

*Leveraging Integrated CNN-Transformer Model and
eXplainable Artificial Intelligence (XAI) Technique
for Enhanced Breast Ultrasound Report Generation.*

Final Thesis

In Partial Fulfillment
of the Requirements for the Degree of
Master in Artificial Intelligence and Data science

Student Name	:	Brintha Thirunavukkarasu
Student ID	:	2555470
Supervisor	:	Dr. Shaheen Khatoon
Assessor	:	

Abstract

Breast cancer is diagnosed using radiology reports that describe explanations derived from medical images, including breast ultrasound images (BUS). A radiologist's primary responsibility is to interpret the key information concealed within medical images and produce diagnostic reports that are highly valuable for future clinical treatment. However, the obstacle in evaluating medical imaging will result in inaccuracies during subsequent medical diagnosis. To address these issues, deep learning models can fully transform radiology practice by lessening the workload for radiologists and time costs. This study proposed a hybrid model consisting of a Convolutional Neural Network (CNN), which is ResNet101, and transformer-based encoder-decoder architecture for generating ultrasound image reports. The suggested CNN is intended to be efficient in capturing relevant visual features, allowing the transformer-based encoder-decoder to reliably turn these features into well-structured medical reports using attention mechanism. The research also presents a curated public dataset of 256 breast ultrasound scans for assessing the performance of the model in reports. The black-box characteristic of models has become more prevalent in medical imaging since physicians never accept an algorithm's decision without first understanding its decision process. Explainable artificial intelligence (XAI) approaches have recently focused on these medical imaging issues, which aim to improve their transparency before being employed for patient care. This study aims to assess the effectiveness of the XAI technique. To visualise the data, the Gradient-weighted Class Activation Mapping (Grad-CAM) method was utilised, and the Attention heat map was applied to improve report interpretability. Our suggested model has been evaluated on the BUS dataset and receives competitive scores with a BLEU-1 score of 0.486, a METEOR score of 0.203, and a CIDER score of 0.118.

Keywords: Radiology Reports; Deep Learning; Transformer Encoder-Decoder; Grad-CAM.

Acknowledgements

I would like to express my deepest gratitude to the **University of East London** and the **School of Architecture, Computing, and Engineering** for providing me with the opportunity and resources to pursue my academic journey in AI and Data Science. The guidance, support, and encouragement I have received from the university have been invaluable throughout my studies.

I am especially grateful to my supervisor, **Dr. Shaheen Khatoon, Ph.D.**, Senior Lecturer in Data Science (Computer Science and Digital Technologies), for her continuous support, insightful feedback, and unwavering encouragement during my research. Her expertise and dedication have been instrumental in helping me shape my thesis and develop a deeper understanding of the subject.

I would also like to extend my thanks to the module leaders and staff members within the School of Architecture, Computing, and Engineering for their dedication to teaching and their contribution to my learning in AI and Data Science. Their knowledge and passion for the subjects have been a constant source of inspiration.

Contents

Abstract	ii
Acknowledgements	iii
List of Tables	vi
List of Figures	vii
List of Acronyms	viii
ix	
Chapter 1 Introduction.....	1
1.1 Background.....	1
1.2 Problem Statement.....	3
1.3 Research Question and Objectives	4
1.3.1 Research Questions	4
1.3.2 Objectives	5
1.4 Expected outcomes	5
Chapter 2 Literature Review	7
2.1 Comprehensive Overview of the Existing Literature	7
2.1.1 Background in Explainable Artificial Intelligence (XAI)	7
2.1.2 Medical report generation	8
2.1.3 Chest X-ray vs Breast Ultrasound Images Dataset	9
2.1.4 Encoder-decoder Framework.....	10
2.1.5 XAI Techniques in Medical Imaging Reports	13
2.2 Critical Analysis of Existing Studies.....	16
20	
Chapter 3 Methodology	21
3.1 Data Collection and Preprocessing.....	22
3.1.1 Data collection	22
3.1.2 BI-RADS Category	23
3.1.3 Ethical Approval and Consent	23
3.1.4 Data labels and annotation	23
3.1.5 Data preprocessing.....	25
3.2 ML/AI Model Development.....	26
3.2.1 ResNet101 for Feature Extraction	26
Residual Block.....	26
3.2.2 Transformer Encoder Block.....	27
3.2.3 Transformer Decoder Block	27
3.2.4 Self-Attention.....	27
3.3 Evaluation of the Proposed System.....	28
3.4 XAI Techniques.....	29
3.4.1 Grad CAM	29
3.4.2 Attention Map	30
Chapter 4 Experimental Results.....	31
4.1 Experimental Setup	31
4.1.1 Hardware.....	31
4.1.2 Software	31
4.2 Dataset Description	32

4.2.1	Breast Ultrasound Images (BUS).....	32
4.2.2	Analysis of Textual Report	33
4.3	Results	34
4.3.1	Hybrid Model Performance	34
4.3.2	Quantitative Results and Discussion.....	36
4.3.3	Qualitative Results & Discussion	37
4.4	Comparison with Baseline Methods.....	42
Chapter 5	Conclusion and Future Work	44
5.1	Conclusion	44
5.1.1	Key Contributions and Findings	44
5.1.2	Limitations	45
5.2	Future Work.....	46
	References	47

List of Tables

Table 1.1 Critical analysis/ Summary of the existing studies.....	17
Table 1.2 Data characteristics for sample images.....	24
Table 1.3 Evaluation Metrics and their formulation	29
Table 1. 4 Experiment Results that combined all images, both normal and abnormal, according to the test sets BLEU-n (n=1,2,3,4), CIDEr (C), METEOR(M), and ROUGE(R) scores generated radiology reports.	37
Table 1.5 Examples of qualitative report generation from medical images.....	38
Table 1.6 Comparison of NLG metrics with baseline methods.....	42

List of Figures

Figure 2.1 Basic Architecture of Encoder-Decoder Framework	11
Figure 2.2 Architecture of the proposed system.	21
Figure 2.3 Pre-processed Breast Ultrasound image with respective text.....	32
Figure 2.4 Distribution of ultrasound report lengths.	33
Figure 2.5 Venn-style word cloud for textual report	34
Figure 2.6 Accuracy and validation accuracy of model performance	35
Figure 2.7 Training loss and validation Loss of model performance	36
Figure 2.8 Heatmaps of Grad-CAM	39
Figure 9 Visualization of Attention heatmap a) Original grayscale image b) Attention Heatmap c) Overlaid image d) Combined Image	40
Figure 10 Overlaying attention heatmap image with ground truth.....	41

List of Acronyms

ACR	American College of Radiology
AI	Artificial Intelligence
AUC	Area Under the Curve
BEiT	Bidirectional Encoder Representations from Transformers
BIRADS	Breast Imaging Reporting and Data System
BLEU	Bilingual Evaluation Understudy
BUS	Breast Ultrasound
BUSI	Breast Cancer Ultrasound Images
CE	Clinical Efficacy
CE Metrics	Cross Entropy Metrics
CIDeR	Consensus-based Image Description Evaluation
CNN	Convolutional Neural Network
CT	Computed Tomography
CSV	Comma-Separated Values
DCIS	Ductal Carcinoma In Situ
FFN	Feed-Forward Network
GLCM	Grey Level Co-occurrence Matrix
GradCAM	Gradient-weighted Class Activation Mapping
GRU	Gated Recurrent Unit
IUXRAY	Indiana University X-Ray Dataset
IoU	Intersection over Union
LIME	Local Interpretable Model-agnostic Explanations
LLM	Large Language Model
LRP	Layer-wise Relevance Propagation
LSTM	Long Short-Term Memory
METransformer	Medical Efficient Transformer

METEOR	Metric for Evaluation of Translation with Explicit ORdering
MIMICCXr	Medical Information Mart for Intensive Care - Chest X-Ray
MRI	Magnetic Resonance Imaging
MSA	Multiheaded Self-Attention
NASNetLarge	Neural Architecture Search Network Large
NLG	Natural Language Generation
NLP	Natural Language Processing
NST	No Special Type
PPKED	Prior Knowledge Embedding and DenseNet Framework
RAM	Random Access Memory
ResNet101	Residual Network 101 Layers
R2GenGPT	Radiology Report Generation Generative Pre-trained Transformer
R2GENCMN	Radiology Report Generation with Concept Mention Networks
RMAP	Radiology Multi-label Abnormality Prediction
ROUGEL	Recall-Oriented Understudy for Gisting Evaluation - Longest Match
SHAP	SHapley Additive exPlanations
TFIDF	Term Frequency-Inverse Document Frequency
TL	Transfer Learning
ViTs	Vision Transformers
VRAM	Video Random Access Memory
WGAM	Weakly Guided Attention Mechanism
KIRN	Knowledge-based Interactive Recurrent Network
WHO	World Health Organization
XAI	Explainable Artificial Intelligence

Chapter 1 Introduction

1.1 Background

The first and most prevalent kind of cancer in women is breast cancer. According to statistics collected by the World Health Organization (WHO), it has a significant role in the high rate of female mortality worldwide (WHO, 2023). Predictive modelling indicates that by 2040, there will be an increase, with an anticipated global annual death rate of one million and an estimated 3 million instances of breast cancer are reported each year (World Health Organization, 2023). Anatomically, the breast is made up of lymph nodes, connective tissue, ductal lobules, and healthy blood arteries and the methods most commonly used to identify breast cancer are computed tomography (CT), mammography, magnetic resonance imaging (MRI) and breast ultrasound (BUS). BUS imaging is an inexpensive and radiation-free screening method, and it offers details on the properties of breast tissue and the presence of malignant tissues.

Radiologists also record their findings from the images in radiology reports. To make the conclusions easy for other physicians to understand, the reports need to be accurate and consistent. The regulations for reporting breast illness that outlines the format of the reports is called BI-RADS stands for Breast Imaging Reporting & Data System. Reports that follow these criteria include clinical information, examination results, and a conclusion with a BI-RADS score (which goes from 0 to 6, with 6 representing the most serious malignancy) (Sickles, D’Orsi and Basset, 2013). Because human perception is subjective, radiologists may overlook small results that could result in incorrect diagnoses. Studies reveal that up to 30% of cases may be misdiagnosed, which explains a significant portion of difficult cases (Berlin, 2001).

Artificial intelligence technology can fully transform radiology practice by lessening the workload for radiologists and its errors. Artificial intelligence (AI) approaches such as machine learning and deep learning are examples of how these techniques have developed for precise breast cancer diagnosis. Enhanced breast cancer categorisation with convolutional neural networks improves patient care by enabling early disease

identification and recommending preventive measures (Bai et al., 2024). A machine learning technique called transfer learning comprises teaching a model on one task and then applying that knowledge to another. It increases a model's efficiency in another domain by utilizing learnt information from the first. When the qualities acquired in the source task are applicable to the target task or when there is not enough data available for training on the target task, transfer learning might be advantageous. A transfer learning model with prior ImageNet training has been used successfully in recent years to distinguish breast cancer from ultrasound scans, making it a good choice for breast cancer detection. Because transfer learning has smaller datasets than deep learning, it can solve the problem of insufficient training samples.

However, some restrictions must be considered. The systems may be unclear, making it difficult for physicians to understand the logic behind the decision-making process of system. In medical contexts, where effective diagnosis depends on human expertise and intuition, this lack of openness could be problematic. By providing clear explanations of the processes for making choices used by artificial intelligence systems, explainable artificial intelligence (XAI) research has expanded exponentially to address this problem (Patrício, Neves and Teixeira, 2023). This clarity promotes patient and medical professional trust in recommendations, facilitating well-informed decision-making. Among the initial XAI approaches, in image captioning research, convolution recurrent architectures (CNN-RNN) are frequently used. Transformer-based models and visual attention mechanisms have been integrated for automated report generation in recent research (Raminedi, Shridevi and Won, 2024). Transformers' superior performance in natural language processing tasks makes them a potential option for medical report writing automation and their ability to be parallelized. An encoder and a decoder, each consisting of numerous layers of feedforward and self-attention neural networks, make up the transformer architecture. The attention layer in a transformer model enables the network to prioritise distinct parts of the input sequence, focussing on relevant information while processing each element. It calculates attention scores to dynamically change each token's influence on the output, hence improving the model's capacity to capture contextual linkages.

1.2 Problem Statement

Explainable artificial intelligence (XAI) is a subfield of artificial intelligence that generates neural network explanations, allowing people to comprehend, relate to, and effectively interact with this new breed of artificially intelligent companions. Convolutional neural networks, in particular, produce results that can be explained by several XAI techniques that have recently been developed. This helps researchers better comprehend and evaluate the decisions that these models make (Gurmessa and Jimma, 2024). An array of advances in cutting-edge were enabled by the debut of a novel artificial intelligence model architecture called "Transformers." The ability of this model to assess each element's relative relevance using attention processes in a data sequence while considering its relationship with other elements is what allowed for these breakthroughs. The creation of vision transformers (ViTs), a computer vision adaption, in 2020 was prompted by the success of transformer models (Dosovitskiy et al., 2024). Transformers' primary notion, the attention mechanism, can replace LSTMs and RNNs. Transformers have made it possible to accomplish this by doing away given that networks need feedback, and that feedforward architecture encourages efficient parallel processing. However, considering dataset bias, lengthy sequence lengths, and variations in radiologists' findings, medical imaging data might be challenging to generate reports from. State-of-the-art outcomes for metrics pertaining to clinical efficacy (CE) and natural language generation (NLG) have been achieved with these transformer-based techniques.

A lot of studies have been conducted on generating reports from chest X-ray images. In these studies, public datasets such as IUX-RAY and MIMIC-CXR are frequently employed (Messina et al., 2022). Researchers can develop and evaluate AI models with the help of datasets that include images and related radiological reports. Ultrasound Scan datasets for breast cancer frequently lack the comprehensive, organized reports required to train and validate automatic models for generating reports. Fewer investigations and published articles on automatic report generation specifically for breast cancer ultrasound imaging have been conducted due to these difficulties. This study could be focused on applying models that have been trained on datasets from chest X-rays and

modifying them for publicly available curated benchmark 256 breast ultrasound image datasets (Pawłowska et al., 2024).

The objective of Explainable AI (XAI) is to develop a collection of machine-learning techniques that would prevent the efficacy of the present generation of AI systems from being limited. Provide models that are easier to grasp while maintaining a high level of learning performance, enabling people to engage, comprehend, and trust the next generation of artificially intelligent partners in appropriate ways. Therefore, the challenge is to use XAI methods to progress the model's outputs' interpretability so that radiologists and other medical professionals can feel confident in the decision-making process of the model. This study employs the Grad-CAM, or gradient-weighted class activation mapping, is a technique for visualization to explicitly address this difficulty.

1.3 Research Question and Objectives

1.3.1 Research Questions

1. What are the potential benefits of using a hybrid model combining ResNet101 transformer-based models to generate radiology reports?
2. How do the 256 breast ultrasound images that make up the curated dataset impact the suggested model's capacity for training and generalization?
3. What is the contribution of the transformer-based Encoder-Decoder framework to the quality and contextual coherence of the medical reports produced from breast ultrasound images?
4. In what way do the Grad-CAM and Attention heat maps make the produced reports easier for radiologists to interpret?
5. What are the potential biases and restrictions of the Grad-CAM and Attention heat maps for utilizing ultrasound images to generate reports?

1.3.2 Objectives

1. To increase openness and trust in model decisions for lesion analysis, analyse explainable artificial intelligence (XAI) techniques, in particular attention heat maps and gradient-weighted class activation mapping (Grad-CAM), by determining their applicability and effectiveness using a dataset of 256 breast ultrasound scans.
2. Build a framework that combines transformer-based encoder-decoder with explainable AI (XAI) methods to improve the readability of automated breast cancer diagnosis and report generation from ultrasound images.
3. The evaluation of transformer-based models' clinical efficacy and natural language generation (NLG) capabilities for generating BI-RADS-compliant reports for breast ultrasound imaging.
4. In generating radiology reports for the identification of cancer using breast ultrasound images, compare the accuracy and dependability of transformer-based models to conventional convolutional neural network-based models.
5. Evaluate the developed models and methodology's clinical applicability and adherence to legal and regulatory requirements by validating them in clinical settings, and dependability for implementation in real-world scenarios.

1.4 Expected outcomes

The principal goal of this research is to develop an efficient framework that uses transformer-based models to produce radiological reports automatically from breast ultrasound (BUS) images. By addressing the main research challenges and objectives mentioned, outcomes are anticipated to enhance the field of medical imaging, notably in breast cancer diagnosis. The goal of the study is to show that the suggested hybrid model, which combines a transformer-based encoder-decoder framework with a ResNet101, can efficiently produce radiology reports from breast ultrasound images that are coherent and comply with BI-RADS. The transformer-based Encoder-Decoder system is projected to considerably improve the quality and contextual coherence of generated medical reports,

ensuring that they are thorough and in line with clinical guidelines. The hybrid model should improve the overall quality of the reports by delivering more detailed and contextually relevant diagnostics. It is anticipated that the model will improve the interpretability and transparency of its choices by incorporating XAI techniques such as Grad-CAM and Attention heat maps, thereby boosting radiologists' trust. The curated 256 breast ultrasound image dataset should allow the model to perform consistently over a wide range of circumstances, increasing its usefulness in real-world scenarios. The evaluation of this dataset should produce competitive natural language generation metrics (e.g., BLEU, ROUGE-L, METEOR and CIDEr) and validate the model's applicability and dependability in clinical contexts. The study will look at potential biases and limits in Grad-CAM and Attention heat maps, particularly in how they handle ultrasound images and report generation. This review will help to refine and guide the future applications of these XAI approaches.

Chapter 2 Literature Review

2.1 Comprehensive Overview of the Existing Literature

2.1.1 Background in Explainable Artificial Intelligence (XAI)

Even though advancements in medical image analysis have made it possible to classify many forms of medical data with accuracy comparable to that of a human being, automated medical imaging still needs to be widely used in clinical settings. Due to a lack of understanding of the algorithm's decision-making process, doctors will never be able to trust its outcome. This truth has increased the need for developing approaches that can explain how artificial intelligence algorithms make decisions, which has given rise to an entirely novel area of study well-known as eXplainable Artificial Intelligence (XAI) (Hassija et al., 2024).

Early artificial intelligence systems were built on explicit rules and logic, making them naturally explainable because their decision-making processes were obvious. The post-hoc techniques were employed when deep learning models were first being explained. These are still in use in many medical imaging domains. It is essential to comprehend them in order to appreciate the developments in interpretable deep learning techniques (Carriero et al., 2024). XAI methods have expanded to include both model-agnostic techniques, which work with any model, and model-specific strategies, built for specific types of models. The most common methods used for interpreting models in medical imaging are LIME, SHAP and saliency maps with conventional techniques like LRP, Integrated Gradients, or Grad-CAM (Biswas, 2024). New techniques emphasise providing explanations that are not only accurate but also helpful and actionable for end users, hence improving human-AI collaboration. Hence, a potential research area in medical imaging has been the development of inherently interpretable models. These models' primary benefit is that they provide separate explanations, which enhances their fidelity and transparency and raises the likelihood that they will be incorporated into clinical practice.

2.1.2 Medical report generation

The advancement of using medical imaging data to automatically generate diagnostic reports is essential for minimizing the doctor's workload. Deep learning models can be effectively and semantically supported by teaching machines to generate diagnostic reports on their own. Therefore, It is crucial to look at the generation of reports and automated diagnosis of images to improve the interpretability of deep learning (Pang, Li and Zhao, 2023).

The automatic development of diagnostic reports is inspired by image captioning, which blends computer vision (CV) and natural language processing (NLP) to provide an extensive interpretation of medical images. Recent developments in deep learning have resulted in notable improvements in image captioning, the encoder-decoder architecture is the foundation of this. A decoder, Recurrent Neural Network (RNN) is used to generate descriptions of specified images by leveraging visual information gathered from an encoder, Convolution Neural Network (CNN). Later, For image captioning applications, RNN variations are frequently utilised to capture the semantics. These include the gated recurrent unit (GRU) and long short-term memory (LSTM), which have unique controlling gates that can recall information from a long time ago (Ahmed, Solis-Oba and Ilie, 2022).

In 2016, (Shin et al., 2016) carried out the first deep-learning application in the creation of medical imaging reports. They developed a CNN-RNN network that was only able to predict with accuracy annotated tags containing chest X-ray pictures. They used a recurrent neural cascade model to account for simultaneous image/text options, which enhanced the outcomes of their tests of both LSTM and GRU. LSTM has produced cutting-edge outcomes and has been employed and researched more extensively in literature. But because of its more straightforward architecture and quicker training period than LSTM, GRU is becoming more and more well-liked. LSTM will thereafter be used as the fundamental structure of RNN in subsequent studies on medical image captioning.

In latest days, the exploration of attention mechanisms has advanced significantly in fields like image processing and natural language processing. The human brain's perceptual and attentional systems are similar. The attention map, or matrix that represents the significance of each component of the image for a given task, is computed by image

processing attention processes. Later, transformer-based systems included self-attention. In NLP tasks, self-attention is frequently employed as it concentrates on a single context (Jing, Xie and Xing, 2018).

2.1.3 Chest X-ray vs Breast Ultrasound Images Dataset

Most research in automatic report generation has focused on chest X-ray images since big benchmark datasets like IU X-RAY and MIMIC-CXR are readily available (Ahmed, Solis-Oba, and Ilie, 2022). These datasets offer many annotated image-text combinations, making it easier to develop and evaluate sophisticated models. However, breast ultrasound image datasets especially those with thorough annotations and reports are less common and frequently smaller in size. Largely annotated datasets are more challenging to develop because anatomical structure localisation requires expert knowledge and restricts the development of automated ultrasound report generation models.

The Breast Cancer Ultrasound images (BUSI) dataset, which includes 780 ultrasound images from the Baheya Hospital in Cairo, Egypt, is a valuable and demanding resource for breast lesion segmentation research. A significant portion of the ultrasound images are categorised as benign, malignant, and normal cases based on the size and shape of the nodules (Hekal et al., 2024). The UDIAT dataset was made available by the Parc Tauli Corporation's UDIAT Diagnostic Centre in Sabadell, Spain. The dataset includes 163 US scans together with the corresponding ground truth images, including 109 images in the benign class and 54 in the malignant class (Yap et al., 2017). The Thammasat dataset from The Department of Radiology at Thammasat University in Thailand. A count of 201 ultrasound images together with the accompanying ground truth pictures were gathered of which 106 are classified as malignant and 95 as benign (Rodtook et al., 2018).

These datasets are vital for breast cancer diagnosis because they offer a diverse and annotated set of ultrasound images, which are needed to build and evaluate machine learning models that they are designed to detect and categorise breast lesion. The BUSI dataset includes a wide spectrum of normal, benign, and malignant cases, allowing models to learn from a variety of nodule sizes and forms. The UDIAT and Thammasat datasets

provide variation by categorising images as benign or malignant, which improves the resilience of diagnostic algorithms. The development of more precise, reliable, and broadly applicable methods for early breast cancer diagnosis and detection was aided by these datasets.

2.1.4 Encoder-decoder Framework

Recent advances in generating radiology reports use encoder-decoder frameworks that combine image-processing models with natural language generation approaches. The following studies use a variety of architectures, including transfer learning approaches, LSTM, and advanced models such as R2GenGPT, to create precise reports from medical images. Using chest X-ray images such as IU-CXR dataset, the study used to automatically generate radiologist reports. They focused on findings and impression sections in the radiology report. Google's Inception-v3 model provides the underlying framework for processing chest X-ray images. The study used Glove and RadGlove (Radiology Glove) embeddings with GRU (Gated Recurrent Unit) and LSTM (Long Short-Term Memory) networks to create textual reports using an encoder-decoder framework. The study did not employ particular Explainable AI (XAI) techniques for interpretability; however, these methods aid in capturing the sequential flow of text and comprehending the medical context. The challenges are because of their shortcomings in tracking long-range dependencies, existing RNNs are inadequate for modelling lengthy radiology reports and precisely characterizing anomalies indicate regions that require more investigation (Singh et al., 2019).

A team of researchers focused on accurately extracting BI-RADS scores by creating a hybrid model for automatically producing findings in radiology reports related to breast cancer (Nguyen et al., 2020). They employed an encoder-decoder-attention (EDA) model in comparison to a baseline encoder-decoder model using a text summarisation approach, and then trained an additional BI-RADS classifier to improve report accuracy. The hybrid model combines two approaches, resulting in a significant ROUGE-L F1 with a 0.515 score, exceeding comparable studies because of the simpler and shorter input texts.

However, obstacles included dataset restrictions, since the Dutch language dataset required model change and encountered issues with varied reporting styles.

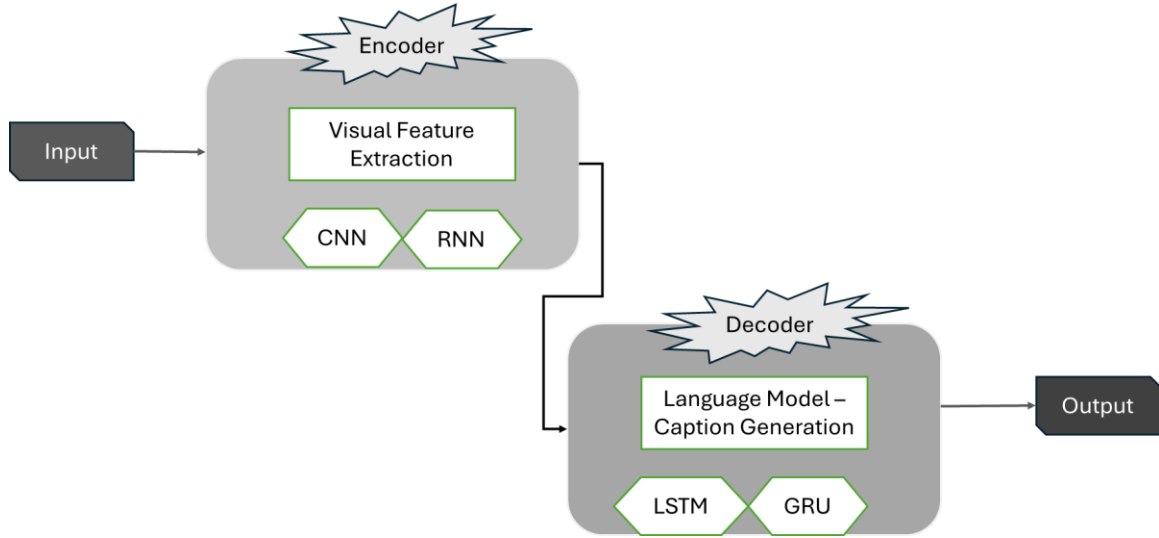


Figure 2.1 Basic Architecture of Encoder-Decoder Framework

A study investigated ultrasound pictures of the liver and gallbladder utilising a Faster RCNN model for lesion detection, linked with an LSTM model (encoder-decoder) improved by a Weight Selection Mechanism (Tsai, Chu and Li, 2023). The encoder-decoder framework's fundamental architecture is depicted in Figure 2.1 Basic Architecture of Encoder-Decoder Framework. While the model worked reasonably well with an IoU larger than 0.5, it struggled to generate reliable diagnostic findings using tiny sample quantities. When utilising prior knowledge instead of depending only on image properties for report generation, performance was better in terms of prior text evaluation measures.

Large Language Models (LLMs) have demonstrated outstanding generalisation capabilities across multiple language challenges. LLMs excel at long text generation due to their strong awareness of syntax, semantic coherence, and grammar. This makes it ideal for approach like medical reporting. With this idea, (Wang et al., 2023) introduced the R2GenGPT model integrating the MIMIC-CXR and IU X-RAY datasets which generate automated reports from X-ray scans of the chest. A Visual Encoder named Swin Transformer model combines with the Llama2-7B large language model (LLM) to transmit

low-dimensional image characteristics into the high-dimensional space of the LLM. The LLM provides diagnostic findings based on visual features in chest x-ray images. With its absence of ensemble approaches, R2GenGPT's CIDEr score was lower than that of METransformer and R2Gen, but it surpassed them in measures like BLEU, METEOR, and ROUGE-L after being trained on strong GPUs. Despite R2GenGPT's need for human annotations, the study demonstrates its efficacy.

A new research investigation employed ultrasound images from the BUSI, Thammasat, and UDIAT datasets to focus on breast lesion segmentation. The research suggested an architecture for segmentation based on encoder-decoder for lesion segmentation and binary mask generation that makes use of ResNet50V2, NASNetLarge, and EfficientNetB7. Using the BUSI dataset, the model obtained a 70.82% IoU score and an 80.56% Dice score. Despite these encouraging findings, the research also pointed up certain drawbacks, such as the possibility of artefact introduction, a lack of explainability, and difficulties incorporating 3D imagery. Although the method performed better than conventional segmentation models, generalisability and artefact introduction remained challenging (Himel, Chowdhury and Hasan, 2024).

Using the BUSI and Breast Histology datasets, advanced transformer-based models were investigated in recent studies on breast cancer imaging (Chaudhury, Sau and Shelke, 2024). To improve classification accuracy, they used Bidirectional Encoder Representations from Transformers (BEiT) and Long Short-Term Memory (LSTM) models with ResNet for feature extraction. This method relied heavily on Vision Transformers (ViT), which are based on the transformer architecture. Vision Transformers interpret images as a sequence of patches, rather than typical CNNs, which permits the model to record contextual linkages and extended dependencies inside the image. As a result, feature representation and classification performance are enhanced, and the model can concentrate on important features across several patches. By utilising transformer principles, the BEiT model outperformed traditional techniques in diagnosing breast cancer from ultrasound and histological pictures with amazing accuracy. This use of transformers demonstrates how well they handle complicated image data by maintaining contextual information more effectively, which improves overall diagnostic performance.

Transformers were first intended for use in natural language processing. Because of this, each word in the input sequence is treated as a token by its encoder, which groups the words into a single information vector. On the other hand, because of the type of data being processed, computer vision is not the same as natural language processing. To achieve this kind of usage, the researcher employed BioBERT as a decoder, transformer-based model (Tao et al., 2024). The suggested encoder uses built-in convolution biases to capture significant spatial characteristics in an efficient and hybrid design. This allows these features to be robustly converted into coherent medical reports by the transformer-based decoder.

2.1.5 XAI Techniques in Medical Imaging Reports

The use of Explainable AI (XAI) methods to imaging in medical field improves the reliability and interpretability of reports produced by AI. Grad-CAM is an application used in much research that helps validate diagnostic accuracy by producing visual heatmaps that show important regions impacting model predictions. By adjusting input features to evaluate their influence on predictions, LIME improves model transparency and offers local interpretability. Particularly in classification problems, A deeper understanding of model making choices is possible using SHAP values by quantifying the contributions of different features. These XAI techniques are particularly important in medical settings, where the development of accurate, dependable, and therapeutically relevant reports depends on knowing the logic underlying AI predictions. Similarly, to improve model interpretability and visualization, the article (Lanjewar, Panchbhai and Patle, 2024) examined breast ultrasound images using the MobileNetV2, ResNet50, and VGG16 models in conjunction with LSTM, Grad-CAM, and LIME approaches. Grad-CAM performed especially well in the VGG16 model, which distinguished regions of interest more quickly than other models. By changing input features and examining their effect on the model's output, the LIME technique, in the meantime, allowed for local interpretability and improved comprehension of the features that had the greatest influence on predictions. These XAI methods assisted model refining, improved validation, and increased confidence in the model's predictions. Impressive performance measures like as the AUC

and F1 score were accompanied by difficulties with overfitting and computational complexity, particularly when working with bigger and more varied ultrasound image datasets.

Another popular Explainable AI (XAI) technique is attention heat mapping, which is especially helpful for medical imaging applications like radiology report production. Attention heat maps give a clear approach to interpreting sophisticated deep-learning models by graphically illustrating the regions of an image that most influence the predictions made by the model. When analysing chest X-rays, (Jin et al., 2024) studied, these heat maps aid in comprehending how the model associates particular X-ray regions with medical issues. This visual explanation increases confidence in AI-generated diagnoses by helping researchers and radiologists confirm the model's focus and make sure it aligns with clinically important regions. The interpretability of attention heat maps is limited because, while they are useful for determining which areas of an image contribute to a prediction, they do not fully explain why a certain region is judged essential.

The presented study's usage of SHAP values makes it possible to recognize feature significance in the LightGBM classification model in great detail. Cooperative game theory-derived SHAP values provide an open method to determine the impact of each feature on the model's output. This is especially crucial for determining which texture features—like metrics from the GLCM (Grey Level Co-occurrence Matrix) have a major impact on the prediction. The study emphasises the directional sensitivity of the texture analysis by showing that features measured in the 90° direction have the greatest influence. Additionally, the study evaluates how well the LightGBM model performs in comparison to common convolutional neural network designs including VGG, Dense Net and ResNet. Although CNNs are well known for their exceptional performance in image classification tasks, their interpretability is frequently lacking, which makes it difficult to comprehend how they arrive at decisions. On the other hand, the LightGBM model obtains similar classification performance with improved precision, recall, and F1-score when enhanced with SHAP for explainability. This method not only yields competitive results but also has a distinct transparency advantage, which is vital in medical applications where

comprehension of the model's logic is just as crucial as prediction accuracy (Rezazadeh, Jafarian and Kord, 2022).

A popular XAI method called Gradient-weighted Class Activation Mapping (Grad-CAM) facilitates comprehension of deep learning models by highlighting the regions of the image that most affect a model's output. Grad-CAM produces heatmaps that indicate areas of the image where the model concentrates while making a diagnosis pertaining to medical imaging, such as thyroid ultrasound analysis. This makes it simpler for doctors to comprehend the AI's decision-making process.

An advanced Explainer model based on DenseNet-121 and Grad-CAM was compared using a dataset of 19,341 thyroid ultrasound images in the study. The Explainer improves doctors' diagnostic accuracy by offering more accurate and clinically relevant heatmaps, even though it uses Grad-CAM for initial training. Grad-CAM and the Explainer help reduce the "black box" character of model which is deep learning techniques by enabling medical personnel to evaluate AI predictions by visually identifying crucial regions linked with malignancy or benignity. This increases the validity and trustworthiness of AI-assisted diagnosis (Song et al., 2023).

The research focuses on the complex structure of brain CT lesions, which are scattered in 3D space and display morphological heterogeneity and discusses the difficulties of automatically generating reports for brain CT imaging. The authors suggest a hierarchical encoder-decoder network with inadequate guidance that uses Gradient-weighted Class Activation Mapping- Grad-CAM to improve report creation by making the more comprehensible model's assumptions. Grad-CAM is an essential component of this architecture; it produces attention maps that identify the most pertinent regions from the Brain CT scans. During the report production process, these maps are used to direct the model's attention towards possible lesion locations. Grad-CAM's spatial and frame attention maps, in particular, allow the model to recognise and concentrate on lesion regions over consecutive scans, accurately representing the continuous nature of brain CT lesions in three dimensions. By emphasising the most important areas of the scans, this specific focus helps the model produce more accurate and clinically relevant reports, which enhances the reports' interpretability and accuracy.

2.2 Critical Analysis of Existing Studies

Research on automatically generating medical reports has made tremendous progress, especially using encoder-decoder frameworks that combine text formation using recurrent neural networks (RNNs) and image processing using convolutional neural networks (CNNs). As an alternative, transformer-based models have surfaced, which provide better performance than conventional RNNs by better capturing contextual relationships. The absence of sufficient interpretability in AI-generated reports, which is essential for clinical adoption, persists despite these advancements.

To close these gaps, the current study uses Grad-CAM attention mapping in conjunction with a transformer-based model that is integrated into an encoder-decoder architecture for report generation. This methodology guarantees precise report production while augmenting comprehensibility through the graphic representation of the principal domains impacting the model's determinations. The **Table 1.1 Critical analysis/ Summary of the existing studies** below compares existing literature to the proposed model, highlighting methodological improvements. This critical analysis and comparison table shows the incremental improvements recommended in this current research, which emphasise progress in the accuracy, interpretability, and clinical application of AI-generated medical reports.

The **Table 1.1 Critical analysis/ Summary of the existing studies** contains a detailed critical analysis of the literature for a dissertation. It contains columns like Journal, which lists the study source cited in Harvard style, and Dataset, which specifies the datasets used, such as IU X-RAY, ultrasound images with or without reports, and clinical observations. The Image Type column shows the type of medical images analysed, while Encoder-Decoder Framework explains the architecture used, including multiple transfer learning techniques and hybrid models. The Comparison Model lists models for benchmarking. The Ablation Study displays whether different elements of the model were tested independently. Limitations describe weaknesses of the model.

Table 1.1 Critical analysis/ Summary of the existing studies

Journal	Dataset	Image Type	Encoder-Decoder Framework	XAI Technique	Comparison Model	Metric or Evaluation	Ablation Study	Limitations
(Jin <i>et al.</i> , 2024)	IU X-RAY, Chest X-ray MIMIC-CXR with report		ResNet-101-the dual-stream adaptive decoder (DSAD), medical concept detector (MCD), and disease detector (DD).	Attention Heat Map	R2GEN, PPKE, BLEU, and R2GENCMN.	ROUGE-L, METEOR, CE Metrics	Yes	Its low number of disease categories and partial coverage of certain conditions result in overfitting; the concept extraction is also imprecise.
(Wang <i>et al.</i> , 2023)	IU X-RAY, Chest X-ray MIMIC-CXR with report		R2GenGPT with Llama2-7B model-LLM	None	Att2in, Transformer, R2Gen, R2GenCMN	BLEU, METEOR, Rouge-L, and CIDEr,	Yes	Reliance on human annotations
(Rezazadeh, Jafarian Kord, 2022)	Public and dataset without report	Breast ultrasound images	LightGBM classification model	SHAP	VGG, ResNet, DenseNet	Accuracy, Precision, Recall, F1-score, and Area under the ROC Curve (AUC)	No	Expanding to bigger datasets without using ROI masks based on ground truth is limitation
(Singh <i>et al.</i> , 2019)	IU-CXR dataset with report	Chest X-ray	Google's Inception-v3 model –encoder & LSTM and GRU networks along with Glove and RadGlove-decoder	None	None	BLEU, METEOR, Rouge-L, and CIDEr,	Yes	Difficulty in accurately generating reports with abnormalities
(Song <i>et al.</i> , 2023)	2D ultrasound images without report	Ultrasound images	Explainer Model-DenseNet-121	Grad-CAM	None	AUROC, accuracy, sensitivity, specificity	Yes	Variability in the way an illness presents itself can lead to different patient subgroups with distinctive characteristics.

Table 1.1 Critical analysis/ Summary of the existing studies

Journal	Dataset	Image Type	Encoder-Decoder Framework	XAI Technique	Comparison Model	Metric or Evaluation Study	Ablation Study	Limitations
(Tsai, Chu and Li, 2023)	Pathology Education Resource (PEIR) with report	Ultrasound images of liver and gallbladder	Faster RCNN with LSTM, Weight Selection Mechanism	None	None	BLEU, METEOR, Rouge-L, and CIDEr,	No	Limited size of the test set
(D'Orsi <i>et al.</i> , 2013)	Brain CT Scan with report	Brain Computed Tomography	Weakly guided Gradient attention model (WGAM) and (KIRN) Network	Grad-CAM	CNN-RNN, Soft-ATT and Up-Down	BLEU, METEOR, Rouge-L, and CIDEr,	Yes	None Reported
(Nguyen <i>et al.</i> , 2020)	Breast cancer radiology report	Mammography and MRI images	Encoder-Decoder Attention Model (EDA)	None	None	BI RAD Score, ROUGE scores	Yes	Dataset in Dutch, diverse reporting styles, leading to potential inaccuracies in the BI-RADS score extraction
(Chaudhury, Sau and Shelke, 2024)	BUSI dataset, Breast Histology Dataset No report	Ultrasound images	ResNet with (Bidirectional Encoder representation from Image Transformers) and RNN-LSTM	None	ResNet-18, (Incremental Cascaded Extreme Learning Machine), HDGM (Hybrid Dilate Ghost Model)	Accuracy, Sensitivity, Precision, Recall	No	Lower classification performance of histology images
(Hekal <i>et al.</i> , 2024)	Breast Cancer Ultrasound Images (BUSI) dataset- No report	Breast Ultrasound images	The Dual Decoder and Attention Mechanism ResUNet (DDA-AttResUNet)	None	Enhanced Small Tumor-Aware Network (ESTAN), Residual Dilated Attention Gate UNet (RDAU-Net)	Dice coefficient, Jaccard index (IOU), sensitivity, precision, and overall accuracy.		need to test the method on other BUS datasets to assess its generalizability and potential for refinement

Table 1.1 Critical analysis/ Summary of the existing studies

Journal	Dataset	Image Type	Encoder-Decoder Framework	XAI Technique	Comparison Model	Metric or Evaluation Study	Ablation Study	Limitations
(Lanjewar, Panchbhai and Patle, 2024)	Kaggle Breast USIs Dataset -No Report	Ultrasound Images	MobileNetV2, ResNet50, and VGG16 with LSTM	Grad-CAM, LIME	Xception, NASNetMobile, InceptionResNetV2, Kappa MobileNetV2 ResNet50, and VGG16	F1 score, MCC, Kappa Coefficient, AUC, CI	No	Overfitting and computational complexity on diverse and larger ultrasound image datasets.
(Himel, Chowdhury and Hasan, 2024)	BUSI, Thammasat & UDIAT dataset No report	US images	ResNet50V2, NASNetLarge, and EfficientNetB7 with encoder decoder-based lesion segmentation-CycleGAN	None	None	Dice score, IoU score, Accuracy, F1-score, AUC score	No	Limited by the potential introduction of artifacts, a lack of explainability, 3D imaging integration.
(Tao et al., 2024)	IU X-RAY, MIMIC-CXR	Chest X-ray	Semantic alignment module (SAM)-based medical contrastive language-image pre-training (MedCLIP_Resnet)	None	R2GenCMN and XPRONET	BLEU, ROUGE-L, METEOR	Yes	Reliance on human annotations, lack of disease classification integration
(Alqahtani et al., 2024)	Private dataset and IU X-RAY	Chest X-ray	ConvNeXt (ResNext-ify) with BioBERT - transformer based language model (decoder)	None	GPT2, MiniLM	BLEU, ROUGE-L, METEOR, CIDER	NO	Limited dataset balance and need for better report generation.
(Zeng et al., 2024)	Fetal heart, Ultrasound, IU X-Ray and MIMIC-CXR	Chest X-ray, Ultrasound images	ResNet101 Attention-Enhanced Relational Memory Network (AERMNet) model	None	Faster-RCNN+LSTM, AdaAttn, AOANet, SFNet, R2Gen	BLEU, METEOR, Rouge-L, and CIDEr	Yes	Utilization of disease information is not sufficient results in deviation

Table 1.1 Critical analysis/ Summary of the existing studies

Journal	Dataset	Image Type	Encoder-Decoder Framework	XAI Technique	Comparison Model	Metric or Evaluation	Ablation Study	Limitations
Proposed Model	Breast Cancer Ultrasound Images (BUSI) dataset Clinical Report	Ultrasound images	ResNet101 Model-Feature Extractor Transformer Based Encoder-decoder Model-	Grad-CAM Attention Heat Map	RMAP, WGAM-KIRN, CNX-B2, R2GenGPT	BLEU, ROUGE-L, METEOR, CIDER	No	expanding the dataset and exploring advanced strategies like ensemble methods to enhance its applicability, specificity, and accuracy.

Chapter 3 Methodology

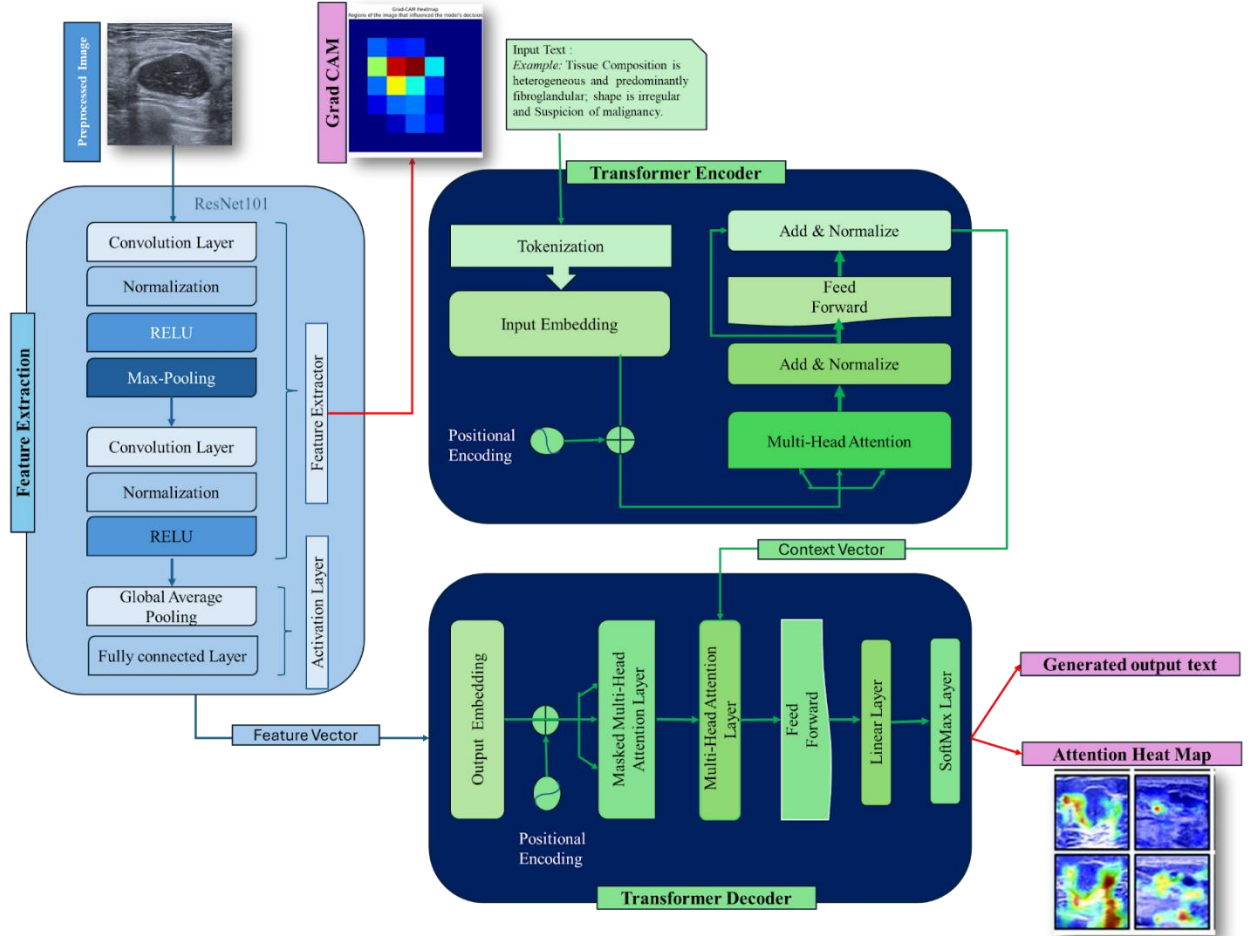


Figure 2.2 Architecture of the proposed system.

The model includes three primary components: a feature extractor, a transformer encoder, and a transformer decoder.

The method includes several critical components for processing ultrasound images and generating textual reports. Here's a full description of each component and how they interact: Initially, a pre-processed ultrasound image is fed into the feature extractor, which is a ResNet101 model. This extractor layer performs convolutional operations to extract high-level features from the image, followed by activation layers (e.g., ReLU) that introduce non-linearity, enhancing the feature representation. The output, a feature map, is visualized using Grad-CAM to highlight important regions of the image. These features

are then passed to the decoder, while the input text (ground truth) is tokenized, embedded, and processed by the encoder, which uses attention mechanisms to generate context vectors. The context vector, along with image features, is passed to the decoder, which generates the report text. The decoder also uses attention weights to produce an attention heatmap, visualizing which parts of the image are most relevant to the generated report.

3.1 Data Collection and Preprocessing

A radiologist possessing breast diagnostic imaging experience must adhere to the BI-RADS guidelines set forth by the American College of Radiology (ACR) to conduct a trustworthy breast ultrasound examination (D’Orsi et al., 2013). Machine learning models were created for various clinical applications in breast examinations, including classification, segmentation and automatic cancer detection of malignant and benign breast tumours.

3.1.1 Data collection

This study presents dataset of 256 expert-annotated breast ultrasound images. The dataset includes images of 154 benign tumours, 98 malignancies, and four normal breasts (Pawłowska et al., 2024). To ensure the dataset's universality, the dataset was assembled from images gathered by five radiologists at various medical centres throughout Poland between 2019 and 2022. Radiologists manually marked and tagged every image using a cloud-based system designed specifically for this purpose. The archive includes labels of the patient, image, and tumour levels and every tumour is verified by the findings of a biopsy or by follow-up care. Specifically, clinical data as patient-level labels were addressed in the initial phase of data collection such as tissue composition, age, symptoms and signs. To identify cancer and other abnormal areas in the image, the following phase was to add a freehand annotation at the image level.

3.1.2 BI-RADS Category

At the moment of the assessment, data were gathered from 256 adult female individuals, ages 18 to 87. There were 197 biopsies performed in all confirming 98 breast cancers representing 77% of the sample. The diagnoses for the BIRADS 3 category which is 12 cases are based on the biopsy results before the follow-up ultrasonography scan. Following this, the BIRADS descriptors and BIRADS category were labelled according to the tumour and imagery, and crucial findings were interpreted as the existence of tumour illnesses. The ultimate labels addressed verification methods, tumour categorisation, and histological diagnosis, with only 33 diagnoses occurring. In contrast to the openly accessible datasets outlined in the literature review (Hekal et al., 2024), (Yap et al., 2017), (Rodtook et al., 2018), this dataset has BIRADS features to support the BIRADS category, core needle biopsy findings, and annotations of multi-lesion images.

3.1.3 Ethical Approval and Consent

This study received ethical approval from the Lower Silesian Chamber of Medicine's Bioethics Committee (no. 2/BNR/2022). The initial clinically non-standard stage in creating the dataset was anonymising, collecting, and transferring the data. All identifying information and any text pertaining to patients within the image such as patient ID were permanently removed to protect patients' privacy. Prior to transfer, each anonymised image was examined in person to ensure that all data had been removed. To make this workflow easier, the study devised and implemented a custom web-based tool for gathering, labelling, and annotating images from breast ultrasounds.

3.1.4 Data labels and annotation

Clinical information on the patient, including age, the composition of the breast tissue, signs, and symptoms labels were obtained by BI-RADS recommendations. Image annotation was followed by labelling by the BI-RADS reporting standards. There were seven features based on B-mode. They were divided into two categories: image-oriented (calcifications, thickening of the skin) and mass-oriented (shape, margin, echogenicity,

posterior characteristics, halo). This BI-RADS categorizing segment concluded by designating one of the following seven groups: BI-RADS 1, 2, 3, 4a, 4b, 4c, and 5. The final set of descriptions pertains to the methods of tumor verification, including follow-up care, biopsy, histological diagnosis, and the eventual classification as benign or malignant. Table 1.2 Data characteristics for sample images shows examples of data characteristics, including all examined features, and data descriptions from the spreadsheet.

Table 1.2 Data characteristics for sample images

S.No	CSV file column name	Example 1	Example 2
1	CaseID	10	46
2	Image_filename	case010.png	case046.png
3	Mask_tumor_filename	case010_tumor.png	case046_tumor.png
4	Mask_other_filename	-	-
5	Pixel_size	0.01078	0.009197324
6	Age	34	73
7	Tissue_composition	not available	homogeneous: fat
8	Signs	not available	no
9	Symptoms	not available	not available
10	Shape	irregular	irregular
11	Margin	not circumscribed - angular&indistinct	not circumscribed - indistinct
12	Echogenicity	Hypoechoic	hypoechoic
13	Posterior_features	enhancement	no
14	Halo	yes	yes
15	Calcifications	no	no
16	Skin_thickening	no	no
17	Interpretation	Suspicion of malignancy&Dysplasia&Fi	Suspicion of malignancy

		broadenoma&Intraductal papilloma	
18	BIRADS	4c	5
19	Verification	confirmed by biopsy	confirmed by biopsy
20	Diagnosis	Invasive carcinoma of no special type (NST)&Ductal carcinoma in situ (DCIS)	Invasive carcinoma of no special type (NST)
21	Classification	Malignant	Malignant

3.1.5 Data preprocessing

The study proceeds with a basic resizing procedure to ensure that our input images are compatible with the ResNet101 model. ResNet101, like the other models in the ResNet family, is intended to accept input images measuring 224x224 pixels (Tao et al., 2024). The size is ideal for the model since the ImageNet dataset served as its pre-training source using this dimension, therefore resizing the images properly is critical for faster processing and improved performance. To prevent the network's older layers from overfitting to the new dataset, freezing the weights of those levels for transfer learning and setting their learning rates to zero is the best option. Freeze the network's weights while leaving the final learnable layer unfrozen. Additionally, data augmentation keeps the network from overfitting and from learning every detail of the training images. To make the training datasets more diverse and increase the generalisation ability of the model, such as random rotation, horizontal flipping, shifting, and zooming are used in image augmenters.

To generate a report, concatenate data from three dataset rows: tissue composition, shape, and interpretation. This combined text provides a thorough analysis description and can be used as input text to generate content using the transformer. Text normalisation is applied. Reduce text to lowercase to remove specified punctuation and special letters. Text vectorization uses a predetermined vocabulary to tokenise the text and turn tokens into integer sequences with a sequence length of 3000. Figure 2.3 Pre-processed Breast

Ultrasound image with respective text. shows the pre-processed sample image and its respective text.

3.2 ML/AI Model Development

In this study, building an image captioning model that generates textual descriptions (captions) from breast images from ultrasound. The architecture comprises a number of essential elements: a CNN (ResNet101) for image feature extraction, a Transformer encoder-decoder for processing these features and generating captions, and various layers and operations to facilitate data preprocessing and augmentation.

3.2.1 ResNet101 for Feature Extraction

The input images are processed using a deep convolutional neural network (CNN) called ResNet101 to extract advanced features. As part of the research by (He et al., 2016), which covers variants ranging from ResNet18 to more complex ones like ResNet152, ResNet101 was first presented in 2015. With 101 layers and residual connections, it mitigates the vanishing gradient issue and aids in the training of very deep networks. Since the purpose of ResNet101 is to use its feature maps rather than classify the images, the top fully connected layers are not employed. This method of feature extraction works well for shallow models (fewer layers), but for deep models (many layers), degradation effects make it hard to assess the original image in the lower layers (Zeng et al., 2024). To tackle the degradation issue, ResNet presents the residual learning framework. Except for the final few layers, which allow for fine-tuning while maintaining learnt features from the ImageNet dataset, the majority of ResNet101's layers are frozen. Convolutional layers, batch normalisation, ReLU activation, and residual connections are essential parts of resnet101 which is explained in the proposed diagram in Figure 2.2 Architecture of the proposed system..

Residual Block

The residual block permits gradients to pass through the shortcut connections directly, which aids in solving the vanishing gradient issue.

$$Output = F(x) + x$$

$F(x)$ is the residual function that is learned through convolutional layers. And X is the input to the residual block.

3.2.2 Transformer Encoder Block

The extracted image characteristics must be encoded by the Transformer encoder into a context-aware representation for captions to be generated. "Attention" is the Transformer model's central concept. The Multi-Head Attention model captures distinct aspects of the visual features by focusing on different portions of the image at the same time. By standardising inputs throughout the batch, Layer Normalisation enhances training stability. A fully connected, two-layer network is called the Feed-Forward Network (FFN) that transforms data in non-linear ways. Skip Connections facilitate gradient flow and deeper model training by adding the input to each block's output before sending it to the subsequent layer.

3.2.3 Transformer Decoder Block

By decoding the encoded image attributes and considering the previously generated words, the Transformer decoder generates captions. Positional Embedding is essential because transformers are not designed to comprehend order; it adds information about each word's location inside the sequence. The remaining layers are identical to the encoder. The last layer predicts the following word in the sequence and provides outputs as a probability distribution over the lexicon.

3.2.4 Self-Attention

A self-attention system that highlights or pulls the most significant information from the text input. The results passed through multiple linear layers after the positional encoding and embedding vector combination to generate the Q Query, K Key, and V Value vectors required to compute self-attention. The attention mechanism computation is represented by the following equation:

$$SelfAttention(Q, K, V) = \text{soft max} ((Q \cdot K^T / \sqrt{dk})) \cdot V$$

Large language models and multiheaded self-attention (MSA) layers can be assembled in Transformers with the help of multiplying the dot product of the Q and K vectors, which also improves computing efficiency.

3.3 Evaluation of the Proposed System

With this segment, outline the approach you will use to assess your system's performance. BLEU, METEOR, ROUGE L, and CIDER are the evaluation metrics used to gauge our suggested method's ability to provide radiological reports (Alqahtani et al., 2024). In literature, BLEU (Bilingual Evaluation Understudy) is utilized for language translation tasks and report creation tasks. Bleu evaluates the degree of n-gram overlap between the generated text and the original text; a higher overlap degree indicates a higher quality created text. A brevity penalty (BP) is incorporated into the calculation to compensate for the resulting text's length.

Metric for Evaluation of Translation with Explicit Ordering (Meteor) processes the link between synonyms and specific sequence matches using WordNet; this method correlates more with manual discrimination. The suggested text's longest shared subsequence with the actual text is determined by Recall-Oriented Understudy for Gisting Evaluation (Rouge-L), in which higher scores are obtained for longer lengths. The Consensus-based Image Description Evaluation-CIDEr metric is not frequently used in the creation of medical reports, but it was designed to reflect the human judgement of consensus in image captioning tasks. To determine how similar the generated text is to the original text, Cider calculates the cosine angle of each phrase's TF-IDF vector by treating each phrase as a document. The greater the score for these assessment indicators, the better the outcome. All the equation to derive the values of BLEU, METEOR, ROUGE L, and CIDER are listed in the **Table 1.3 Evaluation Metrics and their formulation**

Table 1.3 Evaluation Metrics and their formulation

Evaluation Metrics	Formula
	$Pr e cision = \frac{\sum_{i=1}^N Count_{i,n}}{\sum_{i=1}^N Total_{i,n}}$
BLEU	$BLEU = BP \cdot \exp \left(\sum_{n=1}^N w_n \log precision_n \right)$
	$Pr e cision = \frac{Number\ of\ matched\ unigrams}{Number\ of\ unigrams\ in\ generated\ text}$
METEOR	$Recall = \frac{Number\ of\ matched\ unigrams}{Number\ of\ unigrams\ in\ reference\ text}$
	$METEOR = Pr e cision \cdot Recall \cdot F_{mean} \cdot Penalty$
	$Pr e cision = \frac{LCS\ length}{generated\ length}$
ROUGE- L	$Recall = \frac{LCS\ length}{Reference\ length}$
	$ROUGE - L = \frac{(1 + \beta^2) \cdot Recall \cdot Pr e cision}{Recall + \beta^2 \cdot Pr e cision}$
	$TF - IDF_i = TF_i \cdot \log \left(\frac{N}{DF_i} \right)$
CIDEr	$CIDEr = \frac{1}{N} \sum_{i=1}^N \frac{TF - IDF_i \cdot Match_i}{Sum\ of\ TF - IDF}$

3.4 XAI Techniques

3.4.1 Grad CAM

Grad-CAM is used for this study to generate heat maps that show where lesions are located. The spatial attention is then directed by using the localisation heatmaps produced as the inadequate label for lesions by Grad-CAM (Yang et al., 2021). This heatmap to

identify possible model biases and to learn why the model predicted the data. The ResNet101 model must be altered for this implementation in order to extract gradients and feature maps. Grad-CAM can be used with ResNet-101 by first running the input image through the model to provide intermediate feature maps and predictions. Selecting a convolutional layer allows the activations to be seen. Compute the gradients of the output prediction about the feature maps in this layer, then utilise these gradients for assigning weights to the feature maps. Summing the weighted feature maps yields a heatmap. ReLU is then applied to highlight positive contributions. The heatmap is then superimposed over the original image to show the regions that had the most influence on the prediction. A model will be more interpretable and easier for clinicians to comprehend when Grad-CAM and attention maps are included. This can be especially helpful in medical imaging when determining which parts of an ultrasound image of the breast are most representative of a particular problem.

3.4.2 Attention Map

Attention maps are visualized during the text generation process to demonstrate how the model associates different areas of images with text descriptions. The Transformer decoder's attention mechanism, which illustrates how the model pays to various regions of the input feature maps while producing the text descriptions, naturally lends itself to interpretability. At each stage of processing the image and producing text, the Transformer determines the attention weights. These weights show how much emphasis the model puts on certain areas of the image while generating each description word.

Visualize the attention weights graphically to show which areas of the image are being prioritised at each stage of the text generation process. Higher or lower attention is indicated by darker or brighter spots on the map. This can aid in comprehending how the model creates connections between parts of the image and the associated text and providing a clearer understanding of how the model converts visual data into descriptive language.

Chapter 4 Experimental Results

This section provides results of preprocessing techniques, model performance, and evaluation metrics and compares our proposed method against baselines to test its effectiveness using the assumptions stated in the Methodology section.

4.1 Experimental Setup

4.1.1 Hardware

This exploratory setup generates an image captioning model using ResNet-101 and a transformer-based encoder-decoder architecture trained on Google Colab. Google Colab offers free access to GPUs, typically NVIDIA Tesla K80 or P100 with 12-16GB of VRAM. The GPU significantly speeds up training and inference, especially for deep networks such as ResNet-101 and Transformers. RAM has 13 GB, and disks have a size of 108 GB.

4.1.2 Software

Research used Python (3.11.8) to conduct our studies on the Microsoft Windows 10 operating system. The code is written in TensorFlow for developing and training the image captioning model, and Keras (High-level API) for the model architecture and training utilities, with data augmentation, Adam optimiser and custom learning rate scheduling. ResNet101 serves as the pre-trained CNN for image feature extraction. Custom layers for generating text include a Transformer Encoder and Decoder along with data augmentation and text vectorisation modules. Training takes place across 30 epochs with a batch size = 32 with Colab's GPU effectively utilised to manage complicated computations and substantial amounts of data. The model began the warmup phase with a starting learning rate of 0, increased it gradually by 0.06, and finally stabilised it at $1e-4$. With ``tf.data.AUTOTUNE``, you can create effective data pipelines and increase the performance and generalisation of the model. TensorFlow's dataset pipeline consists of loading images and captions, applying data augmentation and preprocessing, and finally

batching and shuffling the data with 'tf.data.Dataset' for efficient training. It optimises the data loading process by parallel processing and prefetching.

4.2 Dataset Description

4.2.1 Breast Ultrasound Images (BUS)

This study presents a collection of 256 ultrasound images with expert annotations from a research Centre in Poland using 5 different ultrasound scanners. In the process of preprocessing data, resize the breast ultrasound scan to 224 by 224 pixels and then apply normalisation and standardisation is seen in Figure 2.3 Pre-processed Breast Ultrasound image with respective text..

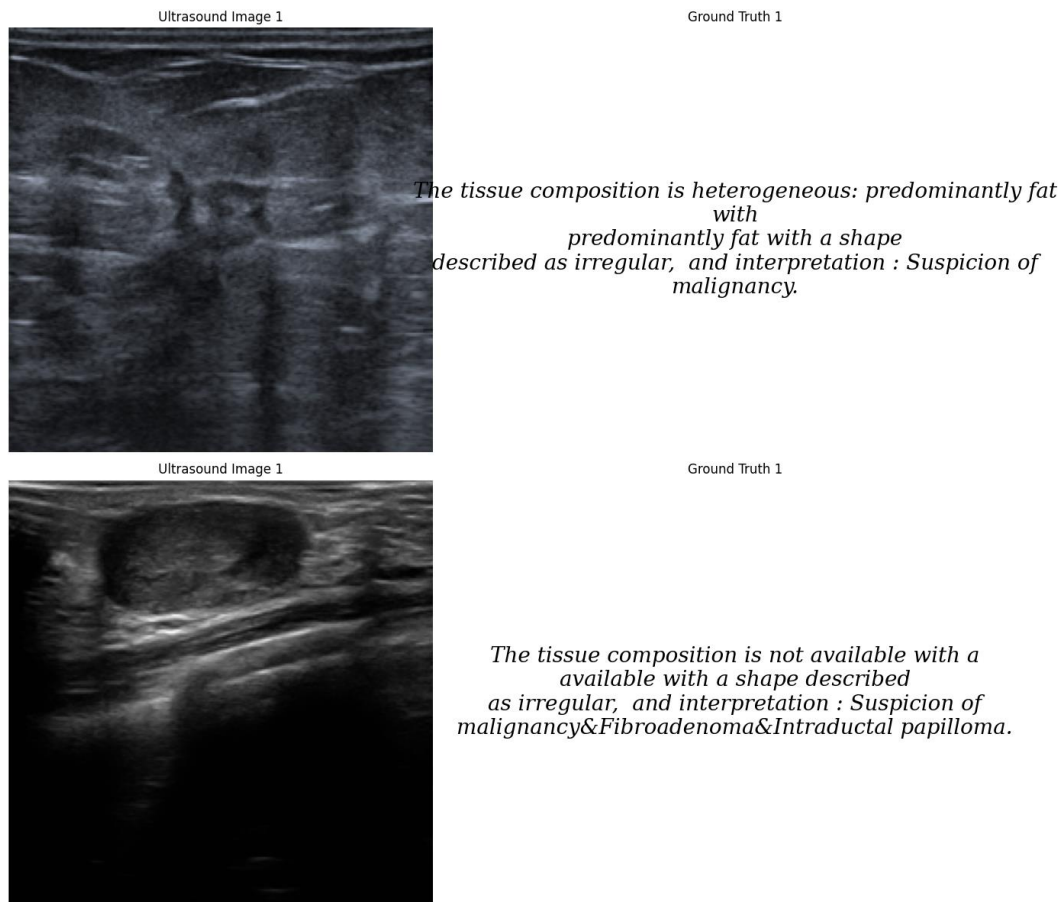


Figure 2.3 Pre-processed Breast Ultrasound image with respective text.

The study used the image data set, dividing the dataset into train, test, and validation sets that are 7:1:2 to provide an equal comparison. Every assessment is completed using the test set. A clinical description for each image is comprised of several sections: Shape, Tissue Composition, Interpretation, etc. In this work, the textual report is created by concatenating three important visual features identified for a breast ultrasound image.

4.2.2 Analysis of Textual Report

The text report has 256 captions and 256 keyword labels for the clinical description and keyword. A few dataset samples are displayed in Figure 2.3 Pre-processed Breast Ultrasound image with respective text. to help readers better comprehend our text dataset with accompanying images. The number of unique reports is 139 which is pretty easy to analyse. The distribution of ultrasound report length was analysed using the histogram which shows a mean of 152.44 as shown in Figure 2.4 Distribution of ultrasound report lengths..

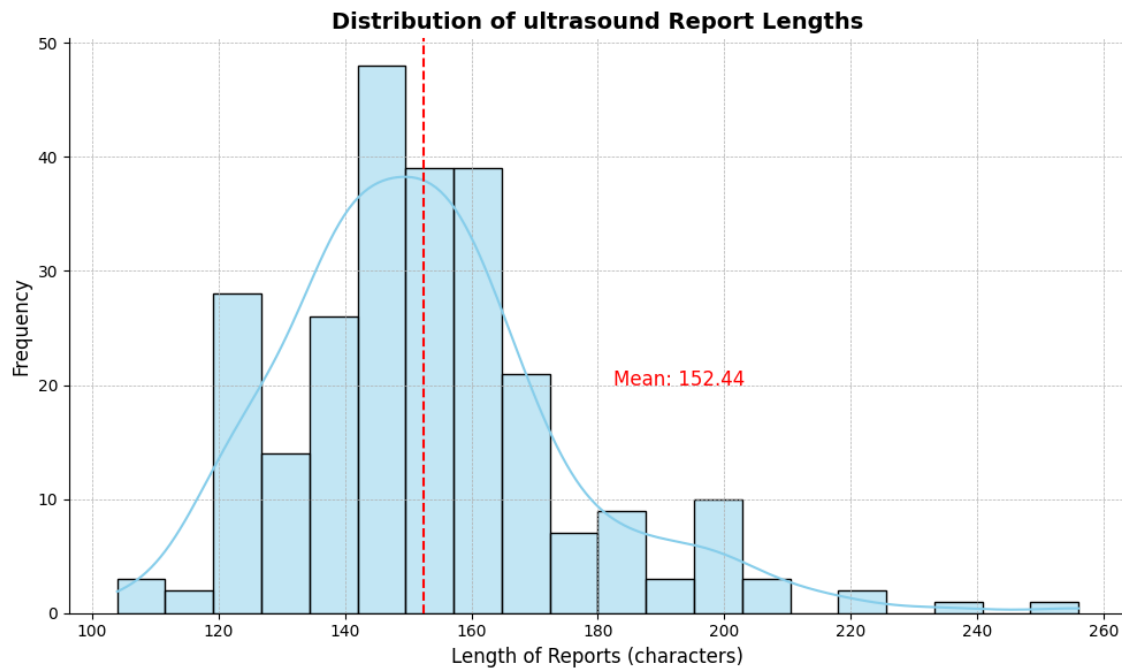


Figure 2.4 Distribution of ultrasound report lengths.

A mean value is indicated by the red dashed line on the histogram, which indicates the average report length's location within the distribution.

that the model generalises effectively to new data is observed in the Figure 2.6 Accuracy and validation accuracy of model performance. The model is not overfitting if there is little difference in the accuracy of training and validation. This is a positive result, indicating that the data augmentation and regularisation approaches were successful. The Resnet101 model can become more specialised in the particular task of generating reports by fine-tuning only the final few layers, which preserves the robustness of the pre-trained network. For sequential data, such as text, the use of a transformer-based architecture is essential. Long-range dependencies and contextual data within the sequence were captured by the model. Its capacity for giving attention to various input sequence segments aids in the production of reports that are coherent and accurate within their context. This indicates that despite the complicated content, the generated report is probably accurate and relevant.

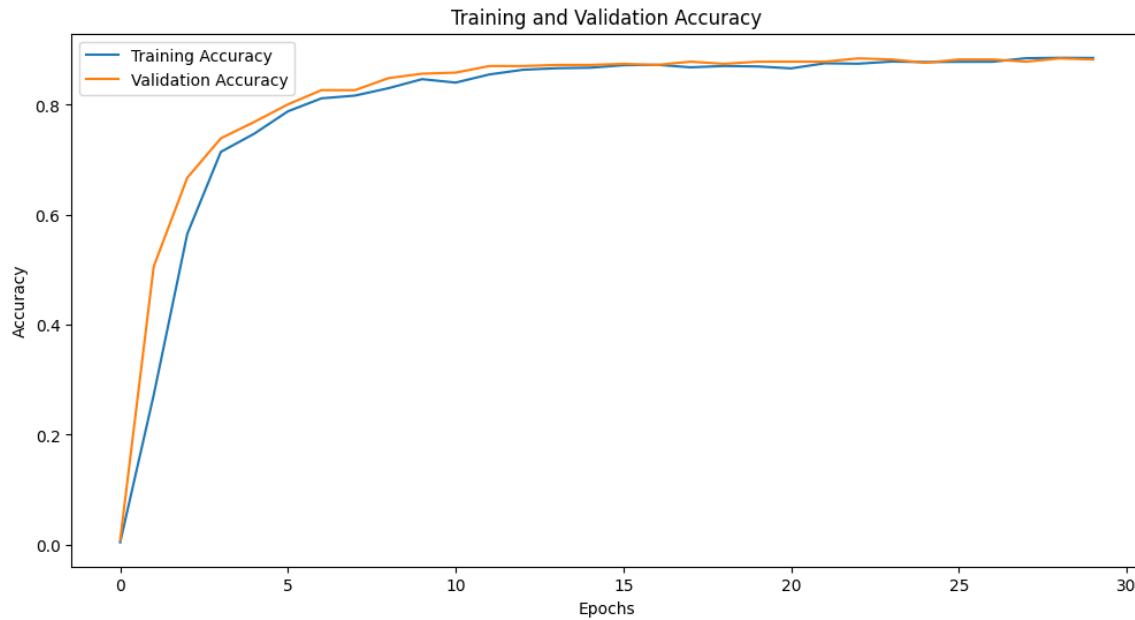


Figure 2.6 Accuracy and validation accuracy of model performance

By adjusting the learning rate from 0 to 0.001, limiting the size of gradients, with all these changes achieved higher accuracy, batch normalization or dropout is not used. Through systematic experimentation, these settings could be changed to get better outcomes. To improve the model's convergence to a better solution, try increasing the total epochs or modifying the learning rate schedule. Enhancing performance could also come from

assembling several models or investigating different network topologies. To evaluate the performance of our model against the common topologies for convolutional neural networks found in (Himel, Chowdhury and Hasan, 2024), which use a similar kind of dataset.

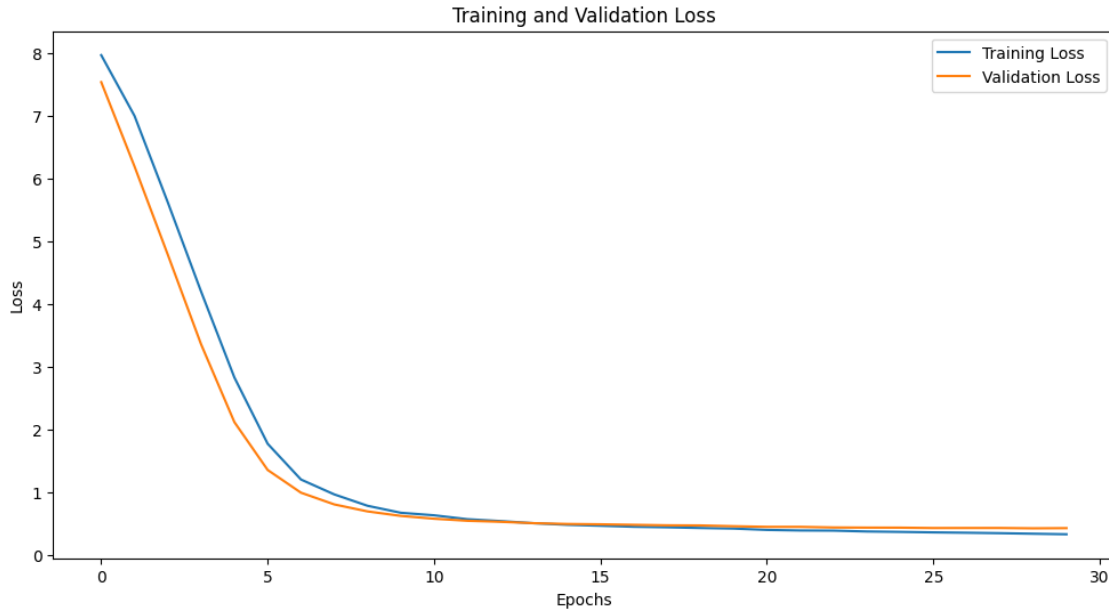


Figure 2.7 Training loss and validation Loss of model performance

4.3.2 Quantitative Results and Discussion

The acquired NLG scores demonstrate an adequate level of proficiency in producing precise and well-written reports. The BLEU scores show that the model performs well at capturing single words and some short phrases, but difficulties at capturing longer, more complicated sequences. These scores decline gradually from BLEU-1 (0.486) to BLEU-4 (0.134). The model produces relatively relevant information relatively relevant, but it lacks richness and specificity, as indicated by the CIDEr score of 0.203, which gauges consensus with reports that are produced by professionals

Table 1. 4 Experiment Results that combined all images, both normal and abnormal, according to the test sets BLEU-n (n=1,2,3,4), CIDEr (C), METEOR(M), and ROUGE(R) scores generated radiology reports.

Method	NLG Metrics						
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	C	M	R
Proposed Method	0.486	0.315	0.224	0.168	0.134	0.203	0.118

The model appears to have problems with semantic similarity, as evidenced by the METEOR score of 0.118, which is below optimal and indicates problems with synonymy and stemming. The generated reports' ROUGE score of 0.168 shows that there is potential for improvement, as they only partially match reference summaries in terms of overlapping phrases. Overall, these scores in the Table 1. 4 Experiment Results that combined all images, both normal and abnormal, according to the test sets BLEU-n (n=1,2,3,4), CIDEr (C), METEOR(M), and ROUGE(R) scores generated radiology reports. suggest that while the model can generate reports, the quality and detail are limited, emphasizing the need for further fine-tuning and possibly more diverse training data.

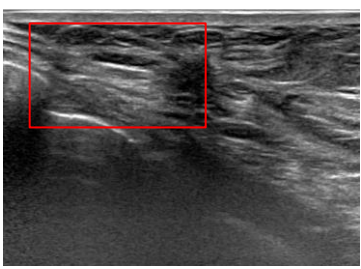
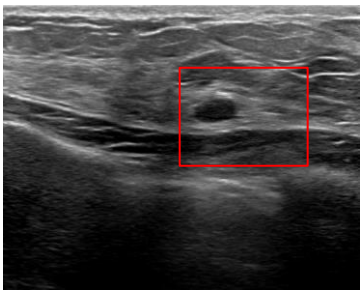
4.3.3 Qualitative Results & Discussion

Typically, the report produced by our proposed model closely resembles the reference report. The baseline model is not able to identify important abnormalities such as fibroglandular tissue in abnormal situations. Nonetheless, our suggested model might effectively draw attention to significant anomalies in addition to the ultrasound image's normal areas. The small discrepancies between the ground truth and proposed model reports illustrate the model's ability to capture crucial details, although with slight errors. For example, **Table 1.5 Examples of qualitative report generation from medical images.** shows the model successfully recognizes "*fibroglandular*" tissue and *irregular*

shapes, which are critical for diagnosis. However, it occasionally misidentifies "homogeneous" as "heterogeneous" or omits certain features, such as "predominantly fat." These inconsistencies indicate that, while the model excels at recognising essential traits, it may struggle with subtle contextual alterations, possibly due to training data limits or the model's ability to generalise complex descriptions across distinct situations.

Table 1.5 Examples of qualitative report generation from medical images.

Red label indicates the keyword term that will be shape or tissue composition

Image	Ground Truth	Proposed Model
	<i>the tissue composition is heterogeneous predominantly fat with a shape described as irregular and interpretation suspicion of malignancy.</i>	<i>the tissue composition is heterogeneous: predominantly fibroglandular with a shape described as irregular and interpretation suspicion of malignancy.</i>
	<i>the tissue composition is homogeneous: fibroglandular with a shape described as oval, and interpretation: Cyst filled with thick fluid</i>	<i>the tissue composition is heterogeneous: predominantly fibroglandular with a shape described as irregular, and interpretation: Cyst filled with thick fluid</i>

A breast ultrasonography image supplied into the framework is shown in the first column. The ground truth is displayed in the second column, with parts for shape, tissue composition, and interpretation concatenated. The report produced by our suggested framework is shown in the last column. Nevertheless, it occasionally misclassifies terms or ignores particular details, indicating difficulties in managing minute contextual differences. These differences imply that the model's sensitivity to subtleties in the training data and its capacity to generalise intricate descriptions both need to be strengthened. The suggested approach successfully captures important anomalies and normal regions, typically producing reports that nearly match reference reports.

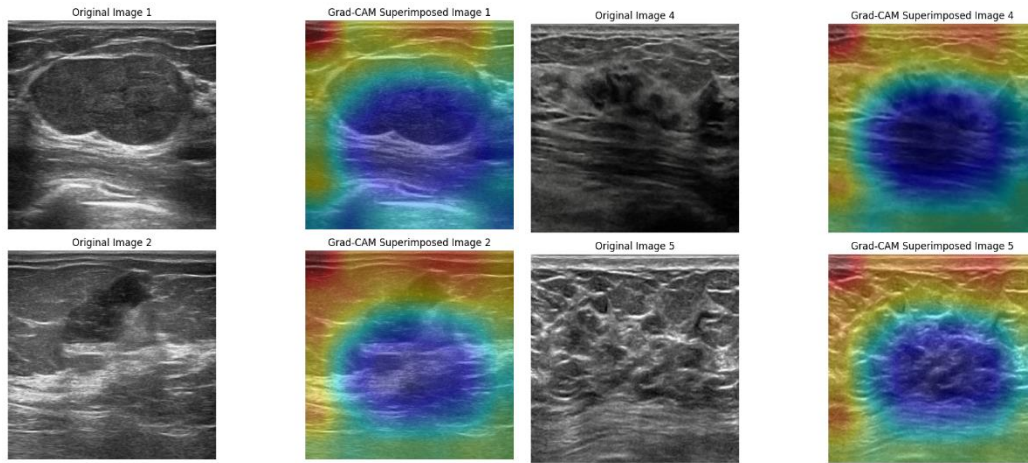


Figure 2.8 Heatmaps of Grad-CAM

Overlaying the Grad-CAM heatmap on the ultrasound images allows us to identify which regions the model focuses on when defining key traits in the report, such as "*fibroglandular*" tissue or "*irregular*" shapes. This visualization aids in the interpretation of the textual report by explicitly correlating visual aspects to the words used in the description, hence supporting the model's emphasis on clinically relevant areas. For example, if the Grad-CAM identifies locations with abnormal tissue composition and the report says, "suspicion of malignancy," it indicates that the model correctly correlates crucial visual signals with relevant medical terms. This correlation is observed in Figure 2.8 Heatmaps of Grad-CAM. According to (Song *et al.*, 2023), the correlation between Grad-CAM and the written report strengthens our belief in the interpretability of the model and its capacity to produce precise and pertinent medical reports. The Figure 9 Visualization of Attention heatmap a) Original grayscale image b) Attention Heatmap c) Overlayed image d) Combined Image displays four images: the original greyscale image of the breast cancer lesion, the associated attention heatmap that highlights areas of interest, the heatmap superimposed on the original image, and the combined image that highlights the locations with the highest concentration of attention. Understanding the model's focal areas in relation to the tumour or lesion is made easier by this visualisation. The areas of the input

image that the model considers important for generating predictions are highlighted in the attention map.

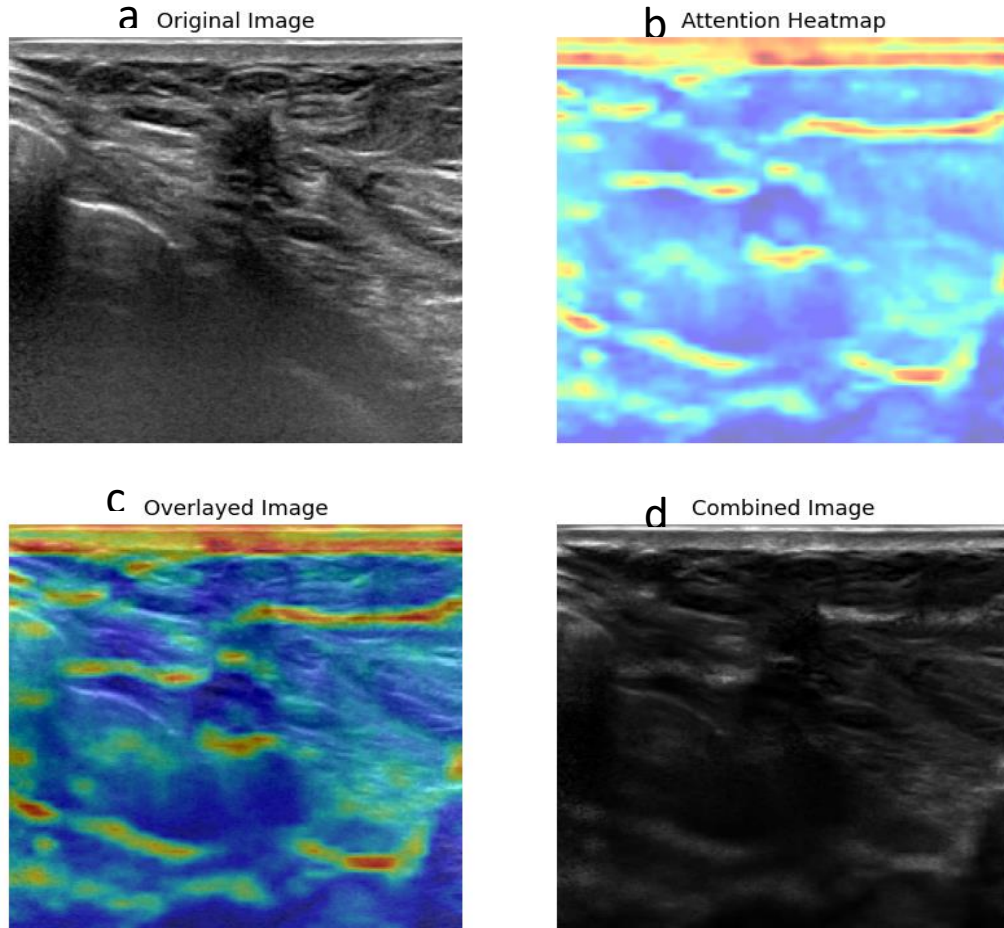
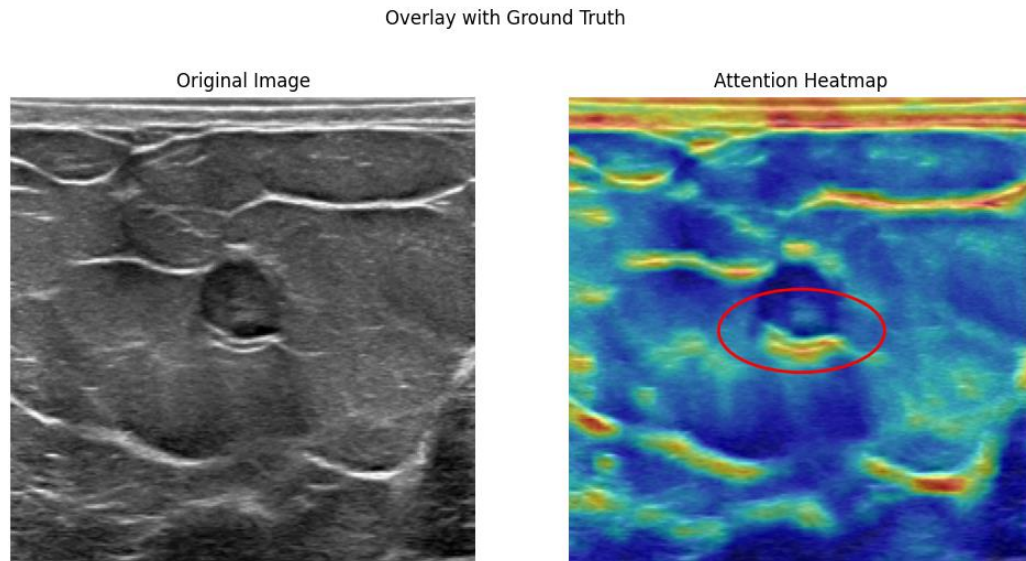


Figure 9 Visualization of Attention heatmap a) Original grayscale image b) Attention Heatmap c) Overlaid image d) Combined Image

High-intensity regions in the attention heatmap, which frequently line up with characteristics like tumour boundaries or lesions, show where the model is focussing like explained in (Jin *et al.*, 2024). The attention map and the original greyscale image are combined to create an overlay image that shows how the lesion's precise locations correspond to the model's focus. The combined image highlights areas of strong attention by multiplying the original image by the attention map; this reveals the exact locations that affect the model's conclusion. Transparency and reliability in clinical contexts are

improved by this visualisation, which makes it easier to understand the behaviour of the model and guarantees that it concentrates on pertinent features. As opposed to Grad-CAM, which highlights areas impacting the final prediction through the use of class activation maps, attention maps directly represent the internal focus of the model throughout processing, potentially indicating distinct or more detailed regions of interest.



The tissue composition is homogeneous: fat with a shape described as oval, and interpretation: Dysplasia & Fibroadenoma.

Figure 10 Overlaying attention heatmap image with ground truth

The traits that the attention map highlights are contextualised by the ground truth text, which details the diagnosis and composition of the tissue. The algorithm visually correlates the individual properties of the tumour with the focus of the model by indicating its location and form with an ellipse is shown in Figure 10 Overlaying attention heatmap image with ground truth. This alignment improves interpretability by confirming whether the model's focus is on clinically significant elements. By making it evident where areas of the image affect the model's predictions and how they correspond with expert comments, this type of visualisation aids in the generation of comprehensive reports.

4.4 Comparison with Baseline Methods

Our suggested ResNet-101-Transformer model's efficacy was assessed by contrasting its NLG metrics performance with several baseline techniques is shown in Table 1.6 Comparison of NLG metrics with baseline methods.

Our model scored 0.486 on BLEU-1, which is much higher than RMAP (0.416) and competitive with WGAM-KIRN (0.488) and R2GenGPT (0.488). This suggests that the baseline approaches may do a better job of capturing the text's short-term dependence. Our model is outperformed by Yang *et al.*, 2021 and Alqahtani *et al.*, 2024 in these metrics, probably because of their more advanced decoding strategies, such KIRN's interactive recurrent network and WGAM's weakly guided attention. Our model scores 0.134 in CIDEr and 0.203 in METEOR, when looking at the CIDEr (C) and METEOR (M) scores, which are more concerned with the content and semantic alignment. R2GenGPT outperforms our model in CIDEr (0.438) but lags in METEOR (0.211). This indicates that even while the text generated by our model is content-wise slightly aligned with the reference, there is still room for growth in terms of reaching higher semantic correctness.

Table 1.6 Comparison of NLG metrics with baseline methods

Journal	Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	C	M	R
(Jin et al., 2024)	RMAP	0.416	0.267	0.190	0.145	-	0.161	0.303
(Yang et al., 2021)	WGAM-KIRN	0.488	0.373	0.300	0.246	0.210	0.298	0.433
(Alqahtani et al., 2024)	CNX-B2	0.445	0.409	0.389	0.375	-	-	-
(Wang et al., 2023)	R2GenGPT	0.488	0.316	0.228	0.173	0.438	0.211	0.377
Our Proposed Model	ResNet-101 – Transformer	0.486	0.315	0.224	0.168	0.134	0.203	0.118

However, compared to Jin et al., 2024 and Wang et al., 2023, our model performs better, producing reports that are more accurate and diversified. RMAP, in particular, performs poorly on several measures. Our ResNet-101-Transformer technique is competitive given the complexity and diversity of the medical report creation task, especially since it does not rely on specialised components such as disease detectors (RMAP) or weakly guided attention (WGAM). The equilibrium between performance and complexity suggests that, with opportunity for greater optimisation, our suggested model provides a reliable but effective substitute for more intricate baseline techniques.

The main drawbacks of this work are the model's potential limited generalisability to a variety of clinical circumstances due to its dependence on a homogeneous and small dataset. The lower CIDEr and METEOR scores suggest that the model also has difficulty representing intricate and nuanced medical descriptions. Additionally, by implementing more sophisticated strategies like ensemble methods or investigating various network designs to improve accuracy and specificity, the performance of the suggested model could be further improved.

Chapter 5 Conclusion and Future Work

5.1 Conclusion

In the current research, investigated how to generate radiology reports from breast ultrasound pictures (BUS) using sophisticated deep-learning techniques. Despite the profusion of improved technology for chest X-ray interpretation, breast cancer detection, particularly through automated report development, is still immature and difficult to obtain. Breast cancer reports are vital since early and precise detection is critical to saving lives; nonetheless, many women still die as a result of undiscovered tumours. My approach seeks to close this gap by using clinical observations and visual details from breast ultrasound images (BUS) to generate reports with less specific input. By simplifying and automating breast cancer diagnosis, this model hopes to make the process more accessible, potentially boosting early detection and lowering mortality rates.

In particular, study suggested a hybrid model that combines an encoder-decoder architecture based on transformers with a convolutional neural network based on ResNet101. This technique tackles the difficulties radiologists encounter when interpreting ultrasound images, which can lead to errors, particularly since manual processing is time-consuming and prone to human bias. Our method leveraged deep learning capabilities to increase productivity and accuracy while addressing the difficulties radiologists have in interpreting ultrasound pictures and producing reliable diagnostic reports.

5.1.1 Key Contributions and Findings

Our hybrid model captures and translates visual data from breast ultrasound images into coherent medical reports by integrating a transformer-based Encoder-Decoder with a CNN (ResNet101). The Transformer-based architecture processes images after the CNN has successfully extracted pertinent visual features, producing reports that are both contextually relevant and in-depth. The transformer model's attention processes aid in ensuring the accuracy and significance of the reports that are produced. The study used a selective dataset of 256 ultrasound scans of the breast to test our model. Clinical descriptions covering fundamental elements like shape, tissue composition, and

interpretation were tagged into the dataset. With competitive scores of 0.203 METEOR, 0.118 CIDEr, and 0.486 BLEU-1, our model demonstrated a respectable degree of competence in producing descriptive reports.

Qualitative outcomes demonstrate that our approach can provide reports that capture important characteristics of ultrasound images of the breast, like the shape and composition of the tissue. A minor difference demonstrates the model's ability to identify important characteristics, but they also point to the need for additional improvement in order to better manage minute contextual alterations. Our model becomes more transparent and interpretable when explainable artificial intelligence (XAI) tools are integrated into it. The study used gradient-weighted class activation mapping which is Grad-CAM and attention maps to show which areas of the ultrasound images affected the reports that were produced. Grad-CAM heatmaps showed that the model generates text descriptions with an emphasis on clinically relevant areas, e.g., aberrant tissue compositions. Attention maps helped to clarify the model's decision-making process by illuminating how various image components relate to particular textual pieces.

5.1.2 Limitations

Although our model performs well, it has several drawbacks. The training and evaluation datasets are small and consistent in size may restrict the applicability of the model to a wider range of clinical circumstances. Future research could concentrate on expanding the dataset to include a greater range of scenarios, improving the model's capacity to handle complex descriptions, in order to overcome these limitations. Furthermore, investigating cutting-edge strategies like ensemble approaches or other network topologies might enhance specificity and accuracy. Using clinical observations as the ground truth for radiology reports has disadvantages, particularly in the absence of comprehensive findings and impression columns, which are generally required in such reports. Clinical observations alone may lack the specificity and depth required to gather the complete spectrum of diagnostic information. Findings describe essential information acquired from medical images, whereas an impression provides a brief diagnostic conclusion that frequently guides subsequent treatment. Without these critical components,

generated reports risk being incomplete or erroneous, perhaps leading to misdiagnoses or insufficient treatment regimens. Thus, depending entirely on clinical observations may jeopardise the report's quality and clinical usefulness.

5.2 Future Work

To summarise, the hybrid model have proposed marks a noteworthy advancement in utilising deep learning to generate radiology reports automatically from breast ultrasound images. To guarantee the model's successful integration into clinical practice, future research should concentrate on boosting its interpretability, broadening its applicability by expanding the model's applicability to more clinical situations and enhancing its functionality.

- Gathering a more comprehensive and varied collection of breast ultrasound images to strengthen the model's resilience and capacity for generalisation.
- Explore newer transformer architectures, such as GPT-based models or Vision Transformers (ViTs), and see if they offer greater interpretability or performance by doing experiments with Advanced Transformers.
- To learn more about the process of making decisions of the model, try experimenting with additional methods for explainable AI, such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). Develop models that can do multiple related tasks at once, such as lesion categorisation, severity evaluation, and prognosis prediction, in order to give more comprehensive diagnostic aid.

References

Books

1. D’Orsi, C., Sickles, E., Mendelson, E. and Morris, E. (2013) *ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System*. Reston, VA: American College of Radiology.
2. World Health Organization. (2023) *Global Breast Cancer Initiative Implementation Framework: Assessing Strengthening and Scaling-up of Services for the Early Detection and Management of Breast Cancer*. Geneva: World Health Organization.

Journals

1. Ahmed, S. bin, Solis-Oba, R. and Ilie, L. (2022) 'Explainable-AI in automated medical report generation using chest X-ray images', *Applied Sciences*, 12(22). doi: 10.3390/app122211750.
2. Alqahtani, F.F., Mohsan, M.M., Alshamrani, K., Zeb, J., Alhamami, S. and Alqarni, D. (2024) 'CNX-B2: A novel CNN-transformer approach for chest X-ray medical report generation', *IEEE Access*, 12, pp. 26626-26635. doi: 10.1109/ACCESS.2024.3367360.
3. Bai, S., Nasir, S., Khan, R.A., Arif, S., Meyer, A. and Konik, H. (2024) 'Breast cancer diagnosis: A comprehensive exploration of explainable artificial intelligence (XAI) techniques', *IEEE Access*. doi: 10.1109/ACCESS.2024.3367360.
4. Berlin, L. (2001) 'Defending the "missed" radiographic diagnosis', *American Journal of Roentgenology*, 176(2), pp. 863–867.
5. Biswas, A.A. (2024) 'A comprehensive review of explainable AI for disease diagnosis', *Array*, 22. doi: 10.1016/j.array.2024.100345.
6. Carriero, A., Groenhoff, L., Vologina, E., Basile, P. and Albera, M. (2024) 'Deep learning in breast cancer imaging: State of the art and recent advancements in early 2024', *Diagnostics*, 14(8). doi: 10.3390/diagnostics14080848.

7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. *et al.*, 2024. An image is worth 16x16 words: Transformers for image recognition at scale.
8. Gurmessa, D.K. and Jimma, W. (2024) 'Explainable machine learning for breast cancer diagnosis from mammography and ultrasound images: a systematic review', *BMJ Health and Care Informatics*, 31(1). doi: 10.1136/bmjhci-2023-100954.
9. Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M. and Hussain, A. (2024) 'Interpreting black-box models: A review on explainable artificial intelligence', *Cognitive Computation*, 16(1), pp. 45–74. doi: 10.1007/s12559-023-10179-8.
10. Hekal, A.A., Elnakib, A., Moustafa, H.E.D. and Amer, H.M. (2024) 'Breast cancer segmentation from ultrasound images using deep dual-decoder technology with attention network', *IEEE Access*, 12, pp. 10087–10101. doi: 10.1109/ACCESS.2024.3351564.
11. Himel, M.H.A.M.H., Chowdhury, P. and Hasan, M.A.M. (2024) 'A robust encoder-decoder based weighted segmentation and dual-staged feature fusion based meta-classification for breast cancer utilizing ultrasound imaging', *Intelligent Systems with Applications*, 22. doi: 10.1016/j.iswa.2024.200367.
12. Jin, Y., Chen, W., Tian, Y., Song, Y. and Yan, C. (2024) 'Improving radiology report generation with multi-grained abnormality prediction', *Neurocomputing*. doi: 10.1016/j.neucom.2024.128122.
13. K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.
14. Lanjewar, M.G., Panchbhai, K.G. and Patle, L.B. (2024) 'Fusion of transfer learning models with LSTM for detection of breast cancer using ultrasound images', *Computers in Biology and Medicine*, 169.
15. Messina, P., Pino, P., Parra, D., Soto, A., Besa, C., Uribe, S., Andía, M., Tejos, C., Prieto, C. and Capurro, D. (2022) 'A survey on deep learning and explainability for

- automatic report generation from medical images', *ACM Computing Surveys*, 54(10S). doi: 10.1145/3522747.
16. Pawłowska, A., Ćwierz-Pieńkowska, A., Domalik, A. *et al.*, 2024. Curated benchmark dataset for ultrasound-based breast lesion analysis. *Scientific Data*, 11, p. 148. doi: 10.1038/s41597-024-02984-z.
 17. Pang, T., Li, P. and Zhao, L. (2023) 'A survey on automatic generation of medical imaging reports based on deep learning', *BioMedical Engineering Online*, 22(1). doi: 10.1186/s12938-023-01113-y.
 18. Patrício, C., Neves, J.C. and Teixeira, L.F. (2023) 'Explainable deep learning methods in medical image classification: A survey', *ACM Computing Surveys*, 56(4). doi: 10.1145/3625287.
 19. Rezazadeh, A., Jafarian, Y. and Kord, A. (2022) 'Explainable ensemble machine learning for breast cancer diagnosis based on ultrasound image texture features', *Forecasting*, 4(1), pp. 262–274. doi: 10.3390/forecast4010015.
 20. Raminedi, S., Shridevi, S. and Won, D., 2024. Multi-modal transformer architecture for medical image analysis and automated report generation. *Scientific Reports*, 14, p. 19281. doi: 10.1038/s41598-024-69981-5.
 21. Rodtook, A., Kirimasthong, K., Lohitvisate, W. and Makhanov, S.S. (2018) 'Automatic initialization of active contours and level set method in ultrasound images of breast abnormalities', *Pattern Recognition*, 79, pp. 172–182.
 22. Song, D., Yao, J., Jiang, Y., Shi, S., Cui, C., Wang, L., Wang, L., Wu, H., Tian, H., Ye, X., Ou, D., Li, W., Feng, N., Pan, W., Song, M., Xu, J., Xu, D., Wu, L. and Dong, F. (2023) 'A new xAI framework with feature explainability for tumors decision-making in ultrasound data: comparing with Grad-CAM', *Computer Methods and Programs in Biomedicine*, 235. doi: 10.1016/j.cmpb.2023.107527.
 23. Tao, Y., Ma, L., Yu, J. and Zhang, H. (2024) 'Memory-based cross-modal semantic alignment network for radiology report generation', *IEEE Journal of Biomedical and Health Informatics*, 28(7), pp. 4145-4156. doi: 10.1109/JBHI.2024.3393018.

24. Wang, Z., Liu, L., Wang, L. and Zhou, L. (2023) 'R2GenGPT: Radiology report generation with frozen LLMs', *Meta-Radiology*, 1(3), p. 100033. doi: 10.1016/j.metrad.2023.100033.
25. Yap, M.H., Pons, G., Marti, J., Ganau, S., Sentis, M., Zwiggelaar, R., Davison, A.K. and Marti, R. (2017) 'Automated breast ultrasound lesions detection using convolutional neural networks', *IEEE Journal of Biomedical and Health Informatics*, 22, pp. 1218–1226.
26. Zeng, X., Liao, T., Xu, L. and Wang, Z. (2024) 'AERMNet: Attention-enhanced relational memory network for medical image report generation', *Computer Methods and Programs in Biomedicine*, 244. doi: 10.1016/j.cmpb.2023.107979.

Conference Technical Articles

1. Chaudhury, S., Sau, K. and Shelke, N. (2024) 'Transforming breast cancer image classification with vision transformers and LSTM integration', in *Proceedings of the 2024 IEEE International Students' Conference on Electrical Electronics and Computer Science (SCEECS 2024)*. doi: 10.1109/SCEECS61402.2024.10482221.
2. Jing, B., Xie, P. and Xing, E.P. (2018) 'On the automatic generation of medical imaging reports', in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 1, pp. 2577–2586. doi: 10.18653/v1/p18-1240.
3. Nguyen, E., Theodorakopoulos, D., Pathak, S., Geerdink, J., Vijlbrief, O., van Keulen, M. and Seifert, C. (2020) 'A hybrid text classification and language generation model for automated summarization of Dutch breast cancer radiology reports', in *Proceedings of the 2020 IEEE 2nd International Conference on Cognitive Machine Intelligence (CogMI 2020)*, pp. 72–81. doi: 10.1109/CogMI50398.2020.00019.
4. Shin, H.C., Roberts, K., Lu, L., Demner-Fushman, D., Yao, J. and Summers, R.M. (2016) 'Learning to read chest X-rays: Recurrent neural cascade model for automated image annotation', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2497–2506.

5. Singh, S., Karimi, S., Ho-Shon, I. and Hamey, L. (2019) 'From chest X-rays to radiology reports: A multimodal machine learning approach', in *Proceedings of the 2019 Digital Image Computing: Techniques and Applications (DICTA)*, Perth, Australia, pp. 1-8. doi: 10.1109
6. Tsai, M.C., Chu, K.C. and Li, Y.X. (2023) 'Building and validating a clinical ultrasound image reporting model', in *Proceedings of the 2023 IEEE 24th International Conference on Information Reuse and Integration for Data Science (IRI 2023)*, pp. 96–100. doi: 10.1109/IRI58017.2023.00024.
7. Yang, S., Ji, J., Zhang, X., Liu, Y. and Wang, Z. (2021) 'Weakly guided hierarchical encoder-decoder network for brain CT report generation', in *Proceedings of the 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Houston, TX, USA, pp. 568-573. doi: 10.1109/BIBM52615.2021.9669626.

Online Sources

1. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. *et al.* (2024) *An image is worth 16x16 words: Transformers for image recognition at scale*. Available at: <http://arxiv.org/abs/2406.00532> (Preprint).
2. WHO. (2023) *Breast Cancer*. Available at: <https://www.who.int/news/item> (Accessed: 21 January 2023)