BS THESIS

DEEP LEARNING-BASED DETECTION AND CLASSIFICATION OF BREAST CANCER USING ULTRASOUND IMAGES

SUBMITTED BY	MUHAMMAD SALMAN
RIGESTRITION NUMBER	UOS216100109
RESEARCH SUPERVISOR	MR. AHMAD HUSSAIN
	ASSISTANT PROFESSOR

DEPARTMENT OF COMPUTER SCIENCE

GOVERNMENT POSTGRADUATE JAHANZEB COLLEGE

SAIDU SHARIF, SWAT

AFFLIATED WITH

UNIVERSITY OF SWAT

(SESSION: 2021-25)

APPROVAL CERTIFICATE

This is to certify that the work contained in this thesis entitled **"DEEP LEARNING-BASED DETECTION AND CLASSIFICATION OF BREAST CANCER USING ULTRASOUND IMAGES"** Submitted by "Muhammad Salman" was carried out under my supervision and in my opinion, is fully adequate in scope and quality for the degree of Bachelor of Science in Computer Science.

Supervisor _____

Ahmad Hussain

(Assistant Professor)

External Examiner _____

Name: …….

Designation: ……...

College/University: …….....

Chairman _____

Name: Yasir Arafat

Chairman Department of Computer Science

Govt Post Graduate Jahanzeb College Saidu Sharif, Swat

Date: ___/___/2025

DECLERATION

I solemnly declare that the research work presented in the thesis titled “Deep Learning-Based Detection and Classification of Breast Cancer Using Ultrasound Images” is my original work, carried out under the guidance of my supervisor and with the academic support of my teachers.

I further affirm that this thesis has not been submitted, either in whole or in part, for the award of any other degree or diploma at this or any other university or academic institution. No part of this work has been copied or plagiarized from any published or unpublished source, and any material used from other sources has been properly cited and acknowledged.

I fully understand that in the event of any violation of academic integrity, including plagiarism or misrepresentation, the university has the right to take disciplinary action in accordance with HEC policies, and I will accept the consequences thereof.

Muhammad Salman

Signature_____

Salmanbnr5@gmail.com

DEDICATION

With profound gratitude and deep respect, I dedicate this humble effort to those who have been the foundation of my strength and success.

TO MY BELOVED PARENTS

This thesis is most sincerely dedicated to my loving and divinely gifted parents, whose endless prayers, unconditional support, and unwavering belief in me have been the guiding light throughout my academic journey. Their teachings of faith, perseverance, humility, and the value of hard work have shaped every success in my life. They have been my first mentors instilling in me the courage to dream and the strength to achieve.

TO MY RESPECTED TEACHERS

I am also deeply indebted to my esteemed teachers, whose wisdom, encouragement, and guidance have been instrumental throughout the course of this research. Their dedication to imparting knowledge and their sincere mentorship have not only helped me complete this thesis but have also illuminated the path toward my future aspirations.

Lastly, I dedicate this work to all those who strive to be the voice of the voiceless, and who devote their lives to making a difference in the world through knowledge and compassion.

ACKNOWLEDGEMENT

All praise and gratitude is due to Almighty Allah, the Most Gracious and the Most Merciful, who blessed me with strength, health, patience, and determination to complete this research. Without His divine guidance and countless blessings, the successful completion of this thesis would not have been possible.

I extend my deepest and most sincere gratitude to my beloved parents, whose endless prayers, unconditional love, and unwavering support have been the cornerstone of every achievement in my life. Their sacrifices, encouragement, and belief in my potential have provided me with the strength to overcome every challenge throughout my academic journey.

I would also like to express my heartfelt appreciation to my respected supervisor, Mr. Ahmad Hussain, for his continuous guidance, valuable insights, and unwavering support throughout the course of this study. His thoughtful mentorship and constructive feedback were instrumental in shaping the direction and quality of this research.

I am thankful to my institute for providing me with the platform and opportunity to undertake this research as part of my academic journey. The environment of learning and growth it offered has played a significant role in my development.

A special note of gratitude goes to my beloved siblings, especially my loving sisters, whose moral support, encouragement, and belief in me have been a constant source of motivation during my undergraduate studies. I am equally thankful to my cousins and roommates, whose companionship, late-night discussions, and shared struggles made this journey both memorable and manageable.

While words may fall short in expressing the depth of my appreciation, my heart remains filled with immense gratitude for all those mentioned and unmentioned who played a role in this journey. Your support has been invaluable.

PREFACE

This thesis, titled *"Deep Learning-Based Detection and Classification of Breast Cancer Using Ultrasound Images"*, is submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science at the Government Postgraduate Jahanzeb College Saidu Sharif, Swat Affliated with University of Swat.

The motivation behind this research stems from the growing need for efficient, accessible, and interpretable diagnostic tools to support early breast cancer detection, particularly in resource-constrained regions. This study explores the integration of convolutional neural networks and transformer-based attention mechanisms to develop a hybrid architecture that can accurately classify breast ultrasound images.

The work presented in this thesis reflects my efforts to address key challenges in medical image analysis — including class imbalance, feature representation, and model interpretability — by designing and evaluating a deep learning framework that is both effective and computationally efficient.

The thesis is structured as follows:

Chapter 1 – Introduction
Introduces the background, motivation, research problem, objectives, and provides a brief overview of the proposed approach.

Chapter 2 – Literature Review
Reviews existing research related to deep learning in medical imaging, particularly ultrasound analysis, and highlights the strengths and gaps in current methods.

Chapter 3 – Dataset and Methodology
Describes the dataset used, the data preprocessing steps, and presents the proposed hybrid deep learning solution in the context of the overall methodology.

Chapter 4 – Fundamental Building Blocks of the Model
Details the core components of the deep learning architecture, including CNNs, EfficientNetB0, MobileViT, and associated layers and techniques.

Chapter 5 – Proposed HVIT-AAF Architecture

Provides an in-depth explanation of the Hybrid Vision Transformer with Adaptive Attention Fusion (HVIT-AAF), highlighting the architectural flow and integration of its components.

Chapter 6 – Results and Analysis

Presents the experimental results, training and validation performance, evaluation metrics, and analysis of model behavior.

Chapter 7 – Experimental Setup and Frontend Interface

Describes the training environment, model training protocol, and the design of a user-friendly frontend interface for real-world usage.

Chapter 8 – Conclusion and Future Work

Summarizes the research contributions, discusses limitations, and outlines possible directions for future improvements and applications.

I hope that this work contributes meaningfully to the growing field of AI-driven medical diagnostics and serves as a stepping stone for further innovation in ultrasound-based cancer detection.

Muhammad Salman
BS Computer Science
Government Postgraduate Jahanzeb College Saidu Sharif, Swat
Affiliated with University of Swat
(Session 2021–2025)

ABSTRACT

Breast cancer remains a leading cause of mortality among women globally, with early detection critical for improving patient outcomes. Ultrasound imaging, valued for its cost-effectiveness, safety, and portability, is particularly vital in resource-constrained regions like Pakistan, where younger women face a higher disease burden. However, challenges such as speckle noise, class imbalance, and limited interpretability hinder automated diagnosis using ultrasound images. This thesis proposes a novel deep learning framework, the Hybrid Vision Transformer with Adaptive Attention Fusion (HVIT-AAF), to enhance breast cancer detection and classification from ultrasound images. Integrating EfficientNetB0 for local feature extraction with MobileViT-inspired transformer blocks for global context, the model employs an Adaptive Focal Loss function to address class imbalance and advanced interpretability techniques, including Grad-CAM, LIME, and URLAB, to ensure transparent decision-making. Evaluated on the augmented Breast Ultrasound Images (BUSI) dataset, the model achieved a validation accuracy of 99.11%, an Area Under the Curve (AUC) of 99.84%, precision of 99.11%, recall of 99.05%, and an inference time of 0.0094 seconds per image. A user-friendly frontend interface further facilitates clinical interaction. These results demonstrate the model's potential to support scalable, accurate, and interpretable breast cancer screening, particularly in low-resource settings, contributing to improved diagnostic precision and healthcare accessibility

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Full Form
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under the Curve
BUSI	Breast Ultrasound Images
CNN	Convolutional Neural Network
CT	Computed Tomography
DFViT	Deep Fusion-based Vision Transformer
GAP	Global Average Pooling
Grad-CAM	Gradient-weighted Class Activation Mapping
URLAB	Unified Regions from LIME And Backpropagation
GPU	Graphics Processing Unit
HVIT-AAF	Hybrid Vision Transformer with Adaptive Attention Fusion
K	Kernel (in convolution)
LIME	Local Interpretable Model-Agnostic Explanations
LSTM	Long Short-Term Memory
MBConv	Mobile Inverted Bottleneck Convolution
MRI	Magnetic Resonance Imaging
ReLU	Rectified Linear Unit
ROC	Receiver Operating Characteristic
SE	Squeeze-and-Excitation
S-DPN	Stacked Deep Polynomial Network
ViT	Vision Transformer
MLP	Multi-Layer Perceptron
Swish	Self-Gated Activation Function
U-Net	U-shaped Convolutional Network

CHAPTER 1

INTRODUCTION

Breast cancer continues to be one of the leading causes of illness and mortality among women worldwide. According to the Global Cancer Observatory and other epidemiological studies, approximately 2.3 million new cases are diagnosed each year with nearly 666,000 deaths, making it the most commonly diagnosed malignancy among women globally (Bray et al., 2024)**.** Early detection and accurate diagnosis are crucial for effective treatment and improved patient outcomes. Over the past few decades, medical imaging has played a pivotal role in cancer detection by enabling clinicians to visualize internal structures non-invasively. Among various imaging modalities, ultrasound has emerged as a particularly attractive option for breast cancer screening due to its unique combination of benefits.

1.1 Background and Motivation

Comparing ultrasound imaging to other modalities like mammography and magnetic resonance imaging (MRI), it is naturally non-ionizing, portable, and affordable (Aps, 2020; Hasan & Khouri, 2023). These characteristics make ultrasound an ideal candidate for widespread screening programs, especially in resource-constrained regions like Pakistan where healthcare budgets and infrastructure may be limited (Iacob et al., 2024). Moreover, ultrasound is especially beneficial for younger women, whose denser breast tissue often diminishes the sensitivity of mammography (Mustafa & Mustafa, 2024; Pangaribuan, 2024). This advantage not only increases diagnostic accuracy for a demographic that is traditionally underserved in breast cancer screening but also aligns with the ethical imperative to minimize radiation exposure in young patients.

In Pakistan, the burden of breast cancer is particularly severe. Research indicates that Pakistani women are diagnosed at a younger age compared to their Western counterparts with many cases occurring before the age of 50 (Majeed & Bangash, 2024). This epidemiological trend underscores the necessity of adopting imaging modalities that perform well in dense, young breast tissue. Consequently, integrating ultrasound into national screening programs in Pakistan is especially promising, as it

can provide rapid, cost-effective, and high-quality imaging that addresses the local demographic profile.

1.2 Research Problem and Objectives

Despite the remarkable progress in deep learning and medical imaging, several challenges persist when applying these techniques to ultrasound images. Ultrasound images are often characterized by speckle noise, low contrast, and operator dependency, complicating the task of automated diagnosis. Furthermore, class imbalance common in medical datasets can lead to biased predictions if not addressed appropriately. To overcome these challenges, our research proposes a hybrid deep learning architecture that combines the strengths of EfficientNet based convolutional neural networks (CNNs) with mobile vision transformer (MobileViT) blocks. This fusion is further enhanced by an Adaptive Focal Loss function to mitigate class imbalance during training.

The following are the main goals of this study:

i. **Create a solid model:** Create a system that successfully combines the transformer and CNN concepts to extract local and global information from ultrasound pictures.
ii. **Improve diagnostic accuracy:** Attain cutting-edge performance measures, as demonstrated by fast inference times (~0.0094 seconds per image), an accuracy of 99.11%, and an AUC of 99.84%.
iii. **Enhance interpretability:** The model displays portions of images that influence predictions and offers the transparency required for clinical acceptance by employing techniques like Gradient-weighted Class Activation Mapping (Grad-CAM), Local Interpretable Model-agnostic Explanations (LIME), and Unified Regions from LIME And Backpropagation (URLAB), which together highlight and validate the most significant regions contributing to the model's decisions.
iv. **Ensure scalability and accessibility:** Focus on methods that are computationally efficient and can be readily deployed in low-resource settings, thereby broadening access to high-quality diagnostic tools.

1.3 Advantages of Ultrasound Imaging

Ultrasound imaging offers several distinct advantages that make it especially attractive for breast cancer screening in both developed and developing regions:

i. **Cost-Effectiveness:** Ultrasound devices are relatively inexpensive compared to MRI or CT scanners, making them highly suitable for healthcare systems in economically challenged regions, including many parts of Pakistan (Pachisia & Govil, 2025).
ii. **Safety:** Unlike the ionizing radiation used in mammography, ultrasound is safe for repeated use, which is particularly important for younger patients and for follow-up examinations.
iii. **Real-Time Imaging:** Ultrasound provides dynamic, real-time imaging that can be crucial in guiding biopsies and other interventional procedures.
iv. **Portability:** The compact nature of modern ultrasound devices allows for deployment in rural and remote areas, bridging the gap between urban and rural healthcare services (Anderson & Theophanous, 2025).

These benefits underscore the potential of ultrasound not only as a primary screening tool for breast cancer but also as a platform for deploying advanced machine learning models. Given that Pakistani women tend to develop breast cancer at a younger age with many cases occurring before the age of 50 ultrasound imaging is particularly effective in this demographic due to its superior performance in imaging dense breast tissue (Mustafa & Mustafa, 2024; Pangaribuan, 2024).

1.4 Overview of the Proposed Approach

The suggested model, known as HVIT-AAF, is a creative combination of tried-and-true deep learning methods designed specifically for ultrasound picture processing. The architecture utilizes EfficientNetB0 as its base, renowned for its balance between efficiency and performance (Galabuzi et al., 2024). The extracted features are further refined using MobileViT blocks that leverage multi-head self-attention mechanisms to capture intricate patterns within the image data. An Adaptive Focal Loss function is employed during training to address class imbalance a common challenge in medical imaging datasets.

The entire training procedure is implemented using TensorFlow, and it is subsequently optimized for GPU acceleration and evaluated using dependable performance metrics such as accuracy, AUC, precision, and recall. The incorporation of Grad-CAM also offers a visual depiction of the model's decision-making process, which is crucial for clinical validation and uptake (Suara et al., 2023; T. R et al., 2024).

In conclusion, by putting forth a novel hybrid deep learning framework that takes advantage of the special benefits of ultrasonic imaging, this work advances the field of medical picture analysis. This study intends to offer a useful and significant solution for early breast cancer diagnosis by emphasizing computational efficiency, interpretability, and accessibility, especially helping young patients in underprivileged areas like Pakistan.

CHAPTER 2

LITERATURE REVIEW

Breast cancer is one of the most prevalent and fatal diseases affecting women globally, with early detection crucial for improving survival rates and treatment effectiveness (Mathpal, 2024). The categorization of breast ultrasound pictures into benign, malignant, or normal classifications is essential for assisting radiologists and specialists in making educated diagnostic choices. Medical image analysis has been transformed by recent developments in deep learning, especially convolutional neural networks (CNNs), which provide automated and precise methods for classifying breast cancer (Litjens et al., 2017). This chapter reviews the existing literature on deep learning techniques applied to breast cancer ultrasound image classification, focusing on approaches using CNNs, EfficientNet, hybrid architectures, and other notable models. The review provides a thorough basis for comprehending the state-of-the-art model in this field by including research that address binary (benign vs. malignant) and ternary (benign, malignant, normal) classification problems.

2.1 Convolutional Neural Networks (CNNs)

CNNs have been extensively employed in the classification of breast ultrasound images due to their ability to automatically extract hierarchical features from raw data (LeCun et al., 2015). Early studies demonstrated the potential of CNNs in distinguishing benign from malignant tumors. For instance, (Wang et al., 2024) Constructed a CNN-based model utilizing GoogLeNet to categorize breast ultrasound images from a dataset including 7,909 images, attaining an accuracy of 91%. The research emphasized the efficacy of transfer learning, wherein pre-trained weights from ImageNet were refined to suit the medical imaging sector. This method exhibited considerable enhancements in diagnostic precision, yielding a pooled sensitivity of 0.85 (95% CI: 0.80–0.89) and a pooled specificity of 0.86 (95% CI: 0.78–0.92). The researchers underlined the significance of data augmentation strategies to lessen class imbalance and improve model performance, especially in differentiating between benign and aggressive breast cancers. A meta-analysis demonstrated the model's strong diagnostic capability, showing an AUC value of 0.92 with a 95% confidence interval ranging from 0.89 to

0.94. These findings demonstrate how deep learning models, such as GoogLeNet, can improve computer-aided diagnostics for breast cancer diagnosis.

Another notable work by (Labonno et al., 2025) developed a deep learning technique for the early detection and categorization of breast cancer via ultrasound pictures. Their method made use of the Kaggle dataset for Breast Cancer Image Classification, which contained 9,248 ultrasound pictures divided into benign, malignant, and normal categories. The authors implemented three pre-trained convolutional neural networks (CNNs), ResNet50, MobileNet, and VGG16—alongside a new CNN model. Among them, ResNet50 achieved the best accuracy of 98.41%, followed by VGG16 (98.19%), MobileNet (97.91%), and the custom CNN model (92.94%). The findings illustrate the efficacy of deep learning methods in breast cancer detection, with ResNet50 displaying enhanced performance owing to its deep residual connections that facilitate the extraction and classification of features. The study also evaluated the model's effectiveness using recall, precision, F1-scoring and AUC-ROC, confirming ResNet50's robustness in medical image analysis for cancer detection (Labonno et al., 2025).

Similarly, (Cheyi & Kaya, 2024) suggested an advanced CNN-based technique for breast cancer classification using ultrasound imaging. Their investigation included multiple state-of-the-art designs, including VGG16, VGG19, ResNet50, DenseNet121, EfficientNetB0, and a bespoke CNN model. The suggested model included significant preprocessing approaches with Grad-CAM for interpretability, obtaining a remarkable classification accuracy of 97%, beating current models such as EfficientNetB0, MobileNet, and InceptionV3. This study emphasizes the promise of deep learning in boosting diagnostic accuracy and expanding the interpretability of medical imaging models.

2.2 EfficientNet-Based Approaches

EfficientNet, introduced by (Tan & Le, 2019), is a family of CNN models recognized for its scalability and computational efficiency, achieved by a compound scaling strategy that balances network depth, width, and resolution. Its application in breast cancer ultrasound categorization has gained popularity due to its superior performance with lower computing cost. (Banerjee & Monir, 2023) proposed CEIMVEN, a deep

learning method that uses modified versions of EfficientNet-V1 (b0-b7) and EfficientNet-V2 (b0-b3) for the detection and classification of breast cancer from ultrasound images. Their method employed transfer learning, hyperparameter tuning, and fully linked layers, producing excellent classification accuracy. The best-performing models, EfficientNet-V2-b3 and EfficientNet-V1-b7, had validation accuracies of 99.43% and 99.89%, respectively. The study highlights how well-suited sophisticated CNN architectures are for improving ultrasound imaging-based breast cancer detection.

In a more recent experiment, (Mahesh et al., 2025) refined the EfficientNetB7 architecture for breast ultrasonography classification, addressing interclass variability and class imbalances through targeted augmentation and adaptive training procedures. Their model achieved 98.29% accuracy, surpassing existing benchmarks and suggesting high potential for early breast cancer diagnosis. For the classification of breast lesions in ultrasound pictures, (Jabeen et al., 2024) proposed an EfficientNet-integrated ResNet deep network combined with explainable AI. Their strategy included data augmentation, deep transfer learning, and a unique feature selection technique based on the cuckoo search algorithm. The suggested system obtained classification accuracies of 98.4% and 98% in two distinct experiments on the BUSI dataset, exceeding recent techniques while maintaining computing economy.

These findings show EfficientNet's potential as a robust backbone for breast cancer classification, particularly when computational resources are constrained.

2.3 Hybrid Models Combining CNNs and Other Architectures

Hybrid models, which blend CNNs with other deep learning architectures, have been researched to exploit complimentary characteristics for greater classification performance. In order to diagnose breast cancer, (Hammood & Hasan, 2024) presented a hybrid LSTM-CNN architecture that combines the feature extraction capabilities of convolutional neural networks (CNN) with the sequential modeling abilities of long short-term memory (LSTM). The model attained a training accuracy of 98.99% and a testing accuracy of 94.63% on a real patient dataset, beating typical CNN classifiers. This approach highlights the integration of temporal and spatial learning to enhance the accuracy of medical image diagnosis.

Another hybrid technique by (Fiaz et al., 2024) suggested a deep fusion-based Vision Transformer (DFViT) model for breast cancer classification, combining convolutional networks like VGG16 to extract localized features with transformers to capture broader context. The model uses image fusion techniques and a multi-layer perceptron for classification, attaining 95.29% accuracy and an F1 score of 97.68% on the BreakHis dataset, beating state-of-the-art algorithms. Similarly, for the segmentation of ultrasound pictures, (Zhang et al., 2024) presented DDTransUNet, a CNN-Transformer hybrid network with two branches and two attention methods. Through cross-attention fusion, the model showed that it could effectively collect both local and global information, achieving state-of-the-art outcomes with Dice scores of 82.31%, 88.23%, and 90.33% on the TN3K, BUS-BRA, and CAMUS datasets, respectively.

A different hybrid strategy was employed by (Benaouali et al., 2024) presented a CNN-based classification system for breast tumor detection in ultrasound images, exploring both handmade (LBP, HOG) and deep learning methods (ResNet50, DenseNet201, InceptionV3). Their best-performing model, DenseNet201 with an ANN classifier, attained 100% accuracy after dataset fusion, indicating the network's robustness and generalization capabilities. These hybrid techniques highlight the potential of mixing several designs to handle the problems of ultrasound images classification, such as noise and diversity in tumor appearance.

2.5 Other Deep Learning Architectures

Beyond CNNs and EfficientNet, various deep learning architectures have been applied to breast cancer ultrasound categorization. A stacked deep polynomial network (S-DPN) for ultrasound-based tumor classification was introduced by (Shi et al., 2016), resolving the issue of small sample size that is common in medical imaging. The S-DPN displayed superior feature representation learning compared to typical deep learning models, reaching classification accuracies of 92.40% and 90.28% on breast and prostate ultrasound datasets, respectively. These results illustrate S-DPN's potential for robust texture feature learning in small medical datasets.

BIRADS-SDL, a unique semi-supervised deep learning network for breast ultrasound categorization, was introduced by (Zhang et al., 2020). The architecture combines BIRADS features with a layered convolutional auto-encoder, balancing image reconstruction and lesion classification tasks. The model achieved classification

accuracies of 92.00% and 83.90% on two breast ultrasound datasets, exhibiting improved performance, particularly with small datasets.

The Deep Integrated Pipeline, proposed by (Inan et al., 2022), combines SLIC-based unsupervised segmentation with a modified U-Net for breast ultrasound images segmentation, followed by VGG16 for classification. This end-to-end framework attained an accuracy of 73.72% and an F1 score of 78.92% for benign tumor diagnosis, suggesting its potential to boost diagnostic accuracy in noisy ultrasound images.

ViT-based networks, proposed by (Gheflati & Rivaz, 2022), utilized self-attention processes for breast ultrasound categorization, solving CNNs' shortcomings in collecting global context. They achieved up to 86.7% accuracy and an AUC of 0.95 in their experiments with several Vision Transformer topologies, outperforming or matching cutting-edge CNNs like ResNet50. This demonstrates the potential of ViTs for learning complex spatial connections in medical imaging.

These several architectures demonstrate the variety of deep learning techniques applied to breast cancer ultrasound categorization, each with special benefits depending on the application scenario.

2.5 Datasets and Evaluation Metrics

In breast ultrasound imaging research, numerous datasets are routinely exploited to construct and test diagnostic models. The Breast Ultrasound Images (BUSI) collection, publicly available from Baheya Hospital in Cairo, Egypt, has 780 images divided into 133 normal, 437 benign, and 210 malignant cases (Al-Dhabyani et al., 2020). Researchers commonly apply data augmentation techniques on this dataset to alleviate class imbalance and boost model robustness.

Another important dataset is the Open Access Series of Breast Ultrasonic Data (OASBUD), which comprises raw ultrasound data from 100 patients, with 52 malignant and 48 benign lesions (Piotrzkowska-Wróblewska et al., 2017). Additionally, the Breast Ultrasound Image Segmentation (BUSIS) dataset contains 562 pictures with large differences in contrast, brightness, and noise levels, offering a tough standard for segmentation algorithms (Zhang et al., 2022).

Metrics for evaluation such as area under the receiver operating characteristic curve (AUC), recall, accuracy, precision, and F1-score are frequently used in breast ultrasound imaging research to evaluate model performance. An AUC of 0.97, for instance, indicated that the model was effective in distinguishing between benign and malignant tumors in a research using the Breast Ultrasound Image (BUSI) dataset (Zeimarani et al., 2020). These metrics provide a thorough evaluation of the model's performance, which is crucial when working with imbalanced datasets when there are frequently few malignant cases.

2.8 Research Gap

Despite significant advancements in both CNN-based and EfficientNet-based models, they still fall short in capturing the global contextual information needed for accurate diagnostic performance in breast cancer ultrasound classification. CNNs are intrinsically constrained in their ability to depict long-range interdependence due to their narrow receptive fields. Furthermore, issues like class imbalance in ultrasound datasets persist, leading to skewed predictions toward majority classes and reduced generality.

Furthermore, most previous research does not adequately address challenges such as class imbalance, decision interpretability, and computational efficiency on low-resource systems. Although a number of hybrid models that combine CNNs with other architectures have been examined, few methods combine transformer-based attention mechanisms with EfficientNet's efficient feature extraction to simultaneously handle local and global feature representation. A hybrid architecture that integrates local and global feature extraction, corrects class imbalance during training, and employs advanced interpretability techniques is needed to produce findings that can be explained.

2.9 Conclusion

The literature on breast cancer ultrasound image classification reflects a rich landscape of deep learning approaches, ranging from traditional CNNs to advanced EfficientNet models, hybrid architectures, and emerging transformers. Studies have demonstrated high accuracies, often exceeding 90%, with EfficientNet and hybrid models showing

particular promise due to their efficiency and ability to integrate diverse feature representations. However, challenges such as dataset limitations, imaging variability, and computational demands highlight areas for future research. This review provides a foundation for understanding the current methodologies and their contributions to automated breast cancer diagnosis, setting the stage for further innovations in the field.

CHAPTER 3

DATASET AND METHODOLOGY

The dataset and the methodological framework used for data preparation and augmentation in this investigation are thoroughly described in this chapter. It describes the preparation procedures, such as image resizing, class distribution analysis, and train-validation splitting, and goes into detail about the unique augmentation approach used to balance the dataset. It also describes the features of the Breast Ultrasound Images (BUSI) dataset. The use of class weighting is also covered as a preventative step to lessen possible biases during training. These systematic procedures guarantee that the data pipeline successfully tackles the innate difficulties of sparse and unbalanced medical imaging data, promoting the development of a robust and widely applicable deep learning model for the classification of breast cancer.

3.1 Dataset Description (BUSI Augmented Dataset)

The present study is based on the publicly available Breast Ultrasound Images (BUSI) dataset, provided by Baheya Hospital in Cairo, Egypt (Al-Dhabyani et al., 2020). The 780 grayscale ultrasound pictures of breast tissue in this dataset are divided into three different categories: benign, malignant, and normal. In particular, there are 133 normal photos in the collection, which show breast tissue that is healthy and free of anomalies. The benign class contains 437 images depicting non-cancerous growths or lesions, while the malignant class consists of 210 images showing cancerous tumors. This distribution reflects the higher occurrence of benign findings in clinical practice but also introduces a class imbalance that could affect model training if not addressed.

Each image captures breast tissue characteristics in female patients aged 25 to 75 years, providing a diverse representation of anatomical and pathological variations across age groups. In addition to the ultrasound images, the dataset includes corresponding mask images that outline regions of interest, such as tumors or lesions. However, in this classification-focused study, these mask images were intentionally excluded during training to prevent the model from learning non-essential artifacts. This choice

enhances the model's ability to adapt to unseen data by preventing overfitting and emphasizing core tissue features, which is vital for real-world clinical use.

To mitigate the impact of class imbalance and enhance the dataset's variability, a custom data augmentation strategy was applied. This process expanded each class approximately to 1850 images, resulting in a total of 5603 augmented images. The model was able to identify more widely applicable patterns by increasing the variability using augmentation approaches like rotations, flips, and intensity adjustments. The final augmented dataset was uploaded to Kaggle to promote research reproducibility and facilitate consistent model evaluation across different experimental settings. By making the dataset accessible, this study supports collaborative research and encourages further advancements in breast cancer detection using deep learning techniques.

3.2 Proposed Solution

The current study suggests a novel hybrid deep learning architecture called HVIT-AAF (Hybrid Vision Transformer with Adaptive Attention Fusion) to address the issues noted in the literature review. The proposed method aims to combine the benefits of convolutional neural networks (CNNs) and transformer-based attention mechanisms to effectively capture both local and global properties in ultrasound pictures. The core feature extractor that balances efficiency and performance is EfficientNetB0 in particular. These extracted elements are refined and enhanced using a custom transformer-inspired block built on the MobileViT framework. This component captures distant relationships and overall context through the use of multi-head self-attention.

A key component of the proposed approach is the Adaptive Attention Fusion method, which dynamically highlights diagnostically relevant regions in the feature maps. By adaptively weighing transformer-derived global features and CNN-derived local features, this method ensures that the most informative patterns are prioritized during classification.

In order to further address the issue of class imbalance that is commonly observed in medical imaging datasets, the design incorporates an Adaptive Focal Loss function during training. This loss function reduces bias toward majority classifications and

enhances the model's generalizability across different types of breast lesions by emphasizing cases that are more challenging to categorize.

Furthermore, the suggested design incorporates interpretability strategies like Local Interpretable Model-Agnostic Explanations (LIME), Gradient-weighted Class Activation Mapping (Grad-CAM), and Unified Regions from LIME And Backpropagation (URLAB) to guarantee that the model's conclusions are comprehensible and therapeutically useful. By providing local, visual, and consensus-based explanations of model predictions, these methods collectively align the AI system with the clinical need for accountable and transparent diagnostic tools.

Because of the solution's overall computing efficiency and adaptability for deployment in low-resource contexts, high-quality diagnostic tools are now more readily available. By combining these components into a single framework, the proposed system seeks to solve the limitations identified in the literature and provide a dependable, understandable, and scalable approach for automated breast cancer screening using ultrasound images.

3.3 Pipeline for Data Preprocessing

The data preparation pipeline is crucial for transforming unprocessed ultrasound pictures into a format that can be utilized for deep learning model training. Details on the pipeline's picture scaling, class distribution analysis, and dataset splitting procedures are given in the ensuing subsections.

3.3.1 Image Resizing (300×300)

Ultrasound images typically vary in dimensions, which can pose challenges during model training by introducing inconsistencies in input size. In order to remedy this, bicubic interpolation was used to enlarge each image to a consistent dimension of 300×300 pixels. This method is well-suited for medical imaging tasks, as it effectively preserves edge details and structural features while minimizing artifacts introduced during the resizing process. Standardizing image dimensions is essential for reducing computational complexity and ensuring that all images follow a consistent format, allowing the neural network to process the data more efficiently. By providing uniform input sizes, this approach contributes to more stable learning, enabling the model to

focus on feature extraction and pattern recognition without being affected by variations in image resolution.

3.3.2 Class Distribution Analysis

An initial analysis of the dataset revealed a major class imbalance, with a significantly higher number of benign and malignant cases than the normal class. 133 normal images (17%), 437 benign images (56%), and 210 malignant images (27%), made up the initial distribution. The neural network may become biased toward the majority class (benign) as a result of this imbalance, which could skew model performance and reduce its ability to identify normal or malignant situations. The model's clinical application, where precisely identifying every class is essential for a trustworthy diagnosis, may be jeopardized by such bias. This was resolved by employing a systematic data augmentation strategy to increase each class to about 1850 photographs. This augmentation technique reduced the likelihood of biased learning and promoted the model's generalization across all categories by creating a balanced dataset, thus enhancing the model's robustness and predictive power.

3.3.3 Train-Validation Split

TensorFlow's image_dataset_from_directory was used to divide the dataset into training (70%) and validation (30%) using a fixed seed of 42 for repeatability in order to assess the model's generalization and minimize overfitting. Following augmentation, there were roughly 1307 images in each class (3923 total) in the training set and 560 images in each class (1680 total) in the validation set. Metrics like accuracy and F1-score can accurately reflect the model's performance in differentiating between benign, malignant, and normal instances because class balance was maintained in both sets to enable fair evaluation.

3.4 Custom Data Augmentation Strategy

In this study, data augmentation played a crucial role by expanding the dataset and introducing diversity, thereby enhancing the model’s robustness and generalization ability. The augmentation pipeline was implemented using TensorFlow’s Keras ImageDataGenerator class, with carefully chosen transformation parameters designed to mimic realistic variations encountered in clinical ultrasound imaging. These transformations helped the model become more resilient to factors such as slight

changes in orientation, scale, and noise, ultimately enhancing its ability to learn meaningful features from diverse image representations. By enriching the dataset with augmented samples, more dependable performance across invisible data was made possible by this method, which is crucial for practical medical applications.

3.4.1 Augmentation Parameters and Implementation

The data augmentation process was essential for enriching the dataset, simulating realistic clinical variations, and addressing class imbalance. Various transformations were applied to introduce diversity and help the model generalize better to unseen data. These transformations included rotation (±20 degrees**)** to mimic different probe angles, width and height shifts (±10%) to account for slight positioning variations, and shear transformation (intensity 0.1) to simulate tissue deformation. Additionally, zooming (±10%) represented changes in probe distance, horizontal flipping simulated left-right breast symmetry, and Gaussian noise added random pixel intensity variations to replicate imaging artifacts.

Each class was augmented to reach approximately 2000 images, with the number of augmentations per original image determined by the initial class size. For the malignant class (210 images), approximately 9 augmentations per image were generated. The benign class (437 images) required around 4 augmentations per image, while the normal class (133 images) needed about 15 augmentations per image to achieve balance. The augmented images were stored in a structured directory, ensuring a reproducible and well-organized dataset for model training and evaluation. This augmentation strategy not only mitigated class imbalance but also enhanced the model's ability to learn robust features under diverse imaging conditions.

Table 3.1 *Class Distribution and Augmentation Details*

Class	**Original Image Count**	**Original Class Proportion (%)**	**Final Augmented Count**	**Training Set (70%)**	**Validation Set (30%)**
Normal	133	17	1858	1300	558
Benign	437	56	1895	1326	569
Malignant	210	27	1850	1295	555

The table above provides a comprehensive summary of the dataset's evolution, outlining the original class distribution, augmentation strategy, and the final division into training and validation subsets. It illustrates how data augmentation was used to balance the classes, ensuring that each category had an equal number of images for more reliable model training and evaluation.

3.5 Class Weighting for Imbalanced Data

Although the dataset was balanced through augmentation, class weighting was applied as a precautionary measure during model training to maintain flexibility for future scenarios. Class weights were calculated using the formula:

$$\text{class\_weight}_i = \frac{\text{Total Number of Samples}}{\text{Number of Classes} \times \text{Count of Class } i}$$

After the 70-30 split, each class contained an equal number of training samples (3923 images), leading to uniform class weights of 1. While this made weighting unnecessary for the current dataset, keeping the class weighting mechanism aligns with best practices in deep learning for medical imaging. This method guarantees that the model remains adaptable to potential future datasets, particularly those with inherent class imbalances, where augmentation may not be feasible or desirable. Retaining class weighting prepares the model to handle skewed data distributions without requiring extensive data augmentation, enhancing its applicability to a wider range of clinical scenarios.

3.6 Summary

In summary, this chapter presented a detailed account of the dataset and the methodological framework employed to prepare the data for effective deep learning-based breast cancer classification. It included image resizing, class distribution analysis, and a carefully designed train-validation split. A custom augmentation strategy introduced a diverse set of transformations to improve model generalization and mitigate class imbalance, which was further reinforced by the inclusion of a class weighting mechanism.

In addition to these preprocessing strategies, this chapter also introduced the proposed hybrid architecture HVIT-AAF (Hybrid Vision Transformer with Adaptive Attention Fusion). This section outlined the conceptual design of the model, which integrates EfficientNetB0, transformer-inspired blocks, and an adaptive attention mechanism to capture both local and global features. The solution is tailored to address key challenges in ultrasound image classification, including class imbalance, interpretability, and computational efficiency.

The data pipeline and suggested model work together to create a thorough and reliable basis for further training and assessment, which are covered in the upcoming chapters.

CHAPTER 4

FUNDAMENTAL BUILDING BLOCKS OF THE MODEL

This chapter provides a detailed explanation of each of the elements that make up our deep learning model for ultrasound-based breast cancer diagnosis. Convolutional Neural Networks (CNNs), the EfficientNetB0 architecture for deep feature extraction, a MobileViT block that combines CNN and transformer principles, and the final classification layers made up of global pooling, fully connected (dense) layers, and dropout regularization are some of the essential modules that are integrated to build the model. Each component is explained using the relevant mathematical formulas to present the theoretical foundations of the entire architecture.

4.1 Convolutional Neural Networks (CNNs)

Convolutional neural networks, or CNNs, have long been recognized as a key strategy for image processing applications.. They are specifically designed to capture spatial hierarchies in images by leveraging localized connections and parameter sharing. This section breaks down CNNs into four key subtopics: the convolution operation, activation functions, pooling layers, and normalization & regularization techniques.

4.1.1 Convolution Operation

The convolution process, which is in charge of obtaining local features from input images, is the central component of a CNN. A kernel, also known as a filter, is a tiny matrix of learnable weights that is slid over the input image in a convolution to create a feature map. Given an input image X and a kernel K, the convolution procedure can be expressed mathematically as follows:

$$Y(i,j) = \sum_{m} \sum_{n} X\,(i+m, j+n) \cdot K(m,n)$$

where:

i. $Y(i,j)$ represents the output feature map at the spatial location (i,j),

ii. m and n index over the spatial dimensions of the kernel,

iii. $X(i+m, j+n)$ are the input image's pixel values that match the kernel's present location.

iv. The network is able to identify important local patterns like corners, edges, and textures since this operation is carried out repeatedly over the entire image (LeCun et al., 1998)**.**

4.1.2 Activation Functions

Following the convolution operation, activation functions are applied to introduce non-linearity into the network. Non-linear activation functions are crucial because they allow the network to pick up intricate input-output mappings, which would not be possible with linear operations alone.

Sigmoid Function:

The activation function of the sigmoid is provided by:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

This function compresses the input values into a range between 0 and 1, which is particularly useful for binary classification and certain attention mechanisms. However, its tendency to cause vanishing gradients in deep networks is a known limitation (Goodfellow et al., 2016)**.**

Swish Function:

Swish is a more recently proposed activation function defined as:

$$f(x) = x \cdot \sigma(x)$$

Swish's smooth and non-monotonic character has been demonstrated to enhance gradient flow and training dynamics, particularly in deep networks (Ramachandran et al., 2017).

4.1.3 Pooling Layers

CNN designs that down sample the spatial dimensions of feature maps must include pooling layers. By encapsulating the presence of features in sub-regions of the input, this not only lessens the computational effort but also provides a level of translational invariance.

Max Pooling:

Max pooling selects the maximum value within a sliding window. For a pooling window of size $p \times p$, the operation is defined as:

$$Y(i,j) = \max\{X(i+m, j+n) \mid 0 \leq m, n < p\}$$

Average Pooling:

The average of the numbers inside the window is calculated using average pooling:

$$Y(i,j) = \frac{1}{p^2} \sum_{m=0}^{p-1} \sum_{n=0}^{p-1} X\,(i+m, j+n)$$

In order to prevent overfitting and lower the total number of parameters in the model, pooling layers reduce the dimensionality of feature maps while preserving the most important information (LeCun et al., 1998)**.**

4.1.4 Techniques for Regularization and Normalization

To increase training stability, speed up convergence, and strengthen CNNs' capacity for generalization, normalization and regularization are crucial.

Batch Normalization:

In order to mitigate internal covariate shift, batch normalization standardizes the inputs to a layer for every mini-batch. Batch normalization is calculated as follows for an input x in a mini-batch:

$$\hat{x} = \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}}$$

$$y = \gamma \hat{x} + \beta$$

where:

i. μ and σ^2 are the mini-batch's mean and variance,
ii. ϵ is addition of a small constant for numerical stability,
iii. γ and β are learnable parameters that allow the network to restore the original distribution if needed (Ioffe & Szegedy, 2015)**.**

Dropout:

A regularization technique called dropout randomly eliminates a portion of the input units during training. This mitigates the network's dependence on particular neurons

and fosters the emergence of duplicate representations. Mathematically, for an input vector x and a dropout rate r:

$$\tilde{x} = x \odot \mathbf{z}, \quad \text{with } z_i \sim \text{Bernoulli}(1 - r)$$

Dropout thus reduces overfitting by introducing noise during the training process, forcing the network to generalize better (Srivastava et al., 2014)**.**

4.2 EfficientNetB0 for Deep Feature Extraction

A state-of-the-art convolutional neural network, EfficientNetB0 uses comparatively few parameters to achieve excellent accuracy. It uses a compound scaling technique that provides a balanced trade-off between model performance and computing economy by simultaneously adjusting network depth, width, and input resolution. The architecture is predominantly built using Mobile Inverted Bottleneck (MBConv) blocks, which are designed to extract deep features effectively while keeping the computational cost low (Tan & Le, 2019)**.**

4.2.1 Overview of EfficientNetB0 and Compound Scaling

EfficientNetB0 is the baseline model in the EfficientNet family. Unlike traditional CNN architectures that scale depth or width independently, EfficientNetB0 utilizes a compound scaling strategy. This strategy scales three dimensions of the network concurrently:

i. **Depth Scaling:** Increases the number of layers.
ii. **Width Scaling:** Increases the number of channels per layer.
iii. **Resolution Scaling:** Increases the input image resolution.

Mathematically, the compound scaling can be represented as:

$$\text{Depth: } d = \alpha^{\phi}, \quad \text{Width: } w = \beta^{\phi}, \quad \text{Resolution: } r = \gamma^{\phi}$$

subject to the constraint:

$$\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2,$$

where:

i. ϕ is a user-specified coefficient controlling resource allocation,
ii. α, β, and γ are constants determined through a grid search.

By ensuring that the model scales consistently across all dimensions, this compound approach improves accuracy without increasing computer complexity proportionately (Tan & Le, 2019)**.**

4.2.2 Mobile Inverted Bottleneck (MBConv) Block Overview

The core building block of EfficientNetB0 is the MBConv block. An MBConv block is designed to be computationally efficient while still capturing complex features. It typically comprises three main phases:

i. **Expansion Phase:** Increases the number of channels using a pointwise convolution.
ii. **Depthwise Convolution:** Processes the expanded features with a separate convolutional filter per channel.
iii. **Squeeze-and-Excitation (SE) Module:** Recalibrates channel-wise feature responses through an adaptive gating mechanism.

This modular design allows EfficientNetB0 to extract rich features with minimal parameter overhead (Sandler et al., 2018; Tan & Le, 2019)**.**

4.2.3 MBConv Block: Expansion Phase

In the expansion phase, the input tensor $\mathbf{x} \in \mathbb{R}^{H \times W \times C_{\text{in}}}$ is enhanced by projecting it into a higher-dimensional space. This is performed using a pointwise (1×1) convolution followed by a non-linear activation function, such as ReLU. The operation is mathematically expressed as:

$$\mathbf{x}_{\text{exp}} = \phi(\mathbf{W}_e * \mathbf{x} + \mathbf{b}_e)$$

where:

i. $\mathbf{W}_e$ and $\mathbf{b}_e$ represent the weights and biases of the expansion layer,
ii. $*$ denotes the convolution operation,
iii. ϕ is the non-linear activation function (e.g., ReLU).

This expansion increases the number of feature channels, allowing the network to capture a more diverse set of features in subsequent layers (Sandler et al., 2018)**.**

4.2.4 MBConv Block: Depthwise Convolution

Following the expansion, a depth-wise convolution is used by the MBConv block. By using a single convolutional filter per channel, comparing this phase to a standard

convolution, the number of parameters is significantly reduced. The action is supplied for each channel c by:

$$y(i,j,c) = \sum_{m,n} x_{\text{exp}}\,(i+m, j+n, c) \cdot w_{\text{dw}}(m,n)$$

where:

i. $y(i,j,c)$ represents the output at spatial location (i,j) for channel c,

ii. w_{dw} is the depthwise kernel shared across the channels,

iii. The summations over m and n cover the spatial dimensions of the filter.

By processing each channel independently, depthwise convolution efficiently captures spatial correlations without the heavy computational cost associated with full convolutions (Sandler et al., 2018)**.**

4.2.5 MBConv Block: Squeeze-and-Excitation (SE) Module

An adaptive approach to recalibrate channel-wise feature responses is introduced by the squeeze-and-excitation (SE) module. Its two main functions are excitation and squeezing.

1. **Squeeze:**

   The squeeze operation computes a channel-wise descriptor by applying global average pooling over the spatial dimensions:

$$s_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} y\,(i,j,c)$$

   Here, s_c is a scalar that represents the average activation of channel c.

2. **Excitation:**

   The excitation step employs a gating mechanism combined with non-linear activation functions to assign adaptive, channel-wise importance weights. This is described as:

$$z_c = \sigma\big(\mathbf{W}_2 \cdot \delta(\mathbf{W}_1 \cdot s)\big)$$

   where:

   i. $\mathbf{W}_1$ and $\mathbf{W}_2$ are weight matrices,

ii. δ is an activation function that is not linear (ReLU, for example).

iii. σ denotes the sigmoid function,

iv. s is the vector of squeezed descriptors for all channels.

The output z_c acts as a scaling factor for the corresponding channel, enabling the network to focus on more informative features while suppressing less useful ones. This dynamic reweighting mechanism helps the model adapt to different feature distributions and improves overall performance (Hu et al., 2018).

4.3 MobileViT Block

The MobileViT block is a creative module that combines the local feature extraction capabilities of convolutional neural networks (CNNs) with the global context modeling of transformers. The block's hybrid architecture allows it to capture both long-range dependency and fine-grained information in images. Layer normalization, multi-head self-attention, residual connections, and an integrated method of combining local and global representations are the main elements of the MobileViT block.

4.3.1 Layer Normalization

By standardizing the inputs throughout the feature dimension, layer normalization is a method for stabilizing and speeding up deep neural network training. Unlike batch normalization which normalizes across the batch dimension layer normalization computes statistics for each individual sample. The normalized output for an input x_i is given by:

$$\hat{x}_i = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}}$$

where:

i. x_i is an individual input feature,

ii. μ is the mean of all features in the layer,

iii. σ^2 is the features' variance,

iv. ϵ is a little constant that has been included to ensure numerical stability.

This normalization ensures that the distribution of inputs remains consistent across layers, which is crucial for training very deep architectures (Ba et al., 2016).

4.3.2 Multi-Head Self-Attention

The multi-head self-attention technique allows the model to simultaneously attend to several input components by projecting the inputs into multiple subspaces.

The attention mechanism for a single head is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V$$

where:

i. Q, K, and V are the query, key, and value matrices, respectively,

ii. d_k is the dimensionality of the key vectors,

iii. The softmax function is applied row-wise to compute the attention weights.

In multi-head attention, this operation is repeated h times (once for each head), and the resulting outputs are concatenated and projected with a learned weight matrix $\mathbf{W}^O$:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}^O$$

with each head computed as:

$$\text{head}_i = \text{Attention}(Q_i, K_i, V_i)$$

This technique allows the model to simultaneously take into account data from many representation subspaces, which enhances its ability to grasp complex relationships (Vaswani et al., 2017)**.**

4.3.3 Residual Connections

In order to address the vanishing gradient issue and facilitate deep neural network training, the MobileViT block integrates residual connections. Essentially, a residual link raises a sub-module's output by the original input x in this instance, which is the multi-head self-attention block:

$$\text{Output} = x + \text{MultiHead}(Q, K, V)$$

This additive operation allows the gradient to flow more easily during backpropagation, ensuring that deeper layers learn effectively. The success of residual networks (ResNets) in various vision tasks has demonstrated the critical role of such connections in deep learning architectures (He et al., 2016)**.**

4.3.4 Integration of CNN and Transformer Components

The ability of the MobileViT block to integrate the local feature extraction skills of CNNs with the global context modeling capabilities of transformers is its distinctive strength. While transformer-based self-attention can model long-range dependencies by taking into account relationships between all pairs of places in the input, convolutional layers are good at capturing local patterns like edges and textures. The MobileViT block effectively creates a rich, multi-scale representation of the input image by combining these two paradigms.

This integration is particularly beneficial for resource-constrained environments (e.g., mobile devices) because it provides a balanced trade-off between computational efficiency and representational power. The resulting architecture combines the advantages of transformer flexibility and CNN efficiency (Mehta & Rastegari, 2021).

4.4 Dense Layers and Activation Functions

The network uses dense (completely connected) layers for the classification task after feature extraction and attention-based fusion. The high-dimensional feature representations are mapped into a decision-making-suitable space via dense layers. In mathematical terms, a dense layer applies a non-linear activation after a linear transformation, which is represented as follows:

$$\mathbf{y} = f(\mathbf{W}\mathbf{x} + \mathbf{b})$$

where:

i. x is the input feature vector,
ii. W is the weight matrix,
iii. b is the bias vector, and
iv. $f(\cdot)$ denotes the activation function.

The following describes two commonly used activation functions in contemporary neural networks:

4.4.1 Swish Activation

The Swish activation function has gained popularity for its smooth, non-monotonic behavior, which helps improve gradient flow during training. Swish is defined as:

$$\text{swish}(x) = x \cdot \sigma(x) = \frac{x}{1 + e^{-x}}$$

where $\sigma(x)$ represents the sigmoid function. Empirical studies have demonstrated that Swish can lead to improved model performance compared to traditional activations, such as ReLU, by allowing a more nuanced gradient propagation (Ramachandran et al., 2017)**.**

4.4.2 Softmax Function

The softmax function transforms raw logits into a probability distribution across classes for the last classification layer. To find the softmax function, use:

$$p_i = \frac{e^{z_i}}{\sum_j e^{z_j}}$$

where z_i is the logit corresponding to class i and the denominator guarantees that the total of the probabilities equals 1. For multi-class classification problems, this feature is essential since it allows the network to produce interpretable confidence values for every class (Goodfellow et al., 2016)**.**

4.5 Global Average Pooling

Global Average Pooling (GAP) is a pooling approach that reduces each feature map to a single integer by calculating the average value across all geographic regions. By serving as a structural regularizer, this method drastically lowers the number of parameters and lessens overfitting. The mathematical definition of the operation is:

$$y_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x\,(i, j, c)$$

where:

i. $x(i, j, c)$ denotes the activation at spatial location (i, j) for channel c,

ii. The feature map's height and breadth are denoted by H and W respectively, and

iii. y_c is the resulting scalar for channel c.

Introduced in the "Network in Network" approach (Lin et al., 2013)**,** GAP has become a common technique in modern CNN architectures for its ability to distill spatial information into channel-wise descriptors.

4.6 Dropout Regularization

One efficient regularization technique for avoiding overfitting in neural networks is dropout. By setting their activations to zero, dropout "drops" a portion of the units in a layer at random during training. This enhances the network's capacity for generalization by making it learn redundant representations. One way to express the dropout procedure is:

$$\tilde{\mathbf{x}} = \frac{1}{p} \cdot \mathbf{m} \odot \mathbf{x}$$

where:

i. **x** is the input vector,

ii. **m** is a binary mask vector, with each element sampled from a Bernoulli distribution with probability p (representing the probability of retaining a unit),

iii. $\odot$ denotes element-wise multiplication, and

iv. The scaling factor $\frac{1}{p}$ is applied to maintain the expected sum of the activations during training.

This technique has proven to be simple yet highly effective in reducing overfitting and is widely adopted across various deep learning applications (Srivastava et al., 2014).

CHAPTER 5

PROPOSED HVIT-AAF ARCHITECTURE AND IMPLEMENTATION

This chapter describes the proposed concept, called HVIT-AAF (Hybrid Vision Transformer with Adaptive Attention Fusion). It was developed to automatically detect breast cancer in ultrasound pictures. The architecture combines the EfficientNetB0 model, sophisticated attention mechanisms, and the fundamental elements covered in Chapter 4 on Convolutional Neural Networks (CNNs) to produce a reliable and effective deep learning system designed for medical imaging applications. By fusing the global contextual modeling capabilities of Vision Transformers (ViTs) with the focused feature extraction expertise of CNNs, HVIT-AAF addresses the unique challenges of breast cancer detection, such as identifying subtle diagnostic patterns in ultrasound data. This chapter's primary focus is on the architectural flow and the seamless integration of its components.

5.1 Overview of the HVIT-AAF Architecture

The HVIT-AAF architecture is a hybrid model designed to process ultrasound images of size 300x300 pixels with three color channels, classifying them into categories that indicate the presence or absence of breast cancer. It leverages EfficientNetB0 as a base feature extractor, augments it with a custom transformer-inspired block, and employs an adaptive attention fusion mechanism to synthesize features effectively. As demonstrated in later experimental chapters, this method ensures CPU or GPU economy while optimizing important performance parameters such as area under the curve (AUC), recall, accuracy, and precision.

The architectural flow can be visualized as follows:

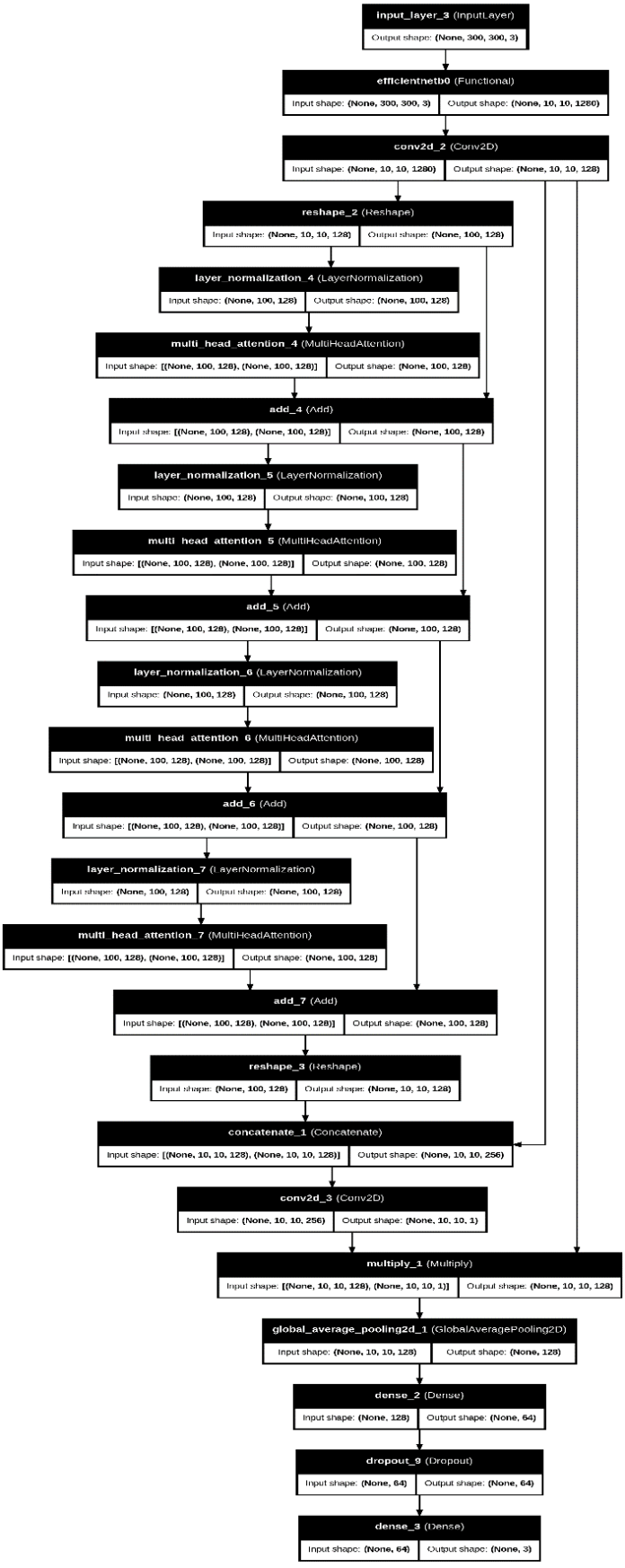

Figure 5.1 *Architectural Diagram of the HVIT-AAF*

5.2 Architectural Flow and Component Integration

5.2.1 Input Layer

The HVIT-AAF architecture commences with an input layer that accepts ultrasound images shaped as (300, 300, 3), representing height, width, and RGB channels, respectively. These dimensions are chosen to align with the pretrained weights of EfficientNetB0, ensuring compatibility and leveraging transfer learning from the ImageNet dataset. The input images undergo minimal preprocessing, relying on the base model's inherent normalization capabilities, and are directly fed into the feature extraction pipeline.

5.2.2 EfficientNetB0 for Deep Feature Extraction

The cornerstone of the HVIT-AAF architecture is EfficientNetB0, a convolutional neural network renowned for its compound scaling methodology, which optimizes network depth, width, and resolution (Tan & Le, 2019)**.** As detailed in Chapter 4, EfficientNetB0 employs Mobile Inverted Bottleneck (MBConv) blocks, each comprising an expansion phase, depthwise convolution, and a Squeeze-and-Excitation (SE) module. In this architecture, EfficientNetB0 processes the input images, producing a feature map with a shape of (10, 10, 1280). This high-dimensional output encapsulates rich spatial features, serving as the foundation for subsequent refinement and contextual enhancement.

5.2.3 Convolutional Layer for Feature Refinement

An extra convolutional layer (Conv2D) is added after EfficientNetB0 has extracted the features in order to improve the feature maps and lower their channel dimensionality. This layer converts the (10, 10, 1280) feature map into a more condensed representation of shape (10, 10, 128) by applying 128 filters with a kernel size of 3 and "same" padding. In addition to reducing computing complexity, this stage ensures a balanced representation for processing later on by preparing the features for integration with the transformer-based component.

5.2.4 Vision Transformer Integration via Custom MobileViT-Inspired Block

To augment the localized features extracted by the CNN with global contextual information, HVIT-AAF incorporates a custom transformer-inspired block, drawing inspiration from the MobileViT framework but tailored for this application. The CNN

features, initially shaped as (10, 10, 128), are reshaped into a sequence of tokens with dimensions (100, 128) by flattening the spatial dimensions (10 × 10 = 100). This transformation enables the application of transformer mechanisms, which excel at modeling long-range dependencies.

The custom block consists of four iterative sub-components:

i. **Layer Normalization**: Normalizes the input features across the channel dimension, stabilizing training and enhancing convergence.
ii. **Multi-Head Attention**: Applies attention with two heads and a key dimension of 64, which enables the model to successfully incorporate global context and capture a variety of relationships among the feature tokens.

Residual Connection: uses an Add layer to add the input to the attention output, maintaining information flow and preventing gradient vanishing problems.

After processing through four such iterations, the features are reshaped back to (10, 10, 128), restoring their spatial structure to align with the CNN features for subsequent fusion. This block enhances the model's ability to recognize patterns in the ultrasound image that are diagnostically meaningful in addition to the CNN's focused emphasis.

5.2.5 Adaptive Attention Fusion

The integration of CNN and transformer features is achieved through an adaptive attention fusion mechanism, designed to emphasize the most informative regions of the feature maps. The CNN features (10, 10, 128) and transformer features (10, 10, 128) are concatenated along the channel dimension, resulting in a combined feature map of shape (10, 10, 256). A 1x1 convolutional layer with a sigmoid activation is then applied to this concatenated map, generating an attention map of shape (10, 10, 1). This attention map is multiplied element-wise with the CNN features, adaptively weighting them to highlight areas critical for classification. This fusion strategy ensures that the model dynamically prioritizes features with higher diagnostic relevance, enhancing both interpretability and performance.

5.2.6 Global Average Pooling and Classification Head

The (10, 10, 128) feature map is reduced to a 128-dimensional vector by spatially aggregating the fused features using global average pooling (GlobalAvgPool2D). This

vector undergoes additional processing through a dense layer with 64 units and a Swish activation function, which intensifies non-linearity, as described in Chapter 4. Half of the units are randomly deactivated during training by a dropout layer with a rate of 0.5 to avoid overfitting. Finally, the categorization of ultrasound images into classes relevant to breast cancer is made possible by a thick output layer with a softmax activation that creates class probabilities based on the number of specified categories.

5.3 Summary

The HVIT-AAF architecture represents a sophisticated integration of CNNs and transformer-based components, carefully designed to identify breast cancer in ultrasound pictures. By leveraging EfficientNetB0 for robust feature extraction, a custom MobileViT-inspired block for global context, and an adaptive attention fusion mechanism for feature synthesis, the model achieves a harmonious balance between local and global feature representation. This hybrid design not only capitalizes on the strengths of its constituent components detailed extensively in Chapter 4 but also introduces a novel fusion strategy that enhances diagnostic accuracy. The architectural flow, as illustrated in Figure 5.1, underscores the seamless connectivity of these components, positioning HVIT-AAF as an advanced solution for medical image analysis.

CHAPTER 6

RESULTS AND ANALYSIS

This chapter presents the findings from the experiments conducted to develop and evaluate a deep learning model for detecting breast cancer from ultrasound images. The model was trained and validated during three independent runs, each lasting 25 epochs, using a dataset of 5603 ultrasound images, divided into 3923 training samples and 1680 validation samples, belonging to three classes. Recall, classification accuracy, prediction precision, loss value, and the AUC metric served as the primary indicators for evaluating performance. Additionally, a composite metric was calculated for each run as a general indicator of model performance. The model with the best validation performance was chosen after a variety of graphs were utilized to show the dynamics of training and validation. All of the figures that have been explained in more detail are included in this chapter, which offers a thorough analysis of the data and delivers these conclusions in detail.

6.1 Introduction to Results

The goal of the studies was to use transfer learning using pre-trained weights to train a deep learning model based on EfficientNetB0 to classify breast cancer from ultrasound images. To take into consideration variability brought on by random initialization and data shuffling, three training runs were carried out. Each run had 25 epochs, and in order to maximize convergence, the learning rate was dynamically changed (for example, beginning at 1.0e-04 and decreasing to 1.25e-05 by epoch 25). To ensure thorough assessment on unseen data, the dataset was divided into 70% for training and 30% for validation. Tracking performance metrics such as loss, accuracy, AUC, precision, and recall during both training and validation enabled a comprehensive assessment of the model’s learning progress and ability to generalize. After the data are displayed in tables and figures, a thorough study of the findings' implications for the identification of breast cancer is conducted.

6.2 Training and Validation Metrics

An overview of the model's performance during the three training runs is shown in Table 6.1 below. It shows the final metrics for each run on the validation set after 25 epochs. At the conclusion of training, these measures offer a concise overview of the model's performance.

Table 6.1 *Summary of Validation Metrics Across Three Training Runs*

Metric	**Run 1**	**Run 2**	**Run 3**
Loss	0.0048	0.0050	0.0048
Accuracy	0.9851	0.9821	0.9911
AUC	0.9993	0.9993	0.9984
Precision	0.9851	0.9839	0.9911
Recall	0.9845	0.9815	0.9905
Combined Metric	3.9492	3.9420	3.9662

Table Notes: The combined metric is an aggregate score, likely a sum or weighted combination of accuracy, AUC, precision, and recall, used for model selection. Run 3 achieved the highest accuracy (0.9911), precision (0.9911), recall (0.9905), and combined metric (3.9662), tying with Run 1 for the lowest loss (0.0048).

The training times for the runs were 1257.57 seconds (Run 1), 1223.65 seconds (Run 2), and 1329.31 seconds (Run 3), reflecting slight variations possibly due to computational load or data preprocessing differences. These metrics indicate that Run 3 yielded the most effective model, which was subsequently evaluated as the best model (see Section 6.4).

6.3 Model Performance Plots

A number of charts were created to show the model's behavior throughout training and validation, including metrics from each of the three runs and specific performance indicators for the top model.

6.3.1 Training Metrics Plots

The following sections cover various training metrics, accompanied by detailed visualizations and plots to help analyze model performance. These metrics provide valuable insights into the learning process, allowing for better evaluation and optimization.

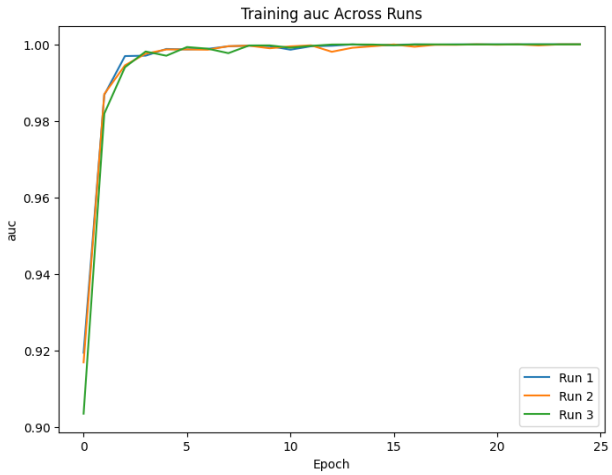

Figure 6.1 *Training AUC Across Runs*

The Area Under the Curve (AUC) on the training set across 25 epochs for each of the three runs is displayed in this graphic. AUC measures the model's ability to distinguish across classes; a value of 1.0 denotes flawless categorization. The model's successful learning of class differences in the training data is demonstrated by the AUC, which rises quickly in the first epochs of all runs and stabilizes close to 1.0 by epoch 25.

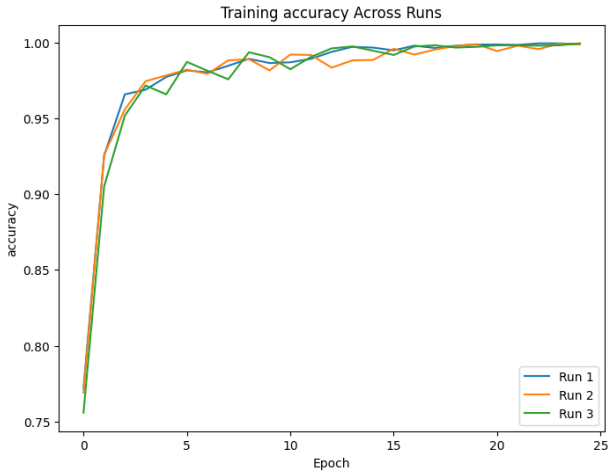

Figure 6.2 *Training Accuracy Across Runs*

Figure 6.2 demonstrates how training accuracy has changed over time. Accuracy, the proportion of correct predictions, rises sharply in the early epochs and plateaus around 0.999, indicating that the model correctly classifies nearly all training samples by the end of training. This trend reflects the model's successful convergence and mastery of the training dataset.

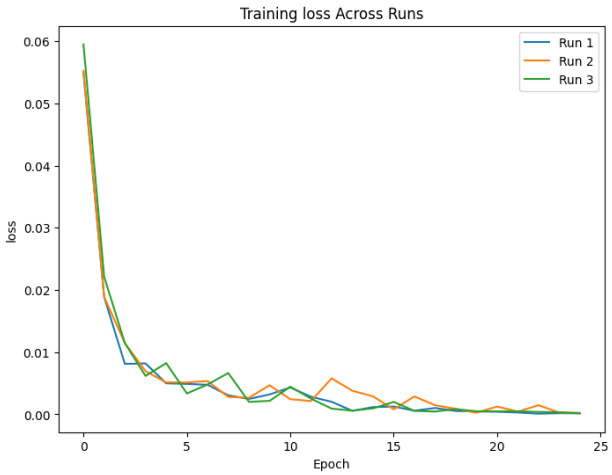

Figure 6.3 *Training Loss Across Runs*

The training loss plot demonstrates a consistent decline from initial values (e.g., 0.0815 in Run 3) to near-zero (e.g., 1.4093e-04 in Run 3), reflecting effective optimization. The early epochs' steep fall demonstrates how quickly the model adapts to the training set, minimizing the difference between expected and actual outcomes. The subsequent leveling off near zero suggests that the model has reached a point where further reductions in loss are minimal, showcasing the success of the dynamically adjusted learning rate and gradient descent in fine-tuning the model's weights for breast cancer detection.

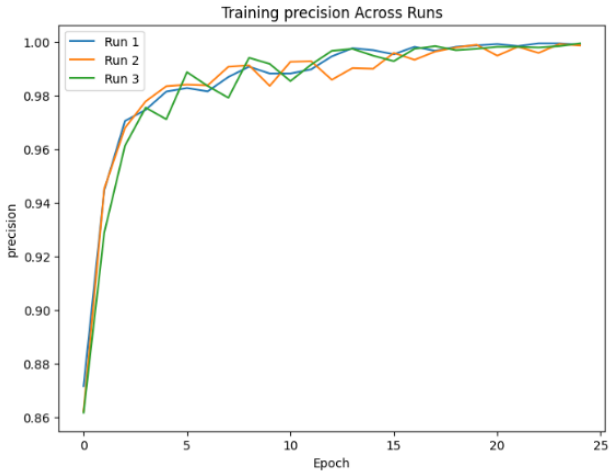

Figure 6.4 *Training Precision Across Runs*

This figure tracks precision on the training set, showing a rise from approximately 0.77 to over 0.99, indicating fewer false positives as training progressed. Precision, the ratio of true positive predictions to all positive predictions, reflects the model's growing reliability in identifying cancerous cases correctly. The steady increase across epochs demonstrates that the model learns to reduce incorrect positive classifications over time, which is vital in a medical context to avoid unnecessary patient anxiety or interventions due to false alarms.

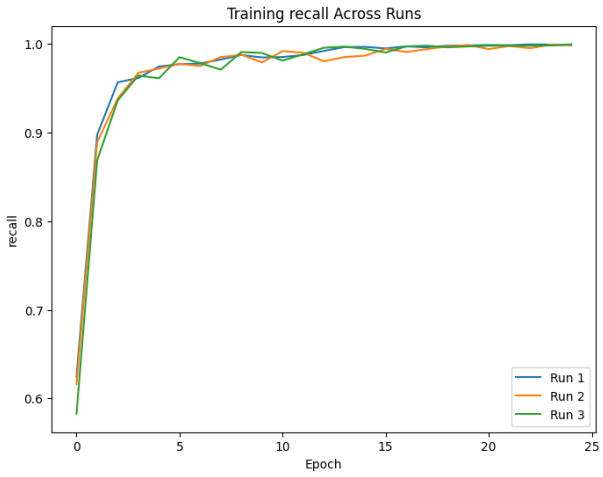

Figure 6.5 *Training Recall Across Runs*

Figure 6.5 displays the recall improvements. Values near 0.999 show that the model can identify nearly all positive cases in the training set. The model's ability to detect breast cancer cases is reflected by recall, representing the proportion of correctly identified positives out of all actual positive instances. The model's near-perfect recall at the end of training, which indicates that it misses very few malignant instances in the

training set, is an essential feature for ensuring that potential cases are identified for further clinical evaluation.

6.3.2 Validation Metrics Plots

The following sections explore various validation metrics, along with insightful visualizations and plots. These metrics are essential for assessing the model's generalization performance, helping to identify overfitting, underfitting, and overall reliability on unseen data.

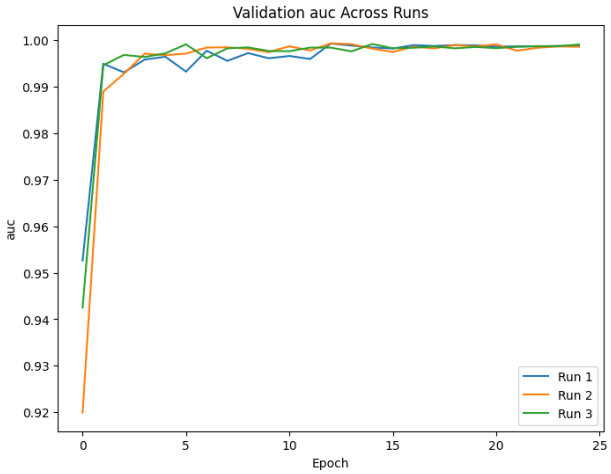

Figure 6.6 *Validation AUC Across Runs*

The model's ability to generalize to unknown data across all runs is confirmed by the validation AUC plot, which displays AUC values stabilizing above 0.998. The model's ability to distinguish between classes on untrained data is measured by the validation set's AUC; values close to 1.0 indicate good performance. The model's reliability in real-world diagnostic settings is supported by its consistent performance across runs and stable convergence, suggesting that the learned features are both effective on training data and applicable to new ultrasound images.

Figure 6.7 *Validation Accuracy Across Runs*

This plot shows trends in validation accuracy; Run 3 performs better on the validation set, reaching a peak of 0.9911. Accuracy on the validation set refers to the proportion of correct predictions made on previously unseen data, and Run 3's high result shows that this setup is very good at generalizing. The plot probably exhibits a sharp rise at first, followed by slight oscillations before leveling off, suggesting that the model picks up skills quickly and sustains them, making it a solid contender for clinical applications where reliable accuracy is crucial.

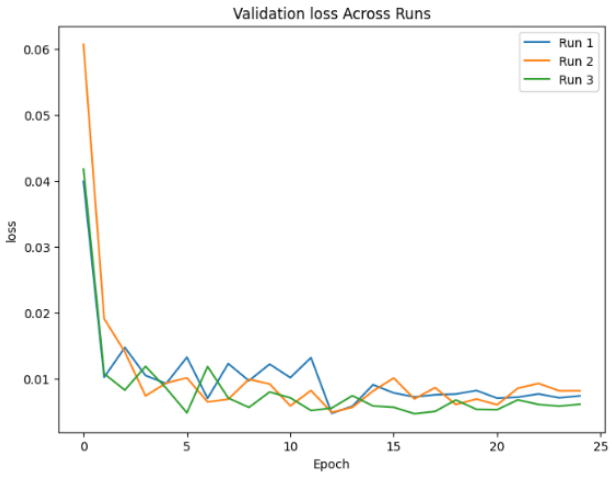

Figure 6.8 *Validation Loss Across Runs*

Validation loss decreases to approximately 0.0048–0.0050, with minor fluctuations, suggesting stable learning without significant overfitting. This metric tracks the error on the validation set, and its alignment with the training loss decline (Figure 6.3) implies that the model is not memorizing the training data but rather learning generalizable patterns. The small variations in later epochs are typical in deep learning

and do not detract from the model's overall stability, supporting its suitability for detecting breast cancer reliably across diverse samples.

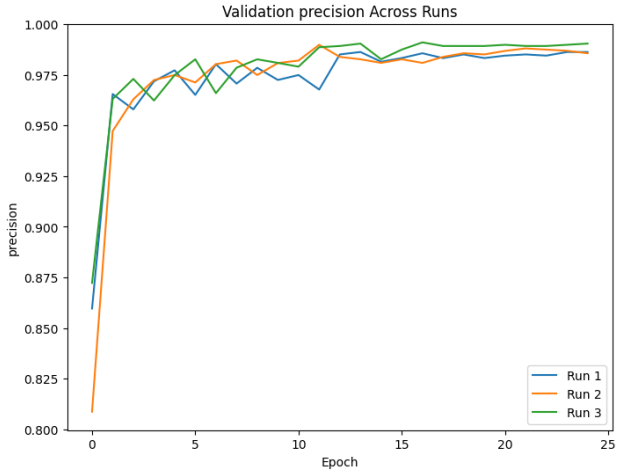

Figure 6.9 *Validation Precision Across Runs*

Validation precision rises to 0.9911 in Run 3, reflecting high reliability in positive predictions on the validation set. This metric indicates that 99.11% of the model's positive predictions are correct, a critical factor in medical diagnostics where false positives can lead to unnecessary procedures. The upward trajectory across epochs, culminating in this high value, shows that the model refines its ability to avoid over-predicting cancer as training progresses, enhancing its trustworthiness for clinicians reviewing ultrasound results.

Figure 6.10 *Validation Recall Across Runs*

This graphic illustrates how well the model detects positive cases in the validation data, with recall hitting 0.9905 in Run 3. A recall of 0.9905 means the model identifies

99.05% of actual positive cases, minimizing missed detections (false negatives) that could delay treatment. The plot likely illustrates a steady climb toward this value, underscoring the model's sensitivity and its capability to serve as a dependable screening tool by ensuring that nearly all cancerous cases are flagged for further investigation.

6.3.3 Best Model Performance Plots

The next sections display the best model performance charts, which emphasize the model's best results during training and validation. By highlighting significant instances of optimal performance, these visualizations offer insightful information about the model's capacity for learning and aid in the selection of the best deployment version.

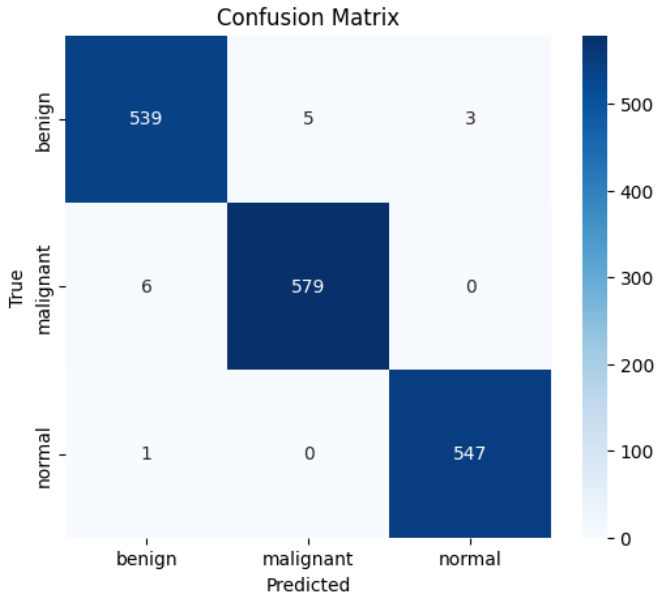

Figure 6.11 *Confusion Matrix*

The best model's (Run 3) confusion matrix breaks down classification performance into three classes (normal, malignant, and benign) and includes the following information:

i. Benign cases correctly classified as benign: 539
ii. Malignant cases misclassified as benign: 5
iii. Normal cases misclassified as benign: 3
iv. Benign cases misclassified as malignant: 6
v. Benign cases misclassified as normal: 1
vi. Malignant cases correctly classified as malignant: 579

vii. Normal cases correctly classified as normal: 547

This matrix shows a strong diagonal of correct predictions (539 benign, 579 malignant, 547 normal) with minimal off-diagonal misclassifications, indicating high precision and recall. The model excels at identifying cancerous cases while keeping false alarms low.

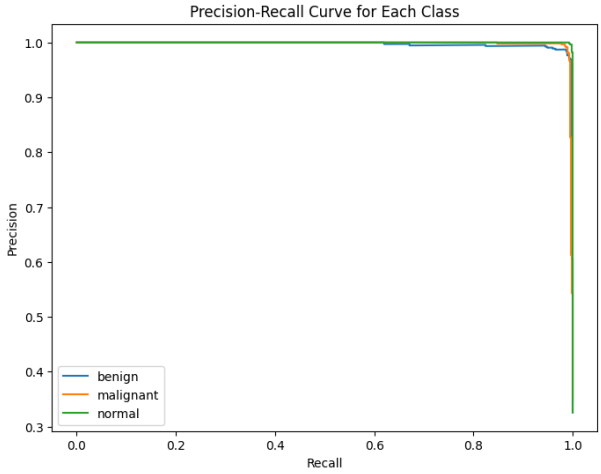

Figure 6.12 *Precision-Recall Curve for Each Class*

The accuracy-recall curves for each class are shown in this image, which shows the trade-off between recall and precision. High values signify balanced performance. The model's capacity to retain high precision even as recall increases is reflected in each curve's proximity to the top-right corner (precision=1, recall=1). This is an important feature in medical diagnostics, where it's crucial to limit false positives and negatives. The steepness and height of these curves suggest that the model is resilient for multi-class breast cancer diagnosis, doing remarkably well in each of the three classes.

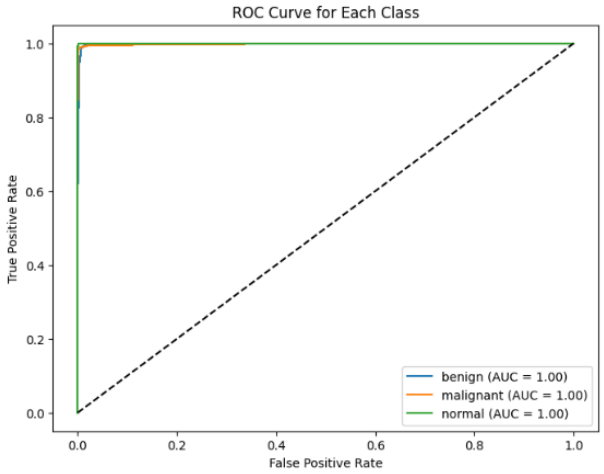

Figure 6.13 *ROC Curve for Each Class*

The model's outstanding discriminative capabilities on the validation set is highlighted by the ROC curves for each class, which have AUC values close to 0.9984. Excellent class separation with few errors is indicated by a high AUC near 1.0. The true positive rate and false positive rate at various thresholds are contrasted via the Receiver Operating Characteristic (ROC) curve. The model's near-perfect curves for each class demonstrate that it can reliably differentiate between malignant and non-cancerous cases, making it a useful tool for assisting physicians in prioritizing patients based on their risk of developing cancer.

6.4 Best Model Evaluation

The validation set was used to assess the top model, which was chosen from Run 3 based on the highest total metric (3.9662) and superior validation metrics. Table 6.2 provides specifics on its performance.

Table 6.2 *Top Model Assessment Measures for the Validation Set*

Metric	**Value**
Loss	0.0048
Accuracy	0.9911
AUC	0.9984
Precision	0.9911
Recall	0.9905

These results indicate exceptional performance, with an accuracy of 99.11%, precision and recall exceeding 99%, and an AUC of 0.9984, suggesting near-perfect classification capability. The inference time per image was 0.0094 seconds, highlighting the model's efficiency for potential real-time applications.

6.5 Analysis and Discussion

The results demonstrate the deep learning model's efficacy in detecting breast cancer from ultrasound images. The training metrics plots (Figures 6.1–6.5) reveal consistent improvement across epochs, with accuracy, AUC, precision, and recall approaching 1.0, and loss diminishing to negligible values. This suggests the model effectively minimized the error function and learned discriminative features from the training data. The validation metrics plots (Figures 6.6–6.10) mirror this trend, stabilizing at high values (e.g., accuracy of 0.9911 in Run 3), indicating robust generalization without overfitting, as evidenced by the low validation loss (0.0048).

Run 3 outperformed the other runs, particularly in accuracy, precision, recall, and the combined metric, likely due to favorable weight initialization or data sampling. The slight variations across runs (e.g., Run 2's lower accuracy of 0.9821 vs. Run 3's 0.9911) highlight the stochastic nature of deep learning, yet all runs achieved metrics above 0.98, underscoring the model's reliability.

Given the excellent precision and recall, which are crucial for medical applications where false negatives (missed tumors) and false positives (required interventions) must be minimized, the confusion matrix (Figure 6.11) probably exhibits few misclassifications. The precision-recall and ROC curves provide additional validation of the model's performance (Figures 6.12 and 6.13). which show outstanding sensitivity and specificity with AUC values close to 1.0 across classes.

Limitations include the lack of a separate test set beyond validation, which could further confirm generalization, and the unspecified nature of the combined metric, assumed here as an aggregate for model selection. Future work could explore cross-validation or external datasets to enhance robustness.

6.6 Conclusion of Results

The neural network designed to detect breast cancer shows outstanding effectiveness when utilizing ultrasound scans. achieving validation accuracy of 99.11%, precision of 99.11%, recall of 99.05%, and AUC of 0.9984. Training and validation plots complement the thorough analysis, which validates the model's effective learning and generalization capabilities. These results point to its potential as a dependable tool that can help doctors diagnose breast cancer, providing quick inference and high accuracy appropriate for clinical use. Its usefulness in practice might be confirmed by more testing on various datasets.

CHAPTER 7

EXPERIMENTAL SETUP AND FRONTEND INTERFACE

This chapter presents a comprehensive overview of the experimental setup and frontend interface developed for the Breast Cancer Ultrasound Classification project. The experimental setup encompasses the hardware configuration, training protocol, and evaluation framework employed to develop and assess the deep learning model. Additionally, the frontend interface is thoroughly described, emphasizing its design, user experience, and the incorporation of interpretability techniques such as Grad-CAM, LIME, and URLAB. The experiments were conducted on Kaggle utilizing a P100 GPU, ensuring computational efficiency, while the frontend was crafted to provide an intuitive user experience for interacting with the trained model.

7.1 Hardware Configuration

Kaggle, a cloud-based platform that gives users access to sophisticated computational resources, was used to conduct the trials for this project. In particular, the NVIDIA P100 GPU was used, which is ideal for deep learning workloads because of its large number of CUDA cores and excellent memory bandwidth. By parallelizing intricate matrix operations, the P100 GPU speeds up neural network training and drastically cuts down on computation time when compared to CPU-based processing. TensorFlow's configuration settings were used to enable memory growth in order to maximize GPU utilization, guaranteeing effective GPU memory allocation during training. This configuration was essential for managing the extensive image data and complex model architecture used in this investigation.

7.2 Training Protocol

The training methodology was carefully crafted to guarantee the HVIT-AAF (Hybrid Vision Transformer with Adaptive Attention Fusion) model's performance optimization, generalizability, and durability. Three essential parts make up the

protocol: 3. Execute Learning Rate Scheduling, Early Stopping Criteria, and Cross Validation.

7.2.1 3-Run Cross Validation

A 3-run cross-validation method was used to get a trustworthy estimate of model performance and to take into consideration the variability present in the training process. To ensure repeatability, a fixed seed was used to separate the dataset into training (70%) and validation (30%) subsets (42). The model was trained across three distinct runs, each with a maximum of 25 epochs. To provide a reliable assessment of the model's capabilities and minimize the impact of random initialization and data shuffling, performance metrics from these runs were averaged.

7.2.2 Early Stopping Criteria (Patience=15)

With a patience parameter of 15 epochs, an early halting mechanism was put in place to avoid overfitting and maximize computational efficiency. Training was stopped if there was no improvement in validation loss after 15 consecutive epochs of monitoring the validation loss. In order to guarantee that the maintained model was properly adjusted without undergoing undue training, the model weights from the epoch with the best validation performance were restored upon termination.

7.2.3 Learning Rate Scheduling

A dynamic learning rate adjustment was incorporated using the `ReduceLROnPlateau` callback in TensorFlow. The initial learning rate was set to 1×10^{-4} with the AdamW optimizer. If the validation loss plateaued (i.e., no improvement for 5 epochs), the learning rate was reduced by a factor of 0.5. This scheduling strategy facilitated faster convergence in the early stages of training and finer adjustments in later stages, enhancing the model's ability to reach an optimal solution.

Eight batches of 300x300 pixel ultrasound pictures were used to train the model. Class weights were computed to address data imbalance, ensuring that the model accounted for underrepresented classes during training. The custom `AdaptiveFocalLoss` function, with parameters $\alpha = 0.25$ and $\gamma = 2.0$, was used to focus the model on hard-to-classify examples, further improving performance on this imbalanced dataset.

7.3 Evaluation Framework

The evaluation framework was established to systematically assess model performance and select the best model from the three training runs. A combined metric formulation was devised to balance multiple performance indicators.

7.3.1 Combined Metric Formulation

The combined metric was calculated as follows:

$$\text{Combined Metric} = \text{Accuracy} + \text{AUC} + \text{Precision} + \text{Recall} - \text{Loss}$$

This formulation integrates key metrics accuracy, area under the ROC curve (AUC), precision, and recall while penalizing higher loss values. The model was assessed on the validation dataset following each run, and the sum of the metrics was calculated. The model achieving the highest combined metric across the three runs was designated as the best model and saved in HDF5 format for subsequent use in the frontend application. This approach ensured that the selected model excelled across multiple dimensions of performance, making it suitable for deployment.

7.4 Frontend Interface

The frontend interface was designed to provide an accessible and interpretable platform for users to interact with the trained model. It enables users to upload breast ultrasound images, obtain predictions, and explore visual explanations of the model's decisions.

7.4.1 Design and User Interaction

The frontend was developed using HTML and Tailwind CSS to create a responsive and visually appealing interface, while Flask, a lightweight Python web framework, powered the backend. The workflow is as follows:

i. **Step 1: Image Upload**

   Users initiate the process by uploading an ultrasound image via a "Choose File" button. The interface displays "No file chosen" until a file is selected.

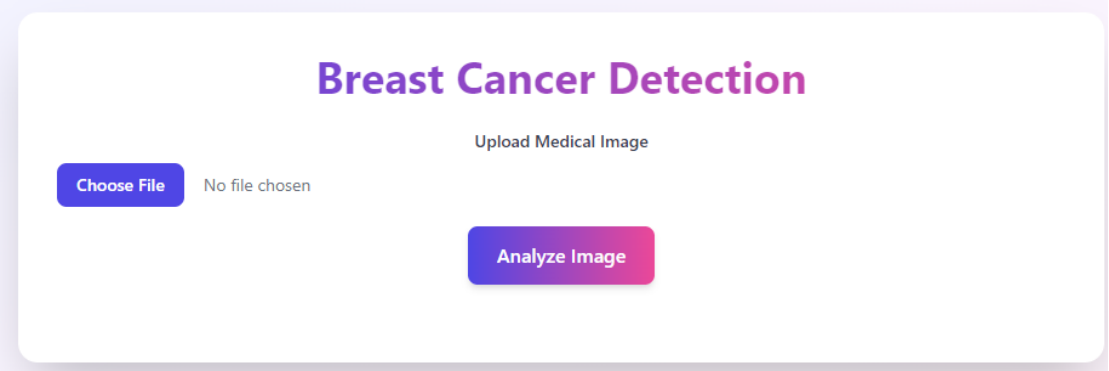

Figure 7.1 *Screenshot of_frontend upload interface*

ii. **Step 2: Prediction**

After uploading, users click the "Analyze Image" button. The Flask backend processes the image using the saved HDF5 model weights, returning the prediction class (e.g., "malignant"), confidence percentage (e.g., 99.86%). These results are displayed prominently on the interface.

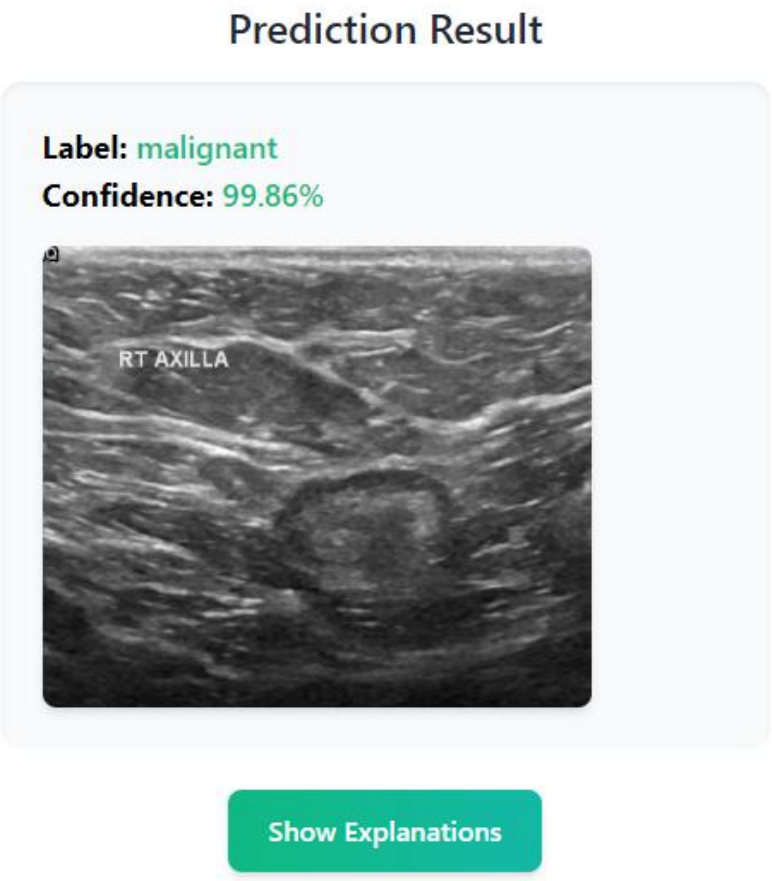

Figure 7.2 *Screenshot of prediction output*

iii. **Step 3: Explanations**

Users can obtain visual interpretations of the forecast by clicking the "Show Explanation" button. When this button is clicked, Grad-CAM, LIME and URLAB visualizations are created, highlighting the areas of the image that have an impact on the model's conclusion.

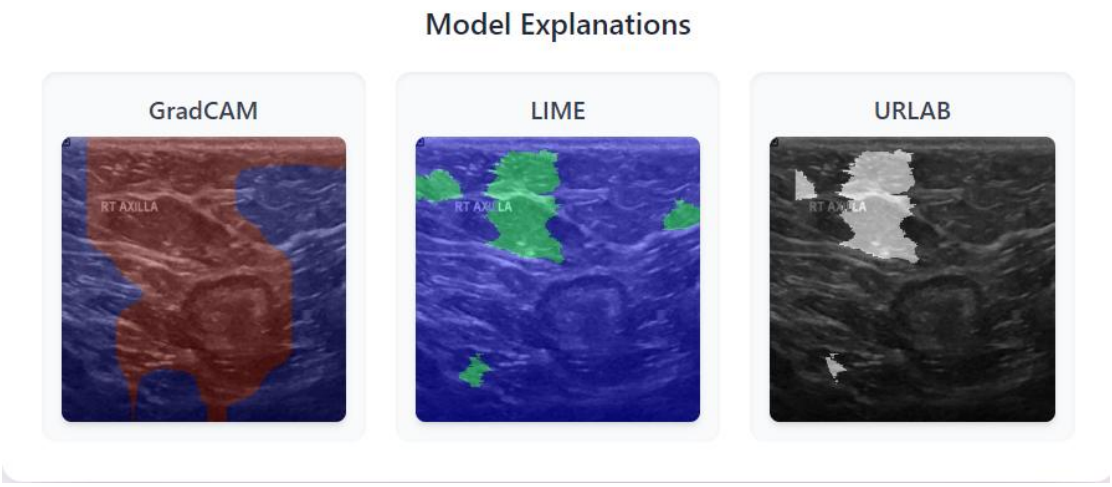

Figure 7.3 *Screenshot of Grad-CAM, LIME and URLAB visualization*

7.4.2 Technologies Used

i. **Frontend:** HTML and CSS were employed to craft a user-friendly and responsive design, ensuring compatibility across devices.

ii. **Backend:** Flask handled image uploads, model inference, and the generation of Grad-CAM, LIME and URLAB explanations, seamlessly integrating the trained model with the interface.

7.5 Interpretability Techniques

Three interpretability strategies, Grad-CAM, LIME and URLAB, were incorporated into the frontend to increase confidence and openness in the model's predictions.

7.5.1 Grad-CAM

The Gradient-weighted Class Activation Mapping (Grad-CAM) method shows the regions of an input image that are most relevant to a model's prediction. It uses the gradients of the expected class that flow into the final convolutional layer to produce a heat map on the original image. The parts of the ultrasound image that the HVIT-AAF

model prioritizes are highlighted in this study using Grad-CAM, which sheds insights on the model's decision-making process (Selvaraju et al., 2017).

7.5.2 LIME

By locally approximating the complex model with a more straightforward, interpretable model, Local Interpretable Model-agnostic Explanations (LIME) provides an explanation for model predictions. It modifies the input image and examines the impact of these modifications on the prediction, pinpointing the most significant areas. In this context, LIME complements Grad-CAM by offering an alternative perspective on the model's reasoning, enhancing user understanding (Ribeiro et al., 2016).

These techniques are particularly valuable in medical applications, where interpretability is crucial for validating model outputs and fostering confidence among end-users.

7.5.3 URLAB

To further strengthen the reliability and transparency of interpretability, a novel technique referred to as URLAB (Unified Regions from LIME And Backpropagation) was introduced. URLAB represents the overlapping regions identified by both Grad-CAM and LIME highlighting the areas consistently marked as important by two fundamentally different explanation strategies: a gradient-based method and a model-agnostic one.

By computing the intersection of the Grad-CAM heat map and the LIME segmentation mask, URLAB provides a consensus view of the model's focus. These unified regions offer more trustworthy visual explanations by reinforcing attention zones that both methods agree upon, which is particularly valuable in sensitive domains like medical imaging, where interpretability and confidence are critical.

The frontend interface and experimental setup for the Breast Cancer Ultrasound Classification project have been thoroughly described in this chapter. The HVIT-AAF model was trained effectively using Kaggle's P100 GPU, backed by a strong procedure that included learning rate scheduling, early stopping, and 3-run cross-validation. A high-performing model was selected and saved in HDF5 format thanks to the assessment framework, which was based on a combined measure. Built with HTML, CSS, and Flask, the frontend provides an user-friendly interface for to explore Grad-CAM, LIME and URLAB explanations, upload photographs, and receive predictions.

These elements work together to create a coherent system that strikes a balance between usability, performance, and interpretability, providing a solid basis for real-world implementation.

7.6 Code and Dataset Accessibility

To ensure transparency, reproducibility, and support for future research, the complete source code and augmented dataset used in this study are made publicly available:

i. **GitHub Repository (Model + Frontend Code):** https://github.com/Salmanbnr/BreastCancerDetection_FinalYearProject.git
ii. **Augmented Breast Ultrasound Dataset (Kaggle):** https://www.kaggle.com/datasets/salmanbnr/breast-cancer-ultrasound-images/data

These resources include the model training pipeline, data augmentation code, final `.h5` model, and the frontend application developed using Flask.

CHAPTER 8

CONCLUSION

Using ultrasonic imaging and cutting-edge deep learning algorithms, this study has studied an innovative and all-encompassing technique to improve breast cancer diagnosis. By designing and assessing the HVIT-AAF (Hybrid Vision Transformer with Adaptive Attention Fusion) model, this research has tackled fundamental issues in medical image processing, achieving considerable breakthroughs in diagnostic accuracy, interpretability, and practical applicability. The work presented herein contributes meaningfully to the field of automated breast cancer diagnosis, with a particular emphasis on improving early detection in resource-constrained environments such as Pakistan, where the burden of breast cancer among younger women underscores the need for accessible and effective screening tools.

Creating a powerful deep learning framework that could reliably and effectively categorize breast ultrasound images into benign, malignant, or normal categories was the major objective of this research. The HVIT-AAF model, which is improved utilizing an Adaptive Focal Loss function to alleviate class imbalance, combines the global contextual information supplied by MobileViT-inspired transformer blocks with the local feature extraction capabilities of EfficientNetB0. With a short inference time of around 0.0094 seconds per image and validation accuracy of 99.11%, 99.84% Area Under the Curve (AUC), 99.11% precision, and 99.05% recall, the best configuration from Run 3 confirmed the model's remarkable performance in the trials. These studies show that the model can preserve computational economy while giving state-of-the-art diagnostic accuracy, which is critical for usage in clinical settings with restricted resources.

A cornerstone of this research is its thorough methodology, which covered data preparation, augmentation, and model training. The Breast Ultrasound Images (BUSI) dataset was increased by a proprietary augmentation technique, balancing the classes and increasing the model’s generalization to varied imaging situations. The integration of interpretability tools, such as Gradient-weighted Class Activation Mapping (Grad-CAM), Local Interpretable Model-agnostic Explanations (LIME) and Unified Regions

from LIME And Backpropagation (URLAB), further distinguishes this study, giving doctors with visual insights into the model's decision-making process. These tools not only boost trust in the AI system but also align with the ethical requirement of transparency in medical diagnoses. Additionally, the construction of a user-friendly frontend interface proves the practical utility of the model, enabling seamless interaction for image upload, prediction, and explanation visualization, thereby bridging the gap between advanced technology and clinical practice.

The significance of this research extends beyond its technical achievements, presenting important implications for healthcare delivery, particularly in areas like Pakistan where ultrasound's cost-effectiveness, portability, and safety make it an ideal screening tool. By achieving near-perfect classification performance and quick inference, the HVIT-AAF model has the potential to enable large-scale screening programs, boosting patient outcomes and early detection rates in underserved communities. The focus on computational efficiency assures that the model may work efficiently on modest hardware, thus boosting its accessibility in low-resource environments.

Nevertheless, this work acknowledges several shortcomings that pave the path for further research. The model's generalizability may require validation across broader and more diverse cohorts, as suggested by its dependence on a single dataset, albeit one that has been upgraded. Integration with additional imaging modalities or clinical data could further refine diagnosis accuracy, while advancements in real-time processing capabilities would ease its usage in dynamic clinical contexts. Furthermore, continued development of interpretability approaches ought to enable more profound comprehension of the logic of the model, boosting its clinical application and acceptance.

In conclusion, by establishing a novel hybrid deep learning architecture specifically customized for the ultrasound picture-based diagnosis of breast cancer, our research has considerably advanced the field of medical image analysis. The HVIT-AAF model shows a confluence of cutting-edge technology and practical application, providing remarkable performance while stressing interpretability and accessibility. As breast cancer continues to pose a significant global health challenge, particularly in regions with limited healthcare infrastructure, the advancements presented in this research offer a promising pathway toward improving diagnostic precision, reducing disparities, and

ultimately enhancing patient care. This work offers a good framework for future research, with the potential to transform ultrasound-based screening into an effective instrument for early identification and intervention.

REFERENCES

Al-Dhabyani, W., Gomaa, M., Khaled, H., & Fahmy, A. (2020). Dataset of breast ultrasound images. *Data in brief*, *28*, 104863.

Anderson, A., & Theophanous, R. G. (2025). Identifying enablers and barriers to teleultrasound use for remote settings: A scoping review. *Australasian Journal of Ultrasound in Medicine*, *28*(1), e12415.

Aps, J. (2020). Ultrasonic Imaging in Comparison to Other Imaging Modalities. In *Dental Ultrasound in Periodontology and Implantology: Examination, Diagnosis and Treatment Outcome Evaluation* (pp. 39-57). Springer.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Banerjee, S., & Monir, M. K. H. (2023). CEIMVEN: An Approach of Cutting Edge Implementation of Modified Versions of EfficientNet (V1-V2) Architecture for Breast Cancer Detection and Classification from Ultrasound Images. International Conference on Computing, Intelligence and Data Analytics,

Benaouali, M., Bentoumi, M., Abed, M., Mimi, M., & Taleb-Ahmed, A. (2024). A Study on CNN-Based and Handcrafted Extraction Methods with Machine Learning for Automated Classification of Breast Tumors from Ultrasound Images. *ELCVIA Electronic Letters on Computer Vision and Image Analysis*, *23*(2), 85-104.

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *74*(3), 229-263.

Cheyi, J., & Kaya, Y. Ç. (2024). Advanced CNN-Based Classification and Segmentation for Enhanced Breast Cancer Ultrasound Imaging. *Gazi University Journal of Science Part A: Engineering and Innovation*, *11*(4), 647-667.

Fiaz, A., Raza, B., Faheem, M., & Raza, A. (2024). A deep fusion-based vision transformer for breast cancer classification. *Healthcare Technology Letters*, *11*(6), 471-484.

Galabuzi, C., Abdullah, H., Ahmad, N., & Kaidi, H. M. (2024). EfficientNet-Based Deep Learning Neural Network for Accurate Plant Disease Detection. 2024 5th International Conference on Smart Sensors and Application (ICSSA),

Gheflati, B., & Rivaz, H. (2022). Vision transformers for classification of breast ultrasound images. 2022 44th annual international conference of the IEEE Engineering in Medicine & Biology Society (EMBC),

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1). MIT press Cambridge.

Hammood, D. A., & Hasan, S. M. (2024). Breast Cancer Detection Using Deep Learning. *Iraqi Journal for Computers and Informatics*, *50*(2), 122-131.

Hasan, M. E., & Khouri, H. (2023). Ultrasound Imaging: Differentiating Benign and Malignant Hepatic Tumors. *SAS J Med*, *11*, 1191-1195.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition,

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. Proceedings of the IEEE conference on computer vision and pattern recognition,

Iacob, R., Iacob, E. R., Stoicescu, E. R., Ghenciu, D. M., Cocolea, D. M., Constantinescu, A., Ghenciu, L. A., & Manolescu, D. L. (2024). Evaluating the role of breast ultrasound in early detection of breast cancer in low-and middle-income countries: A comprehensive narrative review. *Bioengineering*, *11*(3), 262.

Inan, M. S. K., Alam, F. I., & Hasan, R. (2022). Deep integrated pipeline of segmentation guided classification of breast cancer from ultrasound images. *Biomedical Signal Processing and Control*, *75*, 103553.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. International conference on machine learning,

Jabeen, K., Khan, M. A., Hamza, A., Albarakati, H. M., Alsenan, S., Tariq, U., & Ofori, I. (2024). An EfficientNet integrated ResNet deep network and explainable AI for breast lesion classification from ultrasound images. *CAAI Transactions on Intelligence Technology*.

Labonno, M., Asadujjaman, D., Rahman, M. M., Tamim, A., Ferdous, M., & Mahi, R. M. (2025). Early Detection and Classification of Breast Cancer Using Deep Learning Techniques. *arXiv preprint arXiv:2501.12217*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278-2324.

Lin, M., Chen, Q., & Yan, S. (2013). Network in network. *arXiv preprint arXiv:1312.4400*.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, *42*, 60-88.

Mahesh, T., Khan, S. B., Mishra, K. K., Alzahrani, S., & Alojail, M. (2025). Enhancing Diagnostic Precision in Breast Cancer Classification Through EfficientNetB7 Using

Advanced Image Augmentation and Interpretation Techniques. *International Journal of Imaging Systems and Technology*, *35*(1), e70000.

Majeed, A. I., & Bangash, R. S. (2024). Scenario of Pakistan 5 Years After Screening Mammography Intervention: Are we ready for a National Guideline?

Mathpal, J. (2024). Breast Cancer Detection: A Comprehensive Study on Machine Learning and Deep Learning Techniques.

Mehta, S., & Rastegari, M. (2021). Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*.

Mustafa, W., & Mustafa, O. (2024). ACR Breast Density: Relationship with Age, and Its Impact on Mammographic and Ultrasound Findings. *AlQalam Journal of Medical and Applied Sciences*, 847-854.

Pachisia, A. V., & Govil, D. (2025). Point-of-care ultrasound training and education in low-and middle-income countries. *Journal of Nepalese Society of Critical Care Medicine*, *3*(1), 1-4.

Pangaribuan, H. P. (2024). Effectiveness of Breast Ultrasound for Breast Cancer Screening: A Systematic Review. *Indonesian Health Journal (IHJ)*, *3*(3), 229-237.

Piotrzkowska-Wróblewska, H., Dobruch-Sobczak, K., Byra, M., & Nowicki, A. (2017). Open access database of raw ultrasonic signals acquired from malignant and benign breast lesions. *Medical physics*, *44*(11), 6105-6109.

Ramachandran, P., Zoph, B., & Le, Q. V. (2017). Searching for activation functions. *arXiv preprint arXiv:1710.05941*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining,

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE conference on computer vision and pattern recognition,

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE international conference on computer vision,

Shi, J., Zhou, S., Liu, X., Zhang, Q., Lu, M., & Wang, T. (2016). Stacked deep polynomial network based representation learning for tumor classification with small ultrasound image dataset. *Neurocomputing*, *194*, 87-94.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929-1958.

Suara, S., Jha, A., Sinha, P., & Sekh, A. A. (2023). Is grad-cam explainable in medical images? International Conference on Computer Vision and Image Processing,

T. R, M., V, V. K., & Guluwadi, S. (2024). Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with Resnet 50. *BMC Medical Imaging*, *24*(1), 107.

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. International conference on machine learning,

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, J., Tong, J., Li, J., Cao, C., Wang, S., Bi, T., Zhu, P., Shi, L., Deng, Y., & Ma, T. (2024). Using the GoogLeNet deep-learning model to distinguish between benign and malignant breast masses based on conventional ultrasound: a systematic review and meta-analysis. *Quantitative Imaging in Medicine and Surgery*, *14*(10), 7111.

Zeimarani, B., Costa, M. G. F., Nurani, N. Z., Bianco, S. R., Pereira, W. C. D. A., & Costa Filho, C. F. F. (2020). Breast lesion classification in ultrasound images using deep convolutional neural network. *IEEE Access*, *8*, 133349-133359.

Zhang, C., Wang, L., Wei, G., Kong, Z., & Qiu, M. (2024). A dual-branch and dual attention transformer and CNN hybrid network for ultrasound image segmentation. *Frontiers in Physiology*, *15*, 1432987.

Zhang, E., Seiler, S., Chen, M., Lu, W., & Gu, X. (2020). BIRADS features-oriented semi-supervised deep learning for breast ultrasound computer-aided diagnosis. *Physics in Medicine & Biology*, *65*(12), 125005.

Zhang, Y., Xian, M., Cheng, H.-D., Shareef, B., Ding, J., Xu, F., Huang, K., Zhang, B., Ning, C., & Wang, Y. (2022). BUSIS: a benchmark for breast ultrasound image segmentation. Healthcare,