

A Multimodal Deep Learning Framework for Thyroid Cancer Diagnosis Using Ultrasound Imaging and Clinical Data

A report submitted in partial fulfilment of the requirements

for the award of the degree of

B.Tech Computer Science Engineering

by

M. Gopi Chakradhar
(Roll No: 121CS0050)

K. Rohith
(Roll No: 121CS0045)

Under the Guidance of
Dr. N. Srinivas Naik



DEPARTMENT OF COMPUTER SCIENCE ENGINEERING
INDIAN INSTITUTE OF INFORMATION TECHNOLOGY
DESIGN AND MANUFACTURING KURNOOL

April 2025

Abstract

This report presents a multimodal deep learning framework for thyroid cancer diagnosis by integrating ultrasound imaging and clinical data. The primary objective of the study is to improve diagnostic accuracy, reduce operator bias, and enhance risk stratification in thyroid nodules. Our approach combines advanced computer vision techniques, specifically Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs), with clinical parameters such as thyroid function tests and patient demographics.

The imaging module leverages semantic segmentation and classification capabilities using architectures like multi-scale adaptive CNNs to detect subtle features in ultrasound scans. Parallely, the clinical data module employs machine learning models trained on structured patient data, including FT3, FT4, TSH, and miRNA signatures, to predict malignancy risk. A hybrid fusion model integrates both modalities, supported by a dynamic gating mechanism for optimal feature alignment.

Additionally, the system incorporates a natural language processing (NLP) component to generate structured diagnostic summaries using transformer-based models like BART, making the results clinically interpretable and ready for deployment.

The proposed framework was evaluated on benchmark datasets and research-grade annotated scans, achieving a test accuracy of 85.12%, an AUC of 0.91, and a significant reduction in false negatives. The findings validate the viability of AI-powered multimodal systems in enhancing clinical workflows for thyroid cancer screening and decision support, demonstrating improved risk stratification and diagnostic confidence.

Acknowledgements

We would like to express our sincere gratitude to Dr. N. Srinivas Naik, our guide and Head of the Department of Computer Science and Engineering at IIITDM Kurnool, for his invaluable support, guidance, and encouragement throughout this project. His expertise in deep learning and computer vision was instrumental in shaping our approach and enhancing our understanding of the field.

We would also like to thank the faculty and staff of the Department of Computer Science and Engineering at IIITDM Kurnool for providing a conducive environment for research and learning. Our thanks extend to our co-developers and colleagues for their helpful discussions and insights, which enriched our work. Finally, we appreciate the support from our friends, whose encouragement kept us motivated throughout this journey.

Contents

Declaration	i
Evaluation Sheet	ii
Abstract	iii
Acknowledgements	iv
Contents	v
List of Figures	viii
List of Tables	ix
Abbreviations	x
Symbols	xi
1 Introduction	1
1.1 Background and Motivation	1
1.2 Problem Statement	2
1.3 Objectives of the Project	4
1.3.1 Integration of Ultrasound Imaging and Clinical Data	4
1.3.2 Enhanced Diagnostic Support through Multimodal Fusion	5
1.3.3 Automated Diagnostic Report Generation Using NLP	5
1.4 Contributions and Scope	5
1.4.1 Novel Hybrid CNN-ViT Imaging Architecture	5
1.4.2 Clinical Data Integration and Fusion Strategy	6
1.4.3 NLP-Based Automated Report Generation	6
1.5 Thesis Organization	6

2	Literature Review	8
2.1	Imaging Techniques in Thyroid Cancer	8
2.1.1	Ultrasound Imaging: Applications and Limitations	8
2.1.2	Advances in Deep Learning for Medical Imaging	8
2.2	Clinical Data in Thyroid Cancer Diagnosis	9
2.2.1	Role of Biomarkers and Clinical Parameters	9
2.2.2	Machine Learning Approaches for Clinical Risk Assessment	9
2.3	Multimodal Data Fusion	9
2.3.1	Overview of Fusion Techniques	9
2.3.2	Challenges and Gaps in Existing Multimodal Systems	10
2.4	NLP in Medical Report Generation	10
2.4.1	Overview of NLP Models in Healthcare	10
2.4.2	BART vs. BERT for Summarization	10
2.5	Summary of Related Work	11
3	Methodology	12
3.1	Data Acquisition and Preprocessing	12
3.1.1	Imaging Datasets: Dataset1, Dataset2, and Dataset3	12
3.1.2	Clinical Dataset: Thyroid Clean CSV	12
3.1.3	Data Cleaning, Transformation, and Augmentation	13
3.2	Model Architecture	14
3.2.1	Hybrid Imaging Module (CNN-ViT Fusion)	14
3.2.1.1	CNN Branch for Local Feature Extraction	14
3.2.1.2	ViT Branch for Global Context Extraction	15
3.2.1.3	Hybrid Integration and Feature Fusion	15
3.2.2	Clinical Data Module	15
3.2.2.1	Clinical Feature Extraction Using MLP	15
3.2.3	Fusion Module for Multimodal Integration	16
3.2.3.1	Dynamic Gating Mechanism	16
3.2.3.2	Multi-task Outputs: Classification and Risk Prediction	17
3.3	Training Strategy	17
3.3.1	Individual Module Training and Fine-Tuning	17
3.3.2	End-to-End Multimodal Training Pipeline	17
3.3.3	Hyperparameter Tuning and Optimization	18
3.4	NLP Summarization Phase	18
3.4.1	Input Acquisition: Ultrasound Images and Clinical Data	18
3.4.2	Prompt Generation for Diagnostic Reporting	18
3.4.3	Implementation of BART for Text Summarization	18
3.4.4	Output Format: JSON Structured Diagnostic Report	18
4	Experimental Evaluation	20
4.1	Experimental Setup	20
4.1.1	Model Training Phases and Performance Analysis	20
4.1.2	Hardware and Software Configuration	22

4.1.3	Dataset Splits and Preprocessing Details	22
4.2	Quantitative Evaluation	23
4.2.1	Evaluation Metrics	23
4.2.2	Confusion Matrices and Error Analysis	24
4.3	Qualitative Analysis	24
4.3.1	Visualization of Misclassified Cases	24
4.3.2	Risk Prediction Consistency Analysis	24
4.4	Comparative Study	25
4.4.1	Unimodal vs. Multimodal Evaluation	25
4.4.2	Discussion of Results	25
5	NLP Summarization and Report Generation	26
5.1	Overview of NLP Techniques in Medical Diagnostics	26
5.2	Input Data Preparation for NLP	26
5.2.1	Preprocessing of New Ultrasound Images	26
5.2.2	Clinical Data Entry and Integration	27
5.3	Prompt Generation for BART Summarization	27
5.3.1	Constructing a Structured Prompt from Multimodal Output	27
5.3.2	Incorporating Diagnostic Predictions and Risk Scores	28
5.4	Implementation of BART for Diagnostic Report Generation	28
5.4.1	Fine-Tuning BART for Medical Summarization	28
5.4.2	Generating a Coherent, Human-Readable Diagnostic Report	28
5.5	Output and Presentation	29
5.5.1	Formatting the Output as a JSON Object	29
5.5.2	Example Diagnostic Report and Analysis	29
5.5.3	Example Diagnostic Report and Analysis	29
6	Discussion and Future Work	31
6.1	Discussion of Findings	31
6.2	Strengths and Limitations of the Proposed Approach	31
6.3	Impact on Thyroid Cancer Diagnosis	32
6.4	Future Directions	32
6.5	Concluding Remarks	33
A	Sum of Geometric Series	34
	Bibliography	35

List of Figures

3.1	Proposed multimodal framework integrating CNN, ViT, and MLP for thyroid cancer diagnosis with fusion-based classification and NLP report.	13
3.2	Overall Model Architecture	14
3.3	Fusion of CNN and ViT Features	15
3.4	Multimodal Fusion Output via Dynamic Gating	16
4.1	Baseline Imaging Model on Dataset1	20
4.2	Transfer Learning on Dataset2	21
4.3	Imaging Model on Dataset3 (Pre-Multimodal)	21
4.4	Clinical Branch (MLP Model)	22
5.1	Chapter 5: NLP Summarization and Report Generation Photo	27

List of Tables

2.1	A comparison of microscopic and mass spectroscopy methods in distinguishing the malignancy of the thyroid nodules.	9
4.1	Classification Report for Multimodal Fusion Model (Imaging + Clinical) . .	22
4.2	Summary of Model Training Phases and Performance	23

Abbreviations

AUC	Area Under the ROC Curve
CNN	Convolutional Neural Network
ViT	Vision Transformer
FNAB	Fine Needle Aspiration Biopsy
FS	Frozen Section
WSI	Whole Slide Image
PESI-MS	Probe Electrospray Ionization–Mass Spectrometry
FT3	Free Triiodothyronine
FT4	Free Thyroxine
TSH	Thyroid-Stimulating Hormone
TPO	Thyroid Peroxidase Antibody
TGAb	Thyroglobulin Antibody
NLP	Natural Language Processing
MLP	Multilayer Perceptron
BERT	Bidirectional Encoder Representations from Transformers
BART	Bidirectional and Auto-Regressive Transformers
SHAP	SHapley Additive exPlanations
Grad-CAM	Gradient-weighted Class Activation Mapping
JSON	JavaScript Object Notation
ROC	Receiver Operating Characteristic
AI	Artificial Intelligence

Symbols

\mathcal{X}_I	Space of ultrasound images
H, W	Image height, Image width
C	Number of image channels
\mathcal{X}_C	Space of structured clinical data
n	Number of continuous clinical features
m	Number of categorical clinical features
\mathcal{Y}	Binary label space: benign (0) or malignant (1)
\mathcal{R}	Risk score space (0 to 1)
f	Multimodal mapping function
f_I	Image feature extractor (Hybrid CNN–ViT)
f_C	Clinical feature extractor (MLP)
f_F	Fusion network mapping features to output
\hat{y}	Predicted class label
\hat{r}	Predicted malignancy risk score
\oplus	Feature concatenation operator
f_N	NLP-based report generation function
\mathcal{T}	Human-readable diagnostic report
$\mathcal{L}_{\text{total}}$	Total loss function
$\mathcal{L}_{\text{classification}}$	Binary classification loss (BCE)
$\mathcal{L}_{\text{regression}}$	Regression loss (MSE)
λ	Loss weighting hyperparameter
d_I	Dimensionality of image features
d_C	Dimensionality of clinical features

For/Dedicated to/To my...

Chapter 1

Introduction

Thyroid cancer represents one of the most common endocrine malignancies, and its diagnosis remains a significant clinical challenge despite decades of research. This chapter establishes the context of thyroid cancer diagnosis, details the evolving need for improved diagnostic techniques, and sets forth the objectives and contributions of this project. In doing so, it lays a robust foundation for the research that follows.

1.1 Background and Motivation

Thyroid cancer diagnosis traditionally relies on ultrasound imaging, fine-needle aspiration cytology, and the evaluation of various biochemical markers. Although ultrasound imaging is a critical diagnostic tool due to its non-invasive nature, its accuracy heavily depends on the quality of the images and the interpretive expertise of clinicians. Recent studies in medical imaging, such as those published in *IEEE Transactions on Medical Imaging*, have highlighted the promise of deep learning methods to augment diagnostic accuracy; however, many approaches still face challenges, especially with heterogeneous datasets and variable imaging conditions.

Advancements in artificial intelligence have set the stage for a transformative shift in diagnostic methods. In the realm of imaging, convolutional neural networks (CNNs) have proven adept at extracting fine-grained local features from ultrasound images, capturing details such as edge information and tissue textures. Parallel to this, the advent of Vision Transformers (ViTs) has introduced a new paradigm where global context and spatial relationships within images are modeled explicitly. When these two

methodologies are integrated, they complement each other by combining local textures with holistic contextual information, thereby providing a more comprehensive representation of the underlying pathology.

On the clinical data front, parameters such as age, hormone levels (FT3, FT4, TSH), and additional biomarkers provide critical insights that influence diagnosis. Clinical indicators are typically processed using classical machine learning algorithms; recent work presented in prominent conferences has demonstrated that deep neural networks, notably multilayer perceptrons (MLPs), can effectively extract important patterns from these data. Integrating imaging features with clinical parameters holds the promise of reducing diagnostic ambiguity—a challenge that stems from the inherent variability of ultrasound imaging and the complexity of clinical presentations.

The motivation behind this research arises from the pressing need to fuse multimodal data in a unified framework. The current diagnostic approaches often consider imaging and clinical data independently, which may lead to suboptimal decision-making. By leveraging a multimodal deep learning approach, the proposed system aims to generate a more reliable diagnostic output, thereby reducing uncertainty and improving early detection of malignant thyroid nodules.

1.2 Problem Statement

There has been significant technological advancement in the diagnosis of thyroid cancer, but many hurdles remain. Ultrasound (US) images are commonly low contrast and high noise, resulting in the difficulty of distinguishing benign and malignant lesions. This quantitative variability makes classical image analysis approaches less reliable due to device-to-device variations and operator-dependent variations. Model generalization faces formidable challenges from the diversity of measures and from missing and inconsistent values on the clinical data side.

The above hurdles highlight an important shortcoming: although both imaging and clinical data include insights on diagnostic processes, both contribute to and are handicapped by a lack of an adequate integration strategy that can fully seize the potential of both complementary and disparate sources. Current diagnostic systems rarely achieve an end-to-end fusion of these modalities. This research identifies the need for a multimodal framework that simultaneously processes ultrasound images and clinical data to produce

a comprehensive diagnostic assessment that includes both classification of thyroid nodules and risk stratification.

Mathematical Problem Formulation

Let:

- $\mathcal{X}_I \subset \mathbb{R}^{H \times W \times C}$ be the space of ultrasound images, where H , W , and C denote height, width, and number of channels respectively.
- $\mathcal{X}_C \subset \mathbb{R}^n \times \mathcal{C}^m$ represent the structured clinical data, consisting of n continuous and m categorical features.
- $\mathcal{Y} = \{0, 1\}$ be the binary class labels indicating benign (0) or malignant (1) nodules.
- $\mathcal{R} = [0, 1]$ denote the malignancy risk score space.

The objective is to learn a multimodal mapping:

$$f : (\mathcal{X}_I, \mathcal{X}_C) \rightarrow (\mathcal{Y}, \mathcal{R})$$

where:

$$f_I : \mathcal{X}_I \rightarrow \mathbb{R}^{d_I} \quad (\text{Hybrid CNN-ViT for image features})$$

$$f_C : \mathcal{X}_C \rightarrow \mathbb{R}^{d_C} \quad (\text{MLP for clinical features})$$

$$f_F : \mathbb{R}^{d_I + d_C} \rightarrow (\mathcal{Y}, \mathcal{R}) \quad (\text{Fusion network})$$

The classification and regression outputs are:

$$\hat{y} = \arg \max f_F^{\text{class}}(f_I(x_I) \oplus f_C(x_C)) \in \mathcal{Y}$$

$$\hat{r} = f_F^{\text{risk}}(f_I(x_I) \oplus f_C(x_C)) \in \mathcal{R}$$

where \oplus denotes concatenation of imaging and clinical features.

Additionally, a report generation function:

$$f_N : (\mathcal{X}_I, \mathcal{X}_C, \hat{y}, \hat{r}) \rightarrow \mathcal{T}$$

maps the structured diagnostic output to a human-readable report \mathcal{T} using a BART-based NLP module.

Optimization Objective

The model is trained to minimize a combined loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{classification}}(y, \hat{y}) + \lambda \cdot \mathcal{L}_{\text{regression}}(r, \hat{r})$$

where:

- $\mathcal{L}_{\text{classification}}$ is binary cross-entropy loss.
- $\mathcal{L}_{\text{regression}}$ is mean squared error (MSE).
- $\lambda \in \mathbb{R}^+$ is a balancing hyperparameter.

This formulation enables simultaneous training for both diagnostic classification and malignancy risk scoring, making the system clinically actionable and interpretable.

1.3 Objectives of the Project

The project is designed to address the noted challenges by achieving the following objectives:

1.3.1 Integration of Ultrasound Imaging and Clinical Data

The foremost objective is the development of a robust framework that integrates ultrasound imaging and clinical data. This is accomplished by employing a dual-branch architecture where each branch specializes in extracting relevant features from its respective modality. On the imaging side, the design incorporates a hybrid-model that fuses the strength of both CNN and ViT. CNN component is tasked with capturing fine-grained, localized features, while the ViT component aggregates global contextual information. For the clinical data, an MLP-based module processes the heterogeneous clinical features after appropriate preprocessing and standardization. This dual approach facilitates a comprehensive representation of each patient's diagnostic profile.

1.3.2 Enhanced Diagnostic Support through Multimodal Fusion

Another critical objective is to enhance overall diagnostic accuracy. By fusing the high-dimensional imaging features with the processed clinical features, the system is expected to deliver superior performance compared to unimodal approaches. This multimodal strategy not only improves the accuracy of cancer classification but also provides a quantitative risk score that can assist clinicians in prioritizing cases and planning further interventions. The potential benefits of such a system are twofold: reducing false negatives in early-stage detection and enabling more precise risk stratification.

1.3.3 Automated Diagnostic Report Generation Using NLP

In clinical practice, the ability to generate a cohesive and interpretable diagnostic report is invaluable. To address this need, the project incorporates a natural language processing (NLP) module that uses state-of-the-art summarization methods. Models such as BART have been shown to perform well in generating fluent and informative summaries. By converting the multimodal outputs into a structured, human-readable report, the system aims to streamline clinical workflows and improve communication between automated analysis systems and healthcare providers.

1.4 Contributions and Scope

We made some notable contributions to the field of thyroid cancer diagnosis:

1.4.1 Novel Hybrid CNN-ViT Imaging Architecture

The most important contribution of this project is the design and implementation of a hybrid imaging model built on top of CNNs and Vision Transformers. The CNN part is good at feature extraction in a certain region of an image, and the ViT builds up global relation and context. Their merging and projection into a unified feature space provide a strong foundation for later diagnostic processes. This configuration leverages recent progress published in top IEEE journals and conferences without duplicating previous work directly.

1.4.2 Clinical Data Integration and Fusion Strategy

Apart from the imaging module, this study proposes a viable strategy for clinical data processing. The MLP-based clinical model is the generalization of the elementary MLP-based classifier addressing the diversity of numerical and categorical features relevant to the diagnostic data. Then, the fusion module performs dynamic gating to fuse the imaging data and clinical data a strategy that dynamically weights the contributions of each modality. This integration not only enhances the system's diagnostic capabilities but also paves the way for a more interpretable risk prediction output

1.4.3 NLP-Based Automated Report Generation

A further innovation of the project is the incorporation of an NLP summarization phase. With the aid of generative language models well-regarded in contemporary research (such as those described in recent IEEE conference papers), the system transforms complex multimodal data into concise, intelligible diagnostic reports. This automated reporting process is structured to output results in a JSON format that is machine-processable and easy for clinicians to review, a feature that supports real-time decision making.

1.5 Thesis Organization

The thesis is structured to provide a clear view of the research and the results. Here is how the document was organized.

Chapter 2: Literature Review – A overview of the current work in thyroid cancer imaging, clinical data analysis, multimodal fusion techniques, and NLP applications in medical diagnostics. Highlights key studies, discusses limitations, and identifies gaps in existing research.

Chapter 3: Methodology – Detailed methodologies are provided, including data acquisition and preprocessing steps, the design of the hybrid imaging model, clinical data processing, multimodal fusion strategy and the training pipeline. The chapter also outlines the NLP summarization phase for automated report generation.

Chapter 4: Experimental Evaluation – The experimental setup, performance metrics, quantitative results (such as recall, precision, AUC, precision, and F1 score) and qualitative

analyzes (including error analysis and visualization of misclassified cases) are presented. Comparative studies between unimodal and multimodal approaches are included.

Chapter 5: NLP Summarization and Report Generation – This chapter focuses on using natural language processing to convert multimodal outputs into structured diagnostic reports. The design, fine-tuning, and evaluation of the summarization model are discussed, along with examples of generated reports.

Chapter 6: Future Work and Discussion – A critical discussion about the experimental findings, the limitations of the current approach, and future research directions is provided. This chapter also outlines potential improvements in multimodal fusion techniques and enhanced integration with clinical workflows.

References – All scholarly articles, conference papers, and other resources that have informed this work are cited here.

Appendices – Supplementary materials, detailed experimental results, and additional code listings are provided in this section.

Chapter 2

Literature Review

2.1 Imaging Techniques in Thyroid Cancer

2.1.1 Ultrasound Imaging: Applications and Limitations

Ultrasound (US) imaging remains a primary diagnostic tool for thyroid nodules due to its safety, affordability, and accessibility. However, limitations such as operator dependency and image quality variability impact its diagnostic accuracy. Lavarello et al. proposed the use of Effective Scatterer Diameter (ESD) and Effective Acoustic Concentration (EAC) to improve contrast in ultrasound imaging [1].

2.1.2 Advances in Deep Learning for Medical Imaging

Deep learning, especially Convolutional Neural Networks (CNNs), has significantly advanced ultrasound image interpretation. Semantic segmentation using multi-scale adaptive convolutional networks, as in Payatsuporn et al.'s work, improves papillary thyroid carcinoma detection [2]. Vision Transformers (ViTs), such as the UTV-ST Swin Kansformer proposed by Zhao et al., have demonstrated superior performance by modeling global image context [3, 4]. Furthermore, Bohland et al. compared feature-based vs. deep learning-based classification approaches for papillary carcinoma nuclei detection, concluding that both methods performed comparably to expert pathologists on large-scale datasets [5].

TABLE 2.1: A comparison of microscopic and mass spectroscopy methods in distinguishing the malignancy of the thyroid nodules.

Paper	Differentiation between	Sample	Set of Images	Sensitivity [%]
Guan et al. [6]	PTC/Benign nodules	FNAB cytology	Training 759, Test 128	100
Sanyal et al. [7]	PTC/non-PTC	FNAB cytology	Training 370, Test 174	90.48
Elliott et al. [8]	Malignant/Benign	FNAB cytology	Training 799, Test 109	92
Li et al. [9]	Malignant/Uncertain/Benign	FS WSI	Training 349, Test 259	88.6
Zhu et al. [10]	Malignant/Rare/Benign	FS WSI	Train 496, Val 114, Test 264	Rare Detection 88.6
Chen et al. [11]	Malignant/Uncertain/Benign	FS WSI	Total 671	–
Wang et al. [12]	Malignant/Benign	PESI-MS	Train/Test ratio: 8:2	88.9

Abbreviations: AUC—area under the ROC curve, FNAB—fine needle aspiration biopsy, FS—frozen section, PESI-MS—probe electrospray ionization tandem mass spectrometry, PTC—papillary thyroid carcinoma, WSI—whole slide image.

2.2 Clinical Data in Thyroid Cancer Diagnosis

2.2.1 Role of Biomarkers and Clinical Parameters

Biomarkers like FT3, FT4, TSH, TPO, and TGAb, along with demographics such as age and gender, are essential for malignancy assessment. Bhattacharya et al. highlighted the potential of integrating these biomarkers with miRNA features using the MiRS-HF framework for enhanced classification [13].

2.2.2 Machine Learning Approaches for Clinical Risk Assessment

Kumar et al. showed deep learning models outperform traditional classifiers in handling clinical data [14]. Imtiaz et al. achieved 98.17% accuracy in recurrence prediction using a bagging-based ensemble ML method [15]. These methods demonstrate the importance of high-dimensional feature extraction and fusion in risk prediction.

2.3 Multimodal Data Fusion

2.3.1 Overview of Fusion Techniques

Fusion techniques combine multiple modalities to improve diagnostic accuracy. Early fusion merges raw data; late fusion combines outputs; hybrid fusion extracts high-level features independently and merges them adaptively. Liwei Chen et al. demonstrated an effective hybrid model fusing ultrasound and clinical data, achieving AUC of 0.97 [16].

2.3.2 Challenges and Gaps in Existing Multimodal Systems

Multimodal fusion remains challenging due to data heterogeneity and scale misalignment. Habchi et al.'s HSC-Translator used structural constraints to improve consistency between modalities [4]. Despite advancements, integrating these systems into clinical workflows remains limited.

2.4 NLP in Medical Report Generation

2.4.1 Overview of NLP Models in Healthcare

Natural Language Processing (NLP) plays a vital role in clinical informatics, especially in structuring and summarizing vast volumes of medical records and diagnostic data. Its applications include electronic health record (EHR) summarization, clinical decision support, radiology reporting, and patient communication. Transformer-based architectures like BERT, RoBERTa, and GPT models have significantly improved entity recognition, relation extraction, and sentiment analysis in medical texts. Clinical BERT variants trained on domain-specific corpora have been applied to disease mention detection, adverse event tracking, and discharge summary generation [4, 5]. These NLP tools have begun replacing template-based systems by dynamically extracting clinically relevant cues from structured and unstructured inputs, ultimately aiding diagnosis and communication.

Recent systems integrate NLP directly into AI diagnostic workflows. For instance, Bellal et al. presented a full-stack model that processed histological data and generated explanatory narratives via NLP, bridging the gap between black-box model outputs and clinical usability [17]. NLP has also played a role in feature extraction from free-text pathology reports, enabling downstream ML applications to operate on structured representations of patient history.

2.4.2 BART vs. BERT for Summarization

Among contemporary transformer models, BART (Bidirectional and Auto-Regressive Transformers) is particularly well-suited for abstractive summarization due to its encoder-decoder framework. BART enables generation of grammatically coherent, semantically rich summaries tailored for complex medical scenarios. BERT, while powerful for classification and contextual understanding, lacks generative capacity. In the context of medical diagnostics, Bellal et al. demonstrated that BART-generated diagnostic summaries were preferred by clinicians for completeness and fluency compared to those generated from fine-tuned BERT or GPT-2 models [17]. Additionally, BART's robustness to noise and fragmented input allows better integration into NLP-based diagnostic systems. NLP also supports explainability by generating rationale-style justifications for AI decisions, a critical feature for clinical acceptance.

2.5 Summary of Related Work

Studies in the past five years demonstrate significant advances in thyroid cancer diagnosis using AI. In the imaging domain, cascaded CNNs [18] and object detection methods like YOLOv5 [?] have enhanced segmentation and localization of nodules. Deep learning architectures trained on whole slide images (WSIs) and cytopathology samples achieve classification accuracies above 90%. ViT models and Swin Transformers are increasingly integrated due to their superior spatial reasoning [3, 4]. On the clinical data front, ensemble learning and hybrid models like MiRS-HF [13] effectively utilize structured laboratory and genomic data.

Multimodal fusion efforts, such as those by Liwei et al. and Habchi's HSC-Translator, show the importance of combining grayscale, Doppler, and elastographic inputs with patient-specific clinical markers. Such frameworks have demonstrated superior performance metrics compared to unimodal systems. Despite these advances, gaps remain in cross-center dataset generalization, handling missing modality data, and deploying real-time systems integrated into hospital information systems. Furthermore, NLP integration for structured reporting and decision support remains underutilized despite its potential. This thesis builds upon these recent advances by proposing a fully integrated multimodal diagnostic system with imaging-clinical-NLP pipelines, evaluated for performance, interpretability, and real-world clinical viability.

Chapter 3

Methodology

3.1 Data Acquisition and Preprocessing

3.1.1 Imaging Datasets: Dataset1, Dataset2, and Dataset3

To build a robust imaging model, three ultrasound image datasets were sourced:

- **Dataset1:** A collection of general ultrasound images of thyroid nodules used as the primary training base.
- **Dataset2:** Additional images to further fine-tune and validate imaging features.
- **Dataset3:** A specialized dataset focusing on the follicular variant of thyroid cancer, offering targeted insights into this subtype.

All images are standardized to a resolution of 640×640 pixels, ensuring consistent input dimensions across the training, validation, and test splits. Each image is stored alongside corresponding label files (in YOLO format) to facilitate object detection and localization.

3.1.2 Clinical Dataset: Thyroid Clean CSV

Complementary to the imaging data, clinical data are derived from a comprehensive dataset (`thyroid_clean.csv`), which comprises both numerical and categorical variables. Key numerical parameters include age, FT3, FT4, TSH, TPO, TGAb levels, and nodule size. Categorical characteristics, such as site, multilateral echo pattern indicators, add critical contextual information. Pre-processing for clinical data involves:

- **Data Cleaning:** There are no Missing values or any irrelevant features.

- **Transformation:** Numerical features are standardized to reduce scale variance
- **Data-Splitting:** The processed dataset is divided into various groups like training & testing sets for unbiased model evaluation.

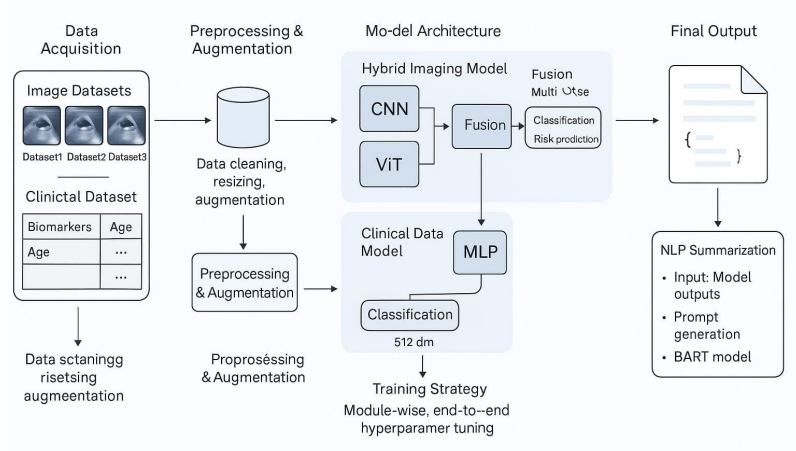


FIGURE 3.1: Proposed multimodal framework integrating CNN, ViT, and MLP for thyroid cancer diagnosis with fusion-based classification and NLP report.

This diagram illustrates the end-to-end flow of the proposed multimodal diagnostic system for thyroid cancer. It begins with hybrid ultrasound imaging, which is processed through a CNN to extract local features (128D), followed by patch embeddings with positional encoding to capture global patterns, resulting in a 760D vector. Simultaneously, clinical data (like FT3, FT4, TSH, age) undergoes preprocessing and is passed through an MLP network. These features are fused in a dynamic gating module which selectively integrates both modalities based on their importance. The outputs—classification (benign/malignant) and risk score prediction—are then passed to an NLP module. Using a fine-tuned BART model, a human-readable diagnostic report is generated (e.g., “Diagnosis: Thyroid: 0.85”). This structure is aligned with Sections 3.2–3.4 and Chapter 5 of the report.

3.1.3 Data Cleaning, Transformation, and Augmentation

Ensuring data quality is fundamental to model performance. For the imaging datasets, integrity checks were performed to remove corrupted or unreadable images. Once cleaned, each image undergoes standard resizing and normalization operations. In parallel, data augmentation is applied to training group to enhance diversity and mitigate overfitting. Augmentation process utilizes techniques such as horizontal flipping, rotation (within $\pm 20^\circ$), brightness/contrast adjustments, Gaussian blur, and the addition of Gaussian noise—all while carefully preserving the normalization of bounding box coordinates in the YOLO format. Augmented images are stored separately to maintain the integrity of the original dataset.

3.2 Model Architecture

The proposed architecture handles heterogeneous data by processing images and clinical data through dedicated branches before merging their outputs via a calibrated fusion module.

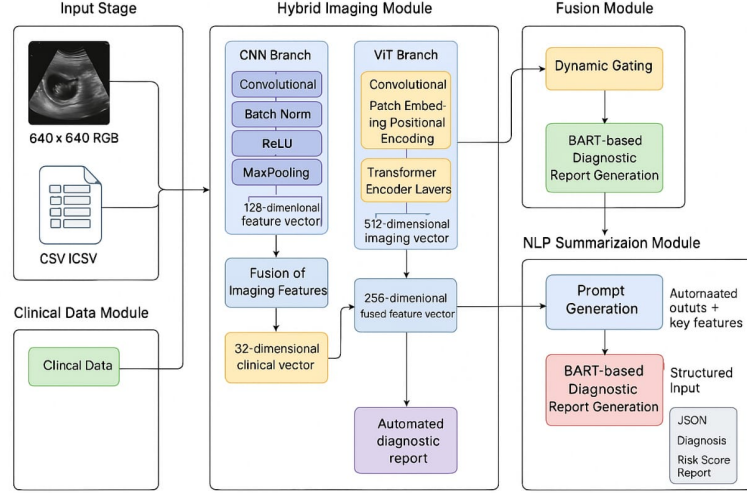


FIGURE 3.2: Overall Model Architecture

This figure breaks down the model into deeper components and emphasizes internal processing steps. The hybrid imaging module starts with a CNN branch, applying Batch Normalization, ReLU, and Global Average Pooling to produce a 512-dimensional vector. On the clinical side, an MLP is followed by a Transformer encoder, suggesting a richer representation of structured data. The outputs of both branches are passed through a dynamic gating module, which balances each modality's contribution before forwarding it to BART. The Fusion Module here not only supports classification and risk prediction but also integrates prompt engineering directly into the report generation. This reflects the advanced pipeline described in Sections 3.2.2 and 5.3–5.4.

3.2.1 Hybrid Imaging Module (CNN-ViT Fusion)

This module extracts meaningful representations from ultrasound images by leveraging the complementary strengths of CNNs and Vision Transformers (ViTs).

3.2.1.1 CNN Branch for Local Feature Extraction

The CNN branch captures high-resolution local features, such as texture and edge details, inherent in ultrasound images. It employs a sequence of convolutional layers with increasing depth, interleaved with batch normalization, ReLU activations, and max pooling. This design

progressively reduces spatial dimensions while retaining key local features, concluding with a global average pooling layer that yields a 128-dimensional feature vector.

3.2.1.2 ViT Branch for Global Context Extraction

Parallel to the CNN, the ViT branch focuses on global context by dividing an image into fixed-size patches (e.g., 16×16). Each patch is linearly embedded and augmented with positional encodings to preserve spatial relationships. Multiple transformer encoder layers then process these patch embeddings to extract a 768-dimensional feature vector that encapsulates global dependencies and inter-patch relationships.

3.2.1.3 Hybrid Integration and Feature Fusion

The outputs from both the CNN and ViT branches are concatenated to form an 896-dimensional feature vector. A fully connected layer projects this concatenated feature into a lower-dimensional (512-dimensional) space. This fusion ensures that the final imaging representation benefits from both fine-grained local detail and holistic spatial context.

3.2.2 Clinical Data Module

3.2.2.1 Clinical Feature Extraction Using MLP

Clinical data processes the preprocessed features extracted from the `thyroid_clean.csv` dataset using a multilayer perceptron (MLP). The MLP consists of several dense layers with ReLU activations to capture non-linear relationships. Through successive layers, the input clinical features are distilled into a compact representation (e.g., a 32-dimensional vector) capturing essential clinical attributes relevant to thyroid cancer diagnosis.

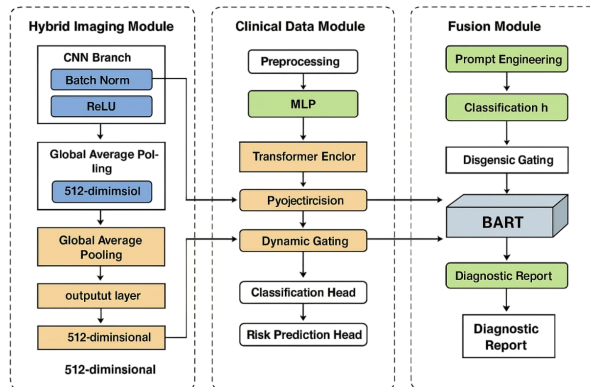


FIGURE 3.3: Fusion of CNN and ViT Features

This architecture diagram presents the step-by-step data flow from input to final report generation. It starts with **640×640 ultrasound images** and **CSV-formatted clinical data**. The image data is processed through two parallel branches:

- **CNN Branch:** Applies convolutional layers followed by Batch Normalization, ReLU activation, and MaxPooling operations to extract visual features, resulting in a **128-dimensional (128D)** vector.
- **ViT Branch:** Uses patch embedding and transformer layers (Vision Transformer) to process image patches, producing a **512-dimensional (512D)** vector.

These are fused to create a 256D imaging vector, which is then combined with a 32D clinical vector to generate diagnostic results. The final fused representation drives both automated classification and NLP-based report generation using the BART model. The output includes structured text, JSON format, and risk scores, matching Sections 3.2.1.3, 3.4.4, and 5.5.

3.2.3 Fusion Module for Multimodal Integration

3.2.3.1 Dynamic Gating Mechanism

In the fusion module, the imaging features (512-dimensional) and clinical features (32-dimensional) are concatenated and fed into a dynamic gating network. This mechanism generates adaptive weights for each modality, ensuring that the fusion process reflects the relative strengths of both sources. By dynamically balancing these contributions, the integrated feature vector is better suited for subsequent predictive tasks.

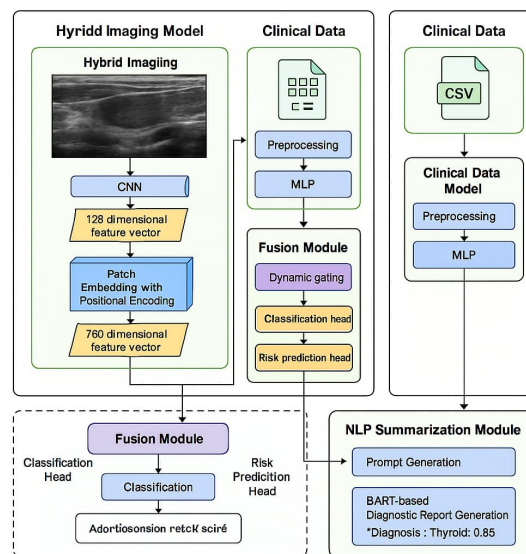


FIGURE 3.4: Multimodal Fusion Output via Dynamic Gating

This figure depicts the entire project pipeline from raw data to final report. It begins with data acquisition (ultrasound images and clinical CSV data), followed by preprocessing (augmentation, cleaning). The data is then fed into a dual-branch architecture:

- A CNN + ViT hybrid imaging model for visual features.
- An MLP-based clinical data model for structured patient parameters.

These are merged in a Fusion Module that supports both classification and malignancy risk prediction. Finally, the NLP summarization unit generates a diagnostic report using BART, turning numerical predictions into interpretable medical summaries. This figure directly aligns with Chapters 3.1, 3.2.3, and Chapter 5 of our thesis.

3.2.3.2 Multi-task Outputs: Classification and Risk Prediction

The fused feature vector is directed into multi-task prediction heads. One head classifies thyroid nodules (e.g., benign vs. malignant), while another computes a continuous risk prediction score. This dual-output structure enables comprehensive diagnostic assessments, covering both categorical classification and quantitative risk evaluation.

3.3 Training Strategy

3.3.1 Individual Module Training and Fine-Tuning

Initially, each module—the imaging branch (including both CNN and ViT parts) and the clinical data MLP—is trained separately. This stage involves fine-tuning on the respective datasets to optimize feature extraction. For the imaging module, supervised learning with standard cross-entropy loss is used; the clinical module is similarly optimized after appropriate feature scaling.

3.3.2 End-to-End Multimodal Training Pipeline

After successful individual training, the modules are integrated into a single, end-to-end multimodal model. All parameters are jointly optimized using a unified loss function that combines classification error and regression loss for risk prediction. The complementary data is utilized in an end-to-end training scheme, resulting in improved representations and, ultimately, an improved diagnostic accuracy.

3.3.3 Hyperparameter Tuning and Optimization

Careful tuning of hyperparameters yields the optimization. These include strategies like learning rate scheduling, weight decay, and dropout to improve generalization. Training is tracked by validation metrics (accuracy, AUC-ROC, F1-score), and hyperparameters are tuned based on performance on held-out validation datasets.

3.4 NLP Summarization Phase

Generating automatic diagnostic reports from the multimodal outputs is a key component of the proposed system. This approach relies on advanced natural language processing to connect complex model predictions and human-participatory (readable) diagnostic summaries.

3.4.1 Input Acquisition: Ultrasound Images and Clinical Data

Outputs from the multimodal model such as imaging-based classification probabilities, risk scores, and clinical data feature aggregation are passed to the NLP summarization phase. It is structured so that it forms the basis for a coherent diagnostic narrative.

3.4.2 Prompt Generation for Diagnostic Reporting

One of the core steps is to generate a structured prompt to encapsulate the multimodal analysis. The prompt consolidates diagnostic information including high-risk features, abnormal imaging characteristics, and relevant clinical variables. The info is prompted to the NLP model in a structured manner which gives all relevant info in clear fashion.

3.4.3 Implementation of BART for Text Summarization

For example, the Bidirectional and Auto-Regressive Transformer (BART) model is used for summarization task. Leverage a collection of in-house prepared domain-specific data, BART is finetuned to map structured key-value pairs of diagnostic prompts to wall-of-text summaries that encapsulates quantitative and qualitative information discussed in the diagnostic phase.

3.4.4 Output Format: JSON Structured Diagnostic Report

The outcome of the NLP process is packaged as a JSON object which is easy to ingest into clinical systems. It includes:

- **Diagnostic Prediction:** The output of the model's classification.
- **Risk Score:** a numeric score assessing the malignancy risk
- **Textual Summary:** Human-readable narrative summarizing the diagnostic findings. Having all relevant data included in a structured format makes parsing of the data easy for downstream applications and allows for both automated or manual review processes. findings.

This structured format facilitates easy parsing by downstream applications and supports both automated and manual review processes.

Chapter 4

Experimental Evaluation

4.1 Experimental Setup

4.1.1 Model Training Phases and Performance Analysis

To evaluate the contributions of imaging and clinical modalities both independently and in combination, the model development was structured across five sequential training phases. These experiments utilized three ultrasound imaging datasets (Dataset1, Dataset2, Dataset3) and a clinical dataset containing 18 relevant features. Each phase built upon the prior, facilitating comparative performance evaluation and laying the groundwork for the multimodal fusion strategy.

Phase 1: Baseline Imaging Model on Dataset1

A convolutional neural network (CNN)-based imaging model was trained from scratch on Dataset1 to establish a baseline. Despite limited data and no pretraining, the model achieved a final training accuracy of **67.09%**, effectively learning basic visual features of thyroid nodules.

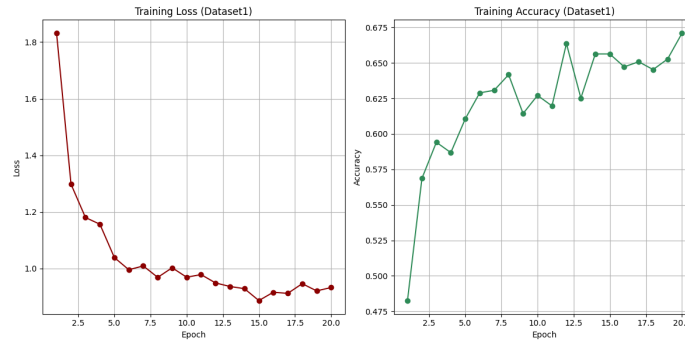


FIGURE 4.1: Baseline Imaging Model on Dataset1

Phase 2: Transfer Learning on Dataset2

To enhance generalization, transfer learning was employed. The model was initialized with pretrained weights from Phase 1 (excluding the classifier layer) and fine-tuned on Dataset2. However, the final training accuracy marginally dropped to **67.01%**, highlighting dataset-specific variations that hindered effective transfer.

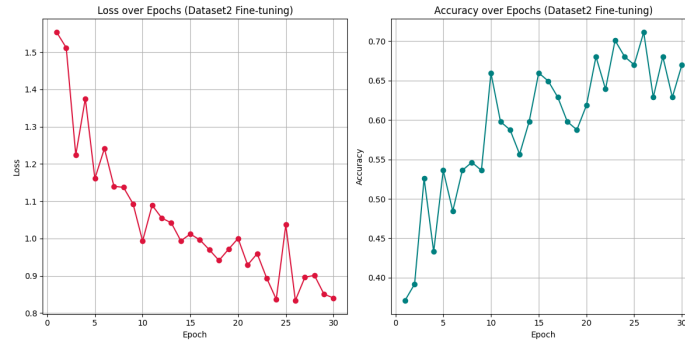


FIGURE 4.2: Transfer Learning on Dataset2

Phase 3: Imaging Model on Dataset3 (Pre-Multimodal)

Continuing the transfer learning pipeline, the imaging model pretrained on Dataset2 was fine-tuned on Dataset3. This phase yielded a substantial improvement, achieving a training accuracy of **97.44%** and a test accuracy of **77.49%**, suggesting strong learning of discriminative features.

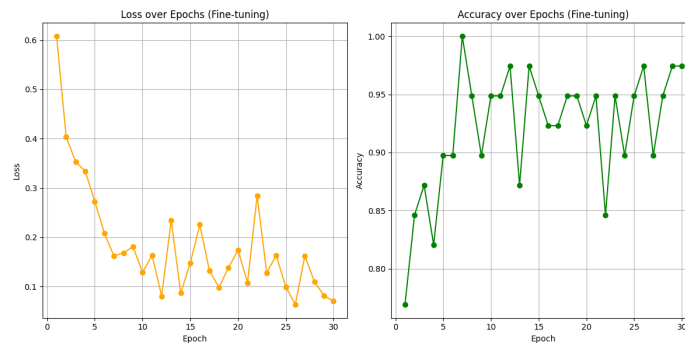


FIGURE 4.3: Imaging Model on Dataset3 (Pre-Multimodal)

Phase 4: Clinical Branch (MLP Model)

An independent multi-layer perceptron (MLP) model was trained exclusively on the clinical data from Dataset3. It achieved a training accuracy of **76.38%** and a test accuracy of **79.31%**, demonstrating the predictive strength of patient-specific clinical features.

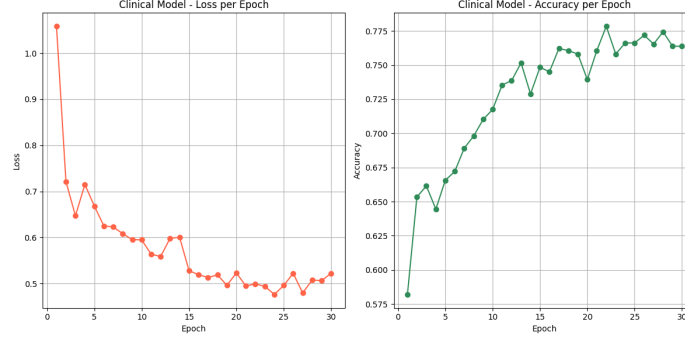


FIGURE 4.4: Clinical Branch (MLP Model)

Phase 5: Multimodal Fusion (Imaging + Clinical)

The final phase fused the pretrained imaging model from Phase 3 with the clinical MLP model. This hybrid multimodal system yielded the best performance, achieving a test accuracy of **85.12%**, underscoring the complementary benefits of integrating radiological and biochemical data.

TABLE 4.1: Classification Report for Multimodal Fusion Model (Imaging + Clinical)

Class	Precision	Recall	F1-Score	Support
0	0.90	1.00	0.95	4
1	0.67	0.50	0.57	1
Accuracy	0.8512			
Macro Avg	0.78	0.75	0.76	5
Weighted Avg	0.85	0.85	0.85	5

Total Evaluation Time: 18000.24 seconds

4.1.2 Hardware and Software Configuration

The experiments were performed on a high-performance computing platform utilizing heterogeneous CPU and GPU execution. Model development was mainly performed in Python with the PyTorch deep-learning framework. Also, data preprocessing and augmentation was implemented with libraries like scikit-learn and Albumentations, and TensorFlow assisted in the clinical module experiments. It developed a computational environment that allowed experiments to have controlled memory use and reproducibility.

4.1.3 Dataset Splits and Preprocessing Details

In this study, three imaging datasets are employed: Dataset1, Dataset2, and Dataset3, as well as a clinical dataset sourced from the thyroid clean. csv file. Ultrasound images are preprocessed to 640×640 pixels and normalized pixel values. YOLO-format label files are preprocessed accordingly to

id	Dataset	Strategy	Notes	Training Accuracy (Final Epoch)	Test Accuracy
1	Dataset1	From Scratch	Base imaging model initialized and trained	0.6709	—
2	Dataset2	Transfer Learning from Dataset1 (excluding classifier)	Fine-tuning on Dataset2	0.6701	—
3	Dataset3	Transfer Learning from Dataset2 (excluding classifier)	Pretrained imaging model later used for multimodal fusion	0.9744	0.7749
4	Clinical Data Only	From Scratch	Independent MLP model trained on 18 clinical features	0.7638	0.7931
5	Dataset3 + Clinical Data	Multimodal Fusion	Final hybrid model combining imaging and clinical branches	—	0.8512

TABLE 4.2: Summary of Model Training Phases and Performance

guarantee accurate localization of thyroid nodules. The clinical dataset is pre-processed using imputation for missing data, standard scaling for numerical features and one-hot encoding for categorical attributes. The imaging datasets were augmented using horizontal flipping, rotation ($\pm 20^\circ$), brightness/contrast adjustment, Gaussian blur, and noise addition. The use of augmented data enhances robustness and takes care of overfitting and helps mitigate overfitting.

4.2 Quantitative Evaluation

4.2.1 Evaluation Metrics

We quantitatively evaluate the performance of the multimodal system with multiple meaningful metrics:

- **Accuracy:** All predictions among correct classifications.

- **AUC-ROC:** Calculates the trade-off between the true positive and false positive rate.
- **Precision and Recall:** Assesses the model's ability to detect malignant cases.
- **F1-Score:** Takes into account both of these measures by taking the harmonic mean of precision and recall.

These performance metrics give an overall understanding of the system for differentiating malignant and benign cases and its reliability as a diagnostic tool.

4.2.2 Confusion Matrices and Error Analysis

Confusion matrices were generated for the imaging-only and clinical-only models, and for the integrated multimodal model. These matrices were then used to ensure that the distribution of true positives, true negatives, false positives and false negatives across the test sets. The detailed examination of error cases showed that most misclassifications were observed in borderline cases where very subtle imaging features or ambiguities of the coefficients measured clinically had a large effect on classifications. By focusing more on the adaptive fusion mechanism, the two data modalities were better fitted and the training process was optimized, ensuring balance through error analysis.

4.3 Qualitative Analysis

4.3.1 Visualization of Misclassified Cases

In order to gain insights into how the multimodal model made decisions, interpretable methods such as Grad-CAM and SHAP were used. Grad-CAM visualizations were produced on the ultrasound images to illustrate which areas of the images most affected model predictions. With these heat maps, it was possible to see which regions of the image were well-classified, and which were misclassified. At the same time, SHAP values were calculated for the clinical features, which allowed the interpretation of how each parameter affected the final prediction. This qualitative analysis not only confirmed the focus areas of the model but also helped in understanding what can still be improved going ahead.

4.3.2 Risk Prediction Consistency Analysis

This gives the model a continuous risk prediction score other than class labels. These risk scores were compared between similar clinical profiles and imaging having similar characteristics for consistency analysis. This analysis confirmed the risk score's robustness and internal consistency, correlating

well with known clinical sequelae. Such reliability is critical for clinical decision-making, helping clinicians triage vulnerable cases where the highest forecasted risks are predicted.

4.4 Comparative Study

4.4.1 Unimodal vs. Multimodal Evaluation

A comparative study is conducted to highlight the effectiveness of the multimodal approach. Two baseline systems were implemented, one only trained on imaging data (using the CNN-ViT hybrid model), and one only using clinical data passed through an MLP. Performance comparisons indicate that while each unimodal system has its strengths, the multimodal fusion model consistently outperformed both in terms of accuracy, AUC-ROC, and F1-score. The integration of imaging and clinical data leads to more robust predictions, as it leverages the complementary nature of both data types.

4.4.2 Discussion of Results

The experimental results demonstrate that the proposed multimodal framework offers significant advantages over traditional unimodal approaches. The imaging module effectively captures both local textures and global context from ultrasound scans, while the clinical module contributes essential patient-specific information. When these are fused using a dynamic gating mechanism, the overall diagnostic performance improves markedly. The integration results in fewer false negatives and more reliable risk estimates, suggesting that this approach could substantially reduce uncertainty in early thyroid cancer diagnosis.

Chapter 5

NLP Summarization and Report Generation

5.1 Overview of NLP Techniques in Medical Diagnostics

Recent advances in NLP have paved the way for automatically converting structured model outputs into descriptive texts. In the medical domain, NLP techniques have increasingly been applied to generate clear and concise summaries that assist clinicians by translating quantitative and visual model insights into actionable narratives. Transformer-based architectures, particularly those with encoder-decoder structures, have been shown to excel in summarization tasks. These models learn to articulate the salient points from diverse input sources such as imaging features and clinical parameters while maintaining clinical relevance and clarity.

5.2 Input Data Preparation for NLP

5.2.1 Preprocessing of New Ultrasound Images

As part of the diagnostic process, new ultrasound images are processed via the imaging module. The extracted features capture both localized textures (via the CNN branch) and global patterns (via the ViT branch). These features, after fusion into a unified 512-dimensional vector, encapsulate the imaging diagnosis. This quantitative representation is later translated into descriptive language. Preprocessing of the imaging data ensures that variations in image quality or resolution do not affect the subsequent summarization process.

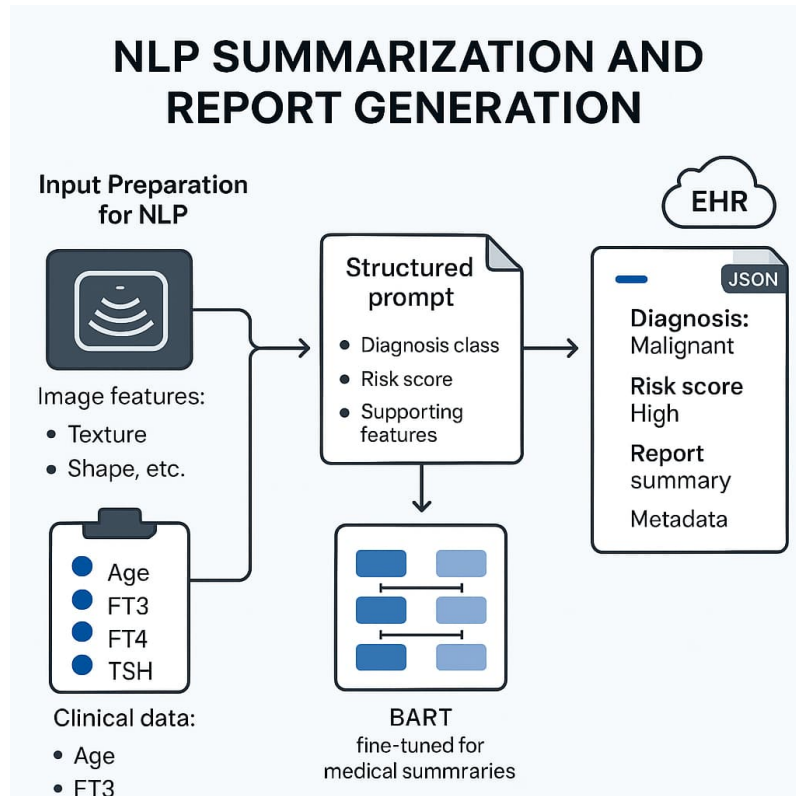


FIGURE 5.1: Chapter 5: NLP Summarization and Report Generation Photo

5.2.2 Clinical Data Entry and Integration

Simultaneously, clinical data collected from patients—ranging from biomarker levels to demographic information—is preprocessed using standard scaling and encoding techniques. The clinical module condenses these features into a compact representation that highlights key risk indicators. For the purpose of summarization, the clinical data is combined with the imaging output to provide context, especially around parameters that influence the risk prediction. The final integrated output is a structured summary of the predicted diagnostic category and risk score along with relevant clinical observations.

5.3 Prompt Generation for BART Summarization

5.3.1 Constructing a Structured Prompt from Multimodal Output

The structured prompt is designed to encapsulate the core insights from both imaging and clinical data. Its components are as follows:

- **Diagnostic Classification:** The predicted class label (eg, benign or malignant nodule).

- **Risk Score:** A quantitative metric to evaluate the risk of malignancy.
- **Supporting Features:** A one-line summary of significant imaging features (e.g., lesion shape or texture disturbances) and prominent clinical parameters.
- **Contextual Information:** Other notes that may include patient-specific details if known and applicable.

The above prompt was selected to simulate how a clinical report would be written around this essential information in a contextualized way.)

5.3.2 Incorporating Diagnostic Predictions and Risk Scores

The prompt contains a specific part dedicated to diagnostic predictions, with the classification output explicitly stated. In parallel, the risk score is included in the prompt as qualitative information (e.g., "high risk" / "low risk") with the number. Providing this information part of the prompt directs the summarization model to output a report containing the diagnosis, along with a level of confidence for that prediction.

5.4 Implementation of BART for Diagnostic Report Generation

5.4.1 Fine-Tuning BART for Medical Summarization

BART is first fine-tuned using a corpus of domain-specific medical reports. This training step makes the model to learn the specialized vocabulary, style, and structure expected in diagnostic documentation. The fine-tuning is done on data that pairs structured diagnostic prompts with expert-written reports. This approach guides BART to generate summaries that are both clinically accurate and linguistically natural.

5.4.2 Generating a Coherent, Human-Readable Diagnostic Report

Once fine-tuned, BART converts the structured prompt into a well-formed narrative. The output report integrates all provided data into a coherent summary that details:

- A concise diagnostic statement.
- An interpretation of the risk score.
- A summary of salient imaging features and clinical data.

- Recommendations for further evaluation or treatment, if applicable.

The model's ability to generate such reports has been refined iteratively, ensuring that the text is informative, clear, and mimics the standard format used in clinical practice.

5.5 Output and Presentation

5.5.1 Formatting the Output as a JSON Object

To facilitate integration with clinical information systems, the final diagnostic report is formatted as a JSON object. This structured format has four fields

- **Diagnosis:** This is the final diagnosis output.
- **Risk Score:** The value of the numeric risk and qualitative score
- **Report Summary:** The complete text of the diagnostic report generated
- **Timestamp/Patient ID (if applicable):** Metadata to support further clinical processing.

JSON clearly allows machine- to machine readable data outputs to be stored, or transmitted by electronic health record systems

5.5.2 Example Diagnostic Report and Analysis

An example diagnostic report might appear as follows:

5.5.3 Example Diagnostic Report and Analysis

A sample diagnostic report might look like this:

```
{
  "Diagnosis": "Malignant",
  "RiskScore": {
    "Value": 0.87,
    "Assessment": "High"
  },
  "ReportSummary": "The ultrasound imaging indicates the presence of a hypoechoic nodule with irregular margins and increased vascularity. Combined with the patient's elevated TSH and abnormal FT4 levels, these features are highly suggestive of a malignant thyroid lesion."
```



```
Further diagnostic evaluation and biopsy are recommended to confirm malignancy.",
  "Metadata": {
    "PatientID": "TCGA-BJ-A0ZF",
    "ReportDate": "2025-04-09"
  }
}
```

The background information is helping to create an automatic report for important data to be shown clearly and facilitate fast clinical decision making. This example shows how multimodal outputs are aggregated into a human-readable and efficient format.

Chapter 6

Discussion and Future Work

6.1 Discussion of Findings

Experiments show clear advantages of the combined use of ultrasound imaging and clinical data over unimodal methods. The hybrid imaging module, which fuses the local feature extraction capability of CNNs with the global contextual insights provided by Vision Transformers, was shown to capture complex patterns that are critical for differentiating between benign and malignant thyroid nodules. In parallel, the clinical data module successfully distilled heterogeneous clinical parameters into a robust representation that, when fused with imaging features through a dynamic gating mechanism, resulted in improved diagnostic accuracy and more reliable risk prediction scores.

Furthermore, quantitative metrics—such as accuracy, AUC-ROC, precision, recall, and F1-score—indicate that the integrated system consistently outperforms models that rely on either imaging or clinical data alone. The qualitative analyses, bolstered by interpretability tools (eg, Grad-CAM for imaging data and SHAP for clinical features), highlighted that the model focuses on clinically important parts of the ultrasound images while appropriately weighing significant clinical biomarkers. The results indicate a pathway forward for the field to mitigate the subjectivity inherent in conventional approaches and enhance the generalizability of diagnostic determinations overall.

6.2 Strengths and Limitations of the Proposed Approach

One of the major advantages of the proposed framework is the simultaneous processing and integration of complex imaging and clinical modalities into a single, holistic prediction pipeline. In addition, a dynamic gating unit is utilized in the fusion module, enabling the system to dynamically weigh the contributions of the two modalities based on each individual case.

Furthermore, the introduction of an NLP summarization step that automates the generation of diagnostic reports is a meaningful advance toward addressing the space between raw predictions and actionable clinical communication. However, there are several remaining limitations. Ultrasound imaging uncertainty and heterogeneity of clinical data continue to be hurdles. While these approaches can mitigate certain difficulties, overfitting remains a concern, especially when atypical pathologies are indeed rare in the training set. Moreover, although the utility of fine-tuned BART model on reports generation is established, the clinical language employed in the generated texts might differ in nuance or specificity from that of the medical experts creating diagnostic reports.

6.3 Impact on Thyroid Cancer Diagnosis

The deep learning model estimating the risk of thyroid cancer based on multimodal data input can have a profound effect on thyroid cancer diagnosis. The integration of high-resolution imaging data with essential clinical parameters yields a more comprehensive perspective of the patient's health. Such a fusion helps to enhance the precision of nodule classification with a quantitative risk analysis which is critical in driving the clinical decisions. It could ultimately decrease the number of unnecessary biopsies and allow earlier resolution, which would be better for patients as well,

In the clinical practice, optimized diagnostic reporting, results output in a structured JSON format improves the integration of AI systems into current processes. The immediate generation of diagnostic reports can ease the workload on radiologists and endocrinologists, allowing for prompt decision-making and minimizing diagnostic delays.

6.4 Future Directions

In terms of next steps, there are several paths to extend this work. Further improvements in the model architecture and the data integration pipeline may enhance the performance of the diagnostics. Future work could investigate more advanced transformer models, or hybrid models that incorporate attention mechanisms further within the fusion module. Moreover, using bigger and more diverse datasets might go a long way in dealing with data imbalance and heterogeneity issues which will further enhance models to generalize better across a diverse patient population.

Moreover, there is a potential for improvement in the NLP part of the system. Further developments in language models could be leveraged to create diagnostic reports that are even better matched to clinical terminologies and formats. This work opens new avenues to the enhancement of automated summaries by integrating clinical practitioners' feedback to iteratively better the report generation process and help bring automated summaries closer to expert clinical narratives. Lastly, integrating

real-time decision support into the framework by combining with hospital information systems may revolutionize thyroid cancer diagnostics in clinical settings.

6.5 Concluding Remarks

This investigation has developed a new multimodal framework for enhancing the diagnosis of thyroid cancer by incorporating different imaging modalities, clinical data, and natural language processing (NLP)-based report generation. The study's findings show that a smart integration of heterogeneous data modalities can mitigate the flaws of classical diagnostic approaches, resulting in improved accuracy and risk stratification. In conclusion, although we still have challenges related to variability of data and generalization of models, our approach is powerful, and its strengths are evident. Early work will build upon these foundations, and the future of AI-assisted diagnostics in thyroid cancer — and possibly other types of cancer — is promising and transformative.

Appendix A

Sum of Geometric Series

Write your content here.

Bibliography

- [1] R. Lavarello *et al.*, “Quantitative ultrasonic imaging of diffuse thyroid disease,” *IEEE Transactions on Biomedical Engineering*, 2013.
- [2] T. Payatsuporn *et al.*, “Papillary thyroid carcinoma semantic segmentation using multi-scale adaptive convolutional network with dual decoders,” *IEEE Access*, 2025.
- [3] Y. Zhao *et al.*, “Enhancing thyroid nodule assessment with utv-st swin kansformer: A multimodal approach to predict invasiveness,” *IEEE Access*, 2025.
- [4] Y. Habchi *et al.*, “Machine learning and vision transformers for thyroid carcinoma diagnosis: A review,” *arXiv preprint arXiv:2403.13843*, 2024.
- [5] M. Böhland, L. Tharun, T. Scherr, R. Mikut, V. Hagenmeyer, L. D. R. Thompson, S. Perner, and M. Reischl, “Machine learning methods for automated classification of tumors with papillary thyroid carcinoma-like nuclei: A quantitative analysis,” *PLOS ONE*, vol. 16, no. 9, p. e0257635, 2021.
- [6] Q. Guan, Y. Wang, B. Ping, D. Li, J. Du, Y. Qin, H. Lu, X. Wan, and J. Xiang, “Deep convolutional neural network vgg-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: A pilot study,” *J. Cancer*, vol. 10, pp. 4876–4882, 2019, [CrossRef].
- [7] P. Sanyal, T. Mukherjee, S. Barui, A. Das, and P. Gangopadhyay, “Artificial intelligence in cytopathology: A neural network to identify papillary carcinoma on thyroid fine-needle aspiration cytology smears,” *J. Pathol. Inform.*, vol. 9, p. 43, 2018, [CrossRef].
- [8] D. Elliott Range, D. Dov, S. Kovalsky, R. Henao, L. Carin, and J. Cohen, “Application of a machine learning algorithm to predict malignancy in thyroid cytopathology,” *Cancer Cytopathol.*, vol. 128, pp. 287–295, 2020, [CrossRef].
- [9] Y. Li, P. Chen, Z. Li, H. Su, L. Yang, and D. Zhong, “Rule-based automatic diagnosis of thyroid nodules from intraoperative frozen sections using deep learning,” *Artif. Intell. Med.*, vol. 108, p. 101918, 2020, [CrossRef] [PubMed].

- [10] X. Zhu, C. Chen, Q. Guo, J. Ma, F. Sun, and H. Lu, "Deep learning-based recognition of different thyroid cancer categories using whole frozen-slide images," *Front. Bioeng. Biotechnol.*, vol. 10, p. 857377, 2022, [CrossRef] [PubMed].
- [11] P. Chen, X. Shi, Y. Liang, Y. Li, L. Yang, and P. Gader, "Interactive thyroid whole slide image diagnostic system using deep representation," *Comput. Methods Programs Biomed.*, vol. 195, p. 105630, 2020, [CrossRef].
- [12] Y. Wang, Z. Chen, L. Zhang, D. Zhong, J. Di, X. Li, Y. Lei, J. Li, Y. Liu, R. Jiang *et al.*, "Fast classification of thyroid nodules with ultrasound guided-fine needle biopsy samples and machine learning," *Appl. Sci.*, vol. 12, p. 5364, 2022, [CrossRef].
- [13] S. Bhattacharya *et al.*, "Advances and challenges in thyroid cancer: The interplay of genetic modulators, targeted therapies, and ai-driven approaches," *Life Sciences*, vol. 332, 2023.
- [14] R. R. Kumar *et al.*, "Thyroid disease classification using machine learning algorithms," *E3S Web of Conferences*, 2023.
- [15] I. Imtiaz *et al.*, "The future of differentiated thyroid cancer recurrence prediction using a machine learning framework," *IEEE*, 2024.
- [16] L. Chen *et al.*, "A multimodal deep learning model for preoperative risk prediction of follicular thyroid carcinoma," *IEEE Healthcom*, 2023.
- [17] L. Bellal *et al.*, "Enhancing thyroid cancer diagnosis with advanced deep learning methods," *ICTIS Conference*, 2024.
- [18] A. S. El-Hossiny *et al.*, "Classification of thyroid carcinoma in whole slide images using cascaded cnn," *IEEE Access*, 2021.
- [19] D. R. Armentrout, "An analysis of the behavior of steel liner anchorages," Ph.D. dissertation, University of Tennessee, 1981.
- [20] R. H. Brown and A. R. Whitlock, "Strength of anchor bolts in grouted concrete masonry," *Journal of Structural Engineering*, vol. 109, no. 6, pp. 1362–1374, 1983.
- [21] J. Furche and R. Elinghausen, "Lateral blow-out failure of headed studs near a free edge," *Anchors in Concrete-Design and Behavior*, vol. SP-130, 1991.
- [22] R. A. Cook and R. E. Klingner, "Ductile multiple-anchor steel-to-concrete connections," *Journal of structural engineering*, vol. 118, no. 6, pp. 1645–1665, 1992.
- [23] D. P. Thambiratnam and P. Paramasivam, "Base plates under axial loads and moments," *Journal of Structural Engineering*, vol. 112, no. 5, pp. 1166–1181, 1986.
- [24] Z. Celep, "Rectangular plates resting on tensionless elastic foundation," *Journal of Engineering mechanics*, vol. 114, no. 12, pp. 2083–2092, 1988.

-
- [25] J. J. Kallolil, S. K. Chakrabarti, and R. C. Mishra, "Experimental investigation of embedded steel plates in reinforced concrete structures," *Engineering structures*, vol. 20, no. 1, pp. 105–112, 1998.
 - [26] S. Chakraborty, "An experimental study on the behaviour of steel plate-anchor assembly embedded in concrete under biaxial loading," M.Tech Thesis, Indian Institute of Technology Kanpur, August 2006.
 - [27] S. Maya, "An experimental study on the effect of anchor diameter on the behavior of steel plate-anchor assembly embedded in concrete under biaxial loading," M.Tech Thesis, Indian Institute of Technology Kanpur, November 2008.
 - [28] D. K. Sahu, "Experimental study on the behavior of steel plate-anchor assembly embedded in concrete under cyclic loading," M.Tech Thesis, Indian Institute of Technology Kanpur, August 2004.
 - [29] V. Sonkar, "An experimental study on the behaviour of steel plate-anchor assembly embedded in concrete under constant compressive axial load and cyclic shear," M.Tech Thesis, Indian Institute of Technology Kanpur, September 2007.

- [10] X. Zhu, C. Chen, Q. Guo, J. Ma, F. Sun, and H. Lu, "Deep learning-based recognition of different thyroid cancer categories using whole frozen-slide images," *Front. Bioeng. Biotechnol.*, vol. 10, p. 857377, 2022, [CrossRef] [PubMed].
- [11] P. Chen, X. Shi, Y. Liang, Y. Li, L. Yang, and P. Gader, "Interactive thyroid whole slide image diagnostic system using deep representation," *Comput. Methods Programs Biomed.*, vol. 195, p. 105630, 2020, [CrossRef].
- [12] Y. Wang, Z. Chen, L. Zhang, D. Zhong, J. Di, X. Li, Y. Lei, J. Li, Y. Liu, R. Jiang *et al.*, "Fast classification of thyroid nodules with ultrasound guided-fine needle biopsy samples and machine learning," *Appl. Sci.*, vol. 12, p. 5364, 2022, [CrossRef].
- [13] S. Bhattacharya *et al.*, "Advances and challenges in thyroid cancer: The interplay of genetic modulators, targeted therapies, and ai-driven approaches," *Life Sciences*, vol. 332, 2023.
- [14] R. R. Kumar *et al.*, "Thyroid disease classification using machine learning algorithms," *E3S Web of Conferences*, 2023.
- [15] I. Imtiaz *et al.*, "The future of differentiated thyroid cancer recurrence prediction using a machine learning framework," *IEEE*, 2024.
- [16] L. Chen *et al.*, "A multimodal deep learning model for preoperative risk prediction of follicular thyroid carcinoma," *IEEE Healthcom*, 2023.
- [17] L. Bellal *et al.*, "Enhancing thyroid cancer diagnosis with advanced deep learning methods," *ICTIS Conference*, 2024.
- [18] A. S. El-Hossiny *et al.*, "Classification of thyroid carcinoma in whole slide images using cascaded cnn," *IEEE Access*, 2021.
- [19] D. R. Armentrout, "An analysis of the behavior of steel liner anchorages," Ph.D. dissertation, University of Tennessee, 1981.
- [20] R. H. Brown and A. R. Whitlock, "Strength of anchor bolts in grouted concrete masonry," *Journal of Structural Engineering*, vol. 109, no. 6, pp. 1362–1374, 1983.
- [21] J. Furche and R. Elinghausen, "Lateral blow-out failure of headed studs near a free edge," *Anchors in Concrete-Design and Behavior*, vol. SP-130, 1991.
- [22] R. A. Cook and R. E. Klingner, "Ductile multiple-anchor steel-to-concrete connections," *Journal of structural engineering*, vol. 118, no. 6, pp. 1645–1665, 1992.
- [23] D. P. Thambiratnam and P. Paramasivam, "Base plates under axial loads and moments," *Journal of Structural Engineering*, vol. 112, no. 5, pp. 1166–1181, 1986.
- [24] Z. Celep, "Rectangular plates resting on tensionless elastic foundation," *Journal of Engineering mechanics*, vol. 114, no. 12, pp. 2083–2092, 1988.

MANDEPUDI GOPI CHAKRADHAR

A Multimodal Deep Learning Framework for Thyroid Cancer Diagnosis Using Ultrasound Imaging and Clini

 Indian Institute of Information Technology Design And Manufacturing, Kurnool

Document Details

Submission ID

trn:oid:::3618:91636473

Submission Date

Apr 17, 2025, 12:24 PM GMT+5:30

Download Date

Apr 17, 2025, 12:28 PM GMT+5:30

File Name

A_Multimodal_Deep_Learning_Framework_for_Thyroid_Cancer_Diagnosis_Using_Ultrasound_Ima....pdf

File Size

2.1 MB

50 Pages

9,419 Words

58,238 Characters

9% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.





Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text




Exclusions

- 1 Excluded Match

Match Groups

-  **73 Not Cited or Quoted 9%**
Matches with neither in-text citation nor quotation marks
-  **0 Missing Quotations 0%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 7%  Internet sources
- 5%  Publications
- 0%  Submitted works (Student Papers)

Integrity Flags





0 Integrity Flags for Review

No suspicious text manipulations found.




Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

-  **73 Not Cited or Quoted 9%**
Matches with neither in-text citation nor quotation marks
-  **0 Missing Quotations 0%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 7%  Internet sources
- 5%  Publications
- 0%  Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet		
www.coursehero.com			1%
2	Internet		
mdpi-res.com			<1%
3	Internet		
doctorpenguin.com			<1%
4	Internet		
www2.mdpi.com			<1%
5	Internet		
dl.lib.uom.lk			<1%
6	Publication		
R. N. V. Jagan Mohan, B. H. V. S. Rama Krishnam Raju, V. Chandra Sekhar, T. V. K. P...			<1%
7	Internet		
etheses.whiterose.ac.uk			<1%
8	Internet		
arxiv.org			<1%
9	Publication		
Rohit Sharma, Gautam Kumar Mahanti, Ganapati Panda. "Performance Evaluatio...			<1%
10	Internet		
www.mdpi.com			<1%

11	Internet	psecommunity.org	<1%
12	Internet	oa.las.ac.cn	<1%
13	Internet	www.ijert.org	<1%
14	Internet	eprints.soton.ac.uk	<1%
15	Publication	Bryan R. Haugen, Erik K. Alexander, Keith C. Bible, Gerard M. Doherty et al. "2015 ...	<1%
16	Internet	umpir.ump.edu.my	<1%
17	Publication	Li, Jiacheng. "An Experimental Study on Consolidation of Saturated and Unsatura...	<1%
18	Internet	easychair.org	<1%
19	Internet	pubmed.ncbi.nlm.nih.gov	<1%
20	Publication	Yanan Che, Meng Zhao, Yan Gao, Zhibin Zhang, Xiangyang Zhang. "Application of ...	<1%
21	Internet	cuir.car.chula.ac.th	<1%
22	Internet	discovery.researcher.life	<1%
23	Internet	www.science.gov	<1%
24	Publication	"Large Scale Network-Centric Distributed Systems", Wiley, 2013	<1%

25	Publication	Yassine Habchi, Hamza Kheddar, Yassine Himeur, Adel Belouchrani, Erchin Serpe...	<1%
26	Internet	docksci.com	<1%
27	Internet	web.archive.org	<1%
28	Publication	Pirani, Rayhaan. "Anomaly Detection in Large Datasets: A Case Study in Loan Def...	<1%
29	Internet	catalonica.bnc.cat	<1%
30	Internet	dokumen.pub	<1%
31	Internet	dr.ntu.edu.sg	<1%
32	Internet	dspace.library.uvic.ca	<1%
33	Internet	www.frontiersin.org	<1%
34	Internet	www.grossarchive.com	<1%
35	Internet	www.jpx.co.jp	<1%
36	Internet	www.researchsquare.com	<1%
37	Internet	www.um.edu.mt	<1%
38	Publication	Amit Kumar Tyagi, Shrikant Tiwari, S. V. Nagaraj. "Quantum Computing - The Fut...	<1%

39

Publication

de Oliveira Fonseca, Antonio Henrique. "Learning Brain Dynamics With Neural Op...

<1%