaly classification, prior work utilized GANs to generate synthetic normal and sinusitis-affected maxillary sinus images [73]. However, these GANs were conditioned at a class level, lacking control over the specific morphology of the sinusitis. A compelling avenue for exploration lies in semantic image synthesis, where sinus generation is conditioned on segmentation maps containing opacification location information. This

approach enables fine-grained control over the generation process, facilitating the creation of opacifications of specific sizes and diverse locations within the sinus walls. Training on a blend of synthetic and real data could potentially enhance generalization to unseen opacifications. To achieve this, diverse methods can be investigated, such as GANs conditioned on segmentation maps [106] or diffusion models [140]. Additionally, exploring both 2D and 3D generation techniques would be valuable in assessing fidelity and realism in the generated images.

6.5 Closing Remarks

The realm of deep learning and artificial intelligence continues to advance rapidly, especially within medical imaging applications. Deep learning's modality-agnostic nature and its innate capacity to discern intricate data patterns have led to its widespread utilization across various medical imaging domains. This study marks a significant stride in developing tailored deep learning models for paranasal anomalies. We extensively explored supervised learning approaches, enhancing them with diverse techniques, introduced a self-supervision task tailored for paranasal anomalies, and scrutinized the potential advantages of unsupervised learning. Furthermore, we applied one of our proposed methods in a clinical context to unveil correlations with clinical variables. We shed light on the challenges and merits inherent in each approach. We hope that this work serves as a catalyst, stimulating further research and advancements in the field of paranasal anomalies.

Chapter 7

Reseach Papers

7.1 Unsupervised anomaly detection of paranasal anomalies in the maxillary sinus

Title of paper: Unsupervised anomaly detection of paranasal anomalies in the maxillary sinus

Conference: Society of Photographic Instrumentation Engineers Medical Imaging

Year: 2023

Topic: Unsupervised Anomaly Detection

PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

Unsupervised anomaly detection of paranasal anomalies in the maxillary sinus

Debayan Bhattacharya, Finn Behrendt, Benjamin Tobias Becker, Dirk Beyersdorff, Elina Petersen, et al.

Debayan Bhattacharya, Finn Behrendt, Benjamin Tobias Becker, Dirk Beyersdorff, Elina Petersen, Marvin Petersen, Bastian Cheng, Dennis Eggert, Christian Betz, Anna Sophie Hoffmann, Alexander Schlaefer, "Unsupervised anomaly detection of paranasal anomalies in the maxillary sinus," Proc. SPIE 12465, Medical Imaging 2023: Computer-Aided Diagnosis, 124651B (7 April 2023); doi: 10.1117/12.2651525

SPIE.

Event: SPIE Medical Imaging, 2023, San Diego, California, United States

Unsupervised Anomaly Detection of Paranasal Anomalies in the Maxillary Sinus

Debayan Bhattacharya[a], Finn Behrendt[b], Benjamin Tobias Becker[c], Dirk Beyersdorff[d], Elina Petersen[e], Marvin Petersen[f], Bastian Cheng[g], Dennis Eggert[h], Christian Betz[i], Anna Sophie Hoffmann*[j], and Alexander Schlaefer*[k]

[a,b,k]Hamburg University of Technology, Hamburg, Germany
[a,c,d,e,f,g,h,i,j] Universitätsklinikum Hamburg-Eppendorf, Hamburg, Germany

ABSTRACT

Deep learning (DL) algorithms can be used to automate paranasal anomaly detection from Magnetic Resonance Imaging (MRI). However, previous works relied on supervised learning techniques to distinguish between normal and abnormal samples. This method limits the type of anomalies that can be classified as the anomalies need to be present in the training data. Further, many data points from normal and anomaly class are needed for the model to achieve satisfactory classification performance. However, experienced clinicians can segregate between normal samples (healthy maxillary sinus) and anomalous samples (anomalous maxillary sinus) after looking at a few normal samples. We mimic the clinicians ability by learning the distribution of healthy maxillary sinuses using a 3D convolutional auto-encoder (cAE) and its variant, a 3D variational autoencoder (VAE) architecture and evaluate cAE and VAE for this task. Concretely, we pose the paranasal anomaly detection as an unsupervised anomaly detection problem. Thereby, we are able to reduce the labelling effort of the clinicians as we only use healthy samples during training. Additionally, we can classify any type of anomaly that differs from the training distribution. We train our 3D cAE and VAE to learn a latent representation of healthy maxillary sinus volumes using L1 reconstruction loss. During inference, we use the reconstruction error to classify between normal and anomalous maxillary sinuses. We extract sub-volumes from larger head and neck MRIs and analyse the effect of different fields of view on the detection performance. Finally, we report which anomalies are easiest and hardest to classify using our approach. Our results demonstrate the feasibility of unsupervised detection of paranasal anomalies from MRIs with an AUPRC of 85% and 80% for cAE and VAE, respectively.

Keywords: paranasal anomaly, unsupervised anomaly detection, autoencoder, VAE

1. INTRODUCTION

Anomalies occurring in the paranasal sinuses are commonly reported in patients who undergo neuroradiological assessment of the head using diagnostic imaging.[1] These incidental findings pose clinical challenges[2] and we have limited knowledge on the importance of these reported findings on the general population. To this end, numerous studies have been done to analyse the significance of these findings.[3–7] Most of these works are population studies involving large sample sizes which are manually annotated by clinicians. In a three year retrospective study, it was observed that malignant tumours and inverted papillomas were classified as nasal polyps with a misdiagnosis rate of 5.63% and 8.45% respectively. Therefore, computer aided diagnosis systems (CADx)[8–10] have been proposed with the idea of working in conjunction with the clinician to reduce the misdiagnosis rate of paranasal anomalies. However, all these works rely on supervised learning and consider at most one anomaly. Apart from requiring large labelled datasets, supervised learning models also need labelled data that are representative of the classes for accurate prediction.[11] Further, the type of anomaly to be classified has to be decided beforehand. In our case, we consider three anomalies, namely: (i) mucosal thickening (ii) polyps (iii) cysts. This is particularly challenging in our case where the considered anomalies are known to have high intra-class morphological variations[12–14] and have unequal occurrences in our dataset.

Further correspondence, please send an email to: debayan.bhattacharya@tuhh.de
* These authors contributed equally.

Medical Imaging 2023: Computer-Aided Diagnosis, edited by Khan M. Iftekharuddin, Weijie Chen, Proc. of SPIE Vol. 12465, 124651B · © 2023 SPIE
1605-7422 · doi: 10.1117/12.2651525

In light of the aforementioned points, we are motivated to perform Unsupervsised Anomaly Detection (UAD) using autoencoders. Autoencoders are a common choice for data compresison and outlier detection.[15] In medical imaging, brain anomaly detection and segmentation[16] has gained popularity over the years. The underlying concept of reconstruction-based UAD is that the autoencoder learns to compress and reconstruct only normal images during training. The assumption is that during testing, reconstruction errors will be low for normal images whereas anomalous images will have a large reconstruction error as the autoencoder will fail to properly reconstruct anomalous regions. In our case, we use cAE and VAE learn to compress and reconstruct healthy maxillary sinus (MS) volumes. Through our approach, we derive multiple benefits, namely: (i) our autoencoder becomes indifferent to the anomaly distribution thereby allowing detection of more than one anomaly, (ii) we reduce the labelling effort of the clinicians as the training dataset does not require anomalous samples, (iii) we are able to generate a heat map based on the reconstruction error between the original volume and the reconstructed volume. The heat maps visualize the region of the potential anomaly and highlight it. This may prove to be beneficial to the clinician when making a diagnosis.

In summary, our contributions are three-fold. First, we pose the paranasal anomaly detection problem as a UAD problem and thereby become indifferent to the anomaly distribution. Second, we systematically evaluate our cAE and VAE approach on MS volumes with different field of view. Third, we report which anomalies are easiest and the hardest to classify using our UAD method.

2. METHODS

2.1 Dataset and Implementation Details

Our labelled dataset consists of head and neck MRIs of 199 patients. Each MRI is a fluid attenuated inversion recovery (FLAIR) MRI. Our labelled dataset is part of the Hamburg City Health Study.[17] Out of the 199 patients, 93 patients exhibit one or multiple anomalies in at least left or right MS. Two Ear, Nose and Throat (ENT) surgeons and one ENT specialised radiologist confirmed the diagnosis of the observed pathology. We group the 3 anomalies into a single class called "anomaly" and the normal MS are categorized into "normal" class. Altogether, we have 269 normal MS volumes and 130 anomalous MS volumes . Each MRI has a resolution of 173 x 319 x 319 voxels along the saggital, coronal and axial directions respectively with each voxel of size 0.53 mm x 0.75 mm x 0.75 mm.

Preprocessing : We randomly selected a FLAIR MRI as the fixed MRI and performed rigid registration on the remaining MRIs. This was followed by resampling to a dimension of 128 $\times$ 128 $\times$ 128. We extracted two sub-volumes from the resampled head and neck MRIs, one for each MS. The extracted sub-volumes were of sizes: 33 $\times$ 47 $\times$ 45 (small), 46 $\times$ 57 $\times$ 55 (medium) and 53 $\times$ 67 $\times$65 (large). Owing to the symmetry of left and right MS, we horizontally flipped the coronal planes of right MS to give it the appearance of left MS for each patient. Finally, these sub-volumes were reshaped to a standard size of 64 $\times$ 64 $\times$ 64 voxels for the 3D cAE and VAE. All the MS volumes were normalised to a range of 0 to 1. Our preprocessing pipeline is illustrated in figure 1 (a,b)

Data split: Our training set contains 172 normal MS volumes , validation set contains 43 normal MS volumes and 52 anomalous MS volumes and test set contains 54 normal MS volumes and 78 anomalous MS volumes. We perform a three-fold cross validation split for all our experiments.

Implementation Details: We use PyTorch[18] and PyTorch Lightning[19] for all our experiments. We use batch size $N = 16$ and latent dimension $n_z = 512$ for the cAE. We use a learning rate of $1e^{-4}$ and Adam optimizer[20] with default parameters to train our cAE and VAE. We run all our experiments for 100 epochs.

2.2 Deep Learning Methods

Similar to Bengs *et al*,[21] we extend our architecture to 3D as it has shown to improve the detection performance in 3D MRI scans. Our cAE and VAE architectures are shown in figure 1 (i) . There are two stages to our approach. First, we train our cAE and VAE to learn the distribution of X_h where X_h represents healthy MS volumes. For training our cAE, we use the L1 reconstruction loss as described below:

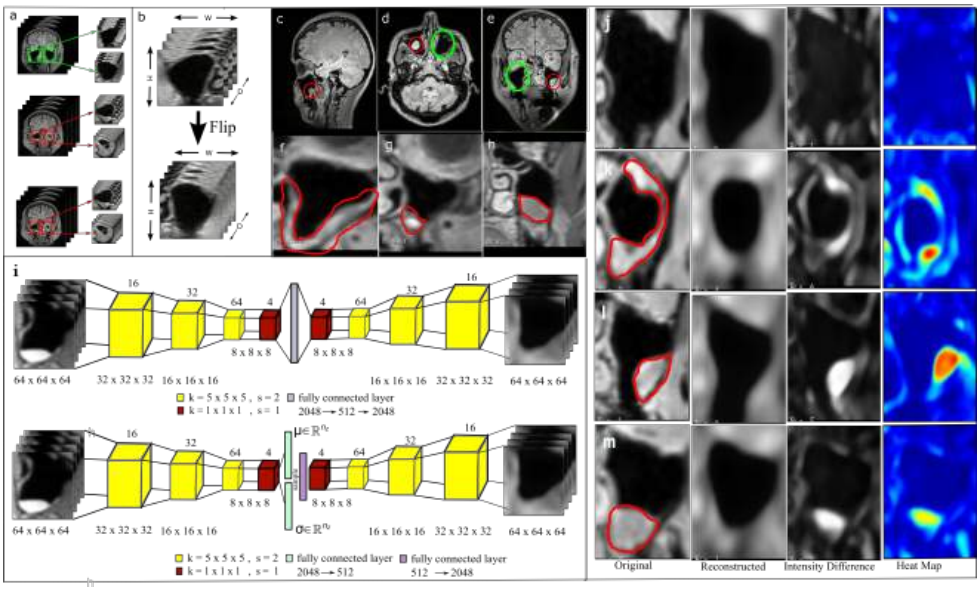

Figure 1. (a) Extraction of left and right MS from head and neck MRI (b) Flipping of the coronal plane of right MS (c) Cyst in the right MS (d) Polyp in the left MS (e) Cyst in the left MS (f) MS (small) showing mucosal thickening (g) MS (medium) showing polyp (h) MS (large) showing cyst (i) TOP: Our cAE architecture with latent vector of size 512 used as bottleneck. BOTTOM: Our VAE architecture. In both the cAE and VAE decoders, we perform 3D convolution followed by trilinear upsampling. ReLU is used as non-linear activation function in both the networks. (j)-(m) Images extracted from original , reconstructed , intensity difference and heat map volumes. (j) is an image extracted from normal MS volume. (k),(l),(m) are mucosal thickening, polyps and cyst anomalies respectively. The red markings denote anomalies and green circles denote normal MS.

$$L_1 = \sum_{k=1}^{N} |x^k - \hat{x}^k| \tag{1}$$

Here, x^k and $\hat{x}^k$ denote the k-th MS volume and reconstructed MS volume respectively. N denotes the mini-batch size. For training our VAE, we use L1 reconstruction loss and KL Divergence. While training, the VAE learns a mean μ_z and variance σ_z from which a sample is drawn and reconstructed. The loss used to train our VAE is shown below:

$$L_{VAE} = L_1 + \lambda_{KL} D_{KL}(q(z|x)||p(z)) \tag{2}$$

Here, $D_{KL}(.||.)$ represents the Kullback–Leibler divergence between the parameterized latent distribution $q(z|x) \sim N(\mu_z, \sigma_z)$ and the prior p(z) which follows a multivariate normal distribution. $z \in \mathbb{R}^{n_z}$ represents the latent vector. λ_{KL} is a Lagrangian multiplier and we have set it to 1 for our experiments. VAE projects the the input MS volume to $q(z|x)$ and KL-Divergence loss attempts to bring it close to a prior $p(z)$.[22]

Second, we use the trained cAE and VAE to reconstruct the MS volumes in the validation set. For each MS volume, we calculate the L1 and L2 reconstruction loss denoted as t_{L1} and t_{L2} respectively. We choose the optimal thresholds by plotting the precision recall curve and select the threshold with the highest F1 score. During inference, $\hat{x}^k$ with $L_1 > t_{L1}$ and $L_2 > t_{L2}$ is classified as anomalous MS volume. Here, L_2 is the L2 reconstruction loss defined as follows:

$$L_2 = \sum_{k=1}^{N} (x^k - \hat{x}^k)^2 \tag{3}$$

Additionally, for the MS volumes classified as anomalous, a further analysis is done by calculating voxel-wise intensity difference $D_k = |x^k - \hat{x}^k|$ after which a median filter of kernel size 5 is applied on it to remove sporadic

Table 1. Anomaly Detection Performance on two thresholds t_{L1} and t_{L2} where positive labels are assigned to the anomalous class.

Method	MS Size	Precision		Recall		F1		AUPRC	
		t_{L1}	t_{L2}	t_{L1}	t_{L2}	t_{L1}	t_{L2}	t_{L1}	t_{L2}
VAE	small	0.69	0.76	0.63	0.62	0.64	0.68	0.76	0.80
VAE	medium	0.63	0.63	0.84	0.91	0.72	0.75	0.70	0.75
VAE	large	0.61	0.64	0.81	0.86	0.70	0.73	0.65	0.69
cAE	small	0.75	0.81	0.74	0.66	0.74	0.73	0.81	0.83
cAE	medium	0.73	0.77	0.62	0.74	0.67	0.75	0.80	0.85
cAE	large	0.68	0.74	0.82	0.73	0.74	0.73	0.73	0.78

reconstruction errors. Finally, a heat map is rendered for the individual slices along the coronal, axial and sagittal planes as shown in figure 1 (j, k, l, m). The regions in red denote the regions which the cAE failed to reconstruct. Our qualitative results indicate an overlap between the anomalous regions and the poorly reconstructed regions of MS volumes.

3. RESULTS

Table 2. Accuracy per anomaly on the test set reported in percentage (%). Here, # refers to the number of correctly classified samples divided by the total samples from a particular category.

Method	MS Size	Normal (%/#)	Mucosal Thickening (%/#)	Polyps (%/#)	Cysts (%/#)
VAE	small	0.48 (26/54)	0.75 (22/29)	0.82 (28/34)	0.73 (11/15)
cAE	medium	0.61 (33/54)	0.62 (18/29)	0.91 (31/34)	0.8 (12/15)

The anomaly detection performance is shown in table 1. We consider the Area Under Precision Recall Curve (AUPRC) to evaluate our classifiers as we have an unbalanced test set. We rank the classifiers based on AUPRC. We report the mean values of the mentioned metrics. In terms of AUPRC, the VAE that uses small MS volume and cAE that uses medium MS volume are the best performing classifiers.

From table 1, we notice that the all the VAEs have relatively lower AUPRC in comparison to cAEs. Furthermore, our results show that considering L2 loss when computing the anomaly score and using t_{L2} as threshold leads to better performance for all cAEs and VAEs. In table 2, we report the accuracy per anomaly in terms of percentage and number of occurances for the best performing VAE and cAE from table 1. We observe that polyps are the easiest to classify, followed by cysts. Mucosal thickening anomaly is the hardest to classify for both VAE and cAE.

4. DISCUSSION AND CONCLUSION

From the results in table 1 we observe that using t_{L2} as threshold leads to better performance. This can be attributed to the fact that L2 loss penalizes voxel-wise intesity outliers more heavily than L1 loss. Therefore, even small regions of poor reconstruction in the MS volume can amount to high overall reconstruction error. Additionally, our accuracy percentage per anomaly reported in table 2 shows that mucosal thickening anomalies are the most difficult to classify. We believe this to be the case because unlike polyps and cysts which occur as visible masses in the MRI (See figure 1 (l),(m)), mucosal thickening anomalies have more subtle appearances as these anomalies mostly cause inflammation of the mucosal walls. Therefore, unless the inflammations are too noticeable, they almost have the appearance of a healthy MS. We also observe that classifiers using small and medium MS volumes have the best AUPRCs. This can be attributed to the fact that cropping large volumes lead to inclusion of unnecessary surrounding anatomical structures outside of the MS. These additional anatomical structures effect the reconstruction error and thereby, we end up selecting sub-optimal thresholds. Our work has

some limitations, one being the accuracy of healthy MS volume detection needs to be higher. We think this can be achieved by better localisation and cropping strategies of the MS in the head and neck MRI and by labelling more healthy MS volumes. Second, we have not experimented with autoencoders with skip connections or with more parameters and studied its effect on reconstruction error.

In conclusion, we evaluate UAD for paranasal anomaly detection. Previous methods[8–10] have used supervised learning methods and as a result are constrained to classify the anomalies that are included in the training distribution. Through our UAD approach, our models learn the healthy MS volume distribution X_h thereby reducing the labelling effort of the clinicians. Also, we are able to detect multiple anomalies. Further, we render heat maps of poor reconstruction. This can provide valuable insights to clinicians while making a diagnosis.

5. ACKNOWLEDGMENTS

This work has not been submitted for publication anywhere else. This work is funded partially by the i3 initiative of the Hamburg University of Technology. The authors also acknowledge the partial funding by the Free and Hanseatic City of Hamburg (Interdisciplinary Graduate School) from University Medical Center Hamburg-Eppendorf. This work was partially funded by Grant Number KK5208101KS0 (Zentrales Innovationsprogramm Mittelstand, Arbeitsgemeinschaft industrieller Forschungsvereinigungen).

REFERENCES

[1] Wilson, R., Kuan Kok, H., Fortescue-Webb, D., Doody, O., Buckley, O., and Torreggiani, W. C., "Prevalence and seasonal variation of incidental mri paranasal inflammatory changes in an asymptomatic irish population," *Irish medical journal* **110**(9), 641 (2017).

[2] Hansen, A. G., Helvik, A.-S., Nordgård, S., Bugten, V., Stovner, L. J., Håberg, A. K., Gårseth, M., and Eggesbø, H. B., "Incidental findings in mri of the paranasal sinuses in adults: a population-based study (hunt mri)," *BMC ear, nose, and throat disorders* **14**(1), 13 (2014).

[3] Tarp, B., Fiirgaard, B., Christensen, T., Jensen, J. J., and Black, F. T., "The prevalence and significance of incidental paranasal sinus abnormalities on mri," *Rhinology* **38**(1), 33–38 (2000).

[4] Rak, K. M., Newell, J. D., Yakes, W. F., Damiano, M. A., and Luethke, J. M., "Paranasal sinuses on mr images of the brain: significance of mucosal thickening," *AJR. American journal of roentgenology* **156**(2), 381–384 (1991).

[5] Stenner, M. and Rudack, C., "Diseases of the nose and paranasal sinuses in child," *GMS current topics in otorhinolaryngology, head and neck surgery* **13**, Doc10 (2014).

[6] Rege, I. C. C., Sousa, T. O., Leles, C. R., and Mendonça, E. F., "Occurrence of maxillary sinus abnormalities detected by cone beam ct in asymptomatic patients," *BMC oral health* **12**, 30 (2012).

[7] Cooke, L. D. and Hadley, D. M., "Mri of the paranasal sinuses: incidental abnormalities and their relationship to symptoms," *The Journal of laryngology and otology* **105**(4), 278–281 (1991).

[8] Kim, Y., Lee, K. J., Sunwoo, L., Choi, D., Nam, C.-M., Cho, J., Kim, J., Bae, Y. J., Yoo, R.-E., Choi, B. S., Jung, C., and Kim, J. H., "Deep learning in diagnosis of maxillary sinusitis using conventional radiography," *Investigative radiology* **54**(1), 7–15 (2019).

[9] Jeon, Y., Lee, K., Sunwoo, L., Choi, D., Oh, D. Y., Lee, K. J., Kim, Y., Kim, J.-W., Cho, S. J., Baik, S. H., Yoo, R.-E., Bae, Y. J., Choi, B. S., Jung, C., and Kim, J. H., "Deep learning for diagnosis of paranasal sinusitis using multi-view radiographs," *Diagnostics (Basel, Switzerland)* **11**(2) (2021).

[10] Liu, G. S., Yang, A., Kim, D., Hojel, A., Voevodsky, D., Wang, J., Tong, C. C. L., Ungerer, H., Palmer, J. N., Kohanski, M. A., Nayak, J. V., Hwang, P. H., Adappa, N. D., and Patel, Z. M., "Deep learning classification of inverted papilloma malignant transformation using 3d convolutional neural networks and magnetic resonance imaging," *International forum of allergy & rhinology* (2022).

[11] Althnian, A., AlSaeed, D., Al-Baity, H., Samha, A., Dris, A. B., Alzakari, N., Abou Elwafa, A., and Kurdi, H., "Impact of dataset size on classification performance: An empirical evaluation in the medical domain," *Applied Sciences* **11**(2) (2021).

[12] Tos, M., Larsen, P. L., Larsen, K., and Cayé-Thomasen, P., [*Nasal Polyps*], 103–125, Springer Berlin Heidelberg, Berlin, Heidelberg (2000).

[13] Janner, S. F. M., Caversaccio, M. D., Dubach, P., Sendi, P., Buser, D., and Bornstein, M. M., "Characteristics and dimensions of the schneiderian membrane: a radiographic analysis using cone beam computed tomography in patients referred for dental implant surgery in the posterior maxilla," *Clin Oral Implants Res* **22**, 1446–1453 (Mar. 2011).

[14] Hung, K., Hui, L., Yeung, A. W. K., Wu, Y., Hsung, R. T.-C., and Bornstein, M. M., "Volumetric analysis of mucous retention cysts in the maxillary sinus: A retrospective study using cone-beam computed tomography," *Imaging Sci Dent* **51**, 117–127 (Jan. 2021).

[15] Pang, G., Shen, C., Cao, L., and Hengel, A. V. D., "Deep learning for anomaly detection: A review," *ACM Comput. Surv.* **54** (mar 2021).

[16] Baur, C., Denner, S., Wiestler, B., Navab, N., and Albarqouni, S., "Autoencoders for unsupervised anomaly segmentation in brain mr images: A comparative study," *Medical Image Analysis* , 101952 (2021).

[17] Jagodzinski, A., Johansen, C., Koch-Gromus, U., Aarabi, G., Adam, G., Anders, S., Augustin, M., der Kellen, R. B., Beikler, T., Behrendt, C.-A., Betz, C. S., Bokemeyer, C., Borof, K., Briken, P., Busch, C.-J., Büchel, C., Brassen, S., Debus, E. S., Eggers, L., Fiehler, J., Gallinat, J., Gellißen, S., Gerloff, C., Girdauskas, E., Gosau, M., Graefen, M., Härter, M., Harth, V., Heidemann, C., Heydecke, G., Huber, T. B., Hussein, Y., Kampf, M. O., von dem Knesebeck, O., Konnopka, A., König, H.-H., Kromer, R., Kubisch, C., Kühn, S., Loges, S., Löwe, B., Lund, G., Meyer, C., Nagel, L., Nienhaus, A., Pantel, K., Petersen, E., Püschel, K., Reichenspurner, H., Sauter, G., Scherer, M., Scherschel, K., Schiffner, U., Schnabel, R. B., Schulz, H., Smeets, R., Sokalskis, V., Spitzer, M. S., Terschüren, C., Thederan, I., Thoma, T., Thomalla, G., Waschki, B., Wegscheider, K., Wenzel, J.-P., Wiese, S., Zyriax, B.-C., Zeller, T., and Blankenberg, S., "Rationale and design of the hamburg city health study," *European Journal of Epidemiology* **35**, 169–181 (Feb 2020).

[18] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S., "Pytorch: An imperative style, high-performance deep learning library."

[19] Falcon et al., W., "Pytorch lightning," *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning* **3** (2019).

[20] Kingma, D. P. and Ba, J., "Adam: A method for stochastic optimization," (2014).

[21] Bengs, M., Behrendt, F., Krüger, J., Opfer, R., and Schlaefer, A., "Three-dimensional deep learning with spatial erasing for unsupervised anomaly segmentation in brain mri," *CARS* , 1–11 (2021).

[22] Kingma, D. P. and Welling, M., "Auto-encoding variational bayes," (2013).

7.2 Self-supervised learning for classifying paranasal anomalies in the maxillary sinus

Title of paper: Self-supervised learning for classifying paranasal anomalies in the maxillary sinus

Journal: International Journal of Computer Assisted Radiology and Surgery

Year: 2024

Topic: Self-Supervised Learning in paranasal anomaly detection

International Journal of Computer Assisted Radiology and Surgery
https://doi.org/10.1007/s11548-024-03172-5

ORIGINAL ARTICLE

Self-supervised learning for classifying paranasal anomalies in the maxillary sinus

Debayan Bhattacharya[1,2] · Finn Behrendt[1] · Benjamin Tobias Becker[2] · Lennart Maack[1] · Dirk Beyersdorff[3] · Elina Petersen[4] · Marvin Petersen[5] · Bastian Cheng[5] · Dennis Eggert[2] · Christian Betz[2] · Anna Sophie Hoffmann[2] · Alexander Schlaefer[1]

Received: 6 December 2023 / Accepted: 1 May 2024

Abstract

Purpose Paranasal anomalies, frequently identified in routine radiological screenings, exhibit diverse morphological characteristics. Due to the diversity of anomalies, supervised learning methods require large labelled dataset exhibiting diverse anomaly morphology. Self-supervised learning (SSL) can be used to learn representations from unlabelled data. However, there are no SSL methods designed for the downstream task of classifying paranasal anomalies in the maxillary sinus (MS).

Methods Our approach uses a 3D convolutional autoencoder (CAE) trained in an unsupervised anomaly detection (UAD) framework. Initially, we train the 3D CAE to reduce reconstruction errors when reconstructing normal maxillary sinus (MS) image. Then, this CAE is applied to an unlabelled dataset to generate coarse anomaly locations by creating residual MS images. Following this, a 3D convolutional neural network (CNN) reconstructs these residual images, which forms our SSL task. Lastly, we fine-tune the encoder part of the 3D CNN on a labelled dataset of normal and anomalous MS images.

Results The proposed SSL technique exhibits superior performance compared to existing generic self-supervised methods, especially in scenarios with limited annotated data. When trained on just 10% of the annotated dataset, our method achieves an area under the precision-recall curve (AUPRC) of 0.79 for the downstream classification task. This performance surpasses other methods, with BYOL attaining an AUPRC of 0.75, SimSiam at 0.74, SimCLR at 0.73 and masked autoencoding using SparK at 0.75.

Conclusion A self-supervised learning approach that inherently focuses on localizing paranasal anomalies proves to be advantageous, particularly when the subsequent task involves differentiating normal from anomalous maxillary sinuses. Access our code at https://github.com/mtec-tuhh/self-supervised-paranasal-anomaly.

Keywords Paranasal anomaly · Self-supervised learning · Maxillary sinus · CNN · Classification

Introduction

The paranasal sinuses, air-filled spaces within the craniofacial complex, vary significantly and include the maxillary, frontal, sphenoid, and ethmoid sinuses [1]. Common pathologies like retention cysts, polyps, and mucosal thickening are identifiable through radiological screenings [2–4]. However, their diagnosis is challenging due to their incidental nature and the variability in sinus appearance [5]. Research

Anna Sophie Hoffmann and Alexander Schlaefer have contributed equally.

Debayan Bhattacharya
debayan.bhattacharya@tuhh.de

1 Institute of Medical Technology and Intelligent Systems, Technische Universitaet Hamburg, Hamburg, Germany

2 Department of Otorhinolaryngology, Head and Neck Surgery and Oncology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

3 Clinic and Polyclinic for Diagnostic and Interventional Radiology and Nuclear Medicine, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

4 Population Health Research Department, University Heart and Vascular Center, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

5 Clinic and Polyclinic for Neurology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Published online: 08 June 2024

underscores their prevalence and the importance of accurate diagnosis in patient care [6]. 3D imaging from computed tomography (CT) and magnetic resonance images (MRI) is vital for precise diagnosis, and misdiagnosis can lead to patient distress and increased healthcare costs [7, 8]. The anatomical variability of the sinuses [9] necessitates careful application of deep learning for reliable diagnoses.

Convolutional neural networks (CNNs) are recognized for diagnosing paranasal pathologies, evidenced in sinusitis classification [10, 11], differentiating inverted papilloma from carcinomas [12], and detecting MS fungal ball and chronic rhinosinusitis in CT scans [13]. Prior studies have explored contrastive learning and cross-entropy loss for MS anomaly classification [14], and MS extraction techniques from MRI [15]. However, all of the aforementioned methods use supervised learning. Given the difficulty in obtaining well-labelled datasets in clinical settings [16], and the relative ease of acquiring unlabelled data, self-supervised learning (SSL), which learns representations from unlabelled data to improve the downstream task, has not yet been explored for paranasal anomaly classification. SSL efficiently utilizes unlabelled data through tasks like nonlinear compression [17, 18], denoising [19], feature alignment from augmented images [20–22] and inpainting masked regions of images [23]. However, these methods are designed to improve the performance of models exposed to 2D natural images. Hence, they lack a specific focus on enhancing MS anomaly classification from 3D MRI. Our aim is to design an SSL task that enables the models trained on it to achieve maximum data efficiency in classifying paranasal anomalies. We hypothesize anomaly segmentation within MS could be a good SSL task. Without ground truth segmentation masks, we use a UAD framework, applied in brain [24, 25] and paranasal anomaly detection [26], to localize MS anomalies. A 3D convolutional autoencoder (CAE) trained on a labelled *normal* dataset is used to reconstruct MS volumes and localize anomalies in an unlabelled dataset by failing to reconstruct anomalies leading to reconstruction errors. These errors, serving as pseudo segmentation masks are used in the SSL task to localize anomalies. We investigate if a 3D CNN, predicting these errors as SSL task, can improve feature discrimination between anomalous and normal MS in our labelled dataset. Our SSL task leverages available normal MS data, essential for supervised downstream task training.

Overall, our main contributions can be summed up as follows:

- We present a self-supervised method that improves the downstream classification of normal vs anomalous MS. Our self-supervision task explicitly learns to coarsely localize anomalies by reconstructing the residual volumes generated through the UAD-trained autoencoder. This distinguishes our approach from the compared methods, where anomaly localization is not a primary focus for the self-supervision task.
- Our self-supervised method effectively utilizes labelled healthy MS data reserved for downstream tasks. Hence, we explore how varying the CAE training set impacts downstream classification performance.
- We investigate post-processing strategies and loss function used in the self-supervision task for learning better transferable features for the downstream task.

Methods

Description of dataset

As part of the Hamburg City Health Study (HCHS) [27], cranial MRI scans were obtained from individuals aged 45-74 years to evaluate neuroradiological parameters. The scans were acquired using fluid attenuated inversion recovery (FLAIR) sequences in the NIfTI format at the University Medical Center Hamburg-Eppendorf. The MRI scans had a resolution of 173 mm x 319 mm x 319 mm. The labelled dataset consisted of 1067 patients. Among the patients, 489 exhibited no pathologies in their left and right MS, while 578 had at least one MS presenting polyp, cyst or mucosal thickening pathology. All these anomalies were grouped into the "anomaly" class. Our unlabelled dataset consists of 1559 patient MRIs. The diagnoses were established by two ENT specialists and one radiologist specialized in ENT. Figure 1b shows coronal slices highlighting the diverse set of anomalies that are present in our dataset.

Dataset preprocessing

In our dataset preprocessing, as outlined in previous work [14, 15], we first align MRIs with a fixed sample from our dataset. Centroid locations of left and right MS regions were recorded for 20 patients, guiding the extraction of MS volumes from larger cranial MRIs. This step isolates the relevant MS volumes for our task of classifying healthy and anomalous MS. We then used the mean centroid location from these 20 recordings to extract left and right MS volumes from all cranial MRIs in our dataset. The extracted volumes, sized 64 mm x 64 mm x 64 mm, cover the entire MS. Figure 1a illustrates this extraction process.

Each cranial MRI yielded one left and one right MS volume. To enhance symmetry, right MS volumes were horizontally flipped to match the left ones. All volumes were normalized to an intensity range of 0 to 1. We employed five-fold cross-validation for evaluation, ensuring diverse labelled datasets (10%, 20%, 40%, 60%, 80%) maintain the anomaly-to-normal ratio. The separation of training, validation, and test sets was strictly maintained, with left or right MS vol-

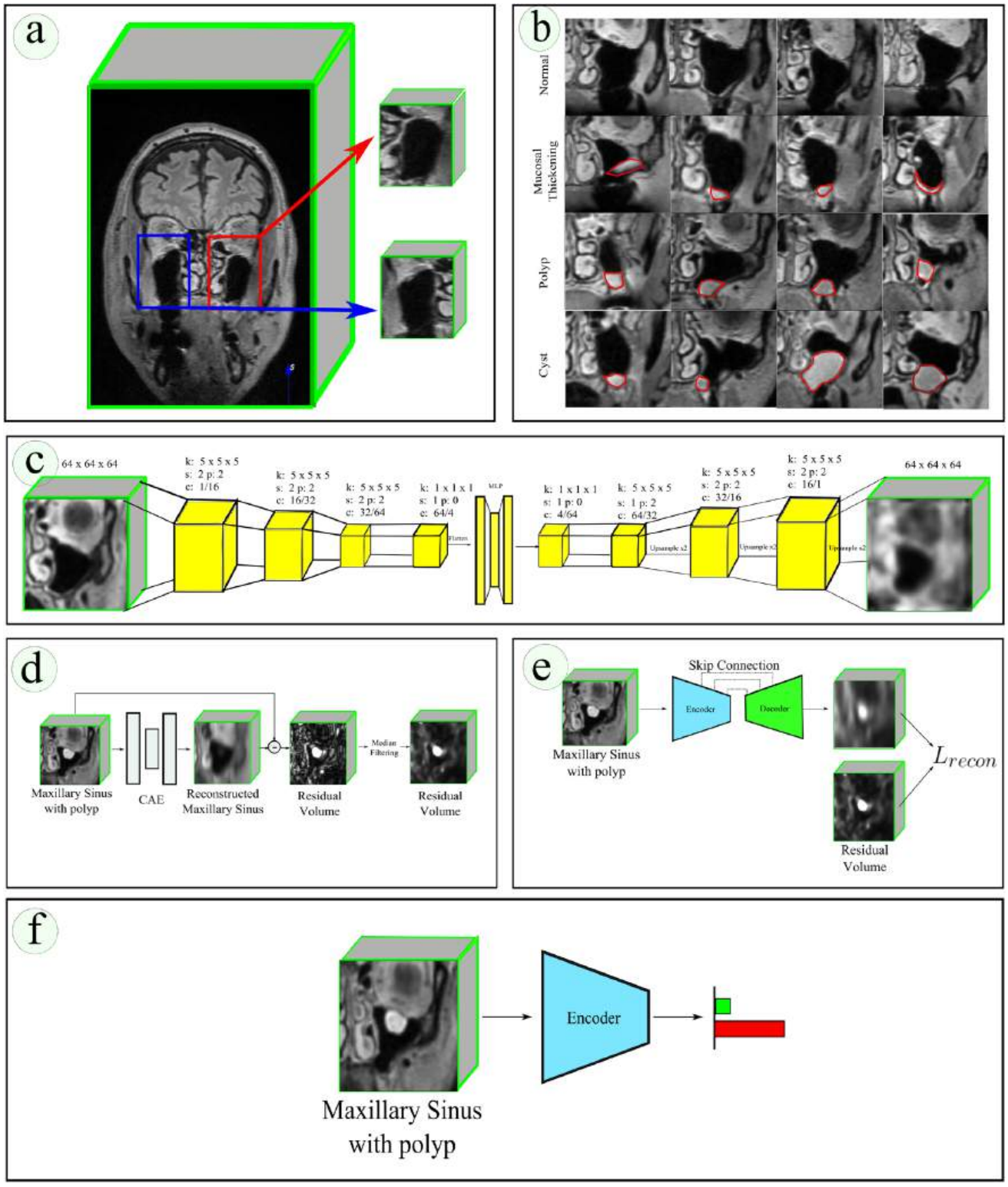

Fig. 1 **a** Extraction of MS volumes from cranial MRI, **b** Exemplary coronal images of normal MS volume and MS with mucosal thickening, polyp and cyst anomaly, **c** Our CAE architecture. Here, *k* refers to kernel size, *s* refers to stride, *p* refers to padding, *c* refers to channel where, for example, 1/16 refers to input channel of 1 and output channel of 16. Each stage of the encoder and decoder is formed using 3D convolution followed by batch normalization and leaky ReLU. Upsample refers to trilinear upsampling. **d** Generation of residual volume required for the self-supervision task using our CAE, **e** Our self-supervision task where the encoder and decoder is trained to reconstruct the residual volume, **f** Downstream task where the self-supervision trained encoder is trained to classify between normal and anomalous MS

Table 1 Statistics of our labelled dataset D_l

Class	Training set	Validation set	Test set
# Normal MS	708	176	380
# Anomalous MS	487	122	261

umes from the same patient assigned to only one set. Table 1 details our dataset division across these sets.

Architecture

Our CAE, depicted in Fig. 1c, uses 3D convolutional operations with a latent bottleneck dimension of 512. The CNN architecture is U-Net inspired, featuring a 3D ResNet18 encoder $E(.)$ [28] with four stages and channel dimensions of 64, 128, 256, and 512. The decoder $D(.)$ mirrors the encoder, with reverse channel dimensions and trilinear upsampling. Skip connections are used to pass encoder features to the decoder. For Bootstrap your own latent (BYOL), SimSiam, and SimCLR training, only the encoder $E(.)$ is used, with an MLP attached to project the final layer features to a dimension of 512.

Autoencoder training and inference on unlabelled dataset

Consider D_l to be our labelled dataset containing normal and anomalous MS and D_u to be our unlabelled dataset. Further, let $D_l^n \subset D_l$ be a dataset consisting of only normal MS volumes. Let $x \in R^{64\times64\times64}$ be an MS volume in D_l. Let the autoencoder be represented as $A(.)$ such that $x' = A(x)$ represents the reconstructed MS volume. We train the autoencoder using L1 reconstruction loss which may be written as $\|x - x'\|$ on D_l^n. Once trained, we use the autoencoder $A(.)$ to generate residual volumes on D_u. Figure 1d illustrates our residual volume generation method.

Transfer learning

Since transfer learning (TL) is a method to achieve data efficiency, we also trained our models initialized with transfer learning weights. However, since our downstream task involves MRI and is in 3D domain, ImageNet [29] weights may not be appropriate. Hence, the model weights we utilized as initial weights were obtained through training on eight diverse public 3D segmentation datasets, covering both MRI and CT modalities. We believe these weights are more suitable than those derived from natural image training and therefore employed them as the basis for our 3D CNN. For further information on the transfer learning model, please see the GitHub repository.[1]

[1] https://github.com/Tencent/MedicalNet.

Self-supervised training

With the residual volumes generated for D_u, we train $E(.)$ and $D(.)$ to reconstruct the residual volumes again. This, in effect, makes the encoder and decoder learn features relevant for anomaly localization within the unlabelled dataset D_u. We train $E(.)$ and $D(.)$ using L_{recon} which in our case is binary cross-entropy (BCE) loss. Figure 1e illustrates our self-supervised training task. We evaluated our self-supervised learning method against autoencoder (AE), denoising autoencoder (DAE), BYOL, SimSiam, SimCLR and sparse masked modelling with hierarchy (SparK). These methods use similar encoders $E(.)$ and decoders $D(.)$, with BYOL, SimSiam, and SimCLR employing an additional MLP for feature projection. Pretraining with the SparK framework requires sparse encoder $E'(.)$ and a special light decoder which contains 3 convolutional blocks and 3 upsampling blocks [23]. Patch size $8 \times 8 \times 8$ and masking ratio of 60% was used during pretraining. Detailed description and implementation details of our state-of-the-art (SOTA) SSL methods is provided in the supplementary material section 1-7. More details about the other masking ratios and patch sizes tested for SparK can be found in the supplementary material section 11.

Finetuning

Having successfully trained the $E(.)$ and $D(.)$ using self-supervision, we move onto the finetuning phase. We discard $D(.)$ and focus on training $E(.)$ by leveraging samples from the labelled dataset D_l. For TL models, we initialize $E(.)$ with transfer learning weights. Next, we introduce a MLP as an additional component, responsible for projecting the encoder features from their original dimension of 512 to an intermediate dimension of 256. Subsequently, the MLP maps these features to a final dimension of 2, corresponding to the number of classes. We finetune $E(.)$ using BCE loss.

Figure 2 illustrates the data processing pipeline and elucidates how the different components fit into our overall method.

Implementation details

Our PyTorch and PyTorch Lightning-based code accommodates a maximum batch size of 256 on NVIDIA A6000 with 48GB VRAM for self-supervised pretraining. We optimize models using LARS [30] with a learning rate of 0.2 across 500 epochs, incorporating a 20-epoch linear warmup and cosine annealing. For finetuning, AdamW [31] is employed with a constant rate of 1e-4 for 100 epochs at a batch size of 16. Models yielding the lowest validation loss are preserved for final evaluation with the test set. The CAE was trained on 708 normal MS volume samples without augmentation. For

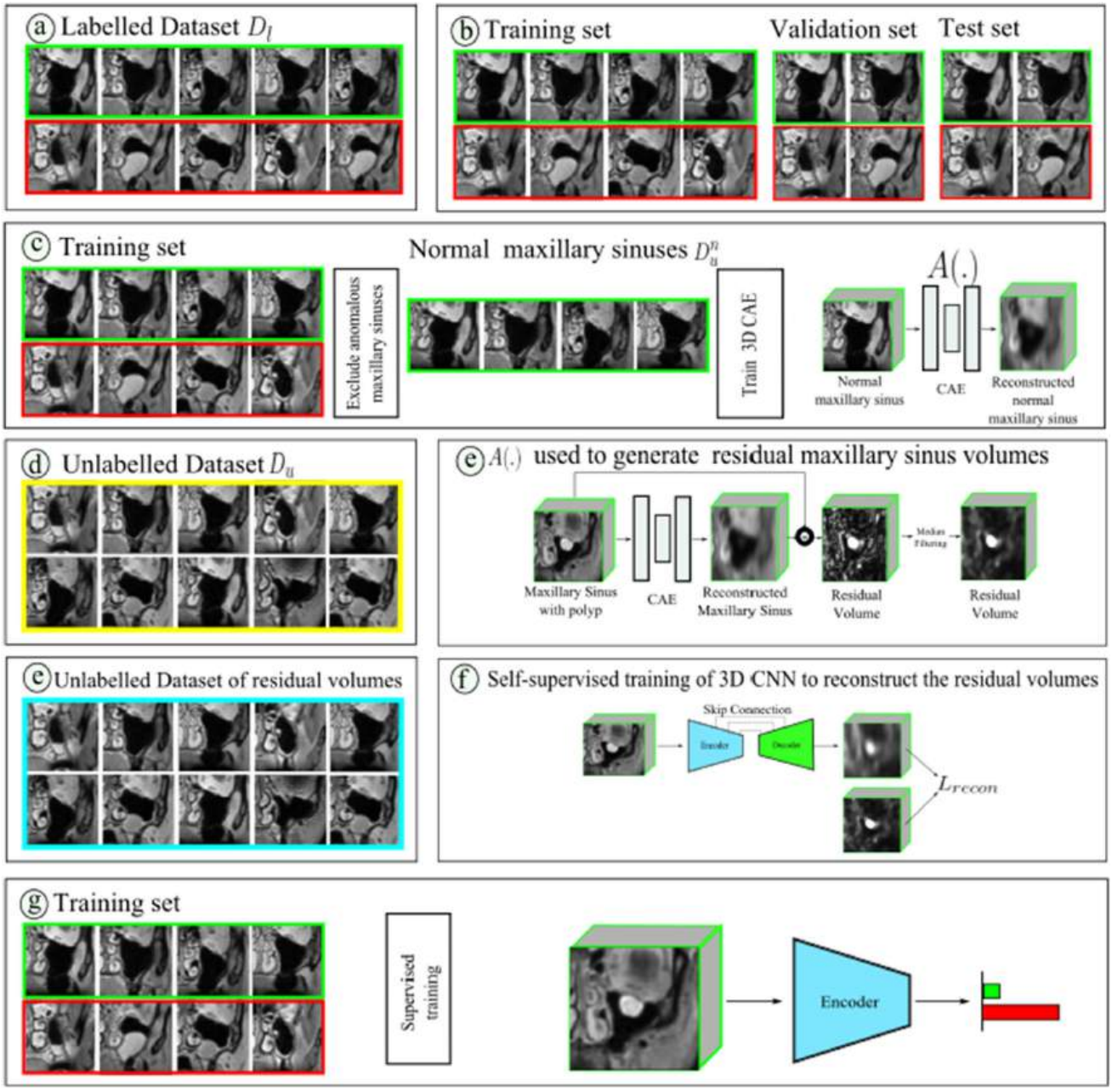

Fig. 2 Our data processing pipeline comprises several steps: **a** The labelled dataset D_l, **b** Splitting D_l into training, validation, and test subsets for downstream classification of normal versus anomalous MS. **c** Normal MS samples from the labelled training set form D_l^n, used to train the 3D CAE $A(.)$ within the UAD framework. **d** Unlabelled dataset D_u, **e** This trained 3D CAE $A(.)$ generates residual volumes from the unlabelled dataset D_u, **e** Unlabelled dataset of residual volumes, **f** The 3D CNN undergoes self-supervised training to reconstruct these residual volumes. **g** The 3D CNN's encoder is initialized with weights from the SSL task and then undergoes supervised training for the final task of classifying normal versus anomalous MS, using the training set created in step (**a**)

self-supervised methods and MS anomaly classification, we applied data augmentations such as random affine transformations, flipping, and Gaussian noise. The DAE specifically used Gaussian noise with a mean of 0 and standard deviation of 0.6 at 100% probability, while other augmentations were applied 50% of the time. Supplementary material offers comprehensive descriptions and visualizations of SOTA SSL methods.

Results

Comparison to state of the art

Results in Table 2 show our method outperforming others in AUROC, AUPRC, and F1 scores across different labelled dataset scenarios (10%, 20%, 100% of D_l). Our method

Table 2 The table displays the mean and 95% confidence intervals of metrics evaluating model performance in the downstream classification task

Method	Training set percentage D_l	AUROC	AUPRC	F1
No pretraining	10%	0.74 (0.64−0.84)	0.69 (0.56−0.82)	0.64 (0.59−0.69)
Transfer Learning	10%	0.77 (0.72−0.82)	0.73 (0.66−0.79)	0.63 (0.57−0.69)
AE	10%	0.73 (0.68−0.79)	0.68 (0.62−0.74)	0.55 (0.43−0.67)
DAE	10%	0.74 (0.73−0.76)	0.68 (0.66−0.69)	0.62 (0.60−0.64)
BYOL	10%	0.79 (0.76−0.81)	0.75 (0.70−0.79)	0.63 (0.59−0.69)
SimSiam	10%	0.77 (0.72−0.83)	0.74 (0.68−0.79)	0.62 (0.53−0.72)
SimCLR	10%	0.78 (0.74−0.81)	0.73 (0.68−0.78)	0.63 (0.59−0.68)
SparK MAE	10%	0.78 (0.77−0.80)	0.75 (0.73−0.76)	0.65 (0.63−0.67)
Ours	10%	**0.81 (0.74−0.88)**	**0.79 (0.71−0.87)**	**0.67 (0.58−0.77)**
No pretraining	20%	0.81 (0.79−0.82)	0.78 (0.76−0.79)	0.67 (0.65−0.69)
Transfer Learning	20%	0.84 (0.79−0.88)	0.81 (0.74−0.88)	0.68 (0.62−0.75)
AE	20%	0.81 (0.76−0.86)	0.78 (0.72−0.83)	0.67 (0.60−0.74)
DAE	20%	0.79 (0.77−0.81)	0.74 (0.70−0.79)	0.67 (0.64−0.70)
BYOL	20%	0.82 (0.80−0.84)	0.79 (0.77−0.82)	0.70 (0.68−0.71)
SimSiam	20%	0.84 (0.82−0.86)	0.81 (0.78−0.84)	0.70 (0.67−0.74)
SimCLR	20%	0.81 (0.79−0.83)	0.77 (0.74−0.81)	0.68 (0.67−0.69)
SparK MAE	20%	0.80 (0.78−0.82)	0.76 (0.73−0.79)	0.67 (0.65−0.68)
Ours	20%	**0.85 (0.83−0.87)**	**0.82 (0.81−0.83)**	**0.72 (0.70−0.75)**
No pretraining	100%	0.90 (0.89−0.91)	0.89 (0.88−0.90)	0.80 (0.78−0.82)
Transfer Learning	100%	0.92 (0.91−0.93)	0.91 (0.90−0.93)	0.82 (0.80−0.83)
AE	100%	0.92 (0.91−0.93)	0.91 (0.90−0.93)	0.82 (0.80−0.84)
DAE	100%	0.90 (0.88−0.92)	0.89 (0.88−0.91)	0.79 (0.77−0.82)
BYOL	100%	0.89 (0.89−0.90)	0.88 (0.87−0.89)	0.78 (0.76−0.81)
SimSiam	100%	0.92 (0.91−0.93)	0.91 (0.90−0.92)	0.81 (0.79−0.83)
SimCLR	100%	0.90 (0.88−0.91)	0.89 (0.87−0.91)	0.79 (0.77−0.80)
SparK MAE	100%	0.87 (0.85−0.88)	0.86 (0.84−0.87)	0.75 (0.73−0.76)
Ours	100%	**0.93 (0.91−0.94)**	**0.92 (0.90−0.93)**	**0.83 (0.80−0.86)**

These models, trained with varying portions of D_l, were initialized using different SSL methods before supervised training
The bold values signify the best/highest values for the given metrics (AUROC, AUPRC, F1)

demonstrated notable improvements in AUROC (3.34% and 4.93% over SimSiam) and AUPRC (5.33% over BYOL and 5.12% over AE) for 10% and 20% dataset scenarios, respectively. SparK trained models perform generally poorer compared to the other SSL and TL methods with the performance gap between SparK MAE and our method widening with increased training set percentage. Our method had AUPRC 8.21% higher than the TL method when finetuned on a 10% training set. Pretraining models using our method significantly boosted AUPRC by 14.49% and AUROC by 9.45% compared to no pretraining when trained on a 10% training dataset. At 100% dataset finetuning, our method achieved the highest scores, with AE and SimSiam showing similar performance. Compared to no pretraining, our method improved AUPRC by 3.33%. Figure 3 illustrates AUPRC and AUROC trends with increasing training set percentages, respectively. Our method excels in settings with 40% or less training data but aligns with SOTA performance beyond that.

Effect of varying the CAE training set

The effectiveness of our self-supervised task is contingent on the CAE's proficiency in reconstructing healthy MS volumes. Inaccurate reconstructions yield unreliable residuals, affecting self-supervision. To assess the impact of training set size, the CAE was trained with different proportions (20%, 40%, 60%, 80%, 100%) of the healthy MS dataset D_l^n. After training, the CAE processed dataset D_u to produce residual volumes, which were refined using a median filter with a kernel size of 5. Subsequent supervised training utilized 10% of our labelled dataset D_l. Table 3 presents improvements in the downstream task metrics correlating with increased healthy MS training set sizes, suggesting that larger *normal* dataset D_l^n enhance normal MS representation learning and improve anomaly localization.

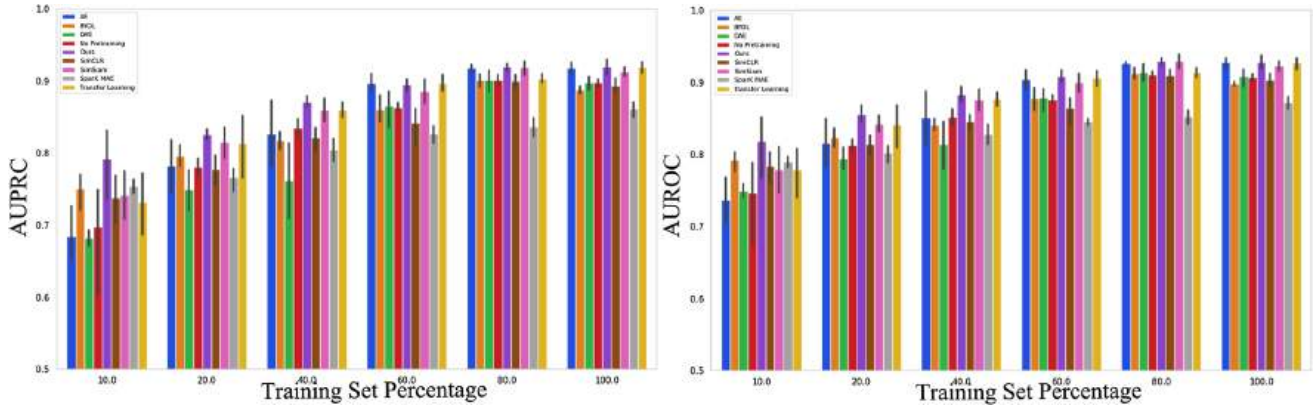

Fig. 3 (LEFT) AUPRC trend vs training set percentage (RIGHT) AUROC trend vs training set percentage

Table 3 The table shows the mean and 95% confidence intervals of metrics for evaluating model performance in downstream classification

Training set percentage D_l^n	AUROC	AUPRC	F1
20%	0.76 (0.70−0.81)	0.72 (0.67−0.77)	0.60 (0.50−0.69)
40%	0.77 (0.73−0.80)	0.72 (0.66−0.78)	0.63 (0.57−0.68)
60%	0.78 (0.75−0.80)	0.74 (0.71−0.77)	0.65 (0.62−0.68)
80%	0.80 (0.76−0.84)	0.76 (0.72−0.81)	0.67 (0.63−0.72)
100%	**0.81 (0.74−0.88)**	**0.79 (0.71−0.87)**	**0.67 (0.58−0.77)**

The CAE was trained on varying proportions of the normal MS volumes dataset (D_l^n) and then used to generate residual volumes from the unlabelled dataset (D_u). Each model was initialized using our proposed SSL method

The bold values signify the best/highest values for the given metrics (AUROC, AUPRC, F1)

Discussion

Tailoring SSL tasks to specific downstream tasks offers distinct advantages [32]. Current SOTA SSL methods [20–22], primarily developed for 2D image classification on datasets like ImageNet, do not address the unique challenges of 3D MRI modalities and the specifics of paranasal anomalies. Our SSL task is specifically tailored to address the challenges associated with 3D environments, MRI modality, and the classification of paranasal anomalies.

We conjecture that segmentation of anomalies as a SSL task, requiring knowledge of anomaly locations, enhances the learning of class-discriminative features for distinguishing normal and anomalous MS. Our SSL task is a segmentation task therefore, it requires segmentation masks highlighting anomalies. To avoid the high costs of annotation, we use a CAE trained in the UAD framework for generating approximate annotations, effective in localizing paranasal anomalies [26]. This CAE training utilizes labelled *normal* datasets, typically accessible in supervised settings. Unlike generic SOTA SSL methods, which do not prioritize anomaly localization, our approach demonstrates improved AUROC and AUPRC (as shown in Table 2), suggesting that effective anomaly localization can enhance classification performance, even with limited labelled data. Methods like BYOL and SimSiam, which aim to maximize agreement between augmented views, are less effective for paranasal anomaly classification. SimCLR's performance shortfall is likely due to smaller batch sizes, a necessity given the impracticality of large batches in 3D settings, despite SimCLR's recommendation of 4096 [33]. Our method is more suited for such constrained computational resources. AE and DAE, focusing on compression-decompression and denoising, do not guarantee discriminative feature learning for downstream classification [34], and were found less effective in our context. When the entire training set is used, our method, AE, and SimSiam yield comparable results, with ours marginally outperforming. We also explored MAE-style pretraining using SparK. However, the results suggest that fine-tuning performance is notably weaker, particularly when fine-tuning with a training set percentage 40% and above. These findings imply that generating masked regions contributes to representation learning; however, the acquired representations do not appear to enhance downstream classification. It is noteworthy that the SparK framework was initially developed and evaluated for 2D natural images. Although we adapted the framework for 3D applications, our findings underscore the necessity for further methodological advancements to effectively support tasks in the 3D domain. Further, TL models exhibit comparable performance to SSL methods when fine-tuning on training sets exceeding 20%. This suggests that transfer learning methods remain viable for paranasal

anomaly classification given an ample supply of labelled samples. However, in the scenario of an extremely limited labelled dataset, such as 10%, our method outperforms TL, indicating that the representations acquired by our approach are especially advantageous in low-data environments. Overall, compared to approaches without pretraining, our tailored SSL task consistently shows superior downstream classification performance, underlining its efficacy.

Our analysis regarding the impact of the CAE training set size shown in Table 3 has demonstrated that the inclusion of a substantial cohort of normal MS volumes yields notable benefits for both the self-supervision task and the subsequent downstream task suggesting that better anomaly localization by the CAE and thereby better representation learning by the CNN in the self-supervision task. We also analysed the influence of the loss function and post-processing used in the self-supervision task which can be found in the supplementary material section 8 and 9.

Our study has limitations that require further investigation. It is based on a single-centre MRI-only study, so multi-centre studies with varied imaging modalities are needed for generalizability. Our methods rely on a cohort of healthy MS volumes, unlike other self-supervised tasks. We focused on convolutional autoencoders, not exploring models like variational autoencoders generative adversarial networks, or transformer-based architectures and diffusion models, which might offer better anomaly localization. We compared L1, L2, and BCE loss functions but not others like the Structural Similarity Index or perceptual loss. Future research should examine these aspects and apply this self-supervision approach to other domains, like brain anomaly detection.

Conclusion

We developed a novel self-supervision task that focuses on anomaly localization to better classify paranasal anomalies in the maxillary sinus, addressing the lack of methods that effectively use unlabelled datasets to learn discriminative features for this purpose. Our approach uses an autoencoder trained on healthy MS volumes to generate residual volumes from an unlabelled dataset. These residuals serve as coarse segmentation masks for localizing anomalies. By training a CNN to reconstruct these volumes, it implicitly learns anomaly localization, thereby developing transferable features for the downstream classification task. Our method outperforms existing self-supervision techniques, proving its effectiveness in this specific domain.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11548-024-03172-5.

Acknowledgements This work has not been submitted for publication anywhere else. This work is funded partially by the i3 initiative of the Hamburg University of Technology. The authors also acknowledge the partial funding by the Free and Hanseatic City of Hamburg (Interdisciplinary Graduate School) from University Medical Center Hamburg-Eppendorf. This work was partially funded by Grant Number KK5208101KS0 (Zentrales Innovationsprogramm Mittelstand, Arbeitsgemeinschaft industrieller Forschungsvereinigungen). Publishing fees supported by Funding Programme Open Access Publishing of Hamburg University of Technology (TUHH).

Funding Open Access funding enabled and organized by Projekt DEAL. Funding was provided by i3 initiative Hamburg University of Technology, Interdisciplinary Graduate School University Medical Center Hamburg-Eppendorf, Zentrales Innovationsprogramm Mittelstand, Arbeitsgemeinschaft industrieller Forschungsvereinigungen (Grant number: K5208101KS0).

Declarations

Conflict of interest The authors declare no conflict of interest.

Ethical approval The study protocol received approval from the local ethics committee (Landesärztekammer Hamburg, PV5131) and was approved by the Data Protection Commissioners for the University Medical Center of the University Hamburg-Eppendorf and the Free and Hanseatic City of Hamburg. It is registered on ClinicalTrial.gov (NCT03934957) and adheres to Good Clinical Practice, Good Epidemiological Practice, and ethical principles outlined in the Declaration of Helsinki.

References

1. Marieb EN (1991) Essentials of Human Anatomy & Physiology. Third edition. Redwood City, Calif., Benjamin/Cummings Pub. Co., 1991. https://search.library.wisc.edu/catalog/9910059601802121
2. Bal M, Berkiten G, Uyanık E (2014) Mucous retention cysts of the paranasal sinuses. Hippokratia 18(4):379
3. Varshney H, Varshney J, Biswas S, Ghosh SK (2015) Importance of CT scan of paranasal sinuses in the evaluation of the anatomical findings in patients suffering from sinonasal polyposis. Indian J Otolaryngol Head Neck Surg 68(2):167–172
4. Van Dis ML, Miles DA (1994) Disorders of the maxillary sinus. Dent Clin North Am 38(1):155–166
5. Hansen AG, Helvik A-S, Nordgård S, Bugten V, Stovner LJ, Håberg AK, Gårseth M, Eggesbø HB (2014) Incidental findings in MRI of the paranasal sinuses in adults: a population-based study (HUNT

MRI). BMC Ear Nose Throat Disord 14(1):13. https://doi.org/10.1186/1472-6815-14-13
6. Tarp B, Fiirgaard B, Christensen T, Jensen JJ, Black FT (2000) The prevalence and significance of incidental paranasal sinus abnormalities on MRI. Rhinology 38(1):33–38
7. Brierley J, Gospodarowicz MK, Wittekind C (eds) (2017) TNM classification of malignant tumours. Eighth edn. John Wiley& Sons Inc, Chichester West Sussex UK and Hoboken NJ
8. Gutmann A (2013) Ethics. The bioethics commission on incidental findings. Science 342(6164):1321–1323. https://doi.org/10.1126/science.1248764
9. Papadopoulou A-M, Chrysikos D, Samolis A, Tsakotos G, Troupis T (2021) Anatomical variations of the nasal cavities and paranasal sinuses: a systematic review. Cureus 13(1):12727
10. Jeon Y, Lee K, Sunwoo L, Choi D, Oh DY, Lee KJ, Kim Y, Kim J-W, Cho SJ, Baik SH, Yoo R-E, Bae YJ, Choi BS, Jung C, Kim JH (2021) Deep learning for diagnosis of paranasal sinusitis using multi-view radiographs. Diagnostics. https://doi.org/10.3390/diagnostics11020250
11. Kim Y, Lee KJ, Sunwoo L, Choi D, Nam C-M, Cho J, Kim J, Bae YJ, Yoo R-E, Choi BS, Jung C, Kim JH (2019) Deep learning in diagnosis of maxillary sinusitis using conventional radiography. Investig Radiol 54(1):7–15. https://doi.org/10.1097/RLI.0000000000000503
12. Liu GS, Yang A, Kim D, Hojel A, Voevodsky D, Wang J, Tong CCL, Ungerer H, Palmer JN, Kohanski MA, Nayak JV, Hwang PH, Adappa ND, Patel ZM (2022) Deep learning classification of inverted papilloma malignant transformation using 3d convolutional neural networks and magnetic resonance imaging. Int Forum Allergy Rhinol. https://doi.org/10.1002/alr.22958
13. Kim K-S, Kim BK, Chung MJ, Cho HB, Cho BH, Jung YG (2022) Detection of maxillary sinus fungal ball via 3-D CNN-based artificial intelligence: Fully automated system and clinical validation. PLoS ONE 17(2):1–19. https://doi.org/10.1371/journal.pone.0263125
14. Bhattacharya D, Becker BT, Behrendt F, Bengs M, Beyersdorff D, Eggert D, Petersen E, Jansen F, Petersen M, Cheng B, Betz C, Schlaefer A, Hoffmann AS (2022) Supervised contrastive learning to classify paranasal anomalies in the maxillary sinus. In: Wang L, Dou Q, Fletcher PT, Speidel S, Li S (eds) Medical image computing and computer assisted intervention-MICCAI 2022. Springer, Cham, pp 429–438
15. Bhattacharya D, Behrendt F, Becker BT, Beyersdorff D, Petersen E, Petersen M, Cheng B, Eggert D, Betz C, Hoffmann AS, Schlaefer A (2023) Multiple instance ensembling for paranasal anomaly classification in the maxillary sinus. Int J Comput Assist Radiol Surg 19(2):223–231
16. Pang G, Shen C, Cao L, Hengel AVD (2021) Deep learning for anomaly detection: a review. ACM Comput Surv. https://doi.org/10.1145/3439950
17. Pihlgren G, Sandin F, Liwicki M (2021) Pretraining image encoders without reconstruction via feature prediction loss. In: 2020 25th international conference on pattern recognition (ICPR), pp 4105–4111. IEEE Computer Society, Los Alamitos, CA, USA. https://doi.org/10.1109/ICPR48806.2021.9412239
18. Xie Y, Thuerey N (2023) Reviving autoencoder pretraining. Neural Comput Appl 35(6):4587–4619. https://doi.org/10.1007/s00521-022-07892-0
19. Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol P-A (2010) Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. J Mach Learn Res 11:3371–3408
20. Grill J-B, Strub F, Altché F, Tallec C, Richemond P, Buchatskaya E, Doersch C, Avila Pires B, Guo Z, Gheshlaghi Azar M, Piot B, kavukcuoglu k, Munos R, Valko M (2020) Bootstrap your own latent-a new approach to self-supervised learning. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds.) Advances in neural information processing systems, vol. 33, pp 21271–21284. Curran Associates, Inc., . https://proceedings.neurips.cc/paper_files/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf
21. Chen X, He K (2021) Exploring simple siamese representation learning. In: 2021 IEEE/CVF Conference on computer vision and pattern recognition (CVPR), pp 15745–15753 . https://doi.org/10.1109/CVPR46437.2021.01549
22. Huang S-C, Pareek A, Jensen M, Lungren MP, Yeung S, Chaudhari AS (2023) Self-supervised learning for medical image classification: a systematic review and implementation guidelines. NPJ Digit Med 6(1):74. https://doi.org/10.1038/s41746-023-00811-0
23. Tian K, Jiang Y, qishuai diao, Lin C, Wang L, Yuan Z (2023) Designing BERT for convolutional networks: sparse and hierarchical masked modeling. In: The eleventh international conference on learning representations. https://openreview.net/forum?id=NRxydtWup1S
24. Baur C, Denner S, Wiestler B, Navab N, Albarqouni S (2021) Autoencoders for unsupervised anomaly segmentation in brain MR images: a comparative study. Med Image Anal 69:101952
25. Behrendt F, Bengs M, Rogge F, Krüger J, Opfer R, Schlaefer A (2022) Unsupervised anomaly detection in 3D brain MRI using deep learning with impured training data. In: 2022 IEEE 19th international symposium on biomedical imaging (ISBI), pp 1–4 . https://doi.org/10.1109/ISBI52829.2022.9761443
26. Bhattacharya D, Behrendt F, Becker BT, Beyersdorff D, Petersen E, Petersen M, Cheng B, Eggert D, Betz C, Hoffmann AS, Schlaefer A (2022) Unsupervised anomaly detection of paranasal anomalies in the maxillary sinus. arXiv. https://doi.org/10.48550/ARXIV.2211.01371. https://arxiv.org/abs/2211.01371
27. Jagodzinski A (2019) Rationale and design of the Hamburg city health study. Eur J Epidemiol 35(2):169–181
28. Tran D, Wang H, Torresani L, Ray J, LeCun Y, Paluri M (2018) A closer look at spatiotemporal convolutions for action recognition. In: 2018 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 6450–6459. IEEE Computer Society, Los Alamitos, CA, USA. https://doi.org/10.1109/CVPR.2018.00675
29. Deng J, Dong W, Socher R, Li L.-J, Li K, Fei-Fei L (2009) ImageNet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, pp 248–255. https://doi.org/10.1109/CVPR.2009.5206848
30. Ginsburg B, Gitman I, You Y (2018) Large batch training of convolutional networks with layer-wise adaptive rate scaling. https://openreview.net/forum?id=rJ4uaX2aW
31. Loshchilov I, Hutter F (2019) Decoupled weight decay regularization. In: International conference on learning representations. https://openreview.net/forum?id=Bkg6RiCqY7
32. Ozbulak U, Lee HJ, Boga B, Anzaku ET, Park H-M, Messem AV, Neve WD, Vankerschaver J (2023) Know your self-supervised learning: a survey on image-based generative and discriminative training. Transactions on Machine Learning Research. Survey Certification
33. Chen T, Kornblith S, Norouzi M, Hinton G (2020) A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th international conference on machine learning. ICML. JMLR.org
34. Raina R, Battle A, Lee H, Packer B, Ng AY (2007) Self-taught learning: transfer learning from unlabeled data. In: Proceedings of the 24th international conference on machine learning. ICML '07, pp. 759–766. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/1273496.1273592

7.3 Convolutional transformer network for paranasal anomaly classification in the maxillary sinus

Title of paper: Convolutional transformer network for paranasal anomaly classification in the maxillary sinus

Conference: Society of Photographic Instrumentation Engineers Medical Imaging

Year: 2024

Topic: Improving supervised learning using architectural modification

Convolutional transformer network for paranasal anomaly classification in the maxillary sinus

Debayan Bhattacharya[a], Finn Behrendt[b], Lennart Maack[l], Benjamin Tobias Becker[c], Dirk Beyersdorff[d], Elina Petersen[e], Marvin Petersen[f], Bastian Cheng[g], Dennis Eggert[h], Christian Betz[i], Anna Sophie Hoffmann*[j], and Alexander Schlaefer*[k]

[a,b,k,l]Hamburg University of Technology, Hamburg, Germany
[a,c,d,e,f,g,h,i,j] Universitätsklinikum Hamburg-Eppendorf, Hamburg, Germany

ABSTRACT

Large-scale population studies have examined the detection of sinus opacities in cranial MRIs. Deep learning methods, specifically 3D convolutional neural networks (CNNs), have been used to classify these anomalies. However, CNNs have limitations in capturing long-range dependencies across the low and high level features, potentially reducing performance. To address this, we propose an end-to-end pipeline using a novel deep learning network called ConTra-Net. ConTra-Net combines the strengths of CNNs and self-attention mechanisms of transformers to classify paranasal anomalies in the maxillary sinuses. Our approach outperforms 3D CNNs and 3D Vision Transformer (ViT), with relative improvements in F1 score of 11.68% and 53.5%, respectively. Our pipeline with ConTra-Net could serve as an alternative to reduce misdiagnosis rates in classifying paranasal anomalies.

Keywords: Paranasal anomaly, maxillary sinus, anomaly classification, CNN, Transformer, Hybrid Network

1 INTRODUCTION

Paranasal sinus anomalies are a common but clinically significant finding in the radiological assessment of the head and neck area.[1] These anomalies present various treatment challenges[2] and have been the subject of numerous studies to analyze their occurrence and progression in the general population.[3] Accurate diagnosis of paranasal inflammations is crucial for patient care and cost reduction. Clinicians rely on cross-sectional views from CT and MRI to assess these conditions. Misdiagnosis can cause unnecessary concern and costs.[4]

Deep learning is widely used for paranasal pathology screening, including sinusitis classification,[5,6] fungal ball detection,[7] and polyp and cyst detection.[8–10] However, CNNs struggle to capture global cues, while transformers excel in computer vision tasks[11,12] because of learning global cues. However, without pretraining, vision transformers fail to learn meaningful representations.[13] Hence, hybrid models combining CNNs and transformers have been proposed[14–16] that benefit from the inductive bias of CNNs. Our novel ConTra-Net network draws inspiration from TransMed[17] and forms dependencies across multiple CNN feature levels.[14] By combining CNNs inductive bias with self-attention mechanisms, we capture global context and increase representation quality across the low and high level features. ConTra-Net classifies healthy and anomalous maxillary sinuses, offering potential for diverse paranasal anomaly identification. We are leveraging CNNs and Multi-head self-attention (MHSA) block to improve the accuracy and efficiency of diagnosing paranasal anomalies in the maxillary sinus. These anomalies, including polyps and cysts, can be differently located along the sinus walls and present in various shapes, sizes, and contrasts. Therefore we hypothesize that extracting features with long-range dependencies, large receptive fields and invariance with respect to the appearance of the anomaly may be beneficial for the classification task.

Our main contributions can be summarised as follows. First, we propose a hybrid convolutional transformer network (ConTra-Net) for classifying maxillary sinus anomalies. This network combines the inductive biases of a CNN with the ability to capture long-range dependencies among multi-level features, resulting in an improved classification performance. Second, we investigate which combination of interaction between low and high level features leads to the best classification performance. Finally, we investigate the influence of the input volume size on the classification performance. Overall, our study aims to leverage ConTra-Net to improve the accuracy and efficiency of diagnosing paranasal anomalies in the maxillary sinus.

2 Materials and Methods

2.1 Dataset

Our population study named Hamburg City Health Study includes MRI images of the head and neck region from participants aged 45 to 74 in the city of Hamburg, Germany. The dataset consists of 299 patients with 174 healthy maxillary sinuses (MS) and 125 MS with anomalies (polyps and cysts) classified as "normal" and "anomaly" classes respectively, confirmed by 2 ENT surgeons and a radiologist. The MRIs of resolution 173 × 319 × 319 voxel were recorded at University Medical Center Hamburg-Eppendorf with FLAIR sequences in NIfTI format. Preprocessing steps were identical to those prescribed by Bhattacharya *et al.*[10] MS volumes of sizes 35 × 35 × 35, 40 × 40 × 40, and 45 × 45 × 45 were extracted from the larger head and neck MRI for the 3D CNN. Our MS volume extraction and dataset samples are shown in figure 1.

Training, validation and test splits: We perform 10-fold patient stratified cross validation set experiments. Altogether, we have 9810 MS volumes in the training set, 1110 MS volumes in the validation set and 1230 MS volumes in the test set. 32% of the MS volumes have anomalies in the training, validation and test sets.

Send correspondence to D.B.)
D.B.: E-mail: debayan.bhattacharya@tuhh.de, * equal contribution

Medical Imaging 2024: Computer-Aided Diagnosis, edited by Weijie Chen, Susan M. Astley, Proc. of SPIE Vol. 12927, 1292717 · © 2024 SPIE · 1605-7422 · doi: 10.1117/12.3005515

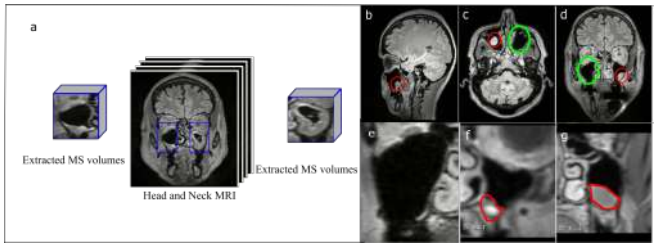

Figure 1. (a) Extracting MS volumes from single head and neck MRI. These MS volumes are passed to the deep learning models. (b) Cyst in the right MS (c) Polyp in the left MS (d) Cyst in the left MS (e) Normal MS of size $35 \times 35 \times 35$ (f) MS of size $40 \times 40 \times 40$ with polyp highlighted in red (g) MS of size $45 \times 45 \times 45$ with cyst highlighted in red

2.2 Convolutional Transformer Network (ConTra-Net)

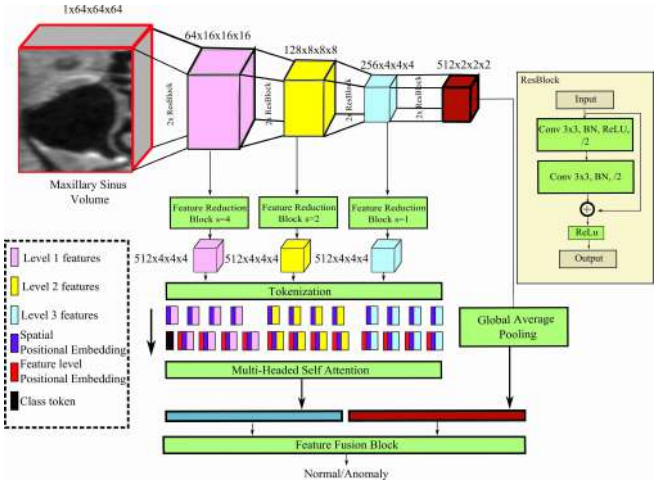

Figure 2. Illustration of ConTra-Net. The s in Feature Reduction Block represents stride of a convolution operation. The pink, yellow and cyan features represent the low, mid and high level features of the CNN.

ConTra-Net improves upon ViT's tokenization process by preserving spatial topology, considering the locality of features through convolutional kernels, and optimizing self-attention. This addresses the drawbacks of destroying spatial topology, insufficient capture of spatial context by feed forward networks, and challenges with optimizing self-attention in ViT. ConTra-Net combines the strengths of CNNs' inductive bias and the ability to capture global cues through the MHSA block. A figure of our proposed ConTra-Net is shown in Figure 2.

Feature extraction: The input MS volume $X \in \mathbb{R}^{H \times W \times D}$ is passed through a 3D CNN $F(.)$ with L stages of 3D residual blocks. Formally, it can be expressed as

$$f_l = \mathcal{F}_l(x;\theta) \mid x \in \mathbb{R}^{C_{l-1} \times \frac{H}{2^{l-1}} \times \frac{W}{2^{l-1}} \times \frac{D}{2^{l-1}}}$$

Here, l is the l-th feature level. $f_l \in \mathbb{R}^{C_l \times \frac{H}{2^l} \times \frac{W}{2^l} \times \frac{D}{2^l}}$ represents the feature maps at level l of 3D CNN $\mathcal{F}_l$ $(.)$. C_l represents the channel dimension at the l-th feature level and θ represents parameters of the CNN.

Feature reduction block: In order to allow our proposed ConTra-Net to capture global context while keeping computational costs low, we downsample the resolution and increase the channel dimension of f_l. The feature transformations are performed using 3D depthwise separable convolutions. The depthwise convolution operation spatially downsamples f_l while the pointwise convolution operation increases the channel dimension. Formally, we can express it as

$$\hat{f}_l = \Psi_l(\Upsilon(f_l;k,s);c_{in},c_{out})$$

Here $\Upsilon(.;k,s)$ is the depthwise convolution with kernel size k and stride s. The stride controls the downsample factor of the feature map f_l. $\Psi_l(.;c_{in},c_{out})$ is a 3D Convolution operation of kernel size 1 with input and output channel dimensions c_{in} and c_{out} respectively. Note, $c_{in} = C_l$.

Multi-head self-attention (MHSA) block: The multi-scale features f_l are transformed to $\hat{f}_l \in \mathbb{R}^{c_{out} \times h \times w \times d}$ by the feature reduction block. The $\hat{f}_l$ represent the tokens for the MHSA block. They are flattened into a sequence $x_l \in \mathbb{R}^{N \times c_{out}}$. Here, x_l is

the sequence of the l-th feature level and $N = h \cdot w \cdot d$ is the sequence length. c_{out} represents the embedding dimension. In order to retain the spatial positional information, we add a learnable positional embedding $p_{spatial} \in \mathbb{R}^{N \times c_{out}}$. In other words, we compute $\hat{f_l}^{pos} = \hat{f_l} + p_{spatial}$. Then, we concatenate the features $\hat{f_l}^{pos}$ arising from different feature levels l into a single large sequence $f_t \in \mathbb{R}^{N_{total} \times c_{out}}$. Here, $N_{total} = N \times (L-1)$ represents the sequence length that spans multiple levels of features. We leave out the features arising from the L-th feature block. To retain the positional information of the low and high level features originating from different parts of the CNN, we further add a learnable positional embedding $p_{level} \in \mathbb{R}^{N_{total} \times c_{out}}$. Formally, $\hat{f_t} = f_t + p_{level}$. Finally, we concatenate the class token $cls \in \mathbb{R}^{1 \times c_{out}}$ to $\hat{f_t}$ such that $\hat{f_t} = (\hat{f_t} \oplus cls) \in \mathbb{R}^{N_{total}+1 \times c_{out}}$. This resulting matrix $\hat{f_t}$ is passed through multiple layers of MHSA blocks and feed forward layers resulting in feature $F_t \in \mathbb{R}^{N_{total}+1 \times c_{out}}$.

Feature Fusion Block The feature fusion block concatenates the MHSA and CNN features and uses the combined feature vector to make class predictions. First, we consider the cls vector from F_t as the representative MHSA feature vector which encodes the global context. Second, we concatenate the CNN feature vector f_L and the MHSA feature vector. The resultant vector is passed through feed forward layers to make the final class prediction. We train ConTra-Net using class weighted cross-entropy loss.

3 Experiments, Results and Discussion

3.1 Implementation Details

For our 3D ViT implementation, we use patch sizes of $p = 4$ and embedding dimension of $C = 512$. The depth and number of attention heads of the MHSA are 2 and 4 respectively. For our 3D CNN, we use 3D variant of ResNet50.[18] Each ResNet has $L = 4$ stages of 3D residual blocks. $C_l = \{64, 128, 256, 512\}$ for our experiments. The depth and number of attention heads for our MHSA block is 2 and 4 respectively. Embedding dimension $c_{out} = 512$ for all our experiments. The feature reduction block downsamples the features to a resolution of $h = 4$, $w = 4$, $d = 4$ using strides 4, 2 and 1 for features arising from layers 1, 2 and 3 of the CNN respectively. This results in $N = 64$ and $N_{total} = 192$. With regards to the training configuration, we run our experiments for 100 epochs with a batch size of 16 for all our experiments. The learning rate was set at 0.0001 with a reduction by a factor of 10 if the validation loss did not improve for 5 epochs. Adam optimisation is used to train our deep learning models.

3.2 Classification performance

In our evaluation, we used Area Under Precision Recall Curve (AUPRC), Precision, Recall, and F1 to assess different methods. The positive class in our analysis was the "anomaly" class. As seen in Table 1, ResNet50 was the second best performing method. 3D ViT performed the worst, likely due to lack of pretraining on large scale dataset. Recall was challenging for all models due to the diverse morphological variations of anomalies. However, ConTra-Net achieved the highest AUPRC, Recall, and F1 scores, outperforming ResNet50 by 2.15%, 22.05%, and 11.68% respectively. Our results indicate that the combination of CNN and MHSA features leads to an improved classification performance.

Method	AUPRC	Precision	Recall	F1
ResNet 50	$\mu = 0.93, \sigma = 0.05$	$\mu = \mathbf{0.94}, \sigma = \mathbf{0.08}$	$\mu = 0.68, \sigma = 0.19$	$\mu = 0.77, \sigma = 0.13$
3D ViT	$\mu = 0.69, \sigma = 0.08$	$\mu = 0.69, \sigma = 0.14$	$\mu = 0.52, \sigma = 0.20$	$\mu = 0.56, \sigma = 0.10$
ConTra-Net	$\mu = \mathbf{0.95}, \sigma = \mathbf{0.04}$	$\mu = 0.92, \sigma = 0.10$	$\mu = \mathbf{0.83}, \sigma = \mathbf{0.15}$	$\mu = \mathbf{0.86}, \sigma = \mathbf{0.07}$

Table 1. Table of performance measures for different methods, with mean and standard deviation. Bold values indicate highest values.

3.3 Influence of low, mid and high level features on classification performance

Contra-Net employs a MHSA to enable interaction between low, mid, and high level features (pink, yellow, and cyan cubes in Figure 2). An evaluation of the importance of each feature level and their combinations was conducted, revealing that the combination of low and high-level features led to the greatest enhancement in recall performance (Table 2). These results suggest that the classification performance is most influenced by the interplay between low and high-level features.

Low-level	Mid-level	High-level	AUPRC	Precision	Recall	F1
		✓	$\mu = 0.93, \sigma = 0.05$	$\mu = \mathbf{0.94}, \sigma = \mathbf{0.06}$	$\mu = 0.78, \sigma = 0.11$	$\mu = 0.85, \sigma = 0.06$
	✓		$\mu = 0.94, \sigma = 0.03$	$\mu = 0.93, \sigma = 0.06$	$\mu = 0.78, \sigma = 0.19$	$\mu = 0.83, \sigma = 0.12$
	✓	✓	$\mu = 0.94, \sigma = 0.06$	$\mu = 0.93, \sigma = 0.07$	$\mu = 0.78, \sigma = 0.19$	$\mu = 0.83, \sigma = 0.12$
✓			$\mu = 0.94, \sigma = 0.03$	$\mu = 0.94, \sigma = 0.06$	$\mu = 0.76, \sigma = 0.15$	$\mu = 0.83, \sigma = 0.08$
✓		✓	$\mu = \mathbf{0.95}, \sigma = \mathbf{0.04}$	$\mu = 0.92, \sigma = 0.10$	$\mu = \mathbf{0.83}, \sigma = \mathbf{0.15}$	$\mu = \mathbf{0.86}, \sigma = \mathbf{0.07}$
✓	✓		$\mu = 0.93, \sigma = 0.04$	$\mu = 0.92, \sigma = 0.08$	$\mu = 0.71, \sigma = 0.24$	$\mu = 0.77, \sigma = 0.17$
✓	✓	✓	$\mu = 0.93, \sigma = 0.05$	$\mu = 0.91, \sigma = 0.08$	$\mu = 0.77, \sigma = 0.21$	$\mu = 0.82, \sigma = 0.09$

Table 2. Table of performance measures for different configurations of feature levels. Bold values indicate highest values.

3.4 Volume size

The size of the extracted MS volume is crucial. If it is too small, pathology may be missed or the sinuses may be only partially extracted. If it is too large, irrelevant anatomical information can hinder paranasal anomaly classification. We evaluated Contra-Net performance on MS volumes of sizes $35 \times 35 \times 35$, $40 \times 40 \times 40$, and $45 \times 45 \times 45$ voxels. Results in table 3 showed that AUPRC of $35 \times 35 \times 35$ and $40 \times 40 \times 40$ were same. AUPRC decreased for $45 \times 45 \times 45$ compared to $35 \times 35 \times 35$ and $40 \times 40 \times 40$, indicating limited impact on model performance beyond $40 \times 40 \times 40$. Our results indicate that including additional surrounding structures in larger volumes negatively affected paranasal anomaly classification. We thereby conclude that careful selection of volume size is vital for optimization.

Volume Size	AUPRC	Precision	Recall	F1
$35 \times 35 \times 35$	$\mu = \mathbf{0.95}, \sigma = \mathbf{0.04}$	$\mu = 0.92, \sigma = 0.10$	$\mu = \mathbf{0.83}, \sigma = \mathbf{0.15}$	$\mu = \mathbf{0.86}, \sigma = \mathbf{0.07}$
$40 \times 40 \times 40$	$\mu = \mathbf{0.95}, \sigma = \mathbf{0.04}$	$\mu = \mathbf{0.93}, \sigma = \mathbf{0.06}$	$\mu = 0.79, \sigma = 0.12$	$\mu = 0.85, \sigma = 0.09$
$45 \times 45 \times 45$	$\mu = 0.93, \sigma = 0.03$	$\mu = 0.90, \sigma = 0.09$	$\mu = 0.78, \sigma = 0.10$	$\mu = 0.83, \sigma = 0.06$

Table 3. Influence of volume size on classification performance. Bold values indicate highest values.

4 Conclusion

In this study, we introduced a novel hybrid CNN transformer, ConTra-Net, for the task of paranasal anomaly classification in the maxillary sinus. We observed that the recall metric has high standard deviation illustrating the difficulty of generalizing to unseen anomaly morphologies.We compared ConTra-Net to ResNet50 and 3D ViT, and also performed an ablation study to investigate the influence of low, mid and high level features towards the classification performance. ConTra-Net outperformed ResNet50 in AUPRC, Recall and F1, suggesting that learning global features using MHSA may be beneficial for this task. Interaction of low and high level features through MHSA proved to be the most beneficial towards paranasal anomaly classification. A limitations to our work is the need for further improvement in the F1 score of ConTra-Net in order to make it applicable to real-world clinical scenarios. Despite this limitation, our results provide a promising deep learning solution for paranasal anomaly classification in the maxillary sinus.

REFERENCES

[1] Wilson, R., Kuan Kok, H., Fortescue-Webb, D., Doody, O., Buckley, O., and Torreggiani, W. C., "Prevalence and seasonal variation of incidental mri paranasal inflammatory changes in an asymptomatic irish population," *Irish medical journal* **110**(9), 641 (2017).

[2] Hansen, A. G., Helvik, A.-S., Nordgård, S., Bugten, V., Stovner, L. J., Håberg, A. K., Gårseth, M., and Eggesbø, H. B., "Incidental findings in mri of the paranasal sinuses in adults: a population-based study (hunt mri)," *BMC ear, nose, and throat disorders* **14**(1), 13 (2014).

[3] Tarp, B., Fiirgaard, B., Christensen, T., Jensen, J. J., and Black, F. T., "The prevalence and significance of incidental paranasal sinus abnormalities on mri," *Rhinology* **38**(1), 33–38 (2000).

[4] Ma, Z. and Yang, X., "Research on misdiagnosis of space occupying lesions in unilateral nasal sinus," *Lin chuang er bi yan hou tou jing wai ke za zhi = Journal of clinical otorhinolaryngology, head, and neck surgery* **26**(2), 59–61 (2012).

[5] Jeon, Y., Lee, K., Sunwoo, L., Choi, D., Oh, D. Y., Lee, K. J., Kim, Y., Kim, J.-W., Cho, S. J., Baik, S. H., Yoo, R.-E., Bae, Y. J., Choi, B. S., Jung, C., and Kim, J. H., "Deep learning for diagnosis of paranasal sinusitis using multi-view radiographs," *Diagnostics (Basel, Switzerland)* **11**(2) (2021).

[6] Kim, Y., Lee, K. J., Sunwoo, L., Choi, D., Nam, C.-M., Cho, J., Kim, J., Bae, Y. J., Yoo, R.-E., Choi, B. S., Jung, C., and Kim, J. H., "Deep learning in diagnosis of maxillary sinusitis using conventional radiography," *Investigative radiology* **54**(1), 7–15 (2019).

[7] Kim, K.-S., Kim, B. K., Chung, M. J., Cho, H. B., Cho, B. H., and Jung, Y. G., "Detection of maxillary sinus fungal ball via 3-d cnn-based artificial intelligence: Fully automated system and clinical validation," *PLOS ONE* **17**, 1–19 (02 2022).

[8] Bhattacharya, D., Behrendt, F., Becker, B. T., Beyersdorff, D., Petersen, E., Petersen, M., Cheng, B., Eggert, D., Betz, C., Hoffmann, A. S., and Schlaefer, A., "Unsupervised anomaly detection of paranasal anomalies in the maxillary sinus," in [*Medical Imaging 2023: Computer-Aided Diagnosis*], Iftekharuddin, K. M. and Chen, W., eds., **12465**, 124651B, International Society for Optics and Photonics, SPIE (2023).

[9] Bhattacharya, D., Becker, B. T., Behrendt, F., Bengs, M., Beyersdorff, D., Eggert, D., Petersen, E., Jansen, F., Petersen, M., Cheng, B., Betz, C., Schlaefer, A., and Hoffmann, A. S., "Supervised contrastive learning to classify paranasal anomalies in the maxillary sinus," in [*Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*], Wang, L., Dou, Q., Fletcher, P. T., Speidel, S., and Li, S., eds., 429–438, Springer Nature Switzerland, Cham (2022).

[10] Bhattacharya, D., Behrendt, F., Becker, B. T., Beyersdorff, D., Petersen, E., Petersen, M., Cheng, B., Eggert, D., Betz, C., Hoffmann, A. S., and Schlaefer, A., "Multiple instance ensembling for paranasal anomaly classification in the maxillary sinus," (2023).

[11] Gao, W., Wan, F., Pan, X., Peng, Z., Tian, Q., Han, Z., Zhou, B., and Ye, Q., "Ts-cam: Token semantic coupled attention map for weakly supervised object localization," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* , 2866–2875 (2021).

[12] Yu, S., Ma, K., Bi, Q., Bian, C., Ning, M., He, N., Li, Y., Liu, H., and Zheng, Y., "Mil-vt: Multiple instance learning enhanced vision transformer for fundus image classification," in [*Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*], de Bruijne, M., Cattin, P. C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., and Essert, C., eds., 45–54, Springer International Publishing, Cham (2021).

[13] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N., "An image is worth 16x16 words: Transformers for image recognition at scale," in [*International Conference on Learning Representations*], (2021).

[14] Jang, J. and Hwang, D., "M3t: three-dimensional medical image classifier using multi-plane and multi-slice transformer," in [*2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 20686–20697 (2022).

[15] Dai, Y., Gao, Y., and Liu, F., "Transmed: Transformers advance multi-modal medical image classification," *Diagnostics* **11**(8) (2021).

[16] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., and Zhou, Y., "Transunet: Transformers make strong encoders for medical image segmentation," *CoRR* **abs/2102.04306** (2021).

[17] Dai, Y., Gao, Y., and Liu, F., "TransMed: Transformers advance Multi-Modal medical image classification," *Diagnostics (Basel)* **11** (July 2021).

[18] Hara, K., Kataoka, H., and Satoh, Y., "Learning spatio-temporal features with 3d residual networks for action recognition," (2017).

7.4 Supervised Contrastive Learning to Classify Paranasal Anomalies in the Maxillary Sinus

Title of paper: Supervised Contrastive Learning to Classify Paranasal Anomalies in the Maxillary Sinus

Conference: Medical Image Computing and Computer Assisted Intervention Society

Year: 2022

Topic: Improving supervised learning using contrastive loss

Supervised Contrastive Learning to Classify Paranasal Anomalies in the Maxillary Sinus

Debayan Bhattacharya[1,2(✉)], Benjamin Tobias Becker[2], Finn Behrendt[1], Marcel Bengs[1], Dirk Beyersdorff[3], Dennis Eggert[2], Elina Petersen[4], Florian Jansen[2], Marvin Petersen[5], Bastian Cheng[5], Christian Betz[2], Alexander Schlaefer[1], and Anna Sophie Hoffmann[2]

[1] Insititute of Medical Technologies and Intelligent Systems, Hamburg University of Technology, Hamburg, Germany
debayan.bhattacharya@tuhh.de, d.bhattacharya@uke.de
[2] Department of Otorhinolaryngology, Head and Neck Surgery and Oncology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany
[3] Clinic and Polyclinic for Diagnostic and Interventional Radiology and Nuclear Medicine, University Medical Center Hamburg-Eppendorf, Hamburg, Germany
[4] Population Health Research Department, University Heart and Vascular Center, University Medical Center Hamburg-Eppendorf, Hamburg, Germany
[5] Clinic and Polyclinic for Neurology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Abstract. Using deep learning techniques, anomalies in the paranasal sinus system can be detected automatically in MRI images and can be further analyzed and classified based on their volume, shape and other parameters like local contrast. However due to limited training data, traditional supervised learning methods often fail to generalize. Existing deep learning methods in paranasal anomaly classification have been used to diagnose at most one anomaly. In our work, we consider three anomalies. Specifically, we employ a 3D CNN to separate maxillary sinus volumes without anomaly from maxillary sinus volumes with anomaly. To learn robust representations from a small labelled dataset, we propose a novel learning paradigm that combines contrastive loss and cross-entropy loss. Particularly, we use a supervised contrastive loss that encourages embeddings of maxillary sinus volumes with and without anomaly to form two distinct clusters while the cross-entropy loss encourages the 3D CNN to maintain its discriminative ability. We report that optimising with both losses is advantageous over optimising with only one loss. We also find that our training strategy leads to label efficiency. With our method, a 3D CNN classifier achieves an AUROC of 0.85 ± 0.03

A. Schlaefer and A. S. Hoffmann—These authors contributed equally to this work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-16437-8_41.

L. Wang et al. (Eds.): MICCAI 2022, LNCS 13433, pp. 429–438, 2022.
https://doi.org/10.1007/978-3-031-16437-8_41

while a 3D CNN classifier optimised with cross-entropy loss achieves an AUROC of 0.66 ± 0.1. Our source code is available at https://github.com/dawnofthedebayan/SupConCE_MICCAI_22.

Keywords: Self-supervised learning · Paranasal pathology · Nasal pathology · Magnetic resonance images

1 Introduction

Paranasal sinus anamolies are common incidental findings reported in patients who undergo diagnostic imaging of the head [21] for neuroradiological assessment. Understanding the different opacifications of the paranasal sinuses is very useful, because the frequency of these findings represent clinical challenges [5], and little is known about the incidence and significance of these morphological changes in the general population. There have been numerous studies on analysing the occurrence and progression of these incidental findings [2,16–19], but mostly in patients with sinunasal symptoms.

In our study, elderly people (45–74 years) received an MRI for neuroradiological assessment [7] in the city of Hamburg. The purpose of our study is to find out what percentage of patients, who do not have sinunasal symptoms, show findings in the paranasal sinus system in MRI images and if it is possible to detect anomalies in the paranasal sinus system using deep learning techniques and further analyze and classify based on their volume, shape and other parameters like local contrast.

A three year retrospective study showed malignant tumors were misdiagnosed as nasal polyps with a misdiagnosis rate of 5.63% while inverted papilloma were misdiagnosed as nasal polyps with a rate of 8.45% [12]. As a first step, it would be beneficial to separate MRI with any paranasal anomaly from normal MRI using Computer Aided Diagnostics (CAD). This would allow the physicians to closely inspect the MRIs containing paranasal anomaly with finer detail. This can reduce chances of misdiagnosis while decreasing the workload of physicians from having to see normal MRIs.

There have been works of paranasal sinus anomaly diagnosis using supervised learning [8,10,11]. However, all these works have been used to classify at most one anomaly. Our work considers three anomalies namely: (i) mucosal thickening (ii) polyps and (iii) cysts. The anomalies are differently located within the maxillary sinus and therefore we use a 3D volume of the maxillary sinus as input to our 3D CNN. Additionally, existing works train on large datasets to achieve good performance. However, labelling is a time consuming task and in some scenarios it requires the supervision of Ears, Nose and Throat (ENT) specialised radiologists who may not be immediately available. In our case, we have a small dataset of 199 MRI volumes. To mitigate the limitations of our small labelled dataset, we use contrastive learning [14] to learn robust representations that does not overfit on the training set. In contrastive learning, an encoder learns to map positive pairs close together in the embedding space while pushing away

negative pairs. In SimCLR [1], an image is transformed twice through random transformations. The transformed "views" of the image constitute a positive pair and every other image in the mini-batch is used to construct negative pairs with the reference image. There has been significant research in finding the best transformations as the chosen transformations dictate the quality of learnt representation [1]. The underlying assumption is that transformations augment the image while preserving the semantic information. However, in our case two or more images in the mini-batch can be semantically similar as they belong to the same class. Particularly, a maxillary sinus of one patient can be semantically similar to another patient's maxillary sinus if both patients do not exhibit any paranasal anomalies. Furthermore, they can be different as well due to the anatomical variations of the maxillary sinus [17,19]. Since our dataset contains anomalies, it is also important that the classifier does not overfit to a particular anomaly. Therefore, it is important for a classifier to learn anatomically invariant representation of the maxillary sinus for volumes that contain no anomaly and learn anomaly invariant representation of the maxillary sinus for volumes exhibiting one of the three anomalies. This can reduce the chances of overfitting on the training set. This motivates us to employ a supervised contrastive learning approach [9] that brings embeddings of maxillary sinus volumes without anomaly closer together while pushing away embeddings of maxillary sinus with anomaly in the embedding space and vice versa. Compared to the original method [9], we propose a training strategy that simultaneously trains our 3D CNN using the supervised contrastive loss and regular cross-entropy loss using two different projection networks. The reasoning behind this is that minimizing the contrastive loss encourages the 3D CNN to learn representations that are robust to anatomical and anomaly variations while the cross-entropy enforces the 3D CNN to preserve discriminative ability.

In summary, our contributions are three-fold. First, we demonstrate the feasibility of a deep learning approach to classify between normal and anomalous maxillary sinuses. Second, we demonstrate through extensive experiments that combining supervised contrastive loss and cross-entropy loss is the better approach to improve the discriminative ability of the 3D CNN classifier. Third, we empirically show that our method is the most label efficient.

2 Method

Our method is shown in Fig. 1. In global contrastive learning methods such as SimCLR [1], we learn a parametric function $F_\theta : X \rightarrow \mathbb{R}^D$ where $\mathbb{R}^D$ is a unit hypersphere. F_θ is trained to map semantically similar samples closer together and semantically dissimilar samples further apart through the InfoNCE loss [20]. The underlying assumption here is that images in the mini-batch are semantically dissimilar. This is untrue in our case as we have maxillary sinus volumes belonging to one of the two classes and there are semantic similarities and dissimilarities in the intra and inter class samples. Therefore, it is not possible to form meaningful clusters from the global contrastive loss described in SimCLR

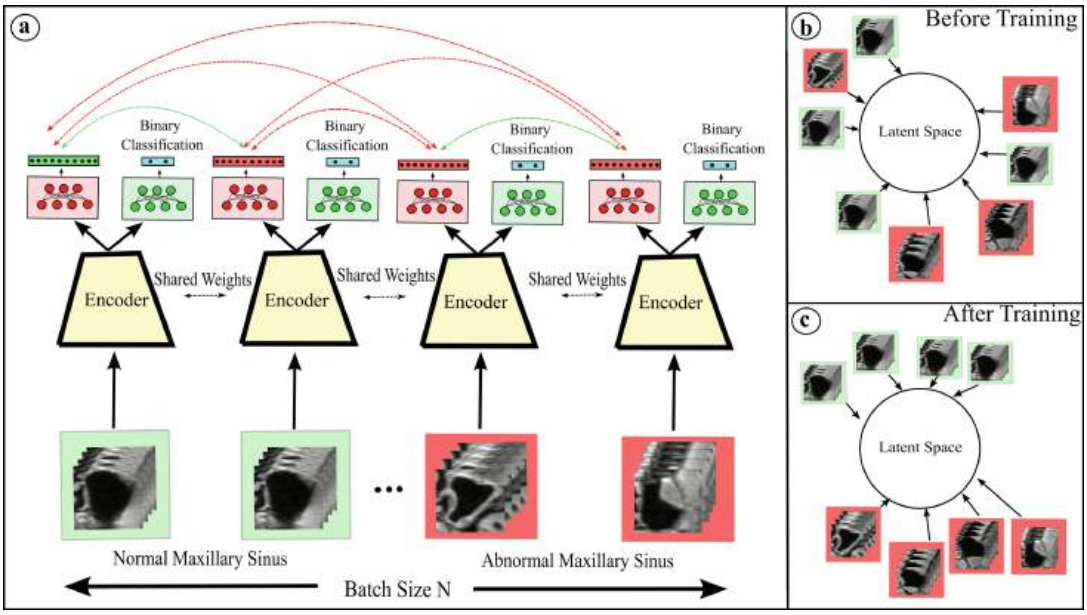

Fig. 1. (a) Our proposed method. The curved green and red lines represent similar and dissimilar representations respectively. (b)–(c) Illustration of the latent space embedding of normal and anomalous maxillary sinuses before and after the encoder is trained respectively. (Color figure online)

[1]. Therefore, to incorporate the class priors, we sample volumes from the two classes explicitly. Given an input mini-batch $B = \{x_1, x_2, ..., x_N\}$ where x_i represent the input 3D volume, a random transformation set T is used to form a pair of augmented volumes. Instead of randomly sampling N samples, we randomly sample $N/2$ maxillary sinus volumes without anomaly and $N/2$ maxillary sinus volumes with anomaly from our dataset. Each of these $N/2$ subsets undergo random transformation twice using T. Let us denote the set containing all the augmented volumes of the C classes as $\mathbb{M} = \bigcup_{c=1}^{C} M_c$. In our case, $C = 2$. Here, M_c is the subset of augmented volumes belonging to a single class and $|M_c| = N$ is its cardinality. Let $m_i, i \in \mathbb{I} = \{1, 2, ..., 2N\}$ represent the augmented volumes in set $\mathbb{M}$ and $m_{k(i)}$ is its corresponding volume augmented from the same volume in B. Furthermore, let I_c represent indices of all the augmented volumes belonging to class c such that $\mathbb{I} = \bigcup_{c=1}^{C} I_c$. Using the above stated assumptions, the InfoNCE loss that takes into consideration the class priors when making positive and negative pairs can be written as:

$$L_{simclr} = -\sum_{c=1}^{C} \frac{1}{|M_c|} \sum_{i \in I_c} log \frac{e^{sim(Z_i, Z_{k(i)})/\tau}}{e^{sim(Z_i, Z_{k(i)})/\tau} + \sum_{j \in \mathbb{I} \setminus I_c} e^{sim(Z_i, Z_j)/\tau}} \quad (1)$$

where $\tau \in \mathbb{R}^+$ is the scalar temperature parameter, $Z_i = F_\theta^{con}(m_i)$ is the normalised feature vector such that $F_\theta^{con}(.) = Proj_1(Enc(.))$, $k(i)$ is the index of the corresponding volume in $\mathbb{M}$ augmented from the same volume in B and $sim(.)$ is the cosine similarity function. Although Eq. 1 only constructs negative pairs

such that the two volumes are from different classes, it still constructs positive pairs by augmenting the same volume using the random augmentation set T. As a result, we are reliant on the transformations to learn meaningful representations. However, the transformations do not guarantee anatomical and anomaly invariance as the encoder is not incentivised to produce similar representations for volumes belonging to the same class in the mini-batch. Therefore, we use a supervised contrastive loss [9] that constructs arbitrary number of positive pairs where each volume in the pair is from the same class but unique in the mini-batch. Formally, the supervised contrastive loss can be described as shown below:

$$L_{sc} = -\sum_{c=1}^{C} \frac{1}{|M_c|} \sum_{i \in I_c} log \frac{\sum_{j \in I_c \setminus \{i\}} e^{sim(Z_i, Z_j)/\tau}}{\sum_{j \in I_c \setminus \{i\}} e^{sim(Z_i, Z_j)/\tau} + \sum_{j \in \mathbb{I} \setminus I_c} e^{sim(Z_i, Z_j)/\tau}} \tag{2}$$

The main differences of Eq. 2 compared to Eq. 1 is that numerous positive pairs are constructed in the numerator by matching every volume with every other volume belonging to the same class in the mini-batch. In this case, $|M_c| = N/2$ as we do not use T and $\mathbb{I} = \{1, 2, ..., N\}$. This incentivises the encoder to give similar representations for volumes belonging to the same class. This leads to learning anatomical and anomaly invariant representations. Apart from using L_{sc} we also use regular cross-entropy loss to preserve the discrimintative ability of our 3D CNN. The cross-entropy loss is formalised as follows:

$$L_{ce} = -\frac{1}{N} \sum_{i \in \mathbb{I}} y_i log(F_\theta^{class}(m_i)) \tag{3}$$

Here, $F_\theta^{class}(.)$ can be decomposed into $Proj_2(Enc(.))$ and y_i is the class label of m_i such that $y_i \in \{0, 1\}$. Therefore, our combined loss function is

$$L_{ours} = L_{sc} + \lambda L_{ce} \tag{4}$$

In our case, we set $\lambda = 1$. In summary, we train models using only L_{ce} and set this as the baseline. We then train models using L_{simclr} to show that transformation invariance does not help in our downstream classification task. Next, we train our models using L_{sc} to show the benefit of clustering based on class priors and the importance of anatomical and anomaly invariant representation learning. Finally, we train our models using L_{ours} to show that contrastive loss and cross-entropy loss improve the discriminative ability of the models and overfit the least on the training set.

2.1 Dataset

As part of the population study [7], MRI of the head and neck area of participants based in the city of Hamburg, Germany were recorded. The MRIs were recorded at University Medical Center Hamburg-Eppendorf. The age group of the

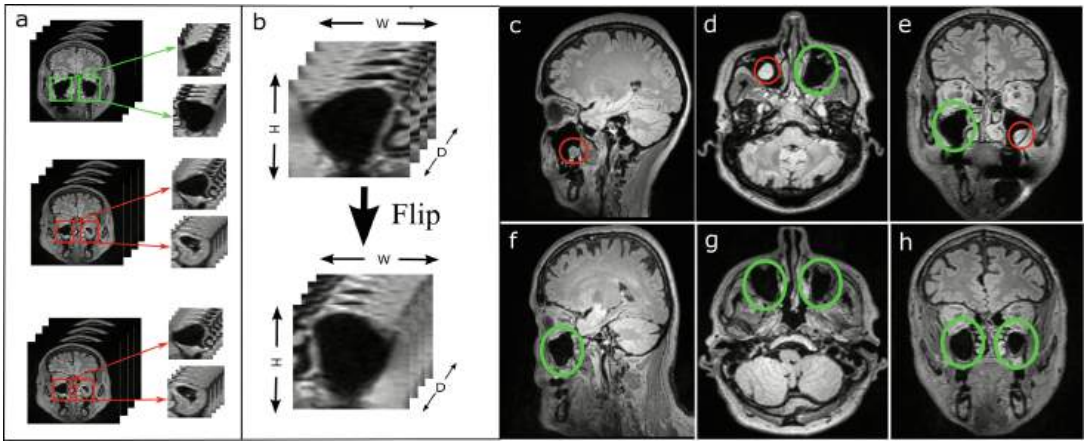

Fig. 2. [LEFT] Pre-processing steps involve (a) Extraction of 3D sub-volumes of left and right maxillary sinus from FLAIR MRI samples. (b) Flipping the coronal plane slices of right maxillary sinus sub-volume to give it the appearance of left maxillary sinus sub-volume. [RIGHT] FLAIR-MRI slices from the sagittal, axial and coronal views illustrating the difference between anomaly and normal class. The red circles denote anomalies and green circles denote normal maxillary sinus. (c) Cyst observed in right maxillary sinus. (d) Polyp observed in the left maxillary sinus (e) Cyst observed in left maxillary sinus. (f)–(h) FLAIR-MRI slices with no pathology. (Color figure online)

participants were between 45 and 74 years. Each participant had T1-weighted and fluid attenuated inversion recovery (FLAIR) sequences stored in the NIfTI[1] format. FLAIR-MRIs were chosen as the imaging modality as the incidental findings are more visible due to the higher contrast relative to T1 weighted MRI. The labelled dataset consists of 199 FLAIR-MRI volumes of which 106 patients exhibit normal maxillary sinuses and 93 patients have maxillary sinuses with anomaly in at least one maxillary sinus. The diagnosis of the observed pathology in 199 FLAIR-MRIs was confirmed by two ENT surgeons and one ENT specialised radiologist. The incidental findings are categorised and defined as follows: (i) mucosal thickening (ii) polyps (iii) cysts. The statistics of the pathology observed is reported in the supplementary material. In this work, all the anomalies are grouped into a single class called "anomaly" and all the normal maxillary sinuses are grouped into a class called "normal". Altogether, there are 269 maxillary sinus volumes without anomaly and 130 abnormal maxillary sinus volumes with anomaly. Each MRI has a resolution of $173 \times 319 \times 319$ voxels along the saggital, coronal and axial directions respectively. The voxel size is $0.53\ \text{mm} \times 0.75\ \text{mm} \times 0.75\ \text{mm}$.

Preprocessing: We performed rigid registration by randomly selecting one FLAIR-MRI sample as a fixed volume followed by resampling to a dimension of $128 \times 128 \times 128$. Of the resampled volumes, we extracted two sub-volumes, one for each maxillary sinus from a single patient. We made sure that the extracted sub-volumes subsumed the maxillary sinus.

[1] https://nifti.nimh.nih.gov/.

Since the maxillary sinus are symmetric, we horizontally flipped the coronal planes of right maxillary sinus volumes to make it look like left maxillary sinus volumes. The decision of which maxillary sinus to flip was arbitrary. Ultimately, these sub-volumes were reshaped to a standard size of $32 \times 32 \times 32$ voxels for the 3D CNN. Finally, all the maxillary sinus volumes were normalised to the range of -1 to 1. Our preprocessing step is shown in Fig. 2.

Training, Validation and Test Split: We perform a nested stratified K-fold with 5 inner and 5 outer folds. The inner fold was used to choose the best hyperparameters. In summary, each experiment has 80 volumes in test set, 64 volumes in cross validation set and 255 samples in training set. The folds are constructed by preserving the percentage of samples for the two classes.

3 Experiments and Discussion

3.1 Implementation Details

All of our experiments are implemented in PyTorch [15] and PyTorch Lightning [4]. We use a batch size of 128 for all our experiments based on hyperparameter tuning (See supplementary material). We use Adam Optimization with a learning rate of 1e-4. Similar to Chen et al. [1], we fix $\tau = 0.1$. Our encoder $Enc(.)$ is the 3DResNet18 [6]. Our projection layer $Proj_1(.)$ is a fully connected layer with input dimension 512 and output dimension 128. A ReLU activation is placed in between the layers. Projection layer $Proj_2(.)$ is a linear fully connected layer with input and output dimensions of 512 and 2 respectively. All our models are trained for 200 epochs.

3.2 Evaluation of Learnt Representations

Table 1. Evaluation of our representations

Method	L_{ce}	L_{simclr}	L_{sc}	Accuracy	F1 (weighted)	AUROC	AUPRC
Scratch	✓			0.68 ± 0.03	0.57 ± 0.04	0.66 ± 0.10	0.53 ± 0.12
Scratch (with aug)	✓			0.69 ± 0.03	0.58 ± 0.05	0.68 ± 0.10	0.56 ± 0.10
SimCLR		✓		0.65 ± 0.03	0.58 ± 0.04	0.54 ± 0.09	0.41 ± 0.09
SupCon [9]			✓	0.72 ± 0.05	0.70 ± 0.06	0.73 ± 0.08	0.61 ± 0.08
Ours	✓		✓	$\mathbf{0.80 \pm 0.02}$	$\mathbf{0.78 \pm 0.03}$	$\mathbf{0.85 \pm 0.03}$	$\mathbf{0.78 \pm 0.03}$

The metrics we have used to evaluate our representations are accuracy, F1 weighted, Area Under Receiver Operator Characteristics (AUROC) and Area Under Precision Recall Curve (AUPRC). We chose F1 weighted and AUPRC as they give a fair assessment of the performance of the models in the presence of class imbalance. Our random transformation set T consists of random affine, flip and gaussian noise as these are semantic preserving transforms. For training models with L_{simclr}, L_{sc} and L_{ours} we followed the training strategy followed

by Khosla et al. [9]. The inference is performed using $Proj_2(.)$ for all the experiments. We test for statistically significant difference in our performance metrics using a permutation test with nP= 10000 samples and a significance level of α = 0.05 [3]. The difference in the AUROC, AUPRC, F1 weighted and accuracy of our method is significant (p $<$ 0.05) compared to the other methods. From the results in Table 1 we observe that the models trained with only L_{ce} overfit and do not generalize well. Models trained with L_{sc} achieve a significant boost in all metrics compared to models trained using L_{ce} and L_{simclr}. We conjecture this to be the case because the supervised contrastive loss clusters the maxillary sinus volumes representations based on its class leading to invariant representations and less overfitting. The absence of L_{ce} causes the clusters to be more spread out (See Fig. 3 SupCon). Our loss L_{ours} shows the best performance due to the increased discriminative ability of the classifier which is reflected by the formation of compact clusters (See Fig. 3 Ours). SimCLR performs the worst as they fail to form meaningful clusters (See Fig. 3 SimCLR).

3.3 Label Efficient Representation Learning

We evaluated the performance of models trained with the different loss functions by supplying 60% and 80% of training samples. We excluded SimCLR approach because it performed very poorly on 60% and 80% of training set. We performed

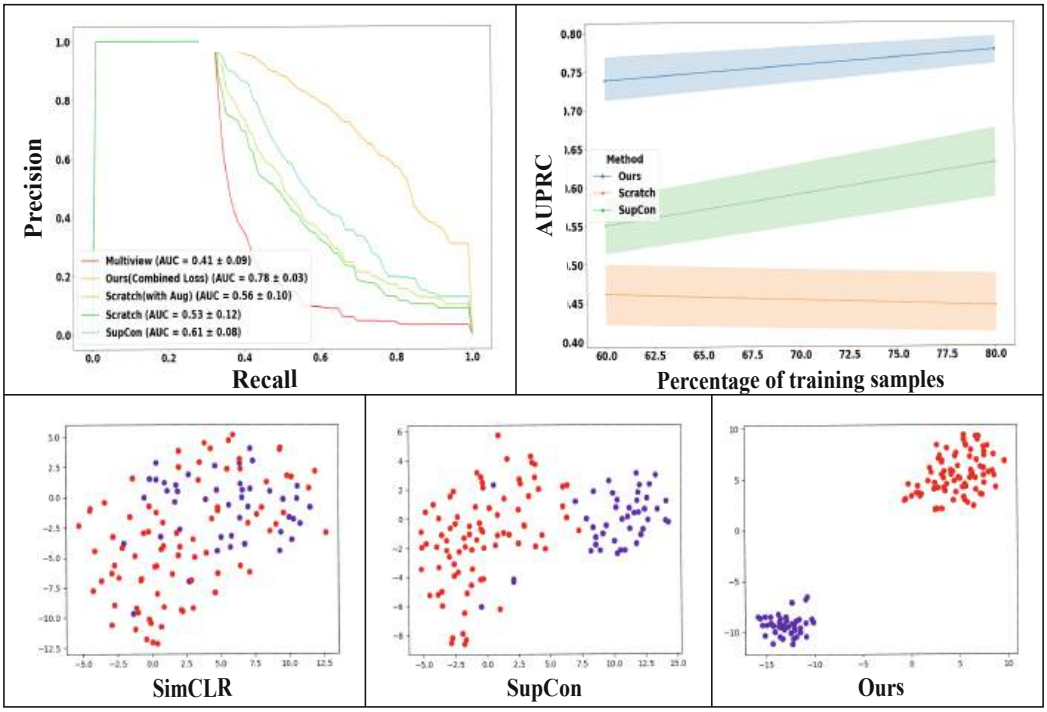

Fig. 3. (UPPER LEFT) Precision Recall Curve. (UPPER RIGHT) Illustrating the label efficiency of our method by plotting the AUPRC against the percentage of training dataset (BOTTOM LEFT, MIDDLE, RIGHT) t-SNE [13] with perplexity = 30, learning rate = 200, iterations=1000 used to visualise the representations learnt by various contrastive losses. The red dots denote normal class and purple dots denote anomaly class.

a five-fold cross validation experiment with the same training strategy. The test set is the same for all the experiments. The AUPRC of the models are displayed in the upper right graphic of Fig. 3. We observe that models trained using L_{sc} outperform the baseline by a significant margin. Even with limited data, both the losses (L_{sc} and L_{ours}) show improved performance with the injection of more training data. We also observe that L_{ours} almost achieves AUPRC of 100% training set and it is also the most label-efficient. These results reveal that our approach can reduce labelling effort of physicians to an extent thereby allowing them to invest more time in the diagnosis and evaluation of difficult clinical cases.

4 Conclusion

Previous works on population studies [2,16–19] have relied on manual analysis by physicians. Our work is a first step towards bringing automation in such studies. We show the benefit of contrastive loss in the classification of anomalies in the maxillary sinus from limited labelled dataset. Furthermore, we report the performance improvements relative to regular cross-entropy. Specifically, clustering based on class priors is helpful to learn representations that overfit less on the training set. We also show that a combinination of cross-entropy loss and supervised contrative loss improve the performance of the model. Finally, compared to other works that use deep learning for paranasal inflammation study [8,10,11], ours is the first work that tries to achieve label efficiency and thus attempts to reduce the workload of physicians. A limitation of our work is that the classification accuracy is still not satisfactory. As future work, we plan to label more MRIs and even perform classification of the type of anomaly observed in the maxillary sinus.

References

1. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. https://arxiv.org/pdf/2002.05709
2. Cooke, L.D., Hadley, D.M.: MRI of the paranasal sinuses: incidental abnormalities and their relationship to symptoms. J. Laryngol. Otol. **105**(4), 278–281 (1991). https://doi.org/10.1017/s0022215100115609
3. Efron, B., Tibshirani, R.: An Introduction to the Bootstrap, Monographs on Statistics and Applied Probability, vol. 57, [nachdr.] edn. Chapman & Hall, Boca Raton (1998)
4. Falcon, F.N., et al.: Pytorch lightning, vol. 3. GitHub (2019). https://github.com/PyTorchLightning/pytorch-lightning
5. Hansen, A.G., et al.: Incidental findings in MRI of the paranasal sinuses in adults: a population-based study (hunt MRI). BMC Ear Nose Throat Disord. **14**(1), 13 (2014). https://doi.org/10.1186/1472-6815-14-13
6. Hara, K., Kataoka, H., Satoh, Y.: Learning spatio-temporal features with 3D residual networks for action recognition. http://arxiv.org/pdf/1708.07632v1

7. Jagodzinski, A., et al.: Rationale and design of the Hamburg city health study. Eur. J. Epidemiol. **35**(2), 169–181 (2019). https://doi.org/10.1007/s10654-019-00577-4
8. Jeon, Y., et al.: Deep learning for diagnosis of paranasal sinusitis using multi-view radiographs. Diagnost. (Basel Switz.) **11**(2) (2021). https://doi.org/10.3390/diagnostics11020250
9. Khosla, P., et al.: Supervised contrastive learning. https://arxiv.org/pdf/2004.11362
10. Kim, Y., et al.: Deep learning in diagnosis of maxillary sinusitis using conventional radiography. Invest. Radiol. **54**(1), 7–15 (2019). https://doi.org/10.1097/RLI.0000000000000503
11. Liu, G.S., et al.: Deep learning classification of inverted papilloma malignant transformation using 3d convolutional neural networks and magnetic resonance imaging. Int. Forum Allergy Rhinol. (2022). https://doi.org/10.1002/alr.22958
12. Ma, Z., Yang, X.: Research on misdiagnosis of space occupying lesions in unilateral nasal sinus. Lin chuang er bi yan hou tou jing wai ke za zhi = J. Clin. Otorhinolaryngol. Head Neck Surg. **26**(2), 59–61 (2012). https://doi.org/10.13201/j.issn.1001-1781.2012.02.005
13. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(86), 2579–2605 (2008). http://jmlr.org/papers/v9/vandermaaten08a.html
14. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. CoRR abs/1807.03748 (2018). http://arxiv.org/abs/1807.03748
15. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. https://arxiv.org/pdf/1912.01703
16. Rak, K.M., Newell, J.D., Yakes, W.F., Damiano, M.A., Luethke, J.M.: Paranasal sinuses on MR images of the brain: significance of mucosal thickening. AJR Am. J. Roentgenol. **156**(2), 381–384 (1991). https://doi.org/10.2214/ajr.156.2.1898819
17. Rege, I.C.C., Sousa, T.O., Leles, C.R., Mendonça, E.F.: Occurrence of maxillary sinus abnormalities detected by cone beam CT in asymptomatic patients. BMC Oral Health **12**, 30 (2012). https://doi.org/10.1186/1472-6831-12-30
18. Stenner, M., Rudack, C.: Diseases of the nose and paranasal sinuses in child. GMS Curr. Top. Otorhinolaryngol. Head Neck Surg. **13**, Doc10 (2014). https://doi.org/10.3205/cto000113
19. Tarp, B., Fiirgaard, B., Christensen, T., Jensen, J.J., Black, F.T.: The prevalence and significance of incidental paranasal sinus abnormalities on MRI. Rhinology **38**(1), 33–38 (2000)
20. den van Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. https://arxiv.org/pdf/1807.03748
21. Wilson, R., Kuan Kok, H., Fortescue-Webb, D., Doody, O., Buckley, O., Torreggiani, W.C.: Prevalence and seasonal variation of incidental MRI paranasal inflammatory changes in an asymptomatic irish population. Ir. Med. J. **110**(9), 641 (2017)

7.5 Multiple instance ensembling for paranasal anomaly classification in the maxillary sinus

Title of paper: Multiple instance ensembling for paranasal anomaly classification in the maxillary sinus

Journal: International Journal of Computer Assisted Radiology and Surgery

Year: 2023

Topic: Improving supervised learning using multiple instance ensembling

International Journal of Computer Assisted Radiology and Surgery
https://doi.org/10.1007/s11548-023-02990-3

ORIGINAL ARTICLE

Multiple instance ensembling for paranasal anomaly classification in the maxillary sinus

Debayan Bhattacharya[1,2] · Finn Behrendt[1] · Benjamin Tobias Becker[2] · Dirk Beyersdorff[3] · Elina Petersen[4] · Marvin Petersen[5] · Bastian Cheng[5] · Dennis Eggert[2] · Christian Betz[2] · Anna Sophie Hoffmann[2] · Alexander Schlaefer[1]

Received: 12 January 2023 / Accepted: 27 June 2023

Abstract

Purpose Paranasal anomalies are commonly discovered during routine radiological screenings and can present with a wide range of morphological features. This diversity can make it difficult for convolutional neural networks (CNNs) to accurately classify these anomalies, especially when working with limited datasets. Additionally, current approaches to paranasal anomaly classification are constrained to identifying a single anomaly at a time. These challenges necessitate the need for further research and development in this area.

Methods We investigate the feasibility of using a 3D convolutional neural network (CNN) to classify healthy maxillary sinuses (MS) and MS with polyps or cysts. The task of accurately localizing the relevant MS volume within larger head and neck Magnetic Resonance Imaging (MRI) scans can be difficult, but we develop a strategy which includes the use of a novel sampling technique that not only effectively localizes the relevant MS volume, but also increases the size of the training dataset and improves classification results. Additionally, we employ a Multiple Instance Ensembling (MIE) prediction method to further boost classification performance.

Results With sampling and MIE, we observe that there is consistent improvement in classification performance of all 3D ResNet and 3D DenseNet architecture with an average AUPRC percentage increase of $21.86 \pm 11.92\%$ and $4.27 \pm 5.04\%$ by sampling and $28.86 \pm 12.80\%$ and $9.85 \pm 4.02\%$ by sampling and MIE, respectively.

Conclusion Sampling and MIE can be effective techniques to improve the generalizability of CNNs for paranasal anomaly classification. We demonstrate the feasibility of classifying anomalies in the MS. We propose a data enlarging strategy through sampling alongside a novel MIE strategy that proves to be beneficial for paranasal anomaly classification in the MS.

Keywords Paranasal anomaly · Maxillary sinus · CNN · Classification

Introduction

Paranasal sinuses, located within specific bones, are prone to pathologies like retention cysts and polyps [1–3]. These anomalies, although often incidental, pose challenges for healthcare professionals, as they are unrelated to the patient's primary clinical indications [4]. Multiple studies emphasize the importance of understanding and addressing the prevalence of these paranasal anomalies in the general population [5–9].

Accurate diagnosis of paranasal inflammations is crucial for effective patient care, with medical professionals relying

Anna Sophie Hoffmann and Alexander Schlaefer have equally contributed to this work.

Debayan Bhattacharya
debayan.bhattacharya@tuhh.de; d.bhattacharya@uke.de

[1] Institute of Medical Technology and Intelligent Systems, Technische Universitaet Hamburg, Hamburg, Germany

[2] Department of Otorhinolaryngology, Head and Neck Surgery and Oncology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

[3] Clinic and Polyclinic for Diagnostic and Interventional Radiology and Nuclear Medicine, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

[4] Population Health Research Department, University Heart and Vascular Center, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

[5] Clinic and Polyclinic for Neurology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Published online: 21 July 2023

on CT and MRI scans to assess the extent of these conditions in the head and neck area [10]. 3D information is essential for identifying paranasal anomalies correctly, as misdiagnosis can lead to patient distress and increased healthcare costs [11]. A retrospective study found misdiagnoses of inverted papillomas and malignant tumors as nasal polyps in a significant percentage of cases [12]. Deep learning (DL) methods offer potential for improving diagnostic accuracy and reducing clinicians' workload, but the highly variable anatomy of paranasal sinuses necessitates cautious consideration when applying these techniques for reliable and accurate diagnoses [13].

DL technologies have shown significant advancements in anomaly detection, particularly in computer vision [14] and medical imaging analysis [15]. CNNs have proven effective in paranasal pathology screening, sinusitis classification, and tumor subtype differentiation. Existing studies typically follow a two-stage approach of localizing sinuses and then classifying them. For instance, one study cropped X-rays and classified anomalies [16], but failed to classify left and right MS anomalies separately. Another study segmented Computed Tomography (CT) images and classified anomalies [17], necessitating pixel-level annotations for localization. A different approach involved using a CNN to detect key slices in CT images containing MS volumes and then classifying MS anomalies [18]. However, two-stage methods relying on specialized annotations for localization and classification pose challenges in terms of increased clinician workload and limited generalization to diverse datasets.

Our proposed end-to-end approach is a non-DL solution for localizing MS volumes and a DL method for classifying MS anomalies. By employing a unique localization strategy using Gaussian sampling of centroid coordinates, we significantly expand the dataset and extract multiple overlapping instances of the MS. Leveraging a 3D CNN in our MIE prediction approach, we achieve boosted classification performance for these volumes. Through rigorous experimentation involving popular DenseNet and ResNet architectures, we ascertain the optimal MS volume size for achieving superior classification performance. Our comprehensive pipeline proves to be advantageous for accurate MS anomaly classification, providing a promising alternative to DL-based methods.

Methods

Dataset: as part of the Hamburg City Health Study (HCHS) [19], cMRIs of participants (45–74 years) were recorded for neuroradiological assessment. These scans were obtained at the University Medical Center Hamburg-Eppendorf and feature fluid attenuated inversion recovery (FLAIR) sequences in the NIfTI format. The dataset comprises 299 patients, with 174 exhibiting healthy left and right MS and 125 exhibiting at least one MS having a polyp or cyst pathology. The diagnoses were confirmed by two ear, nose, and throat (ENT) surgeons and one ENT specialized radiologist. The anomalies under consideration in this study include polyps and cysts. MS exhibiting these anomalies are grouped into "anomalous" class and MS without these anomalies are grouped into "normal" class.

Dataset preprocessing and MS volume extraction: each MRI in the study has a resolution of 173x319x319 voxels, with a voxel size of $0.53 \times 0.75 \times 0.75$ mm. To ensure consistency across all of the head and neck MRI scans in our study, we apply a process of rigid registration using Dipy library [20]. This involves selecting one MRI as a fixed volume and registering other MRIs with respect to the fixed volume.

To increase the size of the dataset and be able to use multiple instances of MS volumes for our ensemble prediction, we extracted multiple MS volumes of left and right MS from individual head and neck MRI scans. This was done by manually recording the centroid locations of the left and right MS of 20 patients, and using these coordinates to compute the mean and standard deviation of the centroid locations. These values are denoted as $\mu(x)$, $\mu(y)$, $\mu(z)$ and $\sigma(x), \sigma(y), \sigma(z)$ for the mean and standard deviation, respectively. We then initialize Gaussian distributions - $\mathcal{N}(\mu(x), \sigma^2(x))$,$\mathcal{N}(\mu(y), \sigma^2(y))$,$\mathcal{N}(\mu(z), \sigma^2(z))$— and use these distributions to sample centroid locations for MS volumes in the head and neck MRI. It is worth noting that the mean and standard deviation of the left and right MS volumes are different, resulting in a total of six Gaussian distributions in practice. In practice, for the left MS the x, y and z μ are 75, 231 and 121 mm and σ are 1.47, 1.56 and 1.76 mm, respectively. Correspondingly for the right MS, the x, y and z μ are 149, 232 and 118 mm and σ are 1.90, 1.66 and 6.47 mm respectively. We sample N left MS volumes and N right MS from each head and neck MRI where N is the sample size. For our experiments, $N \in \{1, 5, 10, 15, 20\}$. An illustration of our sampling method is shown in Fig. 1a. We extract MS volumes of multiple sizes namely, $25 \times 25 \times 25$, $30 \times 30 \times 30$, $35 \times 35 \times 35$, $40 \times 40 \times 40$, $45 \times 45 \times 45$. The extracted MS volumes are finally resampled to a resolution of $64 \times 64 \times \times 64$ for the 3D CNN. To make the right and left MS appear more symmetrical, we horizontally flip the coronal planes of the right MS to give it the appearance of the left MS volume. Figure 2 illustrates our data processing pipeline.

Training, validation and test splits: if we sample with $N = 1$, we end up with 327, 37 and 41 MS volumes in the training, validation and test set, respectively. The training validation and test set size increase by a factor of N with respect to the sample size N. 32% of the MS volumes in the training, validation and test sets are anomalous MS volumes. We perform threefold cross-validation experiments with all the methods.

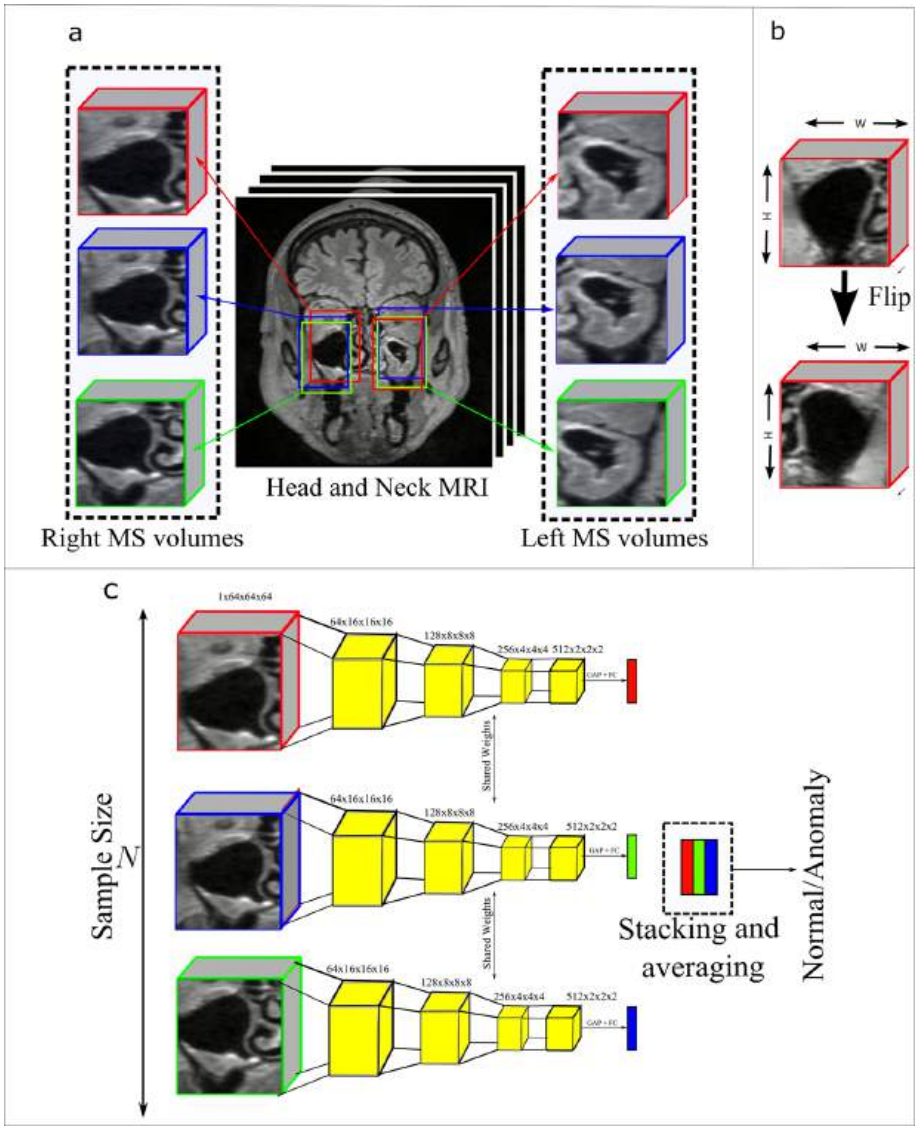

Fig. 1 **a** Illustration of our MS volume extraction strategy showing 3 MS volumes for left and right MS each. **b** Flipping of the coronal plane of the right MS. **c** Illustration of our MIE prediction strategy used during inference. GAP denotes Global Average Pooling and FC denotes Fully Connected Layer

Implementation details We implement a 3D CNN using ResNet18 [21] with four stages of 3D residual blocks (channel dimensions 64, 128, 256, 512). Our models are trained for 100 epochs with a batch size of 16, a learning rate of 0.0001, and Adam optimization. If the validation loss did not improve for 5 epochs, the learning rate is reduced by a factor of 10. We use PyTorch and PyTorch Lightning to build our models.

DL method: to classify the MS volume into normal or anomaly class, we consider multiple 3D ResNet [21][1] and 3D DenseNet[22] architectures.[2] Let us denote the classifier as $f(\cdot)$ and the MRIs as $X \in R^{H\times W\times D}$. From each MRI, we extract N left MS volumes and N right MS volumes. Altogether, we extract $2N$ MS volumes from $X \in R^{H\times W\times D}$. Let us denote the MS volumes as $x \in R^{P\times P\times P}$. Here, P denotes the size of the MS volume such that $P \in \{25, 30, 35, 40, 45\}$. Further, our labels $y \in \{0, 1\}$ represent normal and anomaly class. The anomaly class is the positive class for our use-case. As a baseline, we consider 3DResNet models that do not use our MIE strategy for inferring on the test set.

Multiple instance ensemble prediction strategy: let us denote the extracted MS volumes from a single MRI $x_i \in R^{P\times P\times P}$ where i denotes the $i-th$ MS volume extracted from either the left or right MS area of the MRI. When making a prediction, we average the softmax scores of classifier $f(.)$ from the multiple MS volumes x_i. Formally,

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} softmax(f(x_i))$$

[1] https://github.com/kenshohara/3D-ResNets-PyTorch/blob/master/models/resnet.py.

[2] https://github.com/kenshohara/3D-ResNets-PyTorch/blob/master/models/densenet.py.

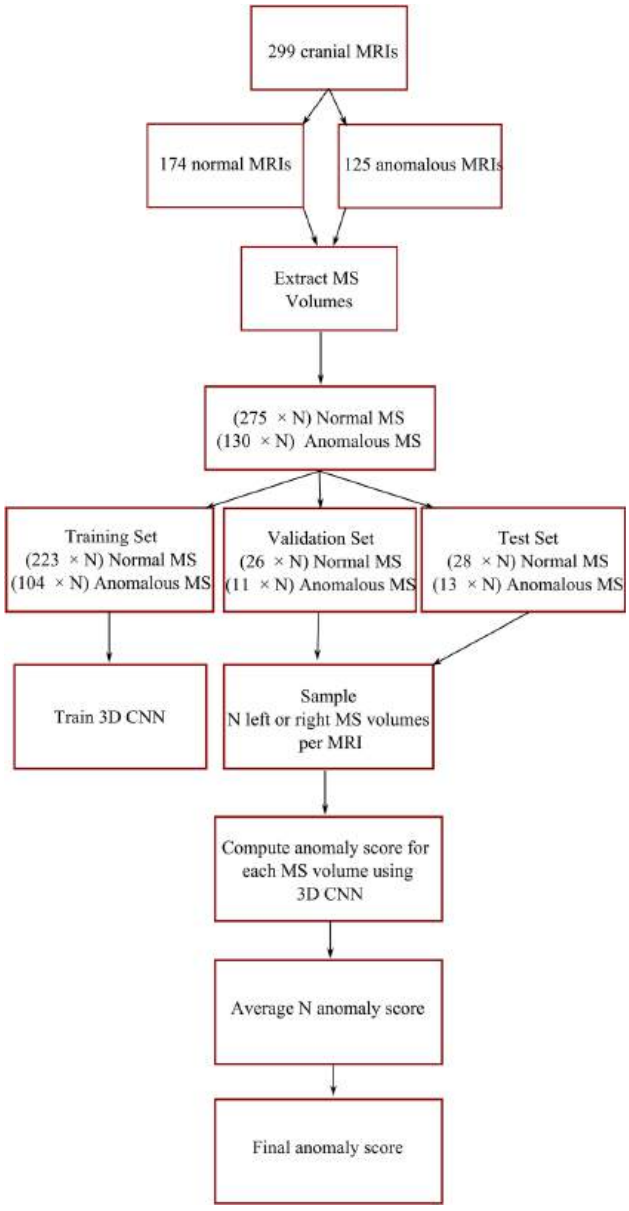

Fig. 2 Illustration of our data processing pipeline

Results

Effect of sampling size

We plot the mean and standard deviation of the Area Under Precision Recall Curve (AUPRC) and F1 score. Both these metrics are useful especially in imbalanced scenarios which is our case. From Table 1, we observe that with the increase in the sample size N, we get a consistent increase in all the reported metrics until $N = 15$ after which we get a decrease in all the metrics. Further, for all the cases, we see that using MIE strategy is beneficial for MS anomaly classification and leads to boost in classification metrics.

Comparison with state-of-the-art 3D CNN architectures

We investigate the benefits of sampling and ensembling techniques on 3D CNN architectures for medical imaging classification. Specifically, we examine their impact on 3D DenseNet and 3D ResNets used in various medical imaging tasks [23–28]. Using $N = 15$, we conduct an ablation study, training the CNNs with sampled data but inferring with a single MS volume. We observe that, for ResNets, the AUPRC decreases as architecture complexity increases, but with N=15 and MIE enabled, the decrease is minimal. Sampling and sampling with MIE consistently improve performance for both ResNets and DenseNets, resulting in percentage increases of $21.86 \pm 11.92\%$ and $4.27 \pm 5.04\%$ (sampling) and $28.86 \pm 12.80\%$ and $9.85 \pm 4.02\%$ (sampling + MIE) in AUPRC, respectively

Effect of sampling type

In our methods, we adopt a Gaussian distribution to model centroid locations and sample random centroids for extracting corresponding MS volumes. However, an alternative approach is to extract equidistant centroids, which allows us to investigate potential advantages of random sampling. For this experiment, we consider the x, y, and z coordinates of the centroid to lie on lines starting at $\mu(x)$ - $\sigma(x)$, $\mu(y)$ - $\sigma(y)$ and $\mu(z)$ - $\sigma(z)$, respectively, and ending at $\mu(x)$ + $\sigma(x)$, $\mu(y)$ + $\sigma(y)$ and $\mu(z)$ + $\sigma(z)$. Figure 3 illustrates the relationship between these axes and the MRI image from which the coordinates are sampled. Utilizing these lines, we sample N equidistant centroid locations, while also exploring variations such as fixing one or two coordinates at their means. Referred to as *equidistant sampling*, we compare this strategy and its variants against the random sampling method we propose. Table 3 presents our findings, demonstrating the most advantageous sampling from the z-axis for classification. Equidistant sampling from all axes yields an AUPRC of 0.89 ± 0.03, while equidistant sampling from x and y axes with a constant z coordinate results in an AUPRC of 0.82 ± 0.16, showcasing an 8.18% difference. Similarly, fixing the y-coordinate incurs a 1.12% AUPRC decrease, while maintaining a constant z-coordinate exhibits no AUPRC decrease. These results suggest varying importance in sampling MS centroid coordinates from each axis, with random sampling proving to be the most effective approach.

Effect of patch size

Further, looking at Fig. 4, we can see the influence of MS volume size to the parnasal classification task. Note, we set $N = 15$ for this experiment and use our multiple instance ensemble prediction strategy. This highlights that patch size

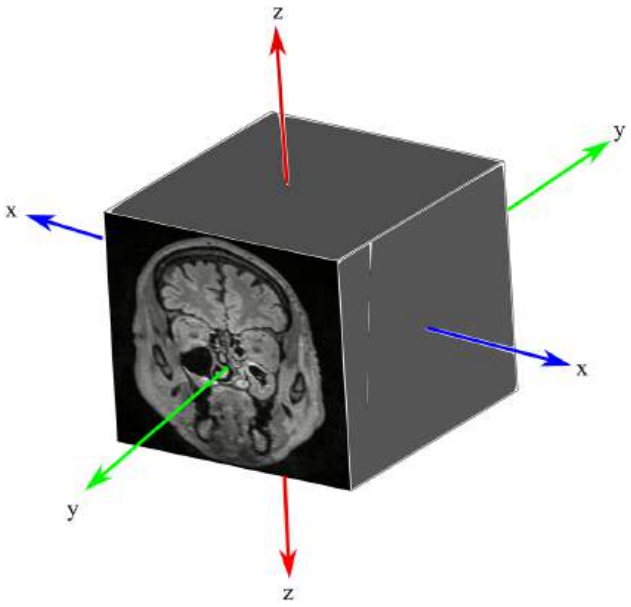

Fig. 3 x, y and z axis for an MRI

Table 1 Result of our experiments

N	MIE	AUPRC	F1
1		0.80 ± 0.12	0.70 ± 0.13
5		0.85 ± 0.03	0.77 ± 0.10
5	✓	0.87 ± 0.04	0.76 ± 0.10
10		0.85 ± 0.04	0.75 ± 0.08
10	✓	0.89 ± 0.05	0.79 ± 0.10
15		0.88 ± 0.07	0.81 ± 0.12
15	✓	**0.92 ± 0.06**	**0.85 ± 0.09**
20		0.87 ± 0.04	0.77 ± 0.05
20	✓	0.91 ± 0.02	0.78 ± 0.07

Patch size $P = 35$ for all the experiments and 3D ResNet18 architecture used

The bold signifies the highest/best metric in each column of a table

Table 2 Result of our experiments

CNN	N	MIE	AUPRC	F1
3D ResNet18	1		0.80 ± 0.12	0.70 ± 0.13
3D ResNet18	15		0.88 ± 0.07	0.81 ± 0.12
3D ResNet18	15	✓	0.92 ± 0.06	0.85 ± 0.09
3D ResNet50	1		0.72 ± 0.13	0.59 ± 0.19
3D ResNet50	15		0.82 ± 0.11	0.71 ± 0.19
3D ResNet50	15	✓	0.85 ± 0.07	0.74 ± 0.13
3D ResNet101	1		0.73 ± 0.10	0.59 ± 0.04
3D ResNet101	15		0.85 ± 0.04	0.69 ± 0.10
3D ResNet101	15	✓	0.90 ± 0.06	0.79 ± 0.14
3D ResNet152	1		0.66 ± 0.06	0.57 ± 0.08
3D ResNet152	15		0.83 ± 0.07	0.76 ± 0.11
3D ResNet152	15	✓	0.89 ± 0.05	0.80 ± 0.08
3D ResNet200	1		0.60 ± 0.21	0.45 ± 0.39
3D ResNet200	15		0.86 ± 0.05	0.79 ± 0.10
3D ResNet200	15	✓	0.90 ± 0.04	0.83 ± 0.07
3D DenseNet121	1		0.86 ± 0.11	0.80 ± 0.07
3D DenseNet121	15		0.86 ± 0.10	0.81 ± 0.06
3D DenseNet121	15	✓	0.92 ± 0.05	0.83 ± 0.12
3D DenseNet169	1		0.81 ± 0.09	0.76 ± 0.11
3D DenseNet169	15		0.91 ± 0.05	0.82 ± 0.04
3D DenseNet169	15	✓	0.94 ± 0.03	0.86 ± 0.09
3D DenseNet201	1		0.88 ± 0.07	0.72 ± 0.07
3D DenseNet201	15		0.88 ± 0.04	0.72 ± 0.08
3D DenseNet201	15	✓	0.93 ± 0.06	0.78 ± 0.07
3D DenseNet264	1		0.84 ± 0.09	0.81 ± 0.07
3D DenseNet264	15		0.88 ± 0.05	0.82 ± 0.12
3D DenseNet264	15	✓	0.93 ± 0.01	0.85 ± 0.09

Patch size $P = 35$ for all the experiments

plays an important role in boosting the paranasal anomaly classification performance. Our experiments indicate that the optimal patch size for our dataset is $P = 35$.

Table 3 Experiment on sampling strategy

x	y	z	AUPRC	F1
$\mu(x)$	$\mu(y) \pm \sigma(y)$	$\mu(z) \pm \sigma(z)$	0.89 ± 0.03	0.74 ± 0.03
$\mu(x) \pm \sigma(x)$	$\mu(y)$	$\mu(z) \pm \sigma(z)$	0.88 ± 0.04	0.77 ± 0.04
$\mu(x) \pm \sigma(x)$	$\mu(y) \pm \sigma(y)$	$\mu(z)$	0.82 ± 0.16	0.73 ± 0.19
$\mu(x)$	$\mu(y)$	$\mu(z) \pm \sigma(z)$	0.89 ± 0.03	0.77 ± 0.04
$\mu(x)$	$\mu(y) \pm \sigma(y)$	$\mu(z)$	0.85 ± 0.08	0.68 ± 0.18
$\mu(x) \pm \sigma(x)$	$\mu(y)$	$\mu(z)$	0.85 ± 0.04	0.75 ± 0.03
$\mu(x) \pm \sigma(x)$	$\mu(y) \pm \sigma(y)$	$\mu(z) \pm \sigma(z)$	0.89 ± 0.03	0.79 ± 0.06
$\mathcal{N}(\mu(x), \sigma^2(x))$	$\mathcal{N}(\mu(y), \sigma^2(y))$	$\mathcal{N}(\mu(z), \sigma^2(z))$	**0.92 ± 0.07**	**0.85 ± 0.09**

Patch size $P = 35$ for all the experiments. $\mu \pm \sigma$ represents equidistant sampling of N points from a line starting at μ - σ and ending at μ + σ. $\mathcal{N}(\mu, \sigma^2)$ represents random sampling of points from a Gaussian distribution parameterized by μ and σ. 3D ResNet18 architecture used

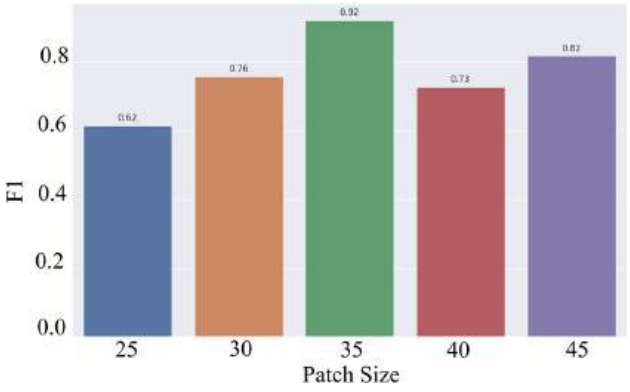

Fig. 4 F1 scores vs patch size P

Discussion

Clinicians classifying paranasal anomalies face the burdensome task of manually searching for MS-containing slices in MRI sequences and then diagnosing, resulting in time-consuming and fatiguing analysis. Additionally, the task of classifying paranasal anomalies is challenging due to the morphological variation of the maxillary sinus as well as the polyp and cyst anomalies inhabiting the sinuses. This can be seen in Fig. 5. Previous approaches [18] used a 2-stage CNN pipeline, learning key MS slices first using a CNN and then classifying anomalies with another CNN [18] or learning to segment the maxillary sinus and then classifying the anomaly [17], but these methods 2 stage CNN pipeline makes it dependent on datasets hindering generalization. To overcome this limitation, we propose a method that extracts multiple MS volumes without DL, using a CNN only once to compute the final anomaly score for each MS. This streamlined approach reduces dataset dependency and facilitates broader applicability to other modalities.

As seen in Table 1, increasing the sample size N improves classification metrics, but a sample size of 20 exhibits a lower F1 score compared to 15, possibly due to overfitting caused by redundant volumes. Thus, careful selection of the appropriate sample size is crucial for optimal performance.

We compared different CNN architectures in Table 2 and found that sampling and MIE are beneficial for our

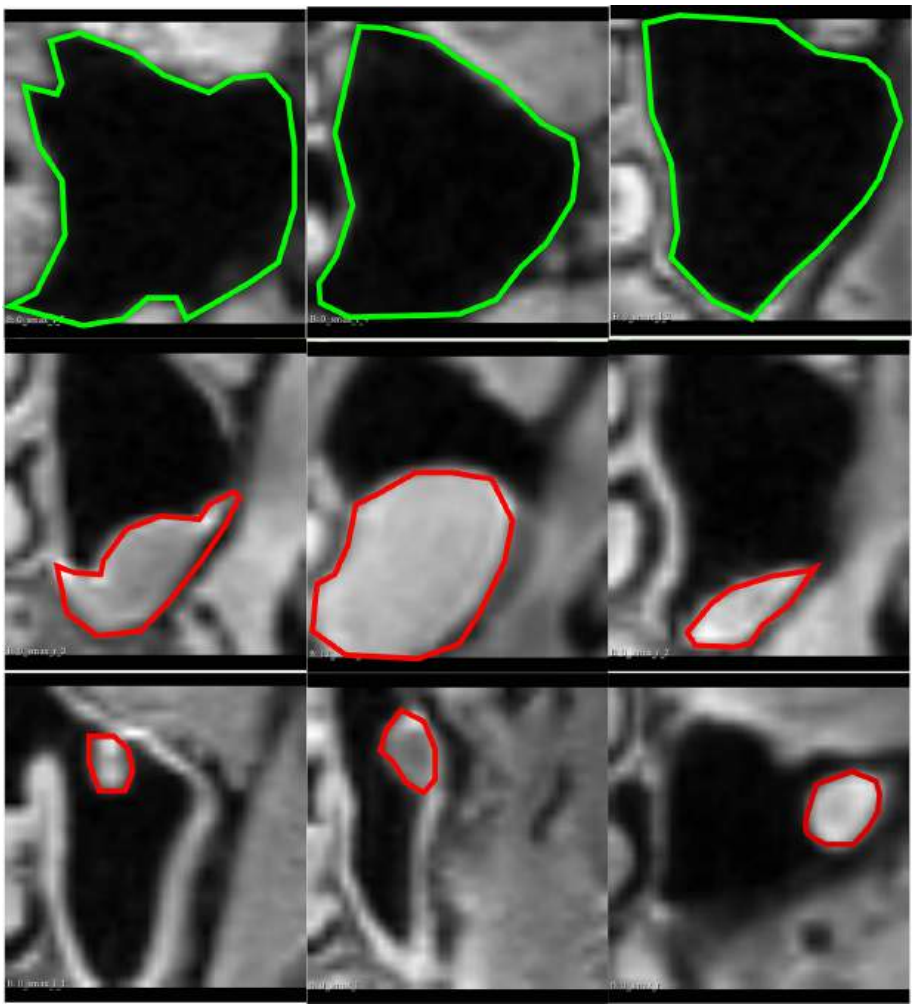

Fig. 5 The coronal planes of the sampled MS volumes. The green contours in the first row represent normal MS anatomy. The red contours enclose masses that represent cysts and polyps in the second and third row, respectively, demonstrating the variety of appearances and morphological variations of these anomalies within the MS

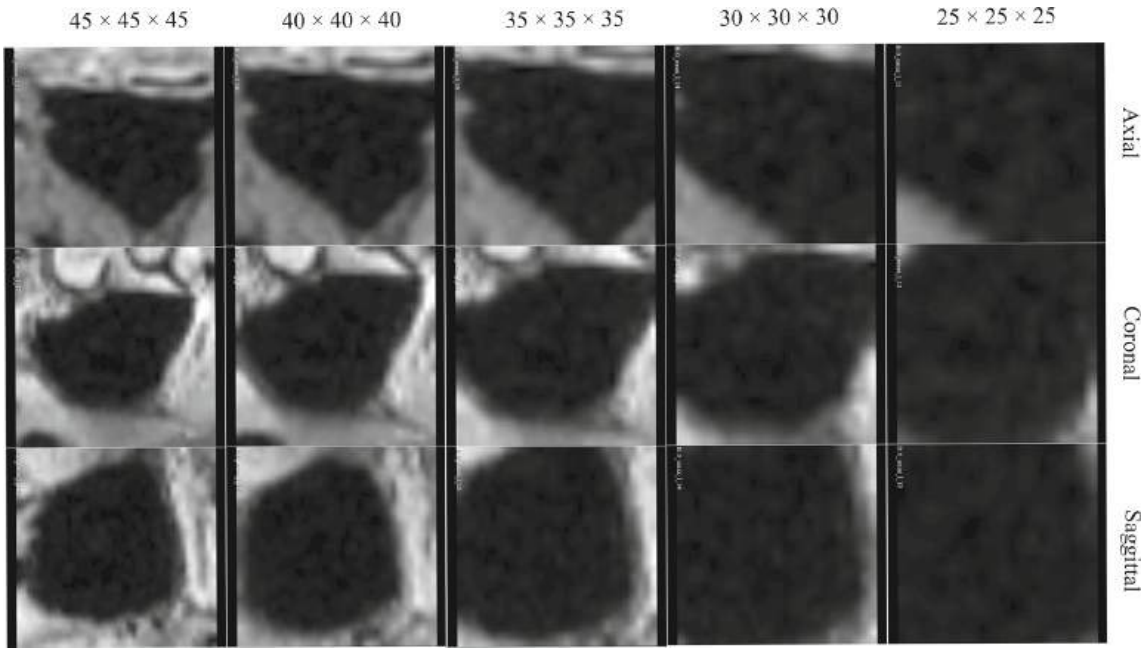

Fig. 6 Slices from the axial, coronal and saggital slices of extracted healthy MS volume with different patch sizes

classification task. The advantages of MIE and sampling are more prominent in ResNet architectures compared to DenseNet. We hypothesize that the multiple skip connections in a dense block [29] contribute to improved gradient flow and optimization in the $N = 1$ sampling scenario. Our method consistently increases AUPRC, with sampling + MIE showing higher efficacy than only sampling and no sampling, demonstrating the effectiveness of our approach.

We compared random sampling and equidistant sampling's impact on classification performance in Table 3. Random sampling yielded better results, likely due to the diverse multi-scale volumes obtained through randomization. Equidistant sampling along the z-axis significantly improved classification performance, possibly due to higher z-coordinate σ, resulting in greater spatial offsets. This increased diversity in sampled volumes facilitated better feature learning and classification performance. Incorporating multi-scale volumes from different z-axis positions enhanced the ability of our 3D CNN to identify patterns, improving overall performance.

Additionally, using an ensemble strategy that averages the scores from multiple instances of the MS leads to a further improvement in classification metrics. The improvement in our classification metrics can be attributed to the incorporation of implicit test-time augmentation during inference on the test set. By sampling multiple overlapping MS volumes, we have MS volumes which have transnational offsets with respect to one another, resulting in better performance. These findings demonstrate the utility of our proposed method for the classification of paranasal anomalies in the MS.

The size of the extracted MS volume is critical for accurate classification. Small volumes may miss important details, while large volumes include irrelevant information as can be seen in Fig. 6. Our evaluation of different sizes ($25 \times 25 \times 25$ to $45 \times 45 \times 45$) found that $35 \times 35 \times 35$ yielded the highest F1 score. This suggests that small volumes miss anomalies, while larger volumes include unnecessary structures. Careful selection of the patch size is crucial for optimal performance.

Conclusion

We propose a DL approach for classifying maxillary sinus anomalies. Our method employs multiple instance ensemble prediction and a sampling strategy to improve classification performance on available dataset. We investigate the optimal sample size and patch size trade-off. Although further improvements are needed for real-world clinical use, our work offers a promising solution for maxillary sinus anomaly classification with DL.

Acknowledgements This work has not been submitted for publication anywhere else. This work is funded partially by the i3 initiative of the Hamburg University of Technology. The authors also acknowledge the partial funding by the Free and Hanseatic City of Hamburg (Interdisciplinary Graduate School) from University Medical Center Hamburg-Eppendorf. This work was partially funded by Grant Number KK5208101KS0 (Zentrales Innovationsprogramm Mittelstand, Arbeitsgemeinschaft industrieller Forschungsvereinigungen).

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest Debayan Bhattacharya states no conflict of interest. Finn Behrendt states no conflict of interest. Benjamin Tobias Becker states no conflict of interest. Dirk Beyersdorff states no conflict of interest. Elina Petersen states no conflict of interest. Marvin Petersen states no conflict of interest. Bastian Cheng states no conflict of interest. Dennis Eggert states no conflict of interest. Christian Betz states no conflict of interest. Anna Sophie Hoffmann states no conflict of interest. Alexander Schlaefer states no conflict of interest.

Ethical approval The study protocol received approval from the local ethics committee (Landesärztekammer Hamburg, PV5131) and was approved by the Data Protection Commissioners for the University Medical Center of the University Hamburg-Eppendorf and the Free and Hanseatic City of Hamburg. It is registered on ClinicalTrial.gov (NCT03934957) and adheres to Good Clinical Practice, Good Epidemiological Practice, and ethical principles outlined in the Declaration of Helsinki.

References

1. Martini, F., Timmons, M.J., Tallitsch, R.B.: Human anatomy. 6th edn. San Francisco, Pearson Benjamin Cummings (2009)
2. Bal M, Berkiten G, Uyanık E (2014) Mucous retention cysts of the paranasal sinuses. Hippokratia 18(4):379
3. Varshney H, Varshney J, Biswas S, Ghosh SK (2015) Importance of CT scan of paranasal sinuses in the evaluation of the anatomical findings in patients suffering from sinonasal polyposis. Indian J Otolaryngol Head Neck Surg 68(2):167–172
4. Hansen AG, Helvik A-S, Nordgård S, Bugten V, Stovner LJ, Håberg AK, Gårseth M, Eggesbø HB (2014) Incidental findings in MRI of the paranasal sinuses in adults: a population-based study (hunt MRI). BMC Ear, Nose, and Throat Disord 14(1):13. https://doi.org/10.1186/1472-6815-14-13
5. Tarp B, Fiirgaard B, Christensen T, Jensen JJ, Black FT (2000) The prevalence and significance of incidental paranasal sinus abnormalities on MRI. Rhinology 38(1):33–38
6. Rak KM, Newell JD, Yakes WF, Damiano MA, Luethke JM (1991) Paranasal sinuses on MR images of the brain: significance of mucosal thickening. AJR Am J Roentgenol 156(2):381–384. https://doi.org/10.2214/ajr.156.2.1898819
7. Stenner M, Rudack C (2014) Diseases of the nose and paranasal sinuses in child. GMS Curr Top Otorhinolaryngol, Head Neck Surg 13:10. https://doi.org/10.3205/cto000113
8. Rege ICC, Sousa TO, Leles CR, Mendonça EF (2012) Occurrence of maxillary sinus abnormalities detected by cone beam CT in asymptomatic patients. BMC Oral Health 12:30. https://doi.org/10.1186/1472-6831-12-30
9. Cooke LD, Hadley DM (1991) MRI of the paranasal sinuses: incidental abnormalities and their relationship to symptoms. J Laryngol Otol 105(4):278–281. https://doi.org/10.1017/s0022215100115609
10. Brierley J, Gospodarowicz MK, Wittekind C (eds) (2017) TNM classification of malignant tumours, 8th edn. Wiley, Chichester, West Sussex, Hoboken
11. Gutmann A (2013) Ethics. The bioethics commission on incidental findings. Science (New York) 342(6164):1321–1323. https://doi.org/10.1126/science.1248764
12. Ma Z, Yang X (2012) Research on misdiagnosis of space occupying lesions in unilateral nasal sinus. Lin chuang er bi yan hou tou jing wai ke za zhi = J Clin Otorhinolaryngol, Head, Neck Surg 26(2):59–61. https://doi.org/10.13201/j.issn.1001-1781.2012.02.005
13. Papadopoulou A-M, Chrysikos D, Samolis A, Tsakotos G, Troupis T (2021) Anatomical variations of the nasal cavities and paranasal sinuses: a systematic review. Cureus 13(1):12727
14. Mohindru V, Singla S (2021) A review of anomaly detection techniques using computer vision. In: Singh PK, Singh Y, Kolekar MH, Kar AK, Chhabra JK, Sen A (eds) Recent innovations in computing. Springer, Singapore, pp 669–677
15. Tschuchnig ME, Gadermayr M (2022) Anomaly detection in medical imaging—a mini review. In: Haber P, Lampoltshammer TJ, Leopold H, Mayr M (eds) Data Sci-Analyt Appl. Springer, Wiesbaden, pp 33–38
16. Kim HG, Lee KM, Kim EJ, Lee JS (2019) Improvement diagnostic accuracy of sinusitis recognition in paranasal sinus X-ray using multiple deep learning models. Quant Imaging Med Surg 9:942–951. https://doi.org/10.21037/QIMS.2019.05.15
17. Ozbay S, Tunc O (2022) Deep learning in analysing paranasal sinuses. Elektron Elektrotech 28:65–70. https://doi.org/10.5755/J02.EIE.31133
18. Kim K-S, Kim BK, Chung MJ, Cho HB, Cho BH, Jung YG (2022) Detection of maxillary sinus fungal ball via 3-d CNN-based artificial intelligence: fully automated system and clinical validation. PLoS One 17(2):1–19. https://doi.org/10.1371/journal.pone.0263125
19. Jagodzinski A, Blankenberg S et al (2020) Rationale and design of the Hamburg city health study. Eur J Epidemiol 35(2):169–181. https://doi.org/10.1007/s10654-019-00577-4
20. Garyfallidis E, Brett M, Amirbekian B, Rokem A, van der Walt S, Descoteaux M, Nimmo-Smith I (2014) Dipy, a library for the analysis of diffusion MRI data. Front Neuroinform 8:8. https://doi.org/10.3389/FNINF.2014.00008/BIBTEX
21. Hara, K., Kataoka, H., Satoh, Y.: Learning spatio-temporal features with 3D residual networks for action recognition. arXiv:1708.07632v1
22. Fu J, Singhrao K, Qi XS, Yang Y, Ruan D, Lewis JH (2021) Three-dimensional multipath densenet for improving automatic segmentation of glioblastoma on pre-operative multimodal MR images. Med Phys 48(6):2859–2866. https://doi.org/10.1002/mp.14800
23. Zhang, G., Lin, L., Wang, J.: Lung nodule classification in ct images using 3d densenet. Journal of Physics: Conference Series **1827** (2021)
24. Liu Z, Zhu Y, Yuan Y, Yang L, Wang K, Wang M, Yang X, Wu X, Tian X, Zhang R, Shen B, Luo H, Feng H, Feng S, Ke Z (2021) 3D DenseNet deep learning based preoperative computed tomography for detecting myasthenia gravis in patients with thymoma. Front Oncol 11:631964
25. Näppi, J.J., Hironaka, T., Yoshida, H.: Detection of colorectal masses in CT colonography: application of deep residual networks

for differentiating masses from normal colon anatomy. In: Medical Imaging (2018)
26. Chen, X., Wang, Z., Zhan, Y., Cheikh, F.A., Ullah, M.: Interpretable learning approaches in structural MRI: 3d-resnet fused attention for autism spectrum disorder classification. In: Medical Imaging (2022)
27. Suryakanth, B., Hari Prasad, S.A.: 3D CNN-residual neural network based multimodal medical image classification. Int J Eng Trends Technol **70**(10), 371–380 (2022). https://doi.org/10.14445/22315381/IJETT-V70I10P236
28. Uemura, T., Näppi, J.J., Hironaka, T., Kim, H., Yoshida, H.: Comparative performance of 3d-densenet, 3d-resnet, and 3d-vgg models in polyp detection for CT colonography. In: Medical Imaging (2020)
29. Huang, G., Liu, Z., Maaten, L.V.D., Weinberger, K.Q.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269. IEEE Computer Society, Los Alamitos (2017). https://doi.org/10.1109/CVPR.2017.243

7.6 Computer-Aided Diagnosis of Maxillary Sinus Anomalies: Validation and Clinical Correlation

Title of paper: Computer-Aided Diagnosis of Maxillary Sinus Anomalies: Validation and Clinical Correlation

Journal: The Laryngoscope

Year: 2024

Topic: Correlation with patient data

The Laryngoscope

Computer-Aided Diagnosis of Maxillary Sinus Anomalies: Validation and Clinical Correlation

Debayan Bhattacharya, MSc; Benjamin Tobias Becker, MD; Finn Behrendt, MSc; Dirk Beyersdorff, MD, PhD; Elina Petersen, MSc; Marvin Petersen, MD; Bastian Cheng, MD, PhD; Dennis Eggert, PhD; Christian Betz, MD, PhD; Alexander Schlaefer, PhD; Anna Sophie Hoffmann, MD, PhD

Objective: Computer aided diagnostics (CAD) systems can automate the differentiation of maxillary sinus (MS) with and without opacification, simplifying the typically laborious process and aiding in clinical insight discovery within large cohorts.

Methods: This study uses Hamburg City Health Study (HCHS) a large, prospective, long-term, population-based cohort study of participants between 45 and 74 years of age. We develop a CAD system using an ensemble of 3D Convolutional Neural Network (CNN) to analyze cranial MRIs, distinguishing MS with opacifications (polyps, cysts, mucosal thickening) from MS without opacifications. The system is used to find correlations of participants with and without MS opacifications with clinical data (smoking, alcohol, BMI, asthma, bronchitis, sex, age, leukocyte count, C-reactive protein, allergies).

Results: The evaluation metrics of CAD system (Area Under Receiver Operator Characteristic: 0.95, sensitivity: 0.85, specificity: 0.90) demonstrated the effectiveness of our approach. MS with opacification group exhibited higher alcohol consumption, higher BMI, higher incidence of intrinsic asthma and extrinsic asthma. Male sex had higher prevalence of MS opacifications. Participants with MS opacifications had higher incidence of hay fever and house dust allergy but lower incidence of bee/wasp venom allergy.

Conclusion: The study demonstrates a 3D CNN's ability to distinguish MS with and without opacifications, improving automated diagnosis and aiding in correlating clinical data in population studies.

Key Words: Convolutional Neural Network, deep learning, maxillary sinus, Paranasal sinus, Population study.

Level of Evidence: 3

***Laryngoscope*, 00:1–8, 2024**

INTRODUCTION

Morphological changes in paranasal sinuses, often seen in magnetic resonance imaging (MRI) scans and computed tomography (CT), have been extensively studied. Research shows a correlation between these changes and factors like allergies and smoking habits.[1] Furthermore, the relationship amongst patients with and without sinus opacifications have been explored in different patient groups: both symptomatic and asymptomatic,[2] exclusively symptomatic,[3] only asymptomatic,[4] and nonselected[5] individuals. However, in all the aforementioned studies, clinicians manually reviewed multiple MRI or CT slices to determine the presence or absence of paranasal opacification.

This manual process can strain clinicians and escalate their workload, leading to fatigue and potential misdiagnoses.[6] Deep learning (DL)-based computer-aided diagnosis system (CAD) presents an opportunity to enhance diagnostic accuracy and alleviate clinician workload by automating the classification of incidental findings. CNNs have demonstrated effectiveness in various aspects of paranasal opacification analysis, including screening, sinusitis classification, and tumor subtype differentiation. Existing studies often employ a two-stage methodology, first localizing sinuses and then classifying anomalies. For example, one study cropped x-ray images to classify anomalies[7] but failed to distinguish between left and right maxillary sinus anomalies. Another study segmented CT images and classified anomalies,[8] demanding pixel-level annotations for localization. Alternatively, a different approach used a CNN to detect key slices within CT images containing maxillary sinus volumes and subsequently classify maxillary sinus anomalies.[9] In our prior work, we explored unsupervised learning[10], contrastive learning[11], and multiple instance ensembling,[12] with multiple instance ensembling yielding the most promising results.

From the Institute of Medical Technology and Intelligent Systems (D.B., F.B., A.S.), Technische Universitaet Hamburg, Hamburg, Germany; Department of Otorhinolaryngology, Head and Neck Surgery and Oncology (D.B., B.T.B., D.E., C.B., A.S.H.), University Medical Center Hamburg-Eppendorf, Hamburg, Germany; Clinic and Polyclinic for Diagnostic and Interventional Radiology and Nuclear Medicine (D.B.), University Medical Center Hamburg-Eppendorf, Hamburg, Germany; Population Health Research Department, University Heart and Vascular Center (E.P.), University Medical Center Hamburg-Eppendorf, Hamburg, Germany; and the Department of Neurology (M.P., B.C.), University Medical Center Hamburg-Eppendorf, Hamburg, Germany.

Additional supporting information may be found in the online version of this article.

Editor's Note: This Manuscript was accepted for publication on March 13, 2024.

The authors have no funding, financial relationships, or conflicts of interest to disclose.

Send correspondence to Debayan Bhattacharya, Gebaude E, Raum 3.086, 3. Etage, Institute of Medical Technology and Intelligent Systems Technische Universitaet Hamburg. Am Schwarzenberg-Campus 3, 21073 Hamburg, Germany. Email: debayan.bhattacharya@tuhh.de

DOI: 10.1002/lary.31413

We implemented an ensemble model (EM) of the multiple instance ensembling approach[12] based on the hypothesis that ensembling will improve model performance even more. We evaluated this EM on an unlabeled cohort within our population dataset to classify the presence or absence of opacification within the MS. Subsequently, we demonstrate a potential application for this EM which is rapid clinical insights of expanding cohorts which is typical in prospective population studies. To this end, we assessed the model's predictions by examining its relationships with various available factors, including smoking habits, alcohol consumption, BMI, chronic asthma, chronic bronchitis, sex, age, leukocyte count, highly sensitive C-reactive protein, and allergies and further assess if the discovered associations using our DL-based CAD are in accordance with previous literature which performed manual diagnosis. Overall, our research contributes to a better understanding of the prevalence of paranasal anomalies and further substantiates the potential of DL-based CAD to automate and enhance the efficiency of population studies, which have traditionally relied on labor-intensive methodologies.

METHODS

Study Design and Participant Collective

Hamburg City Health Study (HCHS)[13] is a single-center, prospective, population-based cohort study. A subsample ($N = 2619$) of the planned 45000 had cranial MRI scans recorded. MRI scans were recorded between February 08, 2016, and November 30, 2018, on individuals aged 45–74 years. Images were acquired using a 3-T Siemens Skyra MRI scanner (Siemens, Erlangen, Germany). 3D T2-weighted fluid attenuated inversion recovery (FLAIR) images were measured with the following sequence parameters: TR = 4700 ms, TE = 392 ms, 192 axial slices, ST = 0.9 mm, and IPR = 0.75 × 0.75 mm. Participant data included laboratory measurements of leukocytes/μL (LK) and high-sensitivity CRP (hCRP). Additionally, participants completed self-reported questionnaires documenting alcohol consumption per day (unit: grams/day [g/day]), smoking habits, diagnosis of chronic bronchitis or chronic obstructive pulmonary disease (COPD), and diagnosis of allergic bronchial asthma. Additionally, BMI, age, sex, and allergies of each participant were also recorded. The dataset comprised 2619 participants (56.05% men, 43.95% women) with a mean age of 63.98 (SD 8.32) years. Among these, 1069 participants (56.64% men, 43.36% women) with a mean age of 63.90 (SD 8.25) years were manually annotated to train a CNN. Among the annotated participants, 489 exhibited no opacifications in both left and right MS, while 580 showed at least one MS with polyp, cyst, or mucosal thickening (mucosa thickening >2 mm) opacification. These diagnoses were established by two ENT specialists and one ENT specialized radiologist. Figure 1 shows the flowchart of our study. Figure S1 shows exemplary MRIs exhibiting different pathologies.

DL Training, Validation and Test Dataset

Detailed explanation of our data processing pipeline is reported in supplementary material sections 1, 2, and 3. We extract 30 MS volumes from each participant. Our dataset used to train the 3D CNN consisted of 19215 (59.91%) MS exhibiting no opacifications, 4815 (15.01%) MS exhibiting mucosal thickening, 6315 (19.69%) MS containing polyps in MS, 1185 (3.69%) MS exhibiting cyst opacification, and 540 (1.68%) MS containing polyps or cysts encompassing the entire MS volume. While constructing the test set, we considered two criteria: (i) accurate representation of opacifications in training, validation, and test set (ii) multiple volumes extracted from each participant does not occur in training, validation, or test set simultaneously. The test set contained 30% of the overall dataset. Table I contains the overall statistics of the training, validation, and test dataset.

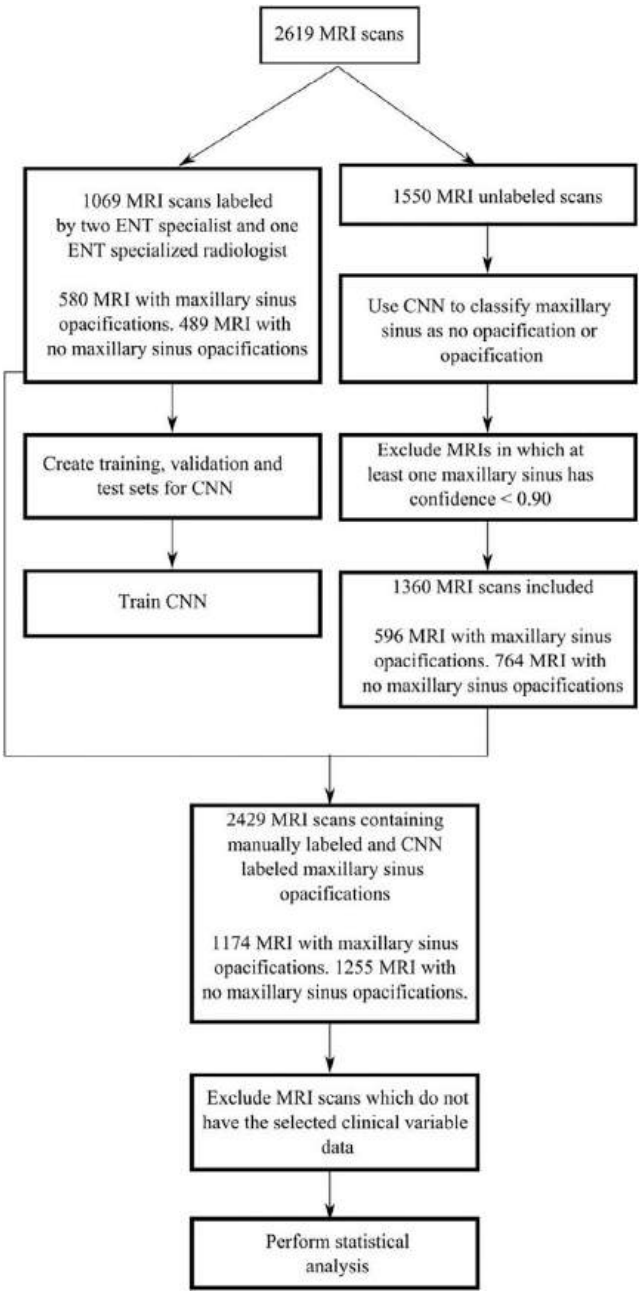

Fig. 1. Flowchart of our study.

Development of 3D Convolutional Neural Network

Our CNN is a 3D implementation of a 264-layer convolutional neural network called DenseNet.[14,15] Figure 2 and Figure S2 shows illustrations of our deep learning method.

TABLE I.
Distribution of Opacifications in Dataset. The Percentages Are Reported Within Parenthesis.

Dataset	Normal	Mucosal thickening	Polyp	Cysts	Fully occupied
Train	10740 (59.86%)	2700 (15.05%)	3540 (19.73%)	660 (3.67%)	300 (1.66%)
Validation	2700 (60%)	675 (15%)	885 (19.66%)	165 (3.66%)	75 (1.66%)
Test	5775 (59.96%)	1440 (14.95%)	1890 (19.62%)	360 (3.73%)	165 (1.71%)

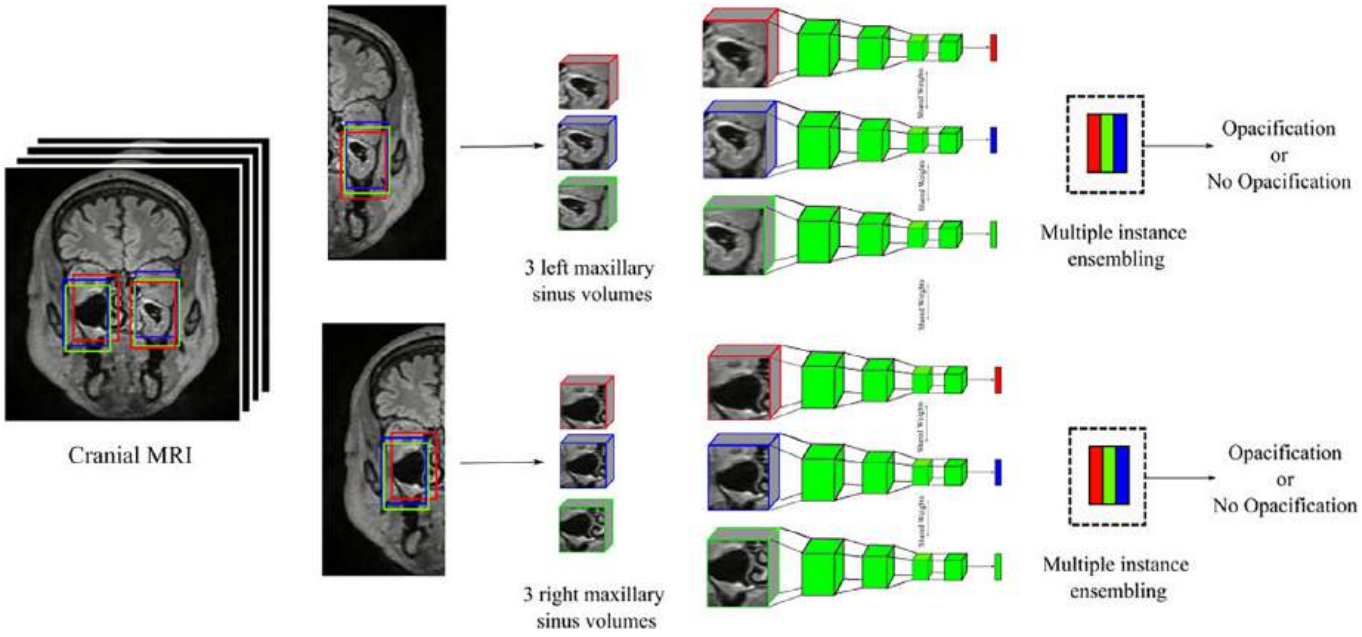

Fig. 2. Data processing pipeline showing how a single MRI is processed for inferring opacifications within the left and right MS of a single patient. As an illustration, 3 MS volumes are extracted from left and right side of MRI. Confidence score of 3 extracted MS volumes from the respective side are ensembled to create a single confidence score. In practice, we extract 15 MS volumes from left and right side of the MRI. The ensembling strategy is termed as multiple instance ensembling. [Color figure can be viewed in the online issue, which is available at www.laryngoscope.com.]

Implementation details are available in supplementary material section 3.1.

Training Protocol

Our deep learning pipeline is discussed in the supplementary material section. The MS volumes are input into a 3D CNN for classification into two classes: "opacification" or "no opacification". The "opacification" class encompasses MS exhibiting mucosal thickening, polyps, or cysts. To ensure robustness, we employ a three-fold cross-validation strategy, training three distinct 3D CNNs. Subsequently, an EM is constructed by aggregating the predictions of these three 3D CNNs, and the final predictions are obtained by averaging the predictions of these models.

Statistical Analysis

The McNemar test was used to test for statistically significant differences between the predictions of two 3D-CNN classifiers. We checked for statistical significance between 3D CNNs trained using single fold against EM. We performed χ^2 test to check for statistically significant associations between categorical variables. We checked the point-biserial correlation, which is a special case of Pearson correlation, to measure the relationship between a continuous variable (BMI, alcohol consumption) and a dichotomous variable (participants with MS opacifications and participants without MS opacifications). A two-sided paired p-value <0.05 was considered significant. McNemar test was performed using MLxtend library[16] version 0.22.0, χ^2 and point-biserial correlation was performed using SciPy[17] version 1.2.1. Python version 3.8.10 was used for our experiments.

RESULTS

Performance Evaluation on Test Set

The outcomes of our analysis are presented in Table II, where "CV" represents the cross-validation set, and all reported results pertain to the test set. The analysis reveals that ensembling contributes noticeably to the classification of MS anomalies, yielding an AUROC of 0.95, precision of 0.85, sensitivity of 0.85, and specificity of 0.90. In contrast, second-best model based on AUROC, a 3D-CNN trained on the first fold, achieves an AUROC of 0.93 in the same task. Using McNemar test, $p = 0.15$, $p = 0.01$, and $p = 6.8 \times e^{-4}$ was computed using the predicted labels of CV 1, CV 2, and CV 3 against the

predicted labels of EM, respectively. We achieve an accuracy of 0.90 for MS without opacification, 0.83 for MS with mucosal thickening, 0.88 for MS with polyp, 0.75 for MS with cyst, and 1.0 for fully occupied MS using EM on the test set.

Class Activation Maps

Figure 3 presents activation maps exemplifying various scenarios of MS with and without associated opacifications. These activation maps generated using Guided Grad CAM[18] predominantly localize in clinically significant regions. Specifically, when the 3D-CNN classifier detects no anomalies, the activation maps are primarily concentrated along the boundaries of mucosal walls. In contrast, for cases of MS with opacifications, the activation maps tend to be focused within the mass of the opacifications.

Characteristics of Participants with Opacification and Without Opacification

The objective of this experiment was to demonstrate a practical application of CNNs in automated diagnosis and the expedited generation of clinical insights through correlation studies between expanding cohorts in population studies, aiming to identify pertinent considerations. To simulate an expanding cohort, we leveraged our unlabeled dataset. The EM, trained on our labeled dataset ($N = 1069$), was applied to our unlabeled dataset ($N = 1550$). We implemented an inclusion criterion to select participants for further analysis by excluding MRIs where the EM confidence was <0.90 for at least one of the MS. From 1550 unlabeled MRI scans, only 1360 MRI scans satisfied the inclusion criteria. To both the groups, participants from the labeled dataset were added. Combining the labeled and unlabeled MRI scans, we considered 2429 MRIs for subsequent processing. A schematic drawing of our data handling strategy is presented in Figure 1. The 2429 MRIs were divided into two cohorts. The first cohort comprised participants for whom left and right MS had "no opacification" which is our *control* group. The second cohort consisted of participants with at least one MS having opacification which is our *case* group. Subsequently, we conducted a comparative analysis of these two groups, utilizing both self-evaluated questionnaires and blood reports available from the HCHS for each participant. Detailed results of these comparisons are presented in Tables III and IV. *case* group showed a higher mean alcohol consumption (20.65 (95% [CI] −34.83–76.15) vs. 15.56 (95% [CI] −30.18–61.31); $p < 0.001$) and BMI (27.11 (95% [CI] 18.99–35.33) vs. 26.26 (95% [CI] 17.33–35.21); $p < 0.001$).

Incidence of intrinsic asthma (9.24% [99 of 1072] vs. 6.75% [78 of 1156]; $p = 0.03$) and extrinsic asthma (8.45% [89 of 1053] vs. 5.68% [66 of 1162]; $p = 0.01$) were higher in the *case* group. Male participants were observed more in the *case* group (68.24% [795 of 1165] vs. 43.55% [544 of 1249]; $p < 0.001$). *Case* group showed higher incidence of hay fever allergy (25.07% [272 of 1085] vs. 19.1% [224 of 1173]; $p < 0.001$) and house dust allergy (11.82%

TABLE II.
Performance Metrics.

Metric	CV 1	CV 2	CV 3	EM
Precision	0.82	0.86	0.83	0.85
Sensitivity	0.85	0.77	0.77	0.85
Specificity	0.88	0.91	0.89	0.90
F1	0.84	0.81	0.80	0.85
AUROC	0.93	0.92	0.92	0.95

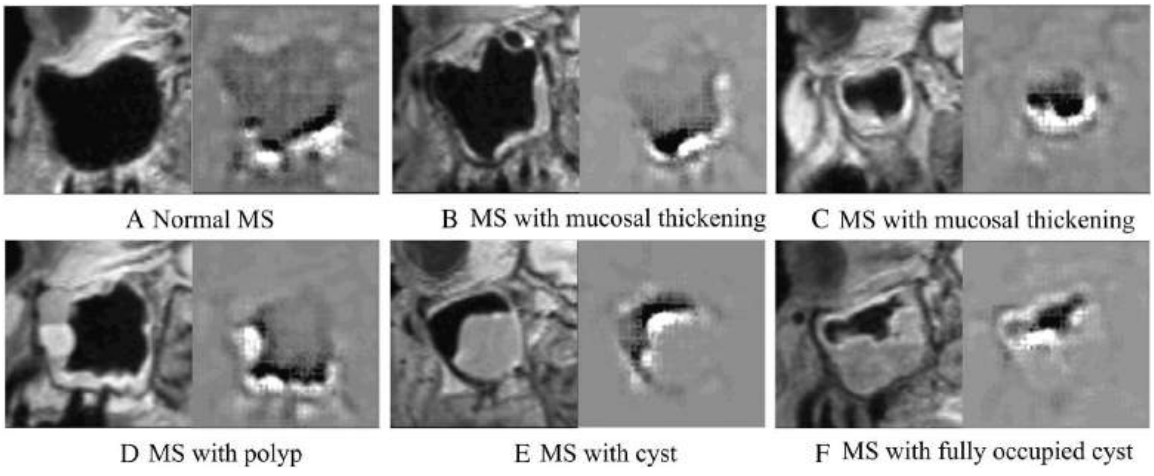

Fig. 3. Sagittal plane images and corresponding activation maps (white pixels meaning high activation and black pixels meaning no activation) of MS exhibiting no opacification, mucosal thickening, polyp, cyst, and fully occupied cyst. In. (A) Normal MS – the high activation is concentrated on the walls of the MS. (B) MS with mucosal thickening – high activation localized on the thickened mucosa. (C) MS with mucosal thickening – high activation localized on the thickened mucosa. (D) MS with polyp – activation inside the polyp mass. (E) MS with cyst – activation inside and on the edges of the cyst mass. (F) MS with fully occupied cyst – activation inside and on the edge of the fully occupied cyst mass.

TABLE III.
Comparison of Participants With No Opacification and Participants With Opacification in at Least One MS with Respect to Health and Lifestyle Factors.

Variable	Participants with no MS opacification	Participants with MS opacification	*p*-value
Smoking habits	$N = 1245$	$N = 1162$	
Yes	218 (17.51%)	202 (17.38%)	0.97
No	1027 (82.49%)	960 (82.62%)	
Alcohol consumption (g/day)	$N = 1170$	$N = 1089$	
Mean (95% CI)	15.56 (−30.18–61.31)	20.65 (−34.83–76.15)	6.79×10^{-8}
BMI	$N = 1210$	$N = 1126$	
Mean (95% CI)	26.26 (17.33–35.21)	27.11 (18.99–35.33)	3.85×10^{-6}
Intrinsic asthma	$N = 1156$	$N = 1072$	
Yes	78 (6.75%)	99 (9.24%)	0.03
No	1078 (93.25%)	973 (90.76%)	
Extrinsic asthma	$N = 1162$	$N = 1053$	
Yes	66 (5.68%)	89 (8.45%)	0.01
No	1096 (94.32%)	964 (91.55%)	
Chronic bronchitis or COPD	$N = 1155$	$N = 1069$	
Yes	67 (5.8%)	74 (6.92%)	0.31
No	1088 (94.2%)	995 (93.08%)	
Sex	$N = 1249$	$N = 1165$	
Male	544 (43.55%)	795 (68.24%)	5.5×10^{-34}
Female	705 (56.45%)	370 (31.76%)	
Age (years)	$N = 1249$	$N = 1165$	
Mean (95% CI)	63.97 (47.5–80.44)	64.01 (47.83–80.2)	0.90
LK	$N = 1220$	$N = 1134$	
Mean (95% CI)	6.19 (2.24–10.15)	6.21 (2.77–9.66)	0.76
hCRP	$N = 1214$	$N = 1126$	
Mean (95% CI)	0.22 (−0.64–1.1)	0.23 (−0.54–1.0)	0.91

[124 of 1049] vs. 8.91% [102 of 1145]; $p = 0.02$) but lower incidence of bee/wasp venom allergy (3.56% [37 of 1038] vs. 5.68% [63 of 1110]; $p = 0.02$). CI denotes confidence interval. No significant statistical correlations were found regarding smoking habits, chronic bronchitis, COPD, LK, hCRP, age, food allergy, animal hair allergy, contact allergy, medication allergy, or other allergies. Figure 4 shows the visualization of all variables with $p < 0.05$ for *control* and *case* group. We analyzed the MRI scans excluded from our study to determine the main factors behind their exclusion. Our findings indicated that borderline cases of mucosal thickening, particularly those around the 2 mm mark, exhibited lower confidence by EM. Additionally, the presence of dental accessories led to the removal of image features around the MS, thereby contributing to decreased confidence in diagnosis. Participant motion during MRI acquisition introduced noise to the data, further lowering confidence. Moreover, albeit rare, certain anatomical variations of the MS, such as those associated with Haller cells, prominent uncinate processes, hypoplasia, or surgical intervention, also contributed to diminished confidence scores. Finally, a few cases involving polyps smaller than 4 mm exhibited low confidence levels. Figure S3 shows bar chart of the image conditions which promotes low confidence score while Figure S4 shows coronal view images of maxillary sinus conditions which promote 3D CNN's low confidence scores. Figure S5 shows maxillary sinus conditions which cause 3D CNN to misclassify. For a more comprehensive understanding of our analysis and potential solutions, please refer to supplementary material section 4.

DISCUSSION

To our knowledge, after Hansen et al. (HUNT-MRI),[5] this is the largest MRI study reporting incidental findings in the paranasal sinuses in an adult, nonselected urban population, recruited for study purposes only. Numerous studies have emphasized the importance of comprehending and addressing the prevalence of paranasal anomalies within the general population. These studies often rely on manual diagnostic methods,[1–5] where clinicians manually record opacification, requiring substantial time and effort. Subsequently, meaningful clinical hypotheses based on available participant data can be tested for statistical significance. In our work, we leverage 3D CNNs to enhance the efficiency of population studies involving large participant cohorts and use it to derive rapid clinical insights, thus reducing the workload of clinicians.

In our study, we focused on the development and validation of an EM, building upon our previous research in

TABLE IV.
Comparison of Participants With No Opacification and Participants with Opacification in at Least One MS with Respect to Different Allergies.

Variable	Participants with no MS opacification	Participants with MS opacification	*p*-value
Hay fever	*N* = 1173	*N* = 1085	
Yes	224 (19.1%)	272 (25.07%)	0.0007
No	949 (80.9%)	813 (74.93%)	
Bee/wasp venom allergy	*N* = 1110	*N* = 1038	
Yes	63 (5.68%)	37 (3.56%)	0.02
No	1047 (3.56%)	1001 (96.44%)	
Food allergy	*N* = 1134	*N* = 1053	
Yes	102 (8.99%)	107 (10.16%)	0.39
No	1032 (91.09%)	946 (89.84%)	
House dust allergy	*N* = 1145	*N* = 1049	
Yes	102 (8.91%)	124 (11.82%)	0.02
No	1043 (91.09%)	925 (88.18%)	
Allergy to animal hair	*N* = 1152	*N* = 1065	
Yes	85 (7.38%)	96 (9.01%)	0.18
No	1067 (92.62%)	969 (90.99%)	
Contact allergy	*N* = 1132	*N* = 1052	
Yes	73 (6.45%)	49 (4.66%)	0.08
No	1059 (93.55%)	1003 (95.34%)	
Medication allergy	*N* = 1110	*N* = 1027	
Yes	154 (13.87%)	132 (12.85%)	0.52
No	956 (86.13%)	895 (87.15%)	
Other allergies	*N* = 1094	*N* = 1006	
Yes	93 (8.5%)	81 (8.05%)	0.76
No	1001 (91.5%)	925 (91.95%)	

multiple instance ensembling for the classification of MS opacifications.[12] To further improve classification accuracy and prediction reliability, we employed a strategy of training three distinct CNNs on various cross-validation folds, coupled with an additional ensemble approach, a technique beneficial to DL-based CAD.[19] The robustness of our prediction model was visually demonstrated through the creation of attention maps, as seen in Figure 3. These maps revealed the CNN's focused activation on clinically significant regions during prediction processes. Notably, for cases lacking opacification, the CNN showed activation predominantly in the bottom wall, while in opacification cases, it highlighted areas such as thickened mucosal walls, polyps, and cyst masses. The precision of these attention maps played a crucial role in enhancing the reliability of our predictions, ensuring that the CNN's accuracy was not compromised by irrelevant image correlations. To show our CNN's effectiveness in deriving rapid clinical insight, we segregated participants into two categories: *control* and *case*. This separation allowed for a detailed analysis of the association between these groups and a range of clinical variables, encompassing both health and lifestyle factors as well as allergic reactions and sensitivities.

Our analysis unveiled that 48.33% (1174 of 2429) exhibited MS opacifications, comprising 13.50% (328 of 2429) with only right MS opacifications, 14.53% (353 of 2429) with only left MS opacifications and 20.29% (493 of 2429) with both left and right MS opacifications. In comparison, Hansen et al.[5], reported 66% (648 of 982) displayed opacifications in their population study. Table III presents associations related to health and lifestyle factors. Males in the *case* group displayed more MS opacifications (68.24% [795 of 1165] vs. 43.55% [544 of 1249]). This trend aligns with other studies finding similiar statistical significance with respect to sex.[1,3,5,20] The higher prevalence of MS opacifications among males may be attributable to an increased risk of allergic rhinitis.[21]

Recognizing the predominant male representation in our *case* cohort, we conducted a sex-controlled case–control analysis as elaborated in supplementary material section 5. Furthermore, participants diagnosed with bronchial asthma exhibited a heightened prevalence of MS opacifications, a trend consistent with findings reported by Hamilos et al.[22] and Zamarron et al.[23] Our study, akin to prior studies[1,3,5,20], did not find a link between smoking and MS opacifications. Our results also did not indicate a significant correlation between age and blood parameters in relation to MS opacifications. No existing literature was found correlating blood parameters with MS opacifications.

With respect to allergy related variables presented in Table IV, participants with MS opacifications also had a higher incidence of hay fever and house dust allergy. Although the literature did not show good prospective studies to address the coexistence of allergic rhinitis and rhinosinusitis (acute and chronic), many studies describe

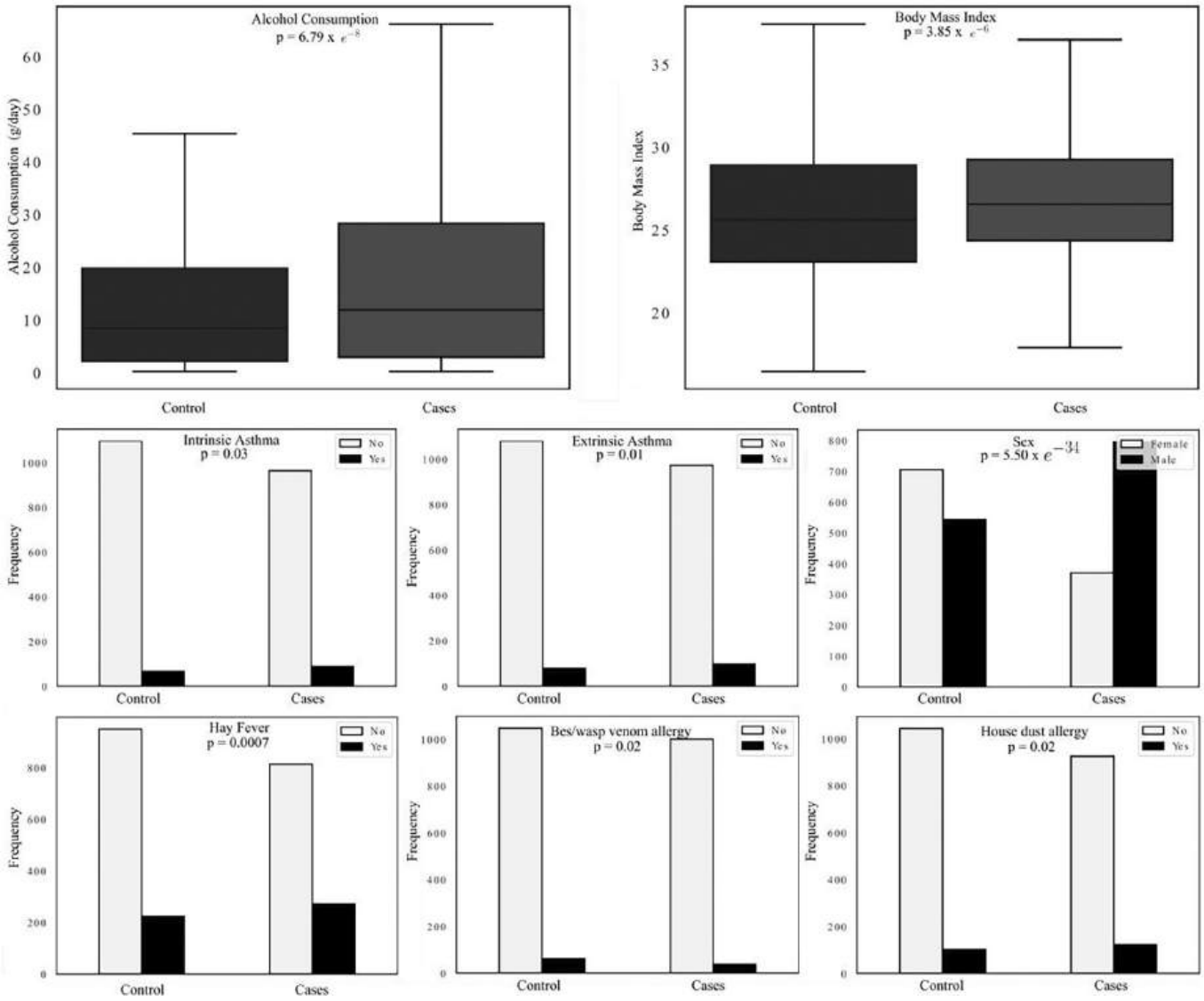

Fig. 4. Box plots and bar charts for continuous and categorical valued variables with $p < 0.05$. The variables compared are mentioned in the respective figure.

a correlation between these two entities.[24–28] Limited research has explored sinus inflammation postallergic nasal inflammation. Baroody et al.[29] and Pelikan et al.[30] demonstrated sinus inflammation through antigen-based nasal challenges. Later, Baroody et al.[31] found MS inflammation correlating with allergy seasons. Allergic rhinitis typically peaks between the second and fourth decades and diminishes thereafter,[32] notable in the context of our study's 45–74 age range participants. Slavin et al.[33] used imaging techniques of the sinuses during the ragweed season and found no changes in the sinuses.

In conclusion, our study shows methodology to develop a CNN for classifying paranasal sinuses and shows a potential application of enhancing case–control analysis in population studies. Our application reaffirms existing correlations (sex, asthma) and refutes any significant link with smoking habits, consistent with prior research. Additionally, we have uncovered previously unexplored associations (alcohol consumption, BMI, hay fever, bee/wasp venom allergy, house dust allergy) not found in the current literature. These findings underscore the potential of deep learning-based CAD, not only in enhancing MS opacifications diagnosis but also in expediting large-scale population studies. Analysis of the excluded MRIs shows the pertinent issues to consider such as noisy data, anatomically diverse MS and challenging pathological morphologies with potential techniques to redress them (see supplementary material section 4. By replacing time-consuming manual diagnosis with rapid automation, our work enables swift access to critical clinical insights. The potential for rapid acquisition of insights may enable real-time monitoring of changes across various variables as the cohort size expands, especially in a prospective study. Such capabilities are essential for effective long-term health monitoring of large populations.

LIMITATIONS

This study has limitations. First, the training data came from one centre, potentially limiting CAD generalizability. Using multicenter, multiethnic data can enhance the reliability of our 3D CNN. Second, we only had MRI and reported clinical data. There were no data on sinonasal symptoms at the time of the MRIs, so we were unable to relate findings to current symptoms. We added questions about symptoms including the Sino-nasal Outcome test (SNOT)[34] to the protocol for the next 10000 participants, to relate the opacities specifically to sinonasal symptoms in the next evaluation. Finally, our model focuses on MS opacifications, excluding other sinuses. Future research should consider a broader sinus opacification classification.

CONCLUSION

We present a CAD system employing CNN to classify MS opacifications and compare them with clinical data. While studies have explored prevalence of MS opacifications,[7–11] they have not been integrated into the broader context of correlating with clinical data using CAD. Our approach offers a less labor-intensive solution for detecting and classifying MS opacifications, leveraging one of the largest datasets available for studying paranasal incidental findings. We demonstrate the effectiveness of our 3D CNN model by generating attention maps, illustrating its ability to focus on clinically significant regions during the diagnostic process.

ACKNOWLEDGMENTS

This work has not been submitted for publication anywhere else. This work is funded partially by the i3 initiative of the Hamburg University of Technology. The authors also acknowledge the partial funding by the Free and Hanseatic City of Hamburg (Interdisciplinary Graduate School) from University Medical Center Hamburg-Eppendorf. This work was partially funded by Grant Number KK5208101KS0 (Zentrales Innovationsprogramm Mittelstand, Arbeitsgemeinschaft industrieller Forschungsvereinigungen). Open Access funding enabled and organized by Projekt DEAL.

BIBLIOGRAPHY

1. Tarp B, Fiirgaard B, Christensen T, Jensen JJ, Black FT. The prevalence and significance of incidental paranasal sinus abnormalities on MRI. *Rhinology*. 2000;38(1):33-38.
2. Rak KM, Newell JD, Yakes WF, Damiano MA, Luethke JM. Paranasal sinuses on MR images of the brain: significance of mucosal thickening. *AJR Am J Roentgenol*. 1991;156(2):381-384. https://doi.org/10.2214/ajr.156.2.1898819.
3. Cooke LD, Hadley DM. MRI of the paranasal sinuses: incidental abnormalities and their relationship to symptoms. *J Laryngol Otol*. 1991;105(4):278-281. https://doi.org/10.1017/s0022215100115609.
4. Rege ICC, Sousa TO, Leles CR, Mendonça EF. Occurrence of maxillary sinus abnormalities detected by cone beam CT in asymptomatic patients. *BMC Oral Health*. 2012;12:30. https://doi.org/10.1186/1472-6831-12-30.
5. Hansen AG, Helvik AS, Nordgård S, et al. Incidental findings in MRI of the paranasal sinuses in adults: a population-based study (HUNT MRI). *BMC Ear Nose Throat Disord*. 2014;14(1):13. https://doi.org/10.1186/1472-6815-14-13.
6. Stec N, Arje D, Moody AR, Krupinski EA, Tyrrell PN. A systematic review of fatigue in radiology: is it a problem? *Am J Roentgenol*. 2018;210(4):799-806. https://doi.org/10.2214/AJR.17.18613.
7. Kim HG, Lee KM, Kim EJ, Lee JS. Improvement diagnostic accuracy of sinusitis recognition in paranasal sinus X-ray using multiple deep learning models. *Quant Imaging Med Surg*. 2019;9:942-951. https://doi.org/10.21037/QIMS.2019.05.15.
8. Ozbay S, Tunc O. Deep Learning in Analysing Paranasal Sinuses. *Elektronika ir Elektrotechnika*. 2022;28:65-70. https://doi.org/10.5755/J02.EIE.31133.
9. Kim KS, Kim BK, Chung MJ, Cho HB, Cho BH, Jung YG. Detection of maxillary sinus fungal ball via 3-D CNN-based artificial intelligence: fully automated system and clinical validation. *PLoS One*. 2022;17(2):1-19. https://doi.org/10.1371/journal.pone.0263125.
10. Bhattacharya D, Behrendt F, Becker BT, et al. Unsupervised anomaly detection of paranasal anomalies in the maxillary sinus. In: Iftekharuddin KM, Chen W, eds. *Medical Imaging 2023: Computer-Aided Diagnosis*. Vol 12465. International Society for Optics and Photonics. SPIE; 2023:124651B.
11. Bhattacharya D, Becker BT, Behrendt F, et al. Supervised contrastive learning to classify paranasal anomalies in the maxillary sinus. In: Wang L, Dou Q, Fletcher PT, Speidel S, Li S, eds. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Cham: Springer Nature Switzerland; 2022:429-438.
12. Bhattacharya D, Behrendt F, Becker BT, et al. Multiple instance ensembling for paranasal anomaly classification in the maxillary sinus. *Int J Comput Assist Radiol Surg*. 2023;19:223-231.
13. Jagodzinski A, Johansen C, Koch-Gromus U, et al. Rationale and design of the Hamburg city health study. *Eur J Epidemiol*. 2019;35(2):169-181.
14. Huang G, Liu Z, Pleiss G, van der Maaten L, Weinberger K. Convolutional networks with dense connectivity. *IEEE Trans Pattern Anal Mach Intell*. 2019;44:8704-8716.
15. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Institute of Electrical and Electronics Engineers Inc (IEEE); 2017.
16. Raschka S. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. *J Open Source Softw*. 2018;3(24):638. https://doi.org/10.21105/joss.00638.
17. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in python. *Nat Methods*. 2020;17:261-272. https://doi.org/10.1038/s41592-019-0686-2.
18. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision*. Institute of Electrical and Electronics Engineers Inc (IEEE); 2017:618-626.
19. Kumar A, Kim J, Lyndon D, Fulham M, Feng D. An ensemble of fine-tuned convolutional neural networks for medical image classification. *IEEE J Biomed Health Inform*. 2016;21(1):31-40.
20. Havas TE, Motbey JA, Gullane PJ. Prevalence of incidental abnormalities on computed tomographic scans of the paranasal sinuses. *Arch Otolaryngol Head Neck Surg*. 1988;114(8):856-859.
21. Tran NP, Vickery J, Blaiss MS. Management of rhinitis: allergic and non-allergic. *Allergy, Asthma Immunol Res*. 2011;3(3):148-156.
22. Hamilos DL. Chronic sinusitis. *J Allergy Clin Immunol*. 2000;106:213-227. https://doi.org/10.1067/mai.2000.109269.
23. Zamarron E, Romero D, Fern'andez-Lahera J, et al. Should we consider paranasal and chest computed tomography in severe asthma patients? *Respir Med*. 2020;169:106013. https://doi.org/10.1016/j.rmed.2020.106013.
24. Savolainen S. Allergy in patients with acute maxillary sinusitis. *Allergy*. 1989;44(2):116-122.
25. Steinke JW, Borish L. The role of allergy in chronic rhinosinusitis. *Immunol Allergy Clin N Am*. 2004;24(1):45-57.
26. Krause HF. Allergy and chronic rhinosinusitis. *Otolaryngol Head Neck Surg*. 2003;128(1):14-16.
27. Slavin RG, Spector SL, Bernstein IL, et al. The diagnosis and management of sinusitis: a practice parameter update. *J Allergy Clin Immunol*. 2005;116(6 Suppl):S13-S47.
28. Gutman M, Torres A, Keen KJ, Houser SM. Prevalence of allergy in patients with chronic rhinosinusitis. *Otolaryngol Head Neck Surg*. 2004;130(5):545-552.
29. Baroody FM, Mucha SM, Detineo M, Naclerio RM. Nasal challenge with allergen leads to maxillary sinus inflammation. *J Allergy Clin Immunol*. 2008;121(5):1126-1132.e7.
30. Pelikan Z, Pelikan-Filipek M. Role of nasal allergy in chronic maxillary sinusitis– diagnostic value of nasal challenge with allergen. *J Allergy Clin Immunol*. 1990;86(4 Pt 1):484-491.
31. Baroody FM, Mucha SM, DeTineo M, Naclerio RM. Evidence of maxillary sinus inflammation in seasonal allergic rhinitis. *Otolaryngol Head Neck Surg*. 2012;146(6):880-886.
32. Wheatley LM, Togias A. Clinical practice. Allergic rhinitis. *N Engl J Med*. 2015;372(5):456-463.
33. Slavin RG, Leipzig JR, Goodgold HM. "Allergic sinusitis" revisited. *Ann Allergy Asthma Immunol*. 2000;85(4):273-276. https://doi.org/10.1016/S1081-1206(10)62529-X.
34. Albrecht T, Beule AG, Hildenbrand T, et al. Cross-cultural adaptation and validation of the 22-item sinonasal outcome test (SNOT-22) in German-speaking patients: a prospective, multicenter cohort study. *Eur Arch Otorhinolaryngol*. 2022;279(5):2433-2439. https://doi.org/10.1007/s00405-021-07019-6.

SUPPLEMENTARY MATERIAL FOR COMPUTER-AIDED DIAGNOSIS OF MAXILLARY SINUS ANOMALIES: VALIDATION AND CLINICAL CORRELATION

Debayan Bhattacharya MSc[1,2] Benjamin Tobias Becker MD[2]
Finn Behrendt MSc[1] Dirk Beyersdorff MD, PhD[3]
Elina Petersen MSc[4] Marvin Petersen MD[5]
Bastian Cheng MD, PhD[5] Dennis Eggert PhD[2]
Christian Betz MD, PhD[2] Alexander Schlaefer PhD[1]
Anna Sophie Hoffmann MD, PhD[2]

[1]

[1]Institute of Medical Technology and Intelligent Systems, Technische Universitaet Hamburg, Germany
[2]Department of Otorhinolaryngology, Head and Neck Surgery and Oncology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany
[3]Clinic and Polyclinic for Diagnostic and Interventional Radiology and Nuclear Medicine, University Medical Center Hamburg-Eppendorf, Hamburg, Germany
[4]Population Health Research Department, University Heart and Vascular Center, University Medical Center Hamburg-Eppendorf, Hamburg, Germany
[5]Department of Neurology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

1 Localisation strategy

In order to increase our dataset and allow for the use of various instances of maxillary sinus (MS) volumes for our collective forecast, we took several sub-volumes from left and right MS from individual MRI scans of the head and neck. We achieved this by recording the centroid locations of the left and right MS from 20 patients manually. We then used these coordinates to calculate the mean and standard deviation of the centroid locations. These values are denoted as $\mu(x), \mu(y), \mu(z)$ and $\sigma(x), \sigma(y), \sigma(z)$ for the mean and standard deviation, respectively. We then initialize Gaussian distributions $\mathcal{N}(\mu(x), \sigma^2(x))$, $\mathcal{N}(\mu(y), \sigma^2(y))$, $\mathcal{N}(\mu(z), \sigma^2(z))$ and use these distributions to sample centroid locations for MS volumes in the head and neck MRI. In practice, we have six Gaussian distributions resulting from the mean and standard deviation of the left and right MS volume. We sample 15 left MS volumes and 15 right MS from each head and neck MRI. An illustration of our sampling method is shown in Figure 1 (a). We extract MS volumes of size 65×65×65.

2 Deep learning method

In order to classify MS volumes as either normal or anomalous, we utilized a 3D DenseNet 264 classifier denoted as $f(.)$, which considered cranial MRI volumes represented by $\mathrm{X} \in R^{H \times W \times D}$. To extract the necessary information, we obtained 15 left MS volumes and 15 right MS volumes from each MRI, resulting in 30 MS volumes being extracted from the MRI, each denoted by $\mathrm{x} \in R^{64 \times 64 \times 64}$. Our goal was to classify each extracted MS volume into one of two classes: no opacification or opacification, with labels $\mathrm{y} \in 0, 1$, where the pathology class is the positive class for our use-case.

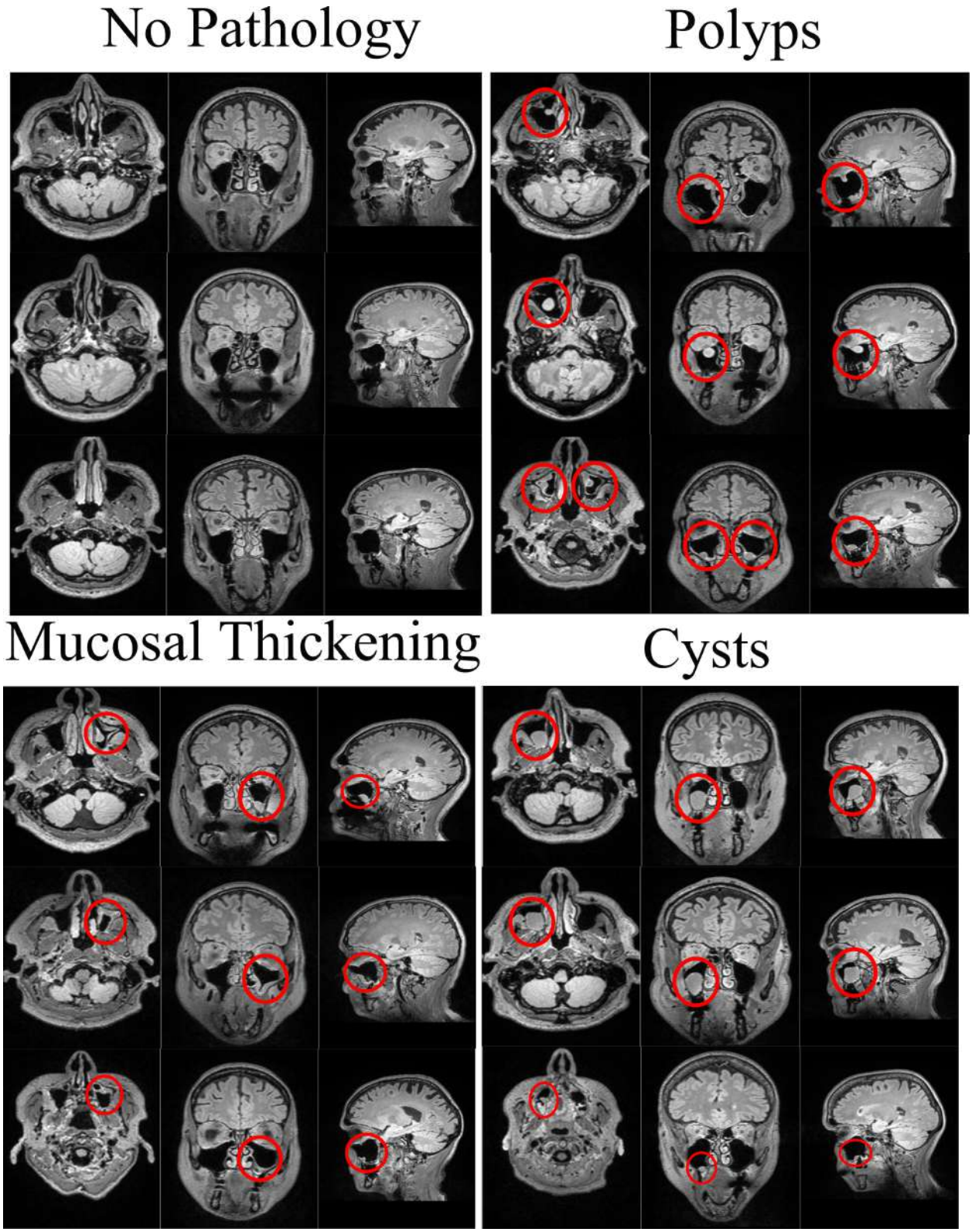

Figure S1: Exemplerary cases showing the axial, coronal, sagittal planes of cranial MRIs exhibiting no pathology, polyp, cyst and mucosal thickening in MS. The MS with pathologies is marked with red contours.

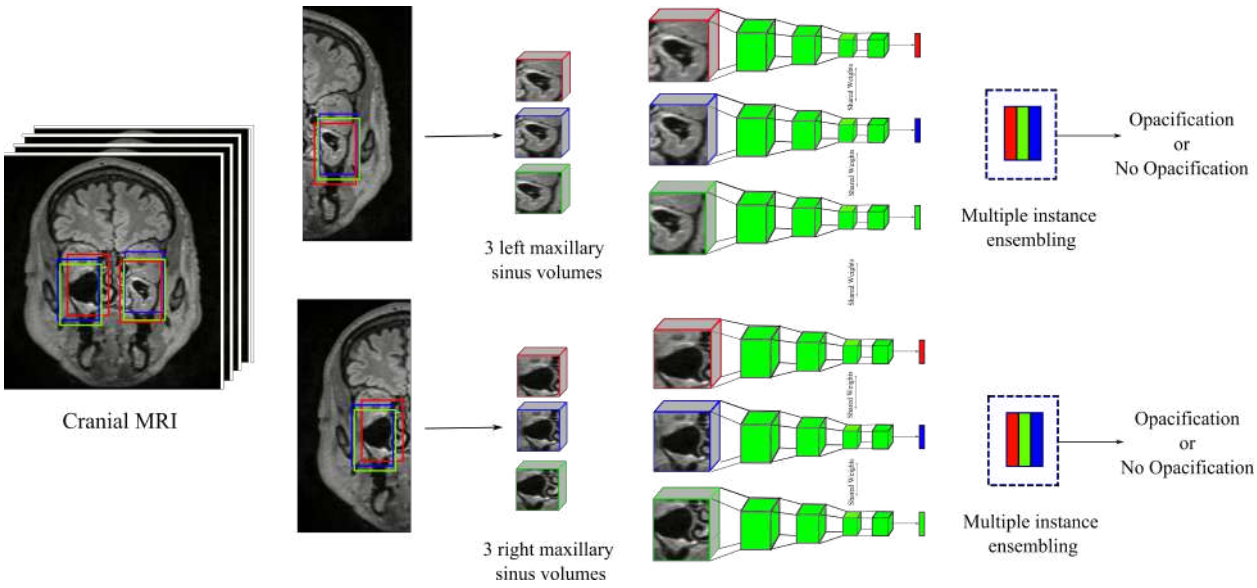

Figure S2: Data processing pipeline showing how a single MRI is processed for inferring opacifications within left and right MS of a single patient. As an illustration, 3 MS volumes are extracted from left and right sides of MRI. Predictions for left or right MS of a patient combines predictions of 3 extracted MS volumes from the respective side and ensembled to create a single prediction. In practice, we extract 15 MS volumes from left and right side of the MRI. This ensembling strategy is termed as multiple instance ensembling.

3 Multiple Instance Ensemble Prediction Strategy

Let us denote each extracted MS volume from the left or right MS area of the MRI as $x_i \in R^{64\times64\times64}$ where i represents the i-th MS volume extracted from either the left or right MS area of the MRI. To obtain an accurate prediction during inference, we employed a multiple instance ensemble strategy by averaging the softmax scores of classifier across all the extracted MS volumes. Specifically, we take the average of the softmax scores from the N MS volumes, resulting in a single prediction score. Mathematically, this is represented as:

$$\hat{y} = \frac{1}{N}\sum_{N}^{i=1} softmax(f(x_i)) \tag{1}$$

Figure S2 shows our data processing pipeline.

3.1 Development of 3D Convolutional Neural Network

Our CNN was trained with a batch size of 16 and an initial learning rate of 0.0001. The learning rate was reduced by a factor of 10 if the validation loss did not improve after 5 epochs. Early stopping with patience 10 was employed to prevent overfitting. Adam optimizer was used to optimize the 3D CNN. Image processing, model development, training and validation were conducted using Python version 3.8.10 and the deep learning framework PyTorch version 1.13.0 with CUDA 11.6. The model was validated on the test set using the precision, sensitivity, specificity, F1 and AUROC. The positive class was the "opacification" class.

4 Analysis of low confidence predictions

From an initial pool of 1550 unlabelled MRI scans, 1360 scans met our predetermined inclusion criteria. Subsequently, we directed our analysis towards the MRI scans that did not meet these criteria, specifically those with at least one MS region identified by our 3D CNN with a confidence score below 0.90. Among the 190 excluded MRI scans, we randomly selected 60 (31.5%) for further investigation, aiming to ascertain the reasons for misclassification compared to our correctly identified MRI scans. Upon examination, we identified several factors contributing to the low confidence scores. Notably, 55% [33 of 60] participants displayed MS with mucosal thickening of 2mm circumferential in at least

one MS. 38.3% [23 of 60] MRI scans had artefacts such as MS image features removed because of dental restorations, such as crowns, fillings, and orthodontic appliances. Participant motion during imaging sessions also resulted in blurry features, further diminishing confidence scores. In total, 50% [30 of 60] MRI scans were affected by motion artefacts. Furthermore, anatomical variations in the MS were observed to contribute to low confidence scores. Notable anatomical anomalies included Haller cell (8.33 % [5 of 60]), uncinate process variations (5% [3 of 60]), and hypoplasia (6.67% [4 of 60]). Surgical interventions also introduced MS deformations leading to reduced confidence in 8.33% [5 of 60]. Additionally, 5% [3 of 60] MRIs exhibited pathological conditions such as MS with polyps under 4 mm. We did not find any new pathology that was not present in our labelled dataset. Figure S3 shows distribution of image conditions which promote low confidence. Figure S4 qualitatively shows different MS conditions for which 3D CNN predicts low confidence scores.

The analysis underscores that the primary factors contributing to low confidence scores were MRI artefacts and instances of maxillary sinus (MS) thickening of 2mm circumferential. These represent ambiguous edge cases, exhibiting characteristics of both normal and mucosal thickening in MS. A potential strategy to mitigate low confidence around the 2mm threshold is to redefine mucosal thickening cases as those with swelling significantly greater than 2mm (for example 5mm). This adjustment aims to accentuate the disparity between normal and anomalous MS within the pixel space, potentially enhancing diagnostic accuracy and confidence levels. Furthermore, artefacts emerged as prominent contributors to anomalies. Proposed strategies include employing data augmentation techniques to simulate motion and dental artefacts and expanding the training dataset to incorporate MRI scans with representative artefacts. However, it is noteworthy that dental artefacts may result in the removal of pertinent image information around the MS, rendering diagnosis unfeasible for such cases. Additionally, anatomical variations stemming from Haller cells, uncinate processes, hypoplasia, and surgical interventions accounted for the remainder of low confidence scores. Although rare, one potential approach is to augment the training dataset to include these variations in training the 3D CNN. Finally, small polypoid masses under 4mm also exhibited low confidence scores. To address this, increasing the inclusion of such cases in the dataset, decreasing the decision threshold or generating synthetic MS images with small polypoid masses using generative deep learning methods could be considered.

5 Analysis of the correlation of alcohol consumption in case-control study

To investigate the relationship between alcohol consumption and MS opacifications while mitigating the influence of sex, we conducted separate analyses on male participants only. This analysis revealed a statistically significant correlation ($N = 1237, p = 0.002$) between alcohol consumption and MS opacifications, with male from *case* cohort consuming 25.13 $\pm$ 31.90 g/day compared to males from *control* cohort consuming 20.01 $\pm$ 26.17 g/day. Subsequently, we examined only female participants within the *case* and *control* cohorts. In this subgroup analysis, we did not find a significant correlation ($N = 1022, p = 0.14$) between alcohol consumption and MS opacifications. Specifically, females from the *case* cohort exhibited an alcohol consumption of 13.96 $\pm$ 22.49 g/day, while females from the *control* cohort had 12.04 $\pm$ 18.95 g/day. Overall, our findings suggest that male participants tend to consume more alcohol than their female counterparts, with *cases* males exhibiting highest consumption levels. Conversely, female participants generally consumed similar and less alcohol in case-control cohort. These results suggest a potential association between alcohol consumption and the development of maxillary sinus opacifications.

We also conducted case-control analyses while controlling for age to check if age influenced the correlation. For this, we determined the median age to be 65 years. In the first analysis, we exclusively examined alcohol consumption in participants under 65 years and found a significant correlation (N=1167, p=6.56 x 10^{-5}). Subsequently, in the second analysis focusing on participants over the median age of 65, we also identified a significant correlation between case-control and alcohol consumption (N=1072, p = 0.001). These results suggest there is potentially no effect of age with respect to alcohol consumption.

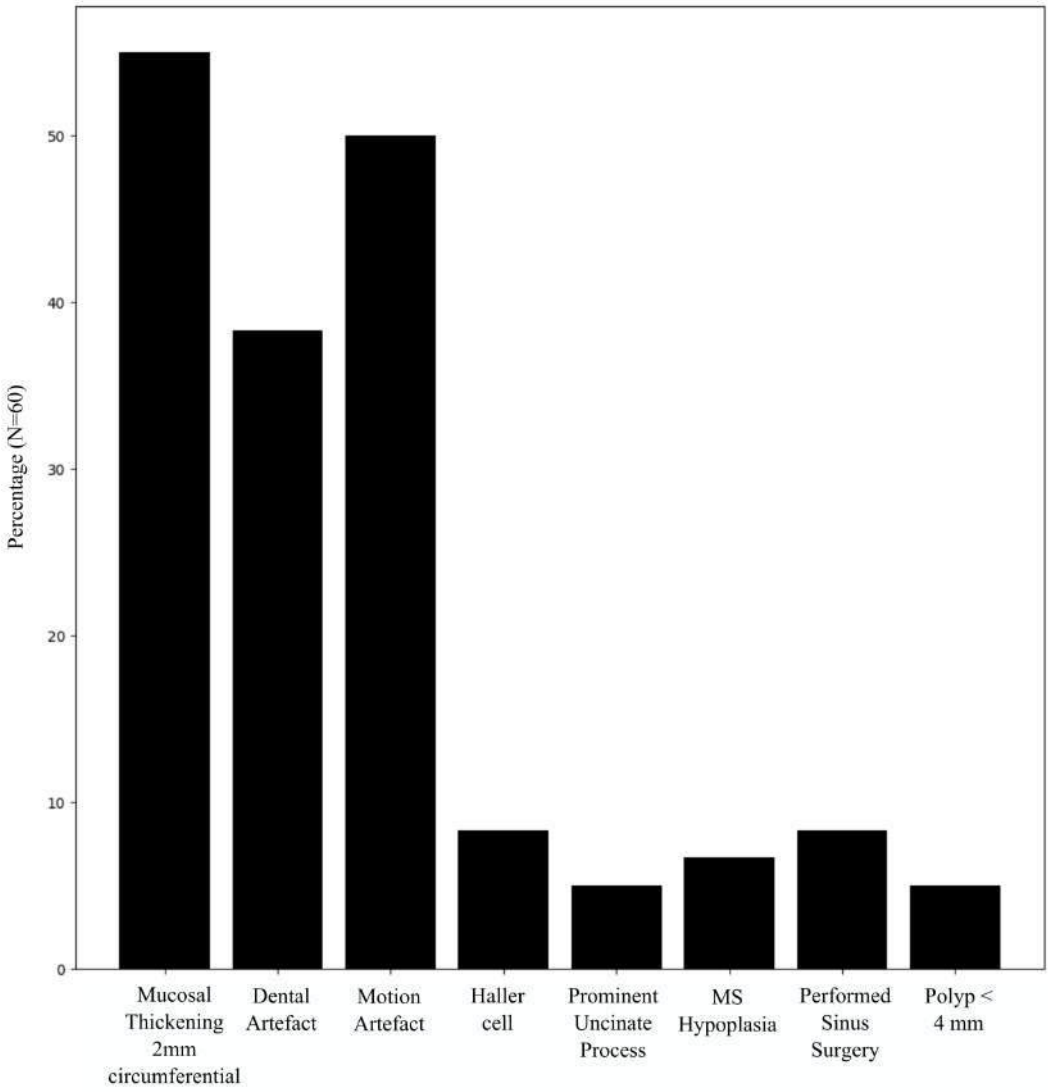

Figure S3: Analysis of image condition which promotes low confidence score of 3D CNN .

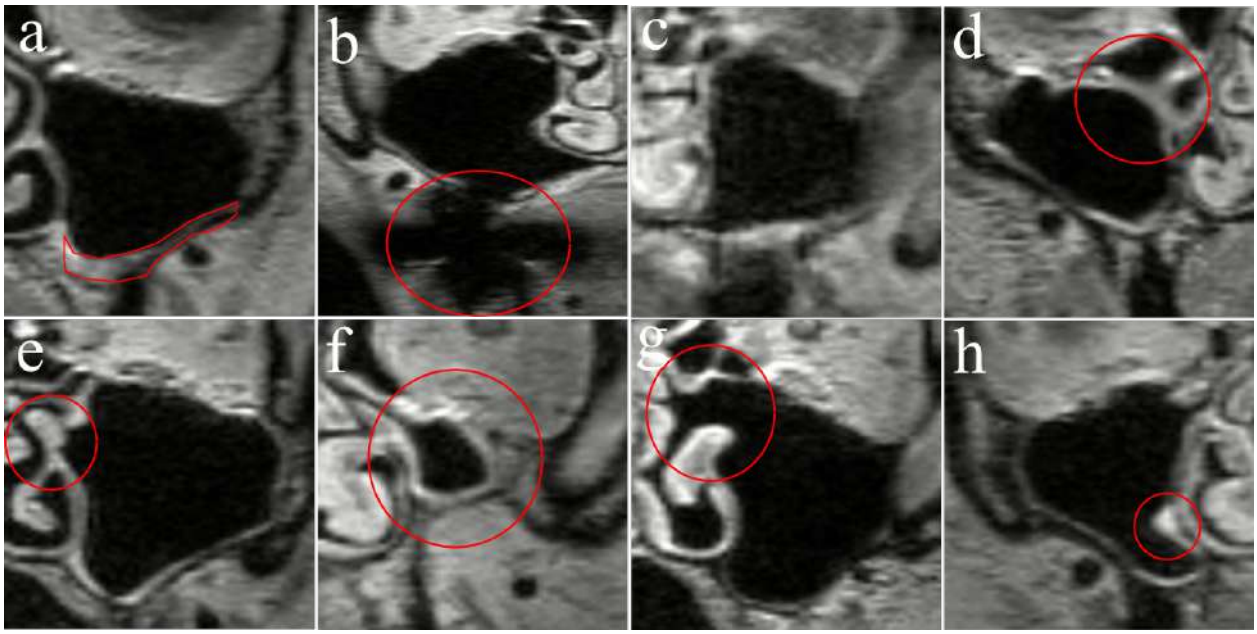

Figure S4: 3D CNN predicts confidence < 0.90 for (a) MS with mucosal thickening around 2mm (b) MS with dental artefact (c) MS with motion artefact (d) MS with Haller cell (e) Prominent uncinate process (f) MS Hypoplasia (g) MS after surgery (h) MS with polyp < 4mm. Regions of interest are marked using red contour.

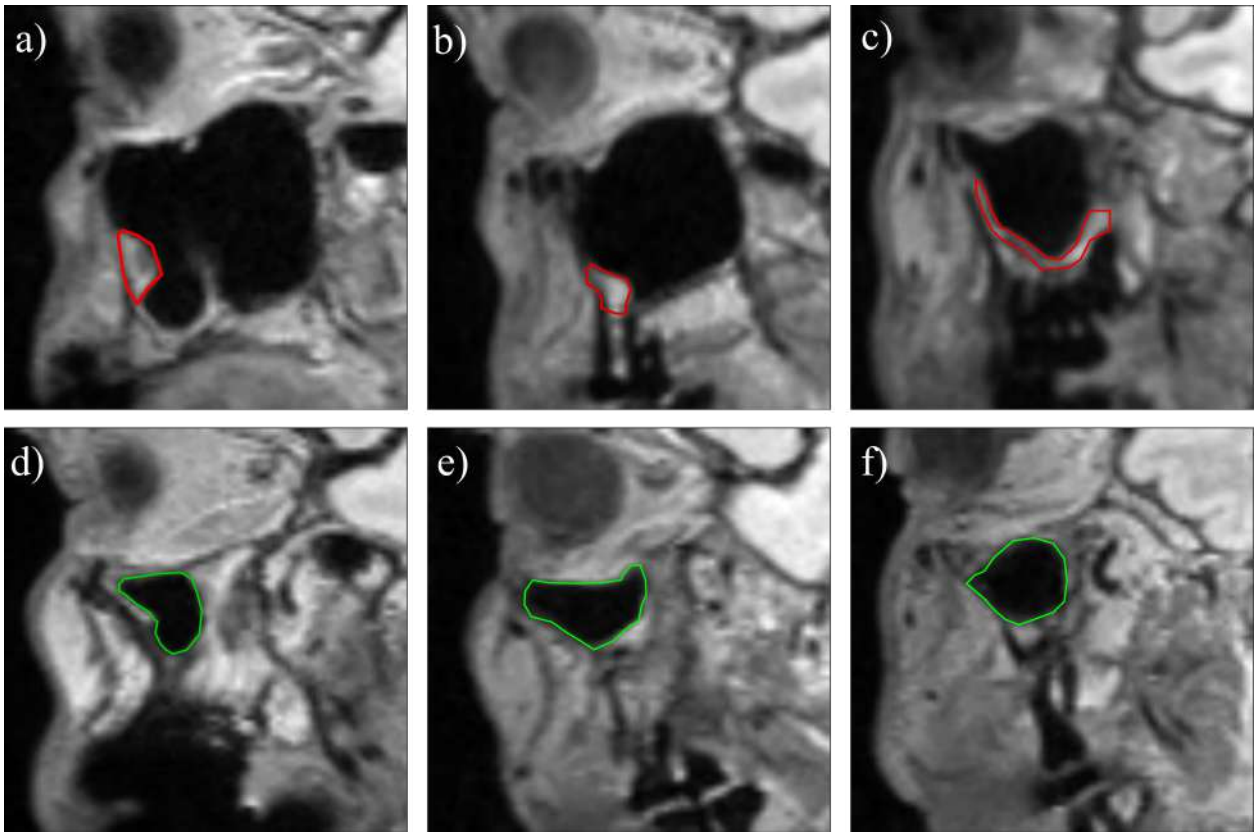

Figure S5: Missclassification of MS condition with confidence < 0.90: Sagitall plane of a) small cyst in the anterior wall of MS misclassified as no-pathology b) small polyp in the bottom wall of MS misclassified as no-pathology c) mucosal thickening slightly larger than 2 mm misclassifed as "no pathology" d,e,f) hypoplastic MS misclassified as pathology. The red contours highlight the anomalous regions. The green contours highlight the boundary walls of hypoplastic MS

LIST OF TABLES

LIST OF FIGURES

REFERENCES

[1] Gülsün Akay, Deniz Yaman, Özge Karadağ, and Kahraman Güngör. Evaluation of the relationship of dimensions of maxillary sinus drainage system with anatomical variations and sinusopathy: Cone-beam computed tomography findings. *Medical Principles and Practice*, 29:354–363, 7 2020.

[2] Mohammed A. Al-masni, Dong Hyun Kim, and Tae Seong Kim. Multiple skin lesions diagnostics via integrated deep convolutional networks for segmentation and classification. *Computer methods and programs in biomedicine*, 190, 7 2020.

[3] Ali Hassan A. Ali, Omar O. Serhan, Mohammed Karrar H. Alsharif, Abubaker Y. Elamin, Sameer Al-Ghamdi, Khaled K. Aldossari, Naif Alrudian, Mansour Alajmi, Bader A. Alhariqi, Mohammad Mokhatrish, and Velmurugan Palanivel. Incidental detection of paranasal sinuses abnormalities on ct imaging of the head in saudi adult population. *PLoS ONE*, 17, 9 2022.

[4] Yoshiko Ariji, Motoki Fukuda, Yoshitaka Kise, Michihito Nozawa, Yudai Yanashita, Hiroshi Fujita, Akitoshi Katsumata, and Eiichiro Ariji. 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 surgery, oral medicine, oral pathology and oral radiology*, 127:458–463, 5 2019.

[5] Aayushi Bansal, Rewa Sharma, and Mamta Kathuria. A systematic review on data scarcity problem in deep learning: Solution and applications. *ACM Computing Surveys (CSUR)*, 54, 9 2022.

[6] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022.

[7] Christoph Baur, Stefan Denner, Benedikt Wiestler, Nassir Navab, and Shadi Albarqouni. Autoencoders for unsupervised anomaly segmentation in brain mr images: A comparative study. *Medical Image Analysis*, 69:101952, 4 2021.

[8] Finn Behrendt, Debayan Bhattacharya, Julia Krüger, Roland Opfer, and Alexander Schlaefer. Patched diffusion models for unsupervised anomaly detection in brain MRI. In *Medical Imaging with Deep Learning*, 2023.

[9] Marcel Bengs, Finn Behrendt, Julia Krüger, Roland Opfer, and Alexander Schlaefer. Three-dimensional deep learning with spatial erasing for unsupervised anomaly segmentation in brain mri. *International Journal of Computer Assisted Radiology and Surgery*, 16:1413–1423, 9 2021.

[10] Marcel Bengs, Finn Behrendt, Max-Heinrich Laves, Julia Krüger, Roland Opfer, and Alexander Schlaefer. Unsupervised anomaly detection in 3d brain mri using deep learning with multi-task brain age prediction. *https://doi.org/10.1117/12.2608120*, 12033:305–309, 4 2022.

[11] Debayan Bhattacharya, Benjamin Tobias Becker, Finn Behrendt, Marcel Bengs, Dirk Beyersdorff, Dennis Eggert, Elina Petersen, Florian Jansen, Marvin Petersen, Bastian Cheng, Christian Betz, Alexander Schlaefer, and Anna Sophie Hoffmann. Supervised contrastive learning to classify paranasal anomalies in the maxillary sinus. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13433 LNCS:429–438, 2022.

[12] Debayan Bhattacharya, Benjamin Tobias Becker, Finn Behrendt, Dirk Beyersdorff, Elina Petersen, Marvin Petersen, Bastian Cheng, Dennis Eggert, Christian Betz, Alexander Schlaefer, and Anna Sophie Hoffmann. Computer-aided diagnosis of maxillary sinus anomalies: Validation and clinical correlation. *The Laryngoscope*, n/a(n/a).

[13] Debayan Bhattacharya, Finn Behrendt, Benjamin Tobias Becker, Dirk Beyersdorff, Elina Petersen, Marvin Petersen, Bastian Cheng, Dennis Eggert, Christian Betz, Anna Sophie Hoffmann, and Alexander Schlaefer. Multiple instance ensembling for paranasal anomaly classification in the maxillary sinus. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–9, 7 2023.

[14] Debayan Bhattacharya, Finn Behrendt, Benjamin Tobias Becker, Lennart Maack, Dirk Beyersdorff, Elina Petersen, Marvin Petersen, Bastian Cheng, Dennis Eggert, Christian Betz, Anna Sophie Hoffmann, and Alexander Schlaefer. Self-supervised learning for classifying paranasal anomalies in the maxillary sinus. *International Journal of Computer Assisted Radiology and Surgery*, Jun 2024.

[15] Debayan Bhattacharya, Dennis Eggert, Christian Betz, and Alexander Schlaefer. Squeeze and multi-context attention for polyp segmentation. *International Journal of Imaging Systems and Technology*, 33:123–142, 1 2023.

[16] PDQ Adult Treatment Editorial Board. Paranasal sinus and nasal cavity cancer treatment (adult) (pdq®). *PDQ Cancer Information Summaries*, 3 2023.

[17] Philippe Jean Bousquet, Pascal Demoly, Philippe Devillier, Kamal Mesbah, and Jean Bousquet. Impact of allergic rhinitis symptoms on quality of life in primary care. *International archives of allergy and immunology*, 160:393–400, 3 2013.

[18] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child,

Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam Mccandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[19] Pascal Vincent@umontreal Ca, Larocheh@cs Toronto Edu, Isabelle Lajoie, Yoshua Bengio@umontreal Ca, and Pierre-Antoine Manzagol@umontreal Ca. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *The Journal of Machine Learning Research*, 11:3371–3408, 12 2010.

[20] Marco Caballo, Domenico R. Pangallo, Ritse M. Mann, and Ioannis Sechopoulos. Deep learning-based segmentation of breast masses in dedicated breast ct imaging: Radiomic feature stability between radiologists and artificial intelligence. *Computers in biology and medicine*, 118, 3 2020.

[21] Bing Cao, Han Zhang, Nannan Wang, Xinbo Gao, and Dinggang Shen. Auto-gan: Self-supervised collaborative learning for medical image synthesis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:10486–10493, 4 2020.

[22] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments.

[23] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *Proceedings of the IEEE International Conference on Computer Vision*, pages 9630–9640, 2021.

[24] Soumick Chatterjee, Alessandro Sciarra, Max Dünnwald, Pavan Tummala, Shubham Kumar Agrawal, Aishwarya Jauhari, Aman Kalra, Steffen Oeltze-Jafra, Oliver Speck, and Andreas Nürnberger. Strega: Unsupervised anomaly detection in brain mris using a compact context-encoding variational autoencoder. *Computers in Biology and Medicine*, 149:106093, 10 2022.

[25] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. 2020.

[26] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 15745–15753, 2021.

[27] Yanbei Chen, Massimiliano Mancini, Xiatian Zhu, and Zeynep Akata. Semi-supervised and unsupervised deep visual learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–23, 8 2022.

[28] Zekai Chen, Devansh Agarwal, Kshitij Aggarwal, Wiem Safta, Mariann Micsinai Balan, and Kevin Brown. Masked image modeling advances 3d medical image analysis. *Proceedings - 2023 IEEE Winter Conference on Applications of Computer Vision, WACV 2023*, pages 1969–1979, 2023.

[29] Shuai Cheng, Qingshan Hou, Peng Cao, Jinzhu Yang, Xiaoli Liu, and Osmar R. Zaiane. Lesion-aware contrastive learning for diabetic retinopathy diagnosis. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 14226 LNCS:671–681, 2023.

[30] Lynn D. Cooke and Donald M. Hadley. Mri of the paranasal sinuses: incidental abnormalities and their relationship to symptoms. *The Journal of laryngology and otology*, 105:278–281, 1991.

[31] J Crystal-Peters, W H Crown, R Z Goetzel, and D C Schutt. The cost of productivity losses associated with allergic rhinitis. *Am J Manag Care*, 6(3):373–378, March 2000.

[32] Yajie Cui, Zhaoxiang Liu, and Shiguo Lian. A survey on unsupervised anomaly detection algorithms for industrial images. *IEEE Access*, 11:55297–55315, 2023.

[33] Rudrajit Das and Subhasis Chaudhuri. On the separability of classes with the cross-entropy loss function. 9 2019.

[34] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. pages 248–255, 3 2010.

[35] P. Devillier, J. Bousquet, H. Salvator, E. Naline, S. Grassin-Delyle, and O. de Beaumont. In allergic rhinitis, work, classroom and activity impairments are weakly related to other outcome measures. *Clinical and experimental allergy : journal of the British Society for Allergy and Clinical Immunology*, 46:1456–1464, 11 2016.

[36] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:4171–4186, 10 2018.

[37] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1422–1430, 2015.

[38] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words:

Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[39] Mark H. Ebell, Brian McKay, Ariella Dale, Ryan Guilbault, and Yokabed Ermias. Accuracy of signs and symptoms for the diagnosis of acute rhinosinusitis and acute bacterial rhinosinusitis. *Annals of Family Medicine*, 17:164, 3 2019.

[40] Davide Farina, Marco Ravanelli, Andrea Borghesi, and Roberto Maroldi. Flying through congested airspaces: imaging of chronic rhinosinusitis. *Insights into Imaging*, 1:155, 7 2010.

[41] Matteo Ferrante, Tommaso Boccato, Simeon Spasov, Andrea Duggento, and Nicola Toschi. Vaesim: A probabilistic approach for self-supervised prototype discovery. *Image and Vision Computing*, 137:104746, 9 2023.

[42] Nils Gessert, Maximilian Nielsen, Mohsin Shaikh, René Werner, and Alexander Schlaefer. Skin lesion classification using ensembles of multi-resolution efficientnets with meta data. *MethodsX*, 7:100864, 1 2020.

[43] Spyros Gidaris, Praveer Singh, and N. Komodakis. Unsupervised representation learning by predicting image rotations. *International Conference on Learning Representations*, 2018.

[44] Lovedeep Gondara. Medical image denoising using convolutional denoising autoencoders. *IEEE International Conference on Data Mining Workshops, ICDMW*, 0:241–246, 7 2016.

[45] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63:139–144, 6 2014.

[46] F. Gordts, P. A.R. Clement, and Th Buisseret. Prevalence of paranasal sinus abnormalities on mri in a non-ent population. *Acta Oto-Rhino-Laryngologica Belgica*, 50:167–170, 1996.

[47] Derek Greene, Pádraig Cunningham, and Rudolf Mayer. Unsupervised learning and clustering. *Cognitive Technologies*, pages 51–90, 2008.

[48] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent a new approach to self-supervised learning.

[49] Aleksander Grande Hansen, Anne Sofie Helvik, Ståle Nordgård, Vegard Bugten, Lars Jacob Stovner, Asta K. Håberg, Mari Gårseth, and Heidi Beate Eggesbø. Incidental findings in mri of the paranasal sinuses in adults: a population-based study (hunt mri). *BMC ear, nose, and throat disorders*, 14, 11 2014.

[50] D. Hastan, W. J. Fokkens, C. Bachert, R. B. Newson, J. Bislimovska, A. Bockelbrink, P. J. Bousquet, G. Brozek, A. Bruno, S. E. Dahlén, B. Forsberg, M. Gunnbjörnsdõttir, L. Kasper, U. Krämer, M. L. Kowalski, B. Lange, B. Lundbäck, E. Salagean, A. Todo-Bom, P. Tomassen, E. Toskala, C. M. Van Drunen, J. Bousquet, T. Zuberbier, D. Jarvis, and P. Burney. Chronic rhinosinusitis in europe–an underestimated disease. a ga^{2}len study. *Allergy*, 66:1216–1223, 9 2011.

[51] Thomas E. Havas, Josephine A. Motbey, and Patrick J. Gullane. Prevalence of incidental abnormalities on computed tomographic scans of the paranasal sinuses. *Archives of otolaryngology–head neck surgery*, 114:856–859, 1988.

[52] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked autoencoders are scalable vision learners. *Proceedings of the IEEE Computer*

Society Conference on Computer Vision and Pattern Recognition, 2022-June:15979–15988, 2022.

[53] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 9726–9735, 2020.

[54] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-December:770–778, 12 2016.

[55] Shaojuan He, Wei Chen, Xuehai Wang, Xinyu Xie, Fangying Liu, Xinyi Ma, Xuezhong Li, Anning Li, and Xin Feng. Deep learning radiomics-based preoperative prediction of recurrence in chronic rhinosinusitis. *iScience*, 26:106527, 4 2023.

[56] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313:504–507, 7 2006.

[57] A. G. Hirsch, W. F. Stewart, A. S. Sundaresan, A. J. Young, T. L. Kennedy, J. Scott Greene, W. Feng, B. K. Tan, R. P. Schleimer, R. C. Kern, A. Lidder, and B. S. Schwartz. Nasal and sinus symptoms and chronic rhinosinusitis in a population-based sample. *Allergy*, 72:274–281, 2 2017.

[58] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models.

[59] Shih Cheng Huang, Anuj Pareek, Malte Jensen, Matthew P. Lungren, Serena Yeung, and Akshay S. Chaudhari. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *npj Digital Medicine 2023 6:1*, 6:1–16, 4 2023.

[60] Stephen M. Humphries, Juan Pablo Centeno, Aleena M. Notary, Justin Gerow, Giuseppe Cicchetti, Rohit K. Katial, Daniel M. Beswick, Vijay R. Ramakrishnan, Rafeul Alam, and David A. Lynch. Volumetric assessment of paranasal sinus opacification on computed tomography can be automated using a convolutional neural network. *International forum of allergy rhinology*, 10:1218–1225, 11 2020.

[61] Dina M. Ibrahim, Nada M. Elshennawy, and Amany M. Sarhan. Deep-chest: Multi-classification deep learning model for diagnosing covid-19, pneumonia, and lung cancer chest diseases. *Computers in Biology and Medicine*, 132:104348, 5 2021.

[62] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:590–597, 7 2019.

[63] Annika Jagodzinski, Christoffer Johansen, Uwe Koch-Gromus, Ghazal Aarabi, Gerhard Adam, Sven Anders, Matthias Augustin, Ramona B. der Kellen, Thomas Beikler, Christian Alexander Behrendt, Christian S. Betz, Carsten Bokemeyer, Katrin Borof, Peer Briken, Chia Jung Busch, Christian Büchel, Stefanie Brassen, Eike S. Debus, Larissa Eggers, Jens Fiehler, Jürgen Gallinat, Simone Gellißen, Christian Gerloff, Evaldas Girdauskas, Martin Gosau, Markus Graefen, Martin Härter, Volker Harth, Christoph Heidemann, Guido Heydecke, Tobias B. Huber, Yassin Hussein, Marvin O. Kampf, Olaf von dem Knesebeck, Alexander Konnopka, Hans Helmut König, Robert Kromer, Christian Kubisch, Simone Kühn, Sonja Loges, Bernd Löwe, Gunnar Lund, Christian Meyer, Lina Nagel, Albert Nienhaus, Klaus Pantel, Elina Petersen, Klaus Püschel, Hermann Reichenspurner, Guido Sauter, Martin Scherer, Katharina Scher-

schel, Ulrich Schiffner, Renate B. Schnabel, Holger Schulz, Ralf Smeets, Vladislavs Sokalskis, Martin S. Spitzer, Claudia Terschüren, Imke Thederan, Tom Thoma, Götz Thomalla, Benjamin Waschki, Karl Wegscheider, Jan Per Wenzel, Susanne Wiese, Birgit Christiane Zyriax, Tanja Zeller, and Stefan Blankenberg. Rationale and design of the hamburg city health study. *European Journal of Epidemiology*, 35:169–181, 2 2020.

[64] Joon Jang, Hyeong Hun Lee, Ji Ae Park, and Hyeonjin Kim. Unsupervised anomaly detection using generative adversarial networks in 1h-mrs of the brain. *Journal of Magnetic Resonance*, 325:106936, 4 2021.

[65] Yejin Jeon, Kyeorye Lee, Leonard Sunwoo, Dongjun Choi, Dong Yul Oh, Kyong Joon Lee, Youngjune Kim, Jeong Whun Kim, Se Jin Cho, Sung Hyun Baik, Roh Eul Yoo, Yun Jung Bae, Byung Se Choi, Cheolkyu Jung, and Jae Hyoung Kim. Deep learning for diagnosis of paranasal sinusitis using multi-view radiographs. *Diagnostics (Basel, Switzerland)*, 11, 2021.

[66] Debesh Jha, Pia H. Smedsrud, Dag Johansen, Thomas De Lange, Havard D. Johansen, Pal Halvorsen, and Michael A. Riegler. A comprehensive study on colorectal polyp segmentation with resunet++, conditional random field and test-time augmentation. *IEEE journal of biomedical and health informatics*, 25:2029–2040, 6 2021.

[67] Antanas Kascenas, Nicolas Pugeault, and Alison Q O'Neil. Denoising autoencoders for unsupervised anomaly detection in brain MRI. In *Medical Imaging with Deep Learning*, 2022.

[68] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Google Research, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

[69] Hee E. Kim, Alejandro Cosa-Linan, Nandhini Santhanam, Mahboubeh Jannesari, Mate E. Maros, and Thomas Ganslandt. Transfer learning for medical image classification: a literature review. *BMC Medical Imaging 2022 22:1*, 22:1–13, 4 2022.

[70] Hyug Gi Kim, Kyung Mi Lee, Eui Jong Kim, and Jin San Lee. Improvement diagnostic accuracy of sinusitis recognition in paranasal sinus x-ray using multiple deep learning models. *Quantitative Imaging in Medicine and Surgery*, 9:942, 2019.

[71] Kyung Su Kim, Byung Kil Kim, Myung Jin Chung, Hyun Bin Cho, Beak Hwan Cho, and Yong Gi Jung. Detection of maxillary sinus fungal ball via 3-d cnn-based artificial intelligence: Fully automated system and clinical validation. *PLoS ONE*, 17, 2 2022.

[72] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes, 12 2013.

[73] Hyoun Joong Kong, Jin Youp Kim, Hye Min Moon, Hae Chan Park, Jeong Whun Kim, Ruth Lim, Jonghye Woo, Georges El Fakhri, Dae Woo Kim, and Sungwan Kim. Automation of generative adversarial network-based synthetic data-augmentation for maximizing the diagnostic performance with paranasal imaging. *Scientific Reports 2022 12:1*, 12:1–12, 10 2022.

[74] Evans Kotei and Ramkumar Thirunavukarasu. A systematic review of transformer-based pre-trained language models through self-supervised learning. *Information 2023, Vol. 14, Page 187*, 14:187, 3 2023.

[75] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[76] Komal Kumar, Snehashis Chakraborty, and Sudipta Roy. Self-supervised diffusion model for anomaly segmentation in medical imaging. pages 359–368, 2023.

[77] Arkadiusz Kwasigroch, Agnieszka Mikołajczyk, and Michał Grochowski. Deep neural networks approach to skin lesions classification - a comparative analysis. *2017 22nd International Conference on Methods and Models in Automation and Robotics, MMAR 2017*, pages 1069–1074, 9 2017.

[78] Qing Li, Weidong Cai, Xiaogang Wang, Yun Zhou, David Dagan Feng, and Mei Chen. Medical image classification with convolutional neural network. *2014 13th International Conference on Control Automation Robotics and Vision, ICARCV 2014*, pages 844–848, 2014.

[79] Wye Keat Lim, Bhaskar Ram, Stephen Fasulakis, and Kevin J. Kane. Incidental magnetic resonance image sinus abnormalities in asymptomatic australian children. *Journal of Laryngology and Otology*, 117:969–972, 12 2003.

[80] Chang Liu, Yuanzhi Cheng, and Shinichi Tamura. Masked image modeling-based boundary reconstruction for 3d medical image segmentation. *Computers in biology and medicine*, 166, 11 2023.

[81] George S. Liu, Angela Yang, Dayoung Kim, Andrew Hojel, Diana Voevodsky, Julia Wang, Charles C.L. Tong, Heather Ungerer, James N. Palmer, Michael A. Kohanski, Jayakar V. Nayak, Peter H. Hwang, Nithin D. Adappa, and Zara M. Patel. Deep learning classification of inverted papilloma malignant transformation using 3d convolutional neural networks and magnetic resonance imaging. *International forum of allergy rhinology*, 12:1025–1033, 8 2022.

[82] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. 12 2016.

[83] Siyuan Lu, Zhihai Lu, and Yu Dong Zhang. Pathological brain detection based on alexnet and transfer learning. *Journal of Computational Science*, 30:41–47, 1 2019.

[84] Guoting Luo, Wei Xie, Ronghui Gao, Tao Zheng, Lei Chen, and Huaiqiang Sun. Unsupervised anomaly detection in brain mri: Learning abstract distribution from massive healthy brains. *Computers in Biology and Medicine*, 154:106610, 3 2023.

[85] Yan Luo, Yongkang Wong, Mohan Kankanhalli, and Qi Zhao. {G}-softmax: Improving intra-class compactness and inter-class separability of features. *IEEE Transactions on Neural Networks and Learning Systems*, 31:685–699, 4 2019.

[86] Omid Nejati Manzari, Hamid Ahmadabadi, Hossein Kashiani, Shahriar B. Shokouhi, and Ahmad Ayatollahi. Medvit: A robust vision transformer for generalized medical image classification. *Computers in Biology and Medicine*, 157:106791, 5 2023.

[87] Jonathan Masci, Ueli Meier, Dan Cireşan, and Jürgen Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6791 LNCS:52–59, 2011.

[88] Ajmal Masood, Ioannis Moumoulidis, and Jaan Panesar. Acute rhinosinusitis in adults: an update on current management. *Postgraduate Medical Journal*, 83:402, 6 2007.

[89] Tojo Mathew, B. Ajith, Jyoti R. Kini, and Jeny Rajan. Deep learning-based automated mitosis detection in histopathology images for breast cancer grading. *International Journal of Imaging Systems and Technology*, 32:1192–1208, 7 2022.

[90] Eli O. Meltzer and Don A. Bukstein. The economic impact of allergic rhinitis and current guidelines for treatment. *Annals of allergy, asthma immunology : official publication of the American College of Allergy, Asthma, Immunology*, 106, 2011.

[91] Sashikala Mishra, Kailash Shaw, Debahuti Mishra, Shruti Patil, Ketan Kotecha, Satish Kumar, and Simi Bajaj. Improving the accuracy of ensemble machine learning classi-

fication models using a novel bit-fusion algorithm for healthcare ai systems. *Frontiers in public health*, 10, 5 2022.

[92] Reza Mohammadi, Iman Shokatian, Mohammad Salehi, Hossein Arabi, Isaac Shiri, and Habib Zaidi. Deep learning-based auto-segmentation of organs at risk in high-dose rate brachytherapy of cervical cancer. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology*, 159:231–240, 6 2021.

[93] Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. Confidence-aware learning for deep neural networks. 2020.

[94] DeepjyotiK Mudgade, PawanC Motghare, GirijaU Kunjir, AshishD Darwade, and AkshayS Raut. Prevalence of anatomical variations in maxillary sinus using cone beam computed tomography. *Journal of Indian Academy of Oral Medicine and Radiology*, 30:18, 2018.

[95] Makoto Murata, Yoshiko Ariji, Yasufumi Ohashi, Taisuke Kawai, Motoki Fukuda, Takuma Funakoshi, Yoshitaka Kise, Michihito Nozawa, Akitoshi Katsumata, Hiroshi Fujita, and Eiichiro Ariji. Deep-learning classification using convolutional neural network for evaluation of maxillary sinusitis on panoramic radiography. *Oral Radiology*, 35:301–307, 9 2019.

[96] Kamil Nar, Orhan Ocal, S. Shankar Sastry, and Kannan Ramchandran. Cross-entropy loss leads to poor margins, 2019.

[97] Alejandro Newell and Jia Deng. How useful is self-supervised pretraining for visual tasks? *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 7343–7352, 2020.

[98] Vahid Noorian and Arya Motaghi. Assessment of the diagnostic accuracy of limited ct scan of paranasal sinuses in the identification of sinusitis. *Iranian Red Crescent Medical Journal*, 14:709, 2012.

[99] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9910 LNCS:69–84, 2016.

[100] Lucy Nwosu, Xiangfang Li, Lijun Qian, Seungchan Kim, and Xishuang Dong. Calibrated bagging deep learning for image semantic segmentation: A case study on covid-19 chest x-ray image. *PloS one*, 17, 11 2022.

[101] Orrett E. Ogle, Robert J. Weinstock, and Ezra Friedman. Surgical anatomy of the nasal cavity and paranasal sinuses. *Oral and maxillofacial surgery clinics of North America*, 24:155–166, 5 2012.

[102] KOLAWOLE S. OKUYEMI and TERANCE T. TSUE. Radiologic imaging in the management of sinusitis. *American Family Physician*, 66:1882–1887, 11 2002.

[103] Serkan Ozbay and Orhan Tunc. Deep learning in analysing paranasal sinuses. *Elektronika ir Elektrotechnika*, 28:65–70, 2022.

[104] Utku Ozbulak, Hyun Jung Lee, Beril Boga, Esla Timothy Anzaku, Ho min Park, Arnout Van Messem, Wesley De Neve, and Joris Vankerschaver. Know your self-supervised learning: A survey on image-based generative and discriminative training. *Transactions on Machine Learning Research*, 2 2023.

[105] James N. Palmer, John C. Messina, Robert Biletch, Kirk Grosel, and Ramy A. Mahmoud. A cross-sectional, population-based survey of u.s. adults with symptoms of chronic rhinosinusitis. *Allergy and asthma proceedings*, 40:48–56, 1 2019.

[106] Taesung Park, Ming Yu Liu, Ting Chun Wang, and Jun Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:2332–2341, 6 2019.

[107] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-December:2536–2544, 12 2016.

[108] Gustav Grund Pihlgren, Fredrik Sandin, and Marcus Liwicki. Pretraining image encoders without reconstruction via feature prediction loss. *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4105–4111, 1 2021.

[109] Walter H.L. Pinaya, Mark S. Graham, Robert Gray, Pedro F. da Costa, Petru Daniel Tudosiu, Paul Wright, Yee H. Mah, Andrew D. MacKinnon, James T. Teo, Rolf Jager, David Werring, Geraint Rees, Parashkev Nachev, Sebastien Ourselin, and M. Jorge Cardoso. Fast unsupervised brain anomaly detection and segmentation with diffusion models. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13438 LNCS:705–714, 2022.

[110] Walter H.L. Pinaya, Petru Daniel Tudosiu, Robert Gray, Geraint Rees, Parashkev Nachev, Sebastien Ourselin, and M. Jorge Cardoso. Unsupervised brain imaging 3d anomaly detection and segmentation with transformers. *Medical Image Analysis*, 79:102475, 7 2022.

[111] Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. Transforming machine translation: a deep

learning system reaches news translation quality comparable to human professionals. *Nature Communications 2020 11:1*, 11:1–15, 9 2020.

[112] K. M. Rak, J. D. Newell, W. F. Yakes, M. A. Damiano, and J. M. Luethke. Paranasal sinuses on mr images of the brain: significance of mucosal thickening. *AJR. American journal of roentgenology*, 156:381–384, 1991.

[113] Prajwal Rao, Nishal Ancelette Fereira, and Raghuram Srinivasan. Convolutional neural networks for lung cancer screening in computed tomography (ct) scans. *Proceedings of the 2016 2nd International Conference on Contemporary Computing and Informatics, IC3I 2016*, pages 489–493, 2016.

[114] Khalid Raza and Nripendra Kumar Singh. A tour of unsupervised deep learning for medical image analysis. *Current medical imaging*, 17:1059–1077, 1 2021.

[115] Inara Carneiro C. Rege, Thiago O. Sousa, Cláudio R. Leles, and Elismauro F. Mendonça. Occurrence of maxillary sinus abnormalities detected by cone beam ct in asymptomatic patients. *BMC Oral Health*, 12:30, 8 2012.

[116] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9351:234–241, 2015.

[117] Christopher R. Roxbury, Mary Qiu, Josef Shargorodsky, and Sandra Y. Lin. Association between allergic rhinitis and poor sleep parameters in u.s. adults. *International forum of allergy rhinology*, 8:1098–1106, 10 2018.

[118] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C.

Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 12 2015.

[119] Madeline C. Schiappa, Yogesh S. Rawat, and Mubarak Shah. Self-supervised learning for videos: A survey. *ACM Computing Surveys*, 55, 7 2023.

[120] J. Urgen Schmidhuber. Neural sequence chunkers. 1991.

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

[122] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-October:618–626, 12 2017.

[123] Hongdeok Seok, Jin Ha Yoon, Jong Uk Won, Wanhyung Lee, June Hee Lee, Pil Kyun Jung, and Jaehoon Roh. Concealing emotions at work is associated with allergic rhinitis in korea. *The Tohoku journal of experimental medicine*, 238:25–32, 12 2016.

[124] Shoaleh Shahidi, Barbad Zamiri, Shahla Momeni Danaei, Setareh Salehi, Shahram Hamedani, and Hamedani Sh. Evaluation of anatomic variations in maxillary sinus with the aid of cone beam computed tomography (cbct) in a population in south of iran. *Journal of Dentistry*, 17:7, 3 2016.

[125] Yuhui Song, Xiuquan Du, Yanping Zhang, and Chenchu Xu. Multi-shot prototype contrastive learning and semantic reasoning for medical image segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 14223 LNCS:578–588, 2023.

[126] Ethan Soudry and Peter H. Hwang. Acute frontal sinusitis. *The Frontal Sinus*, page 63, 1 2016.

[127] Markus Stenner and Claudia Rudack. Diseases of the nose and paranasal sinuses in child. *GMS Current Topics in Otorhinolaryngology, Head and Neck Surgery*, 13:Doc10, 2014.

[128] Joanna C. Stephens and Hesham A. Saleh. Evaluation and treatment of isolated maxillary sinus disease. *Current opinion in otolaryngology head and neck surgery*, 21:50–57, 2 2013.

[129] Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert. 3d self-supervised methods for medical imaging.

[130] Xian Tao, Xinyi Gong, Xin Zhang, Shaohua Yan, and Chandranath Adak. Deep learning for unsupervised anomaly localization in industrial images: A survey. *IEEE Transactions on Instrumentation and Measurement*, 71, 2022.

[131] B Tarp, B Fiirgaard, T Christensen, J J Jensen, and F T Black. The prevalence and significance of incidental paranasal sinus abnormalities on MRI. *Rhinology*, 38(1):33–38, March 2000.

[132] Keyu Tian, Yi Jiang, qishuai diao, Chen Lin, Liwei Wang, and Zehuan Yuan. Designing BERT for convolutional networks: Sparse and hierarchical masked modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

[133] Nenad Tomasev, Ioana Bica, Brian McWilliams, Lars Holger Buesing, Razvan Pascanu, Charles Blundell, and Jovana Mitrovic. Pushing the limits of self-supervised resnets: Can we outperform supervised learning without labels on imagenet?, 7 2022.

[134] Nguyen P Tran, John Vickery, and Michael S Blaiss. Management of rhinitis: allergic and non-allergic. *Allergy Asthma Immunol Res*, 3(3):148–156, May 2011.

[135] K. Trikojat, A. Buske-Kirschbaum, F. Plessow, J. Schmitt, and R. Fischer. Memory and multitasking performance during acute allergic inflammation in seasonal allergic

rhinitis. *Clinical and experimental allergy : journal of the British Society for Allergy and Clinical Immunology*, 47:479–487, 4 2017.

[136] F. Triulzi and S. Zirpoli. Imaging techniques in the diagnosis and management of rhinosinusitis in children. *Pediatric allergy and immunology : official publication of the European Society of Pediatric Allergy and Immunology*, 18 Suppl 18:46–49, 11 2007.

[137] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv.org*, 2018.

[138] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. *Proceedings of the 25th International Conference on Machine Learning*, pages 1096–1103, 2008.

[139] Jing Wang and Xiuping Liu. Medical image recognition and segmentation of pathological slices of gastric cancer based on deeplab v3+ neural network. *Computer Methods and Programs in Biomedicine*, 207:106210, 8 2021.

[140] Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Semantic image synthesis via diffusion models. 6 2022.

[141] Benzheng Wei, Zhongyi Han, Xueying He, and Yilong Yin. Deep learning model based breast cancer histopathological image classification. *2017 2nd IEEE International Conference on Cloud Computing and Big Data Analysis, ICCCBDA 2017*, pages 348–353, 6 2017.

[142] Kristoffer Wickstrøm, Michael Kampffmeyer, Karl Øyvind Mikalsen, and Robert Jenssen. Mixing up contrastive learning: Self-supervised representation learning for time series. *Pattern Recognition Letters*, 155:54–61, 3 2022.

[143] Jhih Ciang Wu, Ding Jie Chen, and Chiou Shann Fuh. Contrastive feature decoupling for weakly-supervised disease detection. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 14224 LNCS:252–261, 2023.

[144] You Xie and Nils Thuerey. Reviving autoencoder pretraining. *Neural Computing and Applications*, 35:4587–4619, 2 2023.

[145] Chen Zhao, Renjun Shuai, Li Ma, Wenjia Liu, and Menglin Wu. Improving cervical cancer classification with imbalanced datasets combining taming transformers with t2t-vit. *Multimedia Tools and Applications*, 81:24265, 7 2022.

[146] Asma Zizaan and Ali Idri. Evaluating and comparing bagging and boosting of hybrid learning for breast cancer screening. *Scientific African*, 23:e01989, 3 2024.

[147] Óscar Gómez, Pablo Mesejo, Óscar Ibáñez, and Óscar Cordón. Deep architectures for the segmentation of frontal sinuses in x-ray images: Towards an automatic forensic identification system in comparative radiography. *Neurocomputing*, 456:575–585, 10 2021.