AnatoFormer: Enhancing Breast Cancer Diagnosis Using Transformer-Based Spatial Analysis and Latent Representation Learning

A PROJECT REPORT

Submitted by

KAWENA M (202109029)

PRAHANYA SELVAKUMAR (202109039)

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

APRIL 2025

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

MEPCO SCHLENK ENGINEERING COLLEGE, SIVAKASI

(An Autonomous Institution affiliated to Anna University Chennai)

APRIL 2025

BONAFIDE CERTIFICATE

Certified that this project report **"AnatoFormer: Enhancing Breast Cancer Diagnosis Using Transformer-Based Spatial Analysis and Latent Representation Learning"** is the bonafide work of **"KAWENA M (Reg. No.: 202109029) , PRAHANYA SELVAKUMAR (Reg. No.: 202109039)"** who carried out the project work under my supervision.

SIGNATURE

Mrs.D.MONICA SELES
Asst. Professor & Supervisor
AI&DS Department
Mepco Schlenk Engg. College,Sivakasi
Virudhunagar Dt. – 626 005

SIGNATURE

Dr. J. ANGELA JENNIFA SUJANA
Asso.Professor(SeniorGrade)&Head
AI&DS Department
Mepco Schlenk Engg. College,Sivakasi
Virudhunagar Dt. – 626 005

Submitted to the Viva-Voce examination held on __/__/___.

INTERNAL EXAMINER

EXTERNAL EXAMINER

ACKNOWLEDGEMENT

First and foremost, we would like to thank the **LORD ALMIGHTY** for his abundant blessings that is showered upon our past, present and future successful endeavors.

We extend our sincere gratitude to our college management and principal **Dr. S. Arivazhagan M.E., Ph.D.,** for providing sufficient working environment such as systems and library facilities.

We would like to extend our heartfelt gratitude to our Head of the Artificial Intelligence and Data Science Department **Dr. J. Angela Jennifa Sujana M. Tech., Ph.D.,** for giving us the golden opportunity to undertake the project of this nature and for her most valuable guidance.

We would also like to extend our gratitude to **Dr.S.Shiny M.E.,Ph.D., Assistant professor(Sr.G),** Department of Artificial Intelligence and Data Science, for being our project coordinator and directing us throughout our project.

We would also like to extend our gratitude and sincere thanks to **Mrs. D. Monica Seles M.Tech., Assistant professor,** Department of Artificial Intelligence and Data Science for being our project guide and for her moral support and suggestions. She has put her valuable experience and expertise in directing, suggesting and supporting us throughout the project to bring our best.

Our sincere thanks to our revered faculty members, lab technicians and beloved family and our friends for their help at right time for making this project a successful one.

ABSTRACT

There are serious risks to public health from the rapid spread of misleading information brought on by our increasing reliance on digital healthcare information. Recognizing such misinformation is essential in the medical industry to ensure that accurate and trustworthy information is disseminated. We suggest an ideal deep learning-based model combining AnatoFormer and LREN-MedNet for ROI-Free Breast Cancer Diagnosis using self-supervised learning and anatomical-aware feature extraction. Our suggested AnatoFormer-LREN system uses Masked Autoencoder (MAE) pretraining and Contrastive Learning (MoCo) to enhance feature extraction from unlabeled ultrasonic images. The AnatoFormer design improves diagnostic accuracy and model interpretability by accurately capturing intra-layer and inter-layer spatial interactions in breast ultrasound images. Because the resulting representations are employed in a fully automated pathway, eliminating the need for manual ROI annotation, the model is more practical for clinical applications. Additionally, optimization techniques like self-supervised training and fine-tuning methodologies have been employed to increase the resilience of our model. The models performance was assessed on the BUSI dataset and it performs better in classification than transformer-based and traditional CNNs. The suggested AnatoFormer-LREN structure offers a new understandable and effective alternative to automated breast cancer diagnosis advancing artificial intelligence in clinical decision-making and medical imaging.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1. BACKGROUND AND CONTEXT:

Artificial intelligence (AI) and deep learning (DL) have revolutionized medical imaging by enabling automated feature extraction and highly accurate disease classification. However, the majority of deep learning models for diagnosing breast ultrasonography still rely on manually segmented ROIs, which restricts their use in real-world clinical settings.Since breast cancer is one of the most common and deadly illnesses in the world early and precise detection is essential to improving patient outcomes. The non-invasiveness radiation-free nature and affordability of breast ultrasound (BUS) make it one of the most popular imaging modalities. Regretfully conventional CAD programs frequently depend on pre-established Regions of Interest (ROIs) necessitating radiologists manual annotations. Clinical efficiency is impacted automation is constrained and radiologists variability is increased by this manual participation.

Manual Region of Interest (ROI) annotations which need expert input to differentiate tumors in ultrasound images have historically been the mainstay of breast cancer detection. ROI-based approaches however are time-consuming arbitrary and might not be transferable across datasets. Thanks to developments in deep learning and self-supervised learning automated ROI-free breast cancer diagnosis is now a dependable and effective substitute. ROI-Free models evaluate breast ultrasound (BUS) images without the need for human ROI selection by utilizing transformer-based topologies and self-supervised learning. These models enable fully automated diagnosis by directly extracting tissue shapes spatial relationships and contextual information from entire ultrasound images. Due to its lack of dependence on pre-established lesion labels this method is more scalable interpretable and therapeutically advantageous.

AnatoFormer (Anatomy Aware Transformer):

Using ROI-Free a transformer-based deep learning model known as AnatoFormer was developed to identify breast cancer in ultrasound pictures. AnatoFormer models the vertical (inter-layer) and horizontal (intra-layer) spatial correlations within breast tissue structures to capture anatomical prior knowledge which sets it apart from traditional CNN-based models. Its use of multi-head self-attention to enhance feature extraction and diagnostic interpretability makes it a very successful classification system for breast ultrasounds.

Latent Representation Extraction Network (LREN):

To enhance feature representation the LRE-Net self-supervised learning framework combines Masked Autoencoder (MAE) pretraining with Contrastive Learning (MoCo). While MAE reconstructs missing image patches to capture context-aware representations MoCo compares similar and dissimilar samples to improve feature discrimination. This hybrid approach enables robust pretraining on unlabeled ultrasound images and improves the accuracy of subsequent classification tests.

Vision Transformer (ViT):

The deep learning architecture Vision Transformer (ViT) uses the self-attention mechanism of transformers to address image processing problems. Unlike Convolutional Neural Networks (CNNs) which depend on local receptive fields ViT divides images into patches and treats each patch as a token much like words in NLP models. ViT is therefore perfect for applications like object detection image classification and medical imaging since it can capture long-range dependencies and global contextual relationships.

1.2. OBJECTIVES:

The objective of the AnatoFormer-LREN model is to use ROI-Free ultrasound data to create a reliable and accurate deep learning model for the diagnosis of breast cancer.To improve tumor classification accuracy without requiring human Region of Interest (ROI) annotations the model makes use of the advantages of LRE-Net (Self-Supervised Learning) and AnatoFormer (Anatomy-Aware Transformer).

The model is meant to receive breast ultrasound (BUS) images which include intricate anatomical features perplexing lesion appearance and domain-specific issues.The two key elements of the proposed model architecture are LRE-Net (Self-Supervised Learning) and AnatoFormer (Anatomy-Aware Transformer).

Using the Masked Autoencoder (MAE) and Contrastive Learning (MoCo) the LRE-Net module learns feature representations from unlabeled ultrasound images while maintaining informative anatomical patterns. The acquired features are transmitted to the AnatoFormer module which uses both horizontal and vertical embeddings to assess them and determine the spatial relationships between and within layers of breast tissue.

The proposed method uses optimization techniques to ensure optimal performance and efficiency. These include weight initialization schemes regularization strategies learning rate schedules and hyperparameter optimization to enhance the generalization and stability of breast ultrasound data.

Precise and automated detection of breast cancer is essential for early detection treatment planning and better patient outcomes. Interpretable scalable and clinically feasible the model enhances medical imaging decision-making and breast cancer screening through transformer-based feature extraction and self-supervised learning.

1.3. PROBLEM STATEMENT:

With increased reliance on digital health platforms and radiology imaging, the importance of accurate, computerized diagnosis of breast cancer has arisen more urgently than at any point in the past. Traditional Region of Interest (ROI)-based detection of breast cancer relies on hand-segmented tumor segmentation, which is time-consuming, subjective, and prone to variability. These limitations hinder the scaling and clinical deployment of computer-aided diagnosis (CAD) schemes for breast ultrasound (BUS) imaging.

Typically supervised learning from annotated data is used in traditional deep learning techniques for breast cancer diagnosis which requires a great deal of expert annotation. Furthermore the incapacity of CNN-based models to understand the complex spatial correlations present in ultrasound images limits their ability to differentiate between benign and malignant tumors. Consequently an automated ROI-free approach is required that could improve classification accuracy and create rich representations without the need for manual segmentation.

By combining AnatoFormer (Anatomy-Aware Transformer) with LRE-Net (Self-Supervised Learning), the AnatoFormer-LREN model tackles this problem. The model learns feature representations from unlabeled ultrasound pictures using Masked Autoencoders (MAE) and Contrastive Learning (MoCo). It then refines the model using identified data. By depicting intra-layer (horizontal) and inter-layer (vertical) spatial interactions, AnatoFormer.
improves diagnostic accuracy and makes it easier to diagnose breast cancer in a way that is more clinically meaningful and interpretable.

1.4 SOCIAL IMPACT:

Social impact is the term used to describe the effects both positive and negative that every activity has on people and communities. The way that natural resources affect peoples livelihoods is also hinted at. The healthcare sector may be positively or negatively impacted by cancer detection which can be explained as follows:

Public Awareness:

1. **Positive Impact:** Increases understanding of the significance of early detection of breast cancer which promotes routine screenings and healthcare decisions that avoid the disease.
2. **Negative Impact:** If an over-reliance on AI-based diagnosis is not adequately validated in addition to professional judgments radiologists credibility may suffer.

Trust in Medical AI:

1. **Positive Impact:** Provides an automated ROI-free approach to breast cancer diagnosis reducing human bias and boosting trust in AI-powered diagnostic tools.
2. **Negative Impact:** Patients and doctors might be reluctant to use AI-assisted diagnostics if the decisions they make are ambiguous or hard to understand.

Healthcare Accessability:

1. **Positive Impact:** Accelerates and simplifies cancer screening especially in places with a lack of resources and qualified radiologists.
2. **Negative Impact:** The high upfront costs of model training and implementation would prevent healthcare facilities with limited funding from fully utilizing AI.

Social Well-being and Patient Outcomes:

1. **Positive Impact:** Self-supervised learning and automated diagnostics can facilitate early detection which may improve treatment results and increase survival.
2. **Negative Impact:** Due to data bias or model limitations misdiagnosis causes patients needless stress and delays in treatment.

CHAPTER 2

LITERATURE SURVEY

2.1. DSMT-Net - Dual Self-Supervised Multi-Operator Transformation for Multi-Source Endoscopic Ultrasound Diagnosis Detection. [1]

The paper proposes a model named DSMT-Net a sophisticated deep learning network intended for multi-source endoscopic ultrasound (EUS) diagnostics mainly for the detection of pancreatic cancer. Device-based variability is decreased by the model by standardizing the region of interest (ROI) from EUS images using Multi-Operator Transformation (MOT). Additionally, Momentum Contrast (MoCo) and Masked Autoencoder (MAE) are used in a Dual Self-Supervised Network (DSN) to help extract local and global features from unlabeled data Several self-supervised models, including CNN, ViT, ResNet, MAE, and MoCo, were compared to the proposed model. The LEPset and BUSI datasets were used to validate the models efficacy and confirm its superiority over other approaches. In order to detect pancreatic cancer the study recommends DSMT-Net a sophisticated deep learning network made for multi-source endoscopic ultrasound (EUS) diagnostics. By employing Multi-Operator Transformation (MOT) to standardize the region of interest (ROI) from EUS images, the model reduces device-based variability. Additionally, Momentum Contrast (MoCo) and Masked Autoencoder (MAE) are used in a Dual Self-Supervised Network (DSN) to help extract local and global features from unlabeled dataThe results showed that DSMT-Net was more successful in detecting pancreatic and breast cancer.

2.2. HoVer-Trans - Anatomy-Aware HoVer-Transformer for ROI-Free Breast Cancer Diagnosis in Ultrasound Images. [2]

According to the paper the HoVer-Trans model is a complex deep learning algorithm designed for ROI-free breast cancer detection in ultrasound images. Using anatomical prior knowledge the model examines the spatial interactions between different layers of breast tissue to differentiate between benign and malignant malignancies.The study recommends the HoVer-Trans model a sophisticated deep learning algorithm made for ultrasound image ROI-free breast cancer detection.In order to differentiate between benign and malignant malignancies the model

uses anatomical prior information to analyze the spatial interactions between different layers of breast tissue. It presents the HoVer-Trans block which uses vertical and horizontal embeddings to extract spatial information both within and between layers. The suggested model was contrasted with models based on CNN and Transformer such as Swin-B VGG16 and ResNet50. HoVer-Trans with an AUC of 0. 924 performed better than these models and offered better interpretability The proposed model was validated on three datasets (UDIAT, BUSI, GDPH & SYUCC) and showed state-of-the-art classification performance. Because of its ROI-free technique, which allows it to operate without known lesion locations, the model is more practical for clinical use.

2.3. Electromechanical Coupling Factor of Breast Tissue as a Biomarker for Breast Cancer.[3]

This research paper proposes the Electromechanical Coupling Factor (ECF) of breast tissue as a biomarker for the diagnosis of breast cancer. The study investigates the effects of differences in collagen density between healthy and malignant breast tissues on the electromechanical response. A piezoresistive sensor layer on a MEMS-based biochip was used to evaluate the mechanical reaction of tissue samples. When the proposed method was tested on both normal and invasive ductal carcinoma (IDC) samples, there was a statistically significant difference in the ECF values ($p < 0.0039$). Because the ECF measurement has a higher sensitivity than conventional mechanical stiffness testing the results indicated that it might be a valuable biomarker for the early diagnosis of breast cancer. The study shows how ECF-based detection can improve early breast cancer diagnosis and validates its efficacy which permits it to function without known lesion locations.

2.4. Spatiotemporal Mammography-based Deep Learning Model for Improved Breast Cancer Risk Prediction. [4]

The Spatiotemporal Deep Learning methodology is a technique proposed in this paper to enhance the prediction of breast cancer risk by analyzing temporal variations in mammography images. The model takes two different times (T1 and T2) of mammograms and uses a Siamese neural network to extract temporal and spatial features. CNN models trained on single-time-point images were contrasted with the suggested model. The Siamese model outperformed CNN

models according to the results (AUC: 0. 75 0. 77) with a higher AUC of 0. 81. By including temporal information cancer risk can be better predicted and possible malignancy can be detected early. In order to improve the estimation of breast cancer risk the study highlights the use of longitudinal mammography data. The model also demonstrated improved prediction stability by reducing false-positive and false-negative rates. The next study will attempt to expand the dataset and look into multimodal imaging in order to enhance risk prediction.

2.5. Using Vision Transformers in 3D Medical Image Classifications. [5]

In order to improve upon conventional Convolutional Neural Networks (CNNs) the article suggests a model for 3D medical picture classification that makes use of Vision Transformers (ViTs). The model analyzes ViTs capacity to identify long-range spatial dependencies in 3D medical images a common CNN shortcoming.We compared the suggested model with CNN models that were built from scratch and CNN models that were pre-trained using ImageNet. The findings demonstrated that ViTs with ImageNet pre-training outperformed CNNs in terms of classification accuracy indicating that they may be useful for 3D medical imaging.In order to enhance model performance the study also employed self-supervised learning and sharpness-aware minimizer (SAM) minimization.The research indicates that Vision Transformers can outperform CNNs in 3D medical picture classification tasks if they are appropriately pre-trained. For more accurate predictions future research will investigate multimodal learning and larger datasets.

2.6. Big Self-Supervised Models Advance Medical Image Classification.[6]

The paper offers a method for enhancing medical image classification that utilizes self-supervised learning (SSL) approaches. The authors present a novel Multi-Instance Contrastive Learning (MICLe) technique that uses multiple images of the same pathology per patient to generate more relevant positive pairs for SSL. Two different tasks were used to assess the suggested method: multi-label chest X-ray classification and dermatology skin condition classification using digital camera photos. The results showed that self-supervised pretraining on ImageNet and subsequent self-supervised learning on unlabeled domain-specific medical images significantly increased classifier accuracy. Specifically it improved top-1 accuracy for

dermatology classification by 6. 7 percent and mean AUC for chest X-ray classification by 1. 1 percent outperforming strong supervised baselines pretrained on ImageNet. Additionally the study found that large self-supervised models are robust to distribution shifts and can train efficiently with a limited number of labeled medical images. Through effective use of both labeled and unlabeled data this study shows how SSL and specifically the MICLe approach can increase the categorization of medical images.

2.7. Deep Feature Representations for Variable-Sized Regions of Interest in Breast Histopathology. [7]

In this paper, a deep convolutional neural network (CNN) framework for modeling variable-sized areas of interest (ROIs) in breast histopathology pictures is proposed. It is difficult to assess ROIs of different sizes and shapes in whole slide images since traditional CNNs require inputs of a fixed size. To address this the authors developed a method that integrates features from multiple fixed-sized patches extracted from nuclei-dense regions within each ROI. The fully connected and convolutional layers of a deep network are used to extract patch-level features which are then weighted based on class predictions and averaged to produce an exhaustive ROI-level representation. The proposed framework was tested on a dataset of 240 slides with 437 ROIs each of which was identified by qualified pathologists and varied in size and shape. In four diagnostic categories the model was able to classify ROIs across the entire histologic spectrum with an accuracy of 72 points and 65 percent. This performance surpasses the mean accuracy of a different group of pathologists as well as existing methods. As a possible method for examining variable-sized ROIs in digital pathology the study shows how effectively the proposed feature representation captures diagnostic relevance as well as local structural information.

2.8. Regions of Interest Extraction for Hyperspectral Small Targets Based on Self-Supervised Learning. [8]

The study presents a paradigm for Regions of Interest (ROI) extraction in hyperspectral images using self-supervised learning (SSL) with a focus on small objects. For conventional methods high-dimensional hyperspectral data and small target detection pose challenges. The proposed model is able to extract ROIs efficiently by using SSL techniques to extract reliable

spectral-spatial features from unlabeled data. In contrast to more traditional supervised methods the model was able to identify small targets more accurately and with fewer false positives. The proposed methodology significantly improves the ability to detect small objects in hyperspectral photography without requiring large labeled datasets. Additionally the model demonstrated reduced processing costs and enhanced generalization ability. Subsequent research will concentrate on expanding the methodology to include real-time processing and data from additional remote sensing sources.

2.9. Self-Supervised Learning Method for SAR Multiinterference Suppression. [9]

In order to lessen the effects of electromagnetic interferences without being aware of their features beforehand, the study suggests a Self-Supervised Learning (SSL) technique for Synthetic Aperture Radar (SAR) multi-interference suppression. The suggested approach analyzes reconstruction mistakes in the time-frequency domain to identify and localize interferences using a Convolutional Autoencoder (CAE) named LocNet. In order to minimize signal loss, a U-Net-based model (RecNet) is also used to reconstruct radar signals following interference suppression. When compared to conventional parametric and nonparametric techniques, the suggested model performed better at identifying and reducing complicated interferences. The outcomes showed that in both simulated and actual SAR data, the SSL approach greatly enhanced signal recovery and interference suppression. Furthermore, the method demonstrated cheap computational cost and great generalization ability, making appropriate for applications involving real-time SAR imaging. Future research will concentrate on improving real-time processing efficiency and expanding the approach to multi-source SAR data.

2.10. IPCL: Iterative Pseudo-Supervised Contrastive Learning to Improve Self Supervised Feature Representation. [10]

In order to enhance self-supervised feature representation the paper suggests an IPCL (Iterative Pseudo-Supervised Contrastive Learning) model that integrates pseudo-supervised signals into the contrastive learning framework. In order to help the network learn more discriminative features the model assigns and improves class labels for unlabeled data using an

iterative pseudo-labeling process. The suggested model performed better on a variety of datasets than baseline self-supervised learning techniques. The findings showed that for contrastive learning tasks IPCL obtained 85. 55 percent accuracy on the STL-10 and 84. 77 percent accuracy on the CIFAR-10. Furthermore the model demonstrated a notable enhancement in unsupervised image classification achieving an accuracy of 80–91 percent on STL-10 and 88–81 percent on CIFAR-10. Pseudo-supervised learning and contrastive learning together improved feature representations and improved the models capacity to detect intra-class differences. Subsequent research will concentrate on expanding the methodology to multimodal learning and large datasets.

2.11. Masked Motion Encoding for Self-Supervised Video Representation Learning. [11]

The paper proposes Masked Motion Encoding (MME), a method that captures both motion and appearance information from movies to train self-supervised video representations. The model uses a masked video modeling technique, which reconstructs masked regions in movies and predicts motion trajectories, to learn fine-grained temporal dynamics.Even the most advanced self-supervised learning techniques were surpassed by the suggested model. With 78point 0 percent top-1 accuracy on HMDB51 and 96point 5 percent top-1 accuracy on UCF101 the results showed that MME outperformed current techniques. After pre-training the models accuracy on the Kinetics-400 dataset was also 81. 8%. Restoring motion trajectories greatly enhanced the models capacity to record temporal data. Subsequent studies aim to extend the approach to multi-view video data and real-time action detection.

2.12. Defensive Patches for Robust Recognition in the Physical World. [12]

In order to strengthen deep learning models resistance to adversarial attacks and outside noise the paper suggests a method known as Defensive Patches. The technique improves target object detection in difficult circumstances by adding class-specific recognizable patterns to localized patches on the object. When compared to current adversarial defensive techniques the suggested model performed better. The findings showed that defensive patches increased accuracy against adversarial perturbations and corruption-based noise by more than 20%. Furthermore the techniques transferability across various models was improved by utilizing

global feature correlation. The method is viable for practical implementation because data-end safeguards guarantee that it can be applied without changing model designs. Expanding the methodology to dynamic real-world scenarios and improving patch adaptability for invisible perturbations will be the main goals of future research.

2.13. Demae: Diffusion-Enhanced Masked Autoencoder for Hyperspectral Image Classification With Few Labeled Samples. [13]

The paper proposes a model named Diffusion-Enhanced Masked Autoencoder (DEMAE) to improve hyperspectral image (HSI) classification when there are few labeled samples available. The model uses diffusion-based representation learning in conjunction with a masked autoencoder (MAE) architecture to enhance feature extraction from HSI data. The asymmetric encoder–decoder structure was constructed using both conditional and conventional Transformer blocks. An auxiliary task focusing on simultaneous denoising and reconstruction is introduced to aid heuristic feature learning. The encoder is pre-trained in a self-supervised manner and then fine-tuned using a limited number of labeled samples. Additionally, a unique signal-to-noise ratio enhanced (SNR-improved) loss function is used to regularize the training process. Four benchmark datasets were used to assess the DEMAE model, which showed better results in terms of classification accuracy and mapping capability than existing state-of-the-art methods under few labeled sample conditions.

2.14. AdamW+: Machine Learning Framework to Detect Domain Generation Algorithms for Malware. [14]

In order to improve the detection of Domain Generation Algorithms (DGAs) which malware uses to circumvent security measures the article suggests a model called AdamW+. In terms of detection accuracy and generalization the model surpasses the current Adam and AdamW optimizers by improving gradient descent techniques. For DGA identification, the proposed model outperformed traditional machine learning frameworks. The results showed that AdamW+ achieved a higher detection accuracy by effectively separating DGA-generated domains from genuine ones. Additionally, the model improved its ability to generalize, making it more resilient to evolving infection strategies. Weight decay optimization ensures better stability

and convergence when used in deep learning model training for cybersecurity applications. Future research will concentrate on implementing the framework in real-time and expanding it to identify other cyber threats.

2.15. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine-Tuning? [15]

The paper examines the effectiveness of fine-tuning pre-trained Convolutional Neural Networks (CNNs) for medical image processing as opposed to building CNNs from scratch. The study evaluates CNN performance in four medical imaging applications: gastrointestinal, cardiology, and radiology for classification, detection, and segmentation tasks. The results demonstrated that fine-tuned CNNs performed better than fully trained CNNs when compared to the proposed technique, especially when there was little labeled data. The study also found that the workload affected the amount of fine-tuning needed and that layer-wise fine-tuning provided an effective way to optimize performance. Additionally, fine-tuning makes pre-trained CNNs more resilient to different dataset sizes, making them a more attractive choice for medical imaging tasks. Future research will concentrate on improving transfer learning techniques and exploring CNN architectures tailored for medical image analysis.

2.16. Understanding How Pretraining Regularizes Deep Learning Algorithms. [16]

This paper investigates the effects of pretraining on model stability and generalization during the regularization phase of deep learning algorithms. In the framework of Tikhonov-regularized batch learning the study examines unsupervised pretraining to show how it helps neural networks acquire meaningful representations. The findings demonstrated that by learning superior Tikhonov matrices compared to non-pretrained models pretraining improves generalization performance. According to the study unsupervised pretraining also increases model stability by reducing overfitting and speeding up convergence. In deep learning optimization pretraining as a regularizer has been shown to have advantages. Subsequent studies will focus on extending the system to incorporate different pretraining strategies and investigating its impact on large datasets.

2.17. Convolutional Recurrent Neural Networks for Text Classification. [17]

In order to improve text classification the study proposes a model called the Convolutional Recurrent Neural Network (CRNN) which combines Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs). A CNN is used to extract key features for classification and a bi-directional RNN is used to capture contextual interactions.Compared to conventional text classification techniques the suggested model outperformed them. The findings showed that CRNN enhances classification accuracy across a variety of datasets by efficiently capturing both local and sequential connections. Additionally the method improves word representation learning which increases its capacity to manage texts long-range dependencies. Better feature extraction and contextual comprehension are ensured by its CNN and RNN fusion which makes it useful for a variety of natural language processing tasks. The models expansion to multilingual text classification and real-time application optimization will be the main goals of future research.

2.18. SPT-Swin: A Shifted Patch Tokenization Swin Transformer for Image Classification.[18]

The paper proposes a model called SPT-Swin, which combines Shifted Patch Tokenization (SPT) and Swin Transformer architecture to enhance picture categorization. The method improves feature extraction and data efficiency by applying spatial shifts to image patches before tokenization. The proposed model outperformed the state-of-the-art photo categorization models. The results showed that SPT-Swin performed better than existing methods, with an accuracy of 89.45% on ImageNet-1K, 95.67% on CIFAR-10, and 92.95% on CIFAR-100. Shifting patch tokenization also ensures improved generalization and computational efficiency in vision tasks. The method may improve classification accuracy while preserving linear computing complexity, which makes it ideal for large-scale image datasets. The adaptation of SPT-Swin for object detection and multimodal learning applications will be the main focus of future research.

2.19. A Deep Learning Based Spatial Dependency Modelling Approach Towards Super-Resolution. [19]

Using a deep learning-based spatial dependency modeling technique the paper suggests a super-resolution model that improves image resolution without the need for labeled data. The model rebuilds high-resolution images from low-resolution inputs by learning spatial correlations using deep autoencoders and convolutional neural networks (CNNs). The suggested model performed better than the current unsupervised super-resolution techniques. The outcomes showed that particularly in complex visual situations the strategy performed better than variogram-based solutions. Furthermore the model guarantees improved image quality and detail retention due to its capacity to capture fine-grained spatial structures. The method works well with sparsely labeled data in real-world applications because it blends deep feature extraction with unsupervised learning. Subsequent studies will focus on extending and modifying the model for supervised settings.

2.20. Object Detection Using Deep Learning, CNNs, and Vision Transformers. [20]

This paper provides an in-depth analysis of object detection methods that make use of deep learning convolutional neural networks (CNNs) and vision transformers (ViTs). The three object identification strategies—transformer-based anchor-based and anchor-free—as well as their benefits drawbacks and performance metrics are examined in this study. The shift from conventional CNN-based detectors to transformer-based structures was observed by comparing the suggested approaches with cutting-edge object detection frameworks. Comparing transformer-based object detectors to conventional techniques the results showed a significant improvement in contextual understanding and detection accuracy. For computer vision researchers the study is a useful resource since it also addresses the speed-accuracy trade-offs in contemporary detection frames. In challenging situations improved performance is ensured by combining Vision Transformers with global context modeling. Future research will concentrate on strengthening resistance to occlusions and hostile attacks as well as improving real-time object identification.

2.21. Exploring Self-Supervised Representation Learning for Low-Resource Medical Image Analysis. [21]

In order to tackle the problem of limited labeled data the study investigates the efficacy of self-supervised learning (SSL) methods for medical image analysis in low-resource scenarios. Using three small-scale medical imaging datasets—BreaKHis400X Colorectal and PneumoniaCXR—the study assesses four cutting-edge SSL techniques: SimCLR DCLW SimSiam and VICReg. When compared to transfer learning from large-scale datasets (like ImageNet) the suggested approach performed competitively. In-domain SSL pretraining on small medical datasets can reduce reliance on large labeled datasets even though it achieves accuracy comparable to or superior to transfer learning according to the results. Additionally the study illustrates how SSL-based pretraining efficiently captures significant representations which makes it ideal for applications involving low data in medical imaging. Enhancing SSL performance for multimodal medical data and adapting these techniques for practical clinical situations are the objectives of future research.

2.22. A VGG Attention Vision Transformer Network for Benign and Malignant Classification of Breast Ultrasound Images. [22]

The paper proposes a model, the VGG Attention Vision Transformer (VGGA-ViT) Network, to differentiate benign and malignant breast cancers from breast ultrasound (BUS) images. The model uses a VGG backbone for local feature extraction and a Vision Transformer (ViT) module to record global contextual interactions. The suggested model outperformed current deep learning techniques. The results showed that VGGA-ViT outperformed previous models in BUS image classification achieving 88. 71 percent accuracy on Dataset A and 81. 72 percent accuracy on Dataset B. Additionally the models capacity to concentrate on significant image regions was improved by the addition of squeeze-and-excitation blocks which raised the diagnostic accuracy. The model can improve clinical decision-making as evidenced by its capacity to distinguish between benign and malignant tumors. For practical medical applications future research will concentrate on growing the dataset and enhancing model generalization.

2.23. On The Impact of Self-Supervised Learning in Skin Cancer Diagnosis. [23]

In order to improve dermoscopy image classification using unlabeled medical data and identify skin cancer the study explores the use of self-supervised learning (SSL). The study intends to ascertain how well two SSL techniques—rotation and SimCLR—pretrain deep neural networks (DNNs) for the purpose of classifying skin lesions.On the other hand models that were trained from scratch performed worse than the suggested approaches. The findings demonstrated that SimCLR pretraining and rotation enhanced classification accuracy and captured complementary data. As per the study employing diverse SSL techniques could potentially improve the precision of diagnosis. Better feature extraction and generalization are made possible by SSL techniques which are highly advantageous for medical imaging applications with sparse labeled data. Future research will concentrate on looking at other SSL methods and modifying them for the diagnosis of multimodal skin cancer.

2.24. A Deep Fusion-Based Vision Transformer for Breast Cancer Classification. [24]

To improve breast cancer classification using histopathology images the study suggests a Deep Fusion-based Vision Transformer (DFViT) model. To increase diagnosis accuracy the model makes use of Vision Transformers (ViTs) for global contextual learning and Convolutional Neural Networks (CNNs) for local feature extraction. The state-of-the-art classification techniques were outperformed by the suggested model. The findings demonstrated that DFViT performed well on several datasets including the UCSC cancer genomics BreakHis and BACH. Better classification outcomes were also obtained by combining RGB and stain-normalized images which improved the models feature extraction capabilities. CNN-ViT fusion is useful for practical clinical applications because it guarantees strong feature representation. Future research will concentrate on enhancing model interpretability for clinical deployment and broadening the strategy for multi-class cancer classification.

2.25. Masked Transformer for Self-Supervised Learning in Medical Imaging. [25]

In order to improve medical image analysis the study suggests a methodology based on Masked Transformers with Self-Supervised Learning (SSL) which uses masked image reconstruction for feature extraction. The approach improves segmentation and classification performance by enabling representation learning from unlabeled medical images. The proposed model outperformed other self-supervised methods and CNN-based methods. The results showed how well Masked Transformers understand long-range relationships, improving medical image processing task accuracy. Additionally, the model's ability to learn from large unlabeled datasets makes it highly effective in real-world clinical applications. Future studies will focus on expanding SSL-based masked transformers to multi-modal medical imaging and increasing computer efficiency for broad use.

2.26. Self-Supervised Learning Guided Transformer for Survival Prediction of Lung Cancer Using Pathological Images. [26]

The paper proposes a Dual Transformer Encoder Model that combines two transformer encoders of different hidden sizes to enhance medical picture classification. The method gets beyond the limitations of traditional Vision Transformers (ViTs) by enabling multi-scale feature extraction. The proposed model outperformed prior transformer-based methods for classifying medical pictures. The results showed that the dual encoder architecture improved classification accuracy by capturing both local and global visual properties. Additionally, the implementation of Layer-wise Class Token Attention (LCA) ensured more effective feature aggregation and enhanced generalization. The model's ability to comprehend token sequences of various sizes makes it incredibly flexible for a variety of medical imaging tasks. Future work will focus on expanding the approach to multi-cancer survival prediction and incorporating additional clinical data for improved prognostic accuracy.

2.27. Self-Supervised Learning Based on StyleGAN for Medical Image Classification on Small Labeled Datasets. [27]

The paper proposes a model that blends StyleGAN with self-supervised learning (SSL) to improve medical picture classification when labeled data is limited. The model uses a StyleGAN generator to extract style-based semantic features, which have been pretrained on large amounts of unlabeled data, for better classification performance. The suggested model performed better than conventional deep learning techniques. Results indicated that adding StyleGAN-generated features increased classification accuracy especially in chest X-ray image datasets with limited labeled samples. The models self-attention mechanism also improved robustness and generalization which made feature fusion more effective. The combination of SSL and StyleGAN ensures efficient feature extraction making the method applicable to low-data medical imaging tasks.

2.28. Multi-Stage Aggregation Transformer for Medical Image Segmentation. [28]

To enhance medical image segmentation the study proposes a Multi-Stage Aggregation Transformer (MA-Transformer) model that effectively captures multi-scale semantic features. CNNs are able to extract local spatial characteristics thanks to a dual-branch encoder that combines CNNs and Transformers. Self-attention processes then provide global context.The best segmentation techniques were surpassed by the suggested model. In terms of segmentation accuracy across publicly available medical imaging datasets the results demonstrated that MA-Transformer performed better than CNN-based and Transformer-based approaches. The model works well for challenging medical picture segmentation tasks thanks to its multi-scale feature aggregation technique which also improves precision and robustness. The models adaptation for multimodal medical imaging and its efficiency enhancement for real-time clinical applications will be the main goals of future research.

2.29. A Multi-Task Self-Supervised Learning Framework for Scopy Images. [29]

The paper recommends ColorMe, a model that enhances medical image analysis for scopy images through the use of self-supervised learning (SSL). The model uses multi-task pretraining by reconstructing color channels and predicting color distributions to learn meaningful representations from unlabeled input. The proposed approach was more effective than traditional supervised learning methods. The results showed that ColorMe performed better than scratch-trained models in terms of cervix type categorization and skin lesion segmentation accuracy. The model's self-supervised pretraining technique also enables better feature extraction, making it applicable to medical imaging challenges with limited data. Future research will expand the framework to include multimodal medical images, and SSL algorithms will be adjusted for real-world clinical use.

2.30. Uncertainty-Aware Transformer Model for Anatomical Landmark Detection in Paraspinal Muscle MRIs. [30]

The paper suggests a model called Uncertainty-Aware Swin Transformer V2 (Swin-V2) that incorporates uncertainty quantification to enhance the recognition of anatomical landmarks in paraspinal muscle MRIs. Monte Carlo (MC) dropout-based uncertainty measures are used in the model to improve the accuracy and dependability of landmark detection.The proposed model performed better than existing deep learning methods. Higher-quality landmark grading was made possible by the results which showed a relationship between uncertainty-aware landmark detection and detection errors. Moreover the models random forest-based grading system provided interpretability and resilience making it a useful instrument for muscle morphometric analysis in low back pain (LBP) research. The systems extension to multi-organ anatomical landmark identification and the integration of domain adaptation strategies for wider clinical applications will be the main goals of future research.

2.31. A Transformer-Based Network for Deformable Medical Image Registration. [31]

In order to enhance deformable medical picture registration, the study suggests a model called Transformer-Based Registration Network, which uses self-attention techniques to collect both local and global spatial characteristics. To improve deformation field estimate and enable more precise alignment of medical pictures, the model uses a bi-level information flow. When compared to deep learning-based and conventional registration techniques, the suggested model performed better. According to the results, the Transformer-based method outperformed current techniques in terms of Dice similarity coefficient and obtained greater registration accuracy on brain MR image datasets (LPBA40, OASIS-1). Better feature extraction is also made possible by the model's self-attention mechanism, which makes it incredibly efficient for challenging medical image alignment tasks.

2.32. CoTrFuse: A Novel Framework by Fusing CNN and Transformer for Medical Image Segmentation. [32]

In order to improve medical image segmentation the research suggests a model called CoTrFuse which combines CNNs and Transformers to capture both local and global information. The model employs Swin Transformer as a Transformer encoder for global context modeling and EfficientNet as a CNN encoder for the extraction of fine-grained spatial features. The suggested model outperformed the most advanced segmentation techniques. CoTrFuse demonstrated superior segmentation accuracy and durability compared to previous methods on the ISIC-2017 and COVID-QU-Ex datasets according to the results. Due to its feature fusion approach which also guarantees improved representation learning the model excels at difficult medical image segmentation tasks. Future research will concentrate on enhancing efficiency for real-time clinical applications and modifying the framework for multimodal medical imaging.

2.33. SegTransVAE: Hybrid CNN–Transformer with Regularization for Medical Image Segmentation. [33]

SegTransVAE a model that improves medical image segmentation by combining CNNs Transformers and Variational Autoencoders (VAEs) is recommended by the study. The model uses a CNN-based encoder for local feature extraction a Transformer module for global context modeling and a VAE branch to enhance generalization and regularization. Instead the proposed model outperformed state-of-the-art segmentation methods. According to the results SegTransVAE outperformed existing techniques in terms of Dice Score and 95 percent-Hausdorff Distance metrics while maintaining an efficient inference time. Furthermore the model is highly effective for difficult medical picture segmentation tasks due to its regularization approach which improves generalization and robustness. Enhancing computational efficiency for real-time applications and expanding the framework to multi-organ segmentation will be the main goals of future research.

2.34. TD-Net: Unsupervised Medical Image Registration Network Based on Transformer and CNN. [34]

The paper proposes a model named TD-Net that combines CNNs and Transformers to gather both local and global information to improve unsupervised medical picture registration. The model employs a framework similar to a U-Net, where CNNs extract local spatial information and Transformers encode global context, to enhance deformation field estimates. The most advanced registration techniques were surpassed by the suggested model. According to the results TD-Net outperformed current methods improving registration accuracy by 1% on brain MRI datasets. The model is very successful at registering deformable medical images because its feature fusion technique guarantees improved representation learning. Subsequent research will concentrate on increasing computational efficiency for real-time applications and expanding the framework to multi-modal medical imaging.

2.35. Dual Transformer Encoder Model for Medical Image Classification. [35]

The paper proposes the Dual Transformer Encoder Model, a model that uses two transformer encoders of different hidden sizes to enhance medical picture classification. The model enhances multi-scale feature extraction by combining both local and global contextual information. The proposed model outperformed the existing transformer-based classification methods. The dual encoder architecture significantly improved classification accuracy according to the results with the Layer-wise Class Token Attention (LCA) technique ensuring superior feature aggregation. Since the model can assess various token sequences it is also incredibly adaptable for medical imaging tasks. Expanding the methodology to multimodal medical picture analysis and optimizing efficacy for real-world clinical settings will be the main goals of future research.

2.36. Cross-Shaped Windows Transformer with Self-Supervised Pretraining for Clinically Significant Prostate Cancer Detection in Bi-Parametric MRI. [36]

In order to enhance the detection of clinically significant prostate cancer (csPCa) using bi-parametric MRI (bpMRI) the study suggests a model called CSwin UNet. For the purpose of capturing both local and global contextual information the model incorporates a Cross-Shaped Windows (CSwin) Transformer into a UNet framework. The suggested model outperformed the most advanced medical image segmentation techniques when compared side by side. According to the results CSwin UNet outperformed other models like Attention UNet Swin UNETR and Former achieving an AUC of 0. 888 on the PI-CAI dataset and 0. 79 on the Prostate158 dataset. The model is also very successful at detecting csPCa because of its self-supervised pretraining approach which enhances feature representation and generalization. Subsequent research will concentrate on improving interpretability for clinical applications and modifying the methodology for multi-center datasets

2.37. SLMT-Net: A Self-Supervised Learning Based Multi-Scale Transformer Network for Cross-Modality MR Image Synthesis. [37]

To improve cross-modality MR image synthesis the study proposes SLMT-Net a model that makes use of self-supervised learning and multi-scale Transformers. During pretraining the model employs an Edge-Preserving Masked Autoencoder (Edge-MAE) for feature extraction and edge preservation and a Multi-Scale Transformer U-Net (MT-UNet) for target modality image synthesis. In contrast to existing MR image synthesis methods the proposed model demonstrated superior performance. The results showed that even when trained with partially unpaired datasets SLMT-Net performed better than conventional deep learning methods in generating high-quality MR images. Additionally the Dual-Scale Selective Fusion (DSF) module of the model improves multi-scale feature integration making it highly valuable for clinical MR imaging applications. Enhancing generalizability across various MR sequences and maximizing computational efficiency for real-time synthesis will be the main goals of future research.

2.38. Self-Supervised 3D Anatomy Segmentation Using Self-Distilled Masked Image Transformer (SMIT). [38]

The Self-Distilled Masked Image Transformer (SMIT) paradigm which employs self-supervised learning to improve 3D multi-organ segmentation is proposed in the study. The model combines self-distillation learning with masked image prediction to pre-train vision transformers on large CT and MRI datasets. Furthermore the proposed model outperformed existing segmentation methods. With an average Dice Similarity Coefficient (DSC) of 0. 875 on MRI data and 0. 878 on CT data the results showed that SMIT demonstrated superior performance to conventional pretext learning techniques. Furthermore because the model uses self-supervised pretraining instead of requiring large labeled datasets it is highly effective for clinical applications. The focus of future studies will be on improving cross-domain generalization and applying the methodology to more anatomical regions

2.39. Residual Vision Transformer (ResViT) Based Self-Supervised Learning Model for Brain Tumor Classification. [39]

The paper proposes the Residual Vision Transformer (ResViT) model, which improves brain tumor classification by utilizing self-supervised learning (SSL) and a hybrid CNN-Transformer architecture. By using CNNs for local feature extraction and Vision Transformers (ViTs) for global context modeling, the approach enhances MRI-based tumor classification. The proposed model outperformed the state-of-the-art deep learning methods. The results showed that ResViT performed better than existing models, achieving accuracy of 98.47% on Kaggle datasets, 98.53% on Figshare, and 90.56% on BraTS 2023. The model's self-supervised MRI synthesis pretraining, which also improved feature representation and generalization, makes it highly valuable for real-world medical applications. Future studies will focus on expanding the system to multi-class tumor grading and increasing model efficiency for clinical implementation.

2.40. Evaluating Self-Supervised Learning in Medical Imaging: A Benchmark for Robustness, Generalizability, and Multi-Domain Impact. [40]

The paper evaluates eight important SSL methods such as masked image modeling and contrastive learning on 11 different medical datasets from the MedMNIST collection to demonstrate the benefits of SSL in data-constrained medical imaging applications. In comparison to supervised learning baselines the results showed that SSL-pretrained models outperformed supervised learning methods in cross-dataset generalization and out-of-distribution detection achieving high accuracy even with only 1% labeled data. The framework also highlights how pretraining techniques are necessary to improve medical picture analysis with minimal supervision. In order to better understand how SSL impacts real-world clinical applications future research will focus on expanding the benchmark to multimodal medical data..

2.41. Evaluating the Robustness of Self-Supervised Learning in Medical Imaging. [41]

In order to assess the generalizability and robustness of self-supervised learning (SSL) techniques in medical imaging the study focuses on multi-organ segmentation from CT scans and pneumonia detection from X-rays. Without labeled data the study learns feature representations through SSL pretraining. The models are then enhanced for particular downstream tasks.Under challenging conditions, the proposed approach outperformed fully supervised learning. The results showed that SSL-pretrained models were more robust to picture disturbances and generalized better to unknown data distributions. The study also demonstrates that, although supervised learning and SSL perform similarly on clean data, SSL models maintain higher accuracy in noisy or disrupted situations.

2.42. Dive into Self-Supervised Learning for Medical Image Analysis: Data, Models, and Tasks. [42]

This paper carefully examines self-supervised learning (SSL) for medical image analysis with an emphasis on data models and task applicability. This study looks into task-specific SSL techniques the efficacy of various network topologies the effect of SSL on class-imbalanced datasets and the incorporation of SSL with common deep learning training rules. In comparison to conventional supervised learning techniques the suggested insights demonstrated notable gains in minority class detection performance. SSL greatly improves performance for unbalanced datasets according to the results however the advantages of SSL differ based on the tasks complexity and the datasets properties. In order to enhance model generalization the paper also emphasizes the necessity of universal pretext tasks and offers helpful recommendations for enhancing SSL in medical imaging applications.

2.43. AI Breakthrough Raises Hopes for Better Cancer Diagnosis. [43]

The Harvard Medical School-developed CHIEF model according to the report improves cancer diagnosis assesses treatment options and predicts survival rates through AI-based histopathology analysis. After being trained on millions of unlabeled whole-slide tissue images the model can differentiate between different cancer types with up to 94% accuracy outperforming existing AI techniques by up to 36%. The suggested model performed better than the most advanced cancer detection methods. The outcomes demonstrate that CHIEF can without DNA sequencing identify potential targets for targeted therapies correlate tumor patterns to genomic alterations and predict patient survival outcomes with high accuracy. Furthermore the models wide clinical relevance and capacity to generalize across a wide range of cancer types make it an essential tool for AI-assisted cancer diagnosis.

2.44. Self-Supervised Learning for Medical Image Analysis Using Transformer-Based Masked Autoencoders. [44]

For the purpose of improving medical image analysis through self-supervised learning (SSL) the study suggests a Transformer-based Masked Autoencoder (MAE) model. In the absence of labeled input the model learns robust feature representations through masked image reconstruction. Both ImageNet transfer learning and the existing SSL techniques were surpassed by the suggested model. The results showed that, particularly in cases involving medical imaging with limited data, MAE pretraining significantly improved segmentation and classification accuracy. Additionally, the model's generalization across 2D and 3D modalities makes it highly effective for a range of medical imaging applications. Future research will primarily focus on expanding MAE-based SSL approaches for multimodal medical data integration and prognosis prediction.

2.45. Self-Supervised Learning with Vision Transformers for Retinal Disease Classification. [45]

Using masked image reconstruction for feature extraction the paper suggests a model based on Vision Transformers (ViTs) with self-supervised learning (SSL) to improve the classification of retinal disorders. The model increases the precision of ophthalmology diagnosis through representation learning from unlabeled retinal images. Compared to CNN-based and fully supervised methods the suggested model performed better. The outcomes demonstrated that SSL-pretrained ViTs were better able to categorize a variety of retinal disorders such as age-related macular degeneration and diabetic retinopathy. Furthermore the model is very advantageous for practical clinical applications due to its broad generalization of the model. Future research will concentrate on enhancing model interpretability for clinical decision support and extending SSL-based ViTs to multimodal ocular imaging.

2.46. Self-Supervised Learning for Medical Image Analysis Using Cross-Modal Transformers. [46]

The paper proposes a model based on Cross-Modal Transformers with Self-Supervised Learning (SSL) to enhance medical image analysis by integrating multi-modal imaging data, such as MRI and CT scans. The method uses masked image reconstruction to retrieve important characteristics without labeled input, which increases the accuracy of classification and segmentation. The recommended model outperformed CNN-based and fully supervised approaches. The results showed that cross-modal transformers effectively capture long-range interactions and improve generalization across many imaging modalities. Additionally, the model's ability to include complementary information from multimodal images makes it highly effective for real-world clinical applications. Future studies will focus on expanding SSL-based cross-modal transformers to more medical imaging modalities and improving computation efficiency for large-scale deployment.

2.47. Self-Supervised Learning with Swin Transformers for Medical Image Analysis. [47]

The paper suggests a model based on Swin Transformers with Self-Supervised Learning (SSL) that uses shifting window attention and hierarchical feature extraction to improve medical image classification and segmentation. By using SSL pretraining the model overcomes the issue of data scarcity and produces reliable feature representations from unlabeled medical images. The proposed model performed better than transformer-based and CNN-based methods. The results showed that Swin Transformers effectively capture both local and global context improving classification and segmentation accuracy across a range of medical imaging datasets. The models scalability and effectiveness make it a great choice for real-world medical imaging applications. The main goals of future research will be to extend SSL-based Swin Transformers to multimodal medical imaging and increase computing efficiency for broad practical application.

2.48. Self-Supervised Learning for 3D Medical Image Analysis Using Vision Transformers. [48]

The paper suggests a paradigm for improving 3D medical image classification and segmentation that combines Vision Transformers (ViTs) with Self-Supervised Learning (SSL). This paradigm captures geographic correlations and long-range interdependence in volumetric data. Masked image reconstruction is used to pretrain the model on unlabeled 3D medical images in order to enhance feature representation and generalization. Compared to CNN-based and other transformer-based approaches the suggested model outperformed them. In segmentation and classification tests the results demonstrated that SSL-pretrained 3D ViTs outperformed conventional methods achieving greater accuracy. Additionally because it can adjust to volumetric data the model is very useful for real-world 3D medical imaging applications. Future research will concentrate on improving computation efficiency for widespread clinical use and extending SSL-based ViTs to multimodal 3D medical imaging.

2.49. Self-Supervised Contrastive Learning for Medical Image Segmentation Using Transformers. [49]

The paper suggests a self-supervised contrastive learning with transformers model that uses unlabeled data for reliable feature representation learning to improve medical image segmentation. By combining transformer topologies and contrastive learning the method improves segmentation accuracy in low-data scenarios. Conversely CNN-based and self-supervised learning techniques were surpassed by the suggested model. The findings demonstrated that contrastive learning improved segmentation accuracy in a variety of medical imaging tasks by aiding in the differentiation of local anatomical features. The model is ideal for medical picture segmentation applications because its local contrastive loss function also guarantees better feature learning. In order to increase computational efficiency for large medical datasets future research will concentrate on extending transformers based on contrastive learning to multi-organ segmentation.

2.50. Self-Supervised Learning for Medical Image Analysis Using Multi-Scale Vision Transformers. [50]

The paper proposes a methodology based on Multi-Scale Vision Transformers (ViTs) with Self-Supervised Learning (SSL) that uses hierarchical feature extraction to enhance medical picture categorization and segmentation. The model leverages SSL pretraining to obtain robust feature representations from unlabeled medical images, thereby overcoming the problem of data scarcity. The proposed model performed better than CNN-based and transformer-based methods. The results showed that multi-scale ViTs effectively capture both local and global context, improving classification and segmentation accuracy across a range of medical imaging datasets. Additionally, the model's effectiveness and scalability make it a great choice for real-world medical imaging applications. Future research will concentrate on improving computational efficiency for widespread clinical implementation and extending SSL-based multi-scale ViTs to multimodal medical imaging.

2.51. Multi-Modal Self-Supervised Learning for Medical Image Analysis. [51]

A multimodal puzzle-solving task is combined with Self-Supervised Learning (SSL) techniques in the studys Multi-Modal Self-Supervised Learning for Medical Image Analysis to achieve the best results. The present model has employed a variety of SSL techniques for representation learning on medical image data pertaining to multiple imaging modalities including contrastive learning cross-modal generation and permutation-based learning. To improve feature learning accuracy in medical image analysis optimization techniques such as multimodal fusion cross-modal prediction and Sinkhorn operator have been developed. These techniques are especially useful when paired with other tasks like segmentation and illness classification. When learning to represent medical images the multimodal approach has outperformed previous self-supervised techniques. SSL methods with assessment on classification and segmentation problems. combining contrastive, generative, and permutation-based methods with multimodal learning. Data augmentation and pretraining steps are emphasized for good results. SSL approaches (contrastive, generative, permutation) are used to learn features from medical images. The multimodal approach showed the highest performance in medical image analysis.

2.52. Self-Supervised Learning with Transformers for Histopathology Image Analysis. [52]

The paper introduced transformer-based self-supervised learning for histopathology image analysis. For optimal results Masked Image Modeling (MIM) is used in conjunction with Self-Supervised Learning (SSL) techniques. The presented model used a range of SSL techniques such as Vision Transformers (ViT) and contrastive learning masked token prediction for representation learning on histopathology image data related to cancer detection. To improve the accuracy of feature learning pre-training techniques like MIM MoCo v2 and ImageNet pre-training have been created for histopathology image analysis. These techniques are especially useful when combined with ViT models. When it comes to histopathology representation learning MIM has outperformed other self-supervised techniques like contrastive learning. Using ViT models in conjunction with MIM-based pre-training to extract features.Constructive MIM and transformer-based SSL techniques are employed to extract features from histopathology images. The best results in medical image analysis were obtained using the MIM approach.