Fetal Plane Classification from 2D-Ultrasound Images Leveraging Squeeze and Excitation Self-attention Mechanism for Feature Recalibration in MedMamba

by

Tanjim Islam Riju
20101403
Tahsin Tanim Ramisha
20101439
Nusrat Billah Aksa
22241116
Johan H Kabir
22241114

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
Bachelor of Science in Computer Science

Department of Computer Science and Engineering
Brac University
May 2024

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Tahsin Tanim Ramisha
20101439

Tanjim Islam Riju
20101403

Nusrat Billah Aksa
22241116

Johan H. Kabir
22241114

Approval

The thesis/project titled "Fetal Plane Classification from 2D-Ultrasound Images Leveraging Squeeze and Excitation Self-attention Mechanism for Feature Recalibration in MedMamba" submitted by

1. Tanjim Islam Riju (20101403)
2. Tahsin Tanim Ramisha (20101439)
3. Nusrat Billah Aksa (22241116)
4. Johan H Kabir (22241114)

Of Spring, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on May 2024.

Examining Committee:

Supervisor:
(Member)

Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

Rafeed Rahman
Lecturer
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, Ph.D.
Chairperson
Department of Computer Science and Engineering
Brac University

Abstract

A fetal ultrasound is a safe pregnancy test that provides an image of the baby's heart, head, and spine while also analyzing various aspects of its anatomy. Maternal-fetal ultrasound imaging is critical during pregnancy, but existing approaches rely on manual interpretation, which can be time-consuming and can overlook irregularities. Thus, the exploration of fetal ultrasound imaging has resulted in the need for accurate and fast medical image classification. However, there have been some limitations with traditional methods such as Convolutional Neural Networks (CNNs) and transformer models. For example, CNNs do not work well when it comes to modeling long-range dependencies that are very important in medical image feature extraction. Also, transformers have a high quadratic complexity hence demanding too much computation despite being good at dealing with long-range interactions. Our thesis is inspired by recent advances made in state space models (SSM) and thus presents an implementation of Vision Mamba called "MedMamba" designed for classifying medical images. This is achieved through integration of SS-Conv-SSM module which combines local feature extraction capabilities brought about by convolution layers together with long range dependency modeling as exhibited by SSMs thus solving the above mentioned problems encountered during CNN usage. In other words we can say that this hybrid method guarantees strong feature extraction across different types of medical imaging modalities while improving on computational efficiency. Furthermore we have presented an enhancement of the architecture MedMamba called "MedMambaSE" by adding a Squeeze and Excitation (SE) block. This addition refines the process of recalibrating features ultimately improving the models sensitivity and accuracy in detecting abnormalities in development. The incorporation of this block boosts MedMambas adaptability and effectiveness, in handling the complexities of ultrasound images. Through experiments on a dataset of ultrasound images we have shown that MedMambaSE not only enhances classification accuracy but also establishes a new standard for automated analysis of fetal images. This study sets a milestone, in diagnostics and opens doors for advancements in AI driven medical imaging that could revolutionize prenatal care with quicker and more precise interpretations.

Keywords: Maternal-fetal, CNN, Transformer, State Space Models, MedMamba, MedMambaSE.

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.
Secondly, to our supervisor Md. Golam Rabiul Alam sir and co-supervisor Mr. Rafeed Rahman sir for their kind support and advice in our work. They helped us whenever we needed help.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Maternal-fetal ultrasound imaging is the foundation of modern pregnancy, which offers a simple way to monitor and assess the health of both the pregnant woman and the developing fetus. Initial detection of possible difficulties, the facilitation of fetal growth evaluation, and finally maintaining the condition of the mother and fetus depend on the accurate and prompt classification of maternal-fetal ultrasound pictures.

The precision of these diagnostics is vital in catching issues before they become major complications associated with pregnancy and delivery. As imaging technologies continue to mature and selection algorithms improve, better detection of anomalies earlier in the process, when possible action is more likely to be successful. This is particularly true now that ultrasounds have experienced regular improvements in technology. The ongoing growth of this product range continues to improve the prospects of prenatal care for the entire world, especially for mothers and intended families at home.

Ultrasound imaging is important for prenatal diagnostics, but the quality of such images can vary greatly, depending on the experience and skill of the operator and the resolution of the machine. Consequently, measurements and diagnostics pulled from the images often are fraught with errors. In addition, ultrasound acquisition inherently creates noise in the images such as speckle noise, shadowing, and attenuation which make identifying or segmenting structures from the images challenging. Automated solutions for ultrasound segmentation are problematic due to the variability of these noise artifacts making homogeneous tissue classification difficult to achieve. Further, although these images can lend key information in determining diagnosis, a complex interpretation may be necessary and often needing a level of expertise that is not always accessible, which is particularly true of low-resource healthcare centers. A discrepancy in expertise creates not only an inconsistent diagnosis of conditions, but one that may be incorrect.

Identifying metrics like fetal dimensions, cardiac activity, and standard scanning views, as well as segmenting anatomical structures and classifying standard planes and anomalies in the fetus, are crucial research areas aimed at enhancing the quality of prenatal assessments [25]. For tracking fetal growth and spotting any potential problems during pregnancy, ultrasonography in gynecological services is essential. One important application of ultrasonography in the medical profession is the clas-

sification of common maternal-fetal imaging planes.

Developing new imaging and machine learning techniques and methods to standardize and automate this process is thus a need. Healthcare professionals can benefit greatly from the work in this area, which will provide meaningful and consistent diagnosis and diagnostics that will increase the safety of prenatal diagnostic methods by working to minimize operator errors. Advances in image processing and machine learning have the potential to greatly improve on these major challenges. Improving image quality and building algorithms to accurately and automatically analyze images and provide standardize diagnoses need to be built. These methods can then be used to standardize the diagnosis of prenatal care, allowing for improved quality of prenatal care and diagnostic quality, regardless of the operator or the practice's level of expertise.

To systematically pinpoint unique markers of atypical fetal growth in ultrasound imagery, employing machine learning techniques for processing and interpreting these images can assist in automated large-scale retrospective analyses. Presently, the automated classification options are limited to either image snippets (cropped sections) or the full image. Cropped sections can result in misidentifying specific organs like the kidneys and abdominal areas, given that many organs in development have similar visual features. On the other hand, using the entire image doesn't offer sufficient localized information to differentiate between various structures based on their location. Consequently, deep learning algorithms have emerged as a groundbreaking approach to enhance the precision and efficiency of ultrasound image classification [13].

Using Convolutional Neural Networks (CNNs), Deep Learning (DL) has made astounding progress in picture identification tasks, which has accelerated the development of artificial intelligence throughout the preceding ten years. CNNs have proven useful in a number of medical fields, including radiography, dermatology, and the classification or segmentation of organs and lesions in computer tomography images. For the purpose of estimating fetal gestational age, Maraci et al. presented a DL-based technique that extracts TC plane frames from point-of-care ultrasound films using a modified CNN from AlexNet [24]. A deep learning model was created by Rasheed et al. to automate fetal head biometry from live ultrasonography. They use CNN ALEXNET to classify headframes, OFD for validation, UNET for segmentation, and LSE to compute HC and BPD for accurate gestational age . CNNs in particular have proven to be adept at image categorization and feature extraction, making them an ideal candidate for improving the interpretation of maternal-fetal ultrasound data [24].

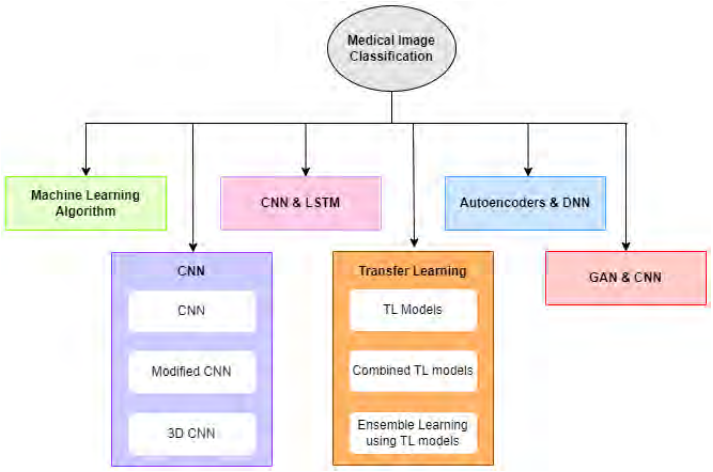

Figure 1.1: Techniques employed in ultrasound fetal plane imaging

1.1 Motivations

"Odisha: Wrong ultrasound reports results in woman delivering a handicapped child; nursing home fined" (Loreng, 2022). This news highlight, therefore, gives the impression that most of the errors in the medical report are riddled with errors in the ultrasound imaging reports. This demonstrates an effect that is so severe in the case of misdiagnosis: the misreporting of conditions that result in children born with anomalies, physical deformities, or multiple gestations being overlooked. These mistakes certainly serve to highlight the importance of getting the image right in ultrasound fetal plane imaging.

With these at the forefront of our minds, this paper will rely on one of the most recent advances in ultrasound fetal plane image classification: MedMamba—an innovative model from recent research. Despite the very important advances in image processing that deep learning has brought, it has its limitations in what can be achieved in the medical field. For example, convolutional neural networks (CNN) do not capture the big picture as they only capture local information. Transformer based architectures are state-of-the-art in many NLP tasks and are well-suited to looking at the big picture but require a colossal amount of computation, therefore limiting their deployment in time-critical and resource-constraint environments. Realizing these constraints, MedMamba offers a new approach that, following help from the state space model (SSM), does better at ultrasound fetal plane image classification. Evidence from other fields suggests that SSMs may lend themselves to large data sets without overpowering computational resources [26]. From this, the MedMamba model maximizes the utility of both CNNs and SSMs to appropriately process and categorize ultrasound fetal plane images.

Our thesis, based on recent developments in the MedMamba framework, provides further improvements for the optimization of ultrasound fetal plane image classification. We made key changes to the original model to improve its ability to identify subtle details within the medical scan. Our ultimate goal is to contribute to a future when ultrasound interpretations will be much more trustworthy.

1.2 Contributions

In this research, we have improved the MedMamba model capabilities—the deep learning architecture—by integrating Squeeze-and-Excitation (SE) blocks within its convolutional self-attention modules. The improvement is particularly remarkable in the model recalibration of the feature channels dynamically after fusing both convolutional and attention-driven features. The SE blocks enhance the ability of the model to emphasize relevant features, suppressing less useful ones. This is very critical for the high variability and often subtle features characteristic of ultrasound imagery. The key contributions of the research are as follows:

- We introduce a novel architecture namely MedMambaSE by adding Squeeze-and-Excitation (SE) blocks to Medmamba's convolutional self-attention modules. We show this improvement increases feature recalibration capabilities within maternal fetal ultrasound images.
- We employ the Medmamba model which combines convolution layers for local feature extraction with state space models for long-range dependency modeling.
- We have evaluated our model performance with convolutional neural network models such as VGG19, SENet and transformer model such as Swin transformer to find useful insights.

1.3 Thesis Organization

This section introduces the structure of our article. First, we review previous studies that have used machine learning for clinical image classification, focusing on fetal ultrasound imaging. In Chapter 3, we describe the various implemented models. Moreover, this chapter includes our dataset description and how we have calibrated the dataset according to the models. Chapter 4 discusses the development of our proposed model MedMambaSE. Chapter 5 summarizes our findings, discusses the limitations of our study. Finally, Chapter 6 suggests areas for future research.

Chapter 2

Literature Review

During prenatal ultrasound screenings, medical professionals manually gather a variety of imaging angles to obtain standard fetal ultrasound views. However, the process can be challenging due to the intricacies involved and the differing levels of expertise among practitioners. As a result, there exists an inconsistency in the images obtained, with minimal variation between classes but significant variation within them [5].

2.1 Traditional Image Processing

Traditional manual feature based approaches for image classification consist of three phases: feature encoding; feature extraction; classification [8]. In a 2012 study [2], Active Appearance Models (AAM) technique was used to accurately identify and position the fetal head's standard plane. In research [1] introduced an innovative automatic positioning strategy for the upper abdomen's standard plane. Leveraging clinical anatomical knowledge, they employed a radial model to illustrate the spatial relationship among vital anatomical elements within the abdominal plane. This approach ensured accurate positioning of the plane. In a subsequent study [3], proposed integrating foundational features with a multi-tiered Fisher Vector (FV) for comprehensive feature encoding to establish holistic image features. To pinpoint the standard fetal plane, they employed an SVM classifier as well. However, it is important to note that this approach has its limitations due to foundational features' restricted capacity in representing features effectively; thus, enhancements are warranted for optimal algorithm performance. In research [4], introduced an innovative approach to recognize the standard plane of fetal faces. They utilized the Root SIFT method to derive the characteristics of images. Subsequently, FV and SVM were used for further classification tasks. This method achieved a remarkable accuracy of 93.27% with a mean average precision (MAP) of 99.19%. In study [6], they employed the Return Woods technique to analyze the visibility, location and orientation of fetal heart ultrasound snapshots. Using this method, they successfully identified the standard plane of the fetal heart from individual video frames with expert-level precision. Moreover, study [7] introduced an improved LBP technique. This technique has the ability to extract both color and texture attributes simultaneously and thus effectively counter impulse noise. It marked a significant advancement in the field of LBP Methodology. In study [15] researchers expanded their work to analyze bark textures with the help of precise classification approach. This approach is grounded

in the enhanced local ternary pattern (ILTP). This work not only showcased advanced iterations of LBP and LTP, but also provided valuable insights for future experimental endeavors. Later, the advancements have moved from simple feature extraction to deep learning models.

2.2 The Integration of CNNs and Transformers in Medical Imaging

Advancements in deep learning approach have begun to emerge since 2012. These advancements have increasingly involved the identification and categorization of standard ultrasound planes. Notably, models like VGG19, SENet, and Swin have emerged as front-runners in this domain, showing significant improvements in image classification and anomaly detectionection within fetal ultrasounds [41]. Because of their sophisticated architectures, subtle features from complex ultrasound images can be extracted effectively. This bridges the gap between technological innovation and clinical practice. By harnessing the prowess of these deep learning tools, maternal-fetal medicine stands on the cusp of yet another breakthrough. Recent studies have demonstrated the exceptional accuracy rates of these models when integrated into clinical scenarios. This proves their potential for broader applications in medical imaging [41]. As advancements in remote and battlefield medicine continue, these models are becoming indispensable, particularly where resources and expert image interpretation are limited [41].

During pregnancy, ultrasound remains a crucial modality for monitoring the fetus. However, it has traditionally been challenging for even expert sonographers to accurately identify anatomical structures [23]. To overcome this obstacle, a deep feature fusion from pretrained models like ResNet-50 and VGG-19-GAP was utilized [23]. In terms of accuracy this method has outperformed many conventional techniques. Moreover, another approach was employed to extract deep features from ultrasound images. In this approach, the AlexNet and VGG-19 models have been used into a multi-layer perceptron for classification[27]. The integration of these diverse convolutional neural networks helped to improve the diagnostic outcomes. Additionally, there has been a recent trend towards automating deep network architectures. Remarkably, this algorithm proved to be competitive with highly acclaimed models such as VGG16 and ResNet50 when applied to maternal-fetal ultrasound images [22]. Together, these advancements highlight the transformative potential of deep learning models in enhancing maternal-fetal ultrasound diagnostics.

Moreover, in study [20], a novel three-dimensional (3D) ultrasound method has been utilized to classify standard fetal planes. The findings demonstrated that this 3D approach exhibited exceptional precision in identifying these essential planes. Another study [10] came up with an automatic way to identify the standard plane of a fetal face in ultrasound scans using deep learning. This was hard because of the differences within the same class and how standard and non-standard fetal faces look similar. They used transfer learning and special ways to increase their data set in their deep learning model, and did better than older methods. They used a set

of 4,849 labeled ultrasound pictures, checked by experienced medical experts. The model did really well, with an average AUC of 0.99 and high scores in Accuracy (0.96), Precision (0.96), Recall (0.97), and F1 (0.97).

In another investigation [21], Zhang et al introduced an automated assessment system for evaluating image quality in fetal sonography. By utilizing multitask learning, the system employed three convolutional neural networks to identify critical anatomical features and assess whether a sonographic image met the desired criteria. The outcomes revealed an impressive success rate of 94.3% and a specificity of 94.6%.

In study [9], a multi-layered dense network was utilized to identify specific areas of the fetal brain, heart, face and abdomen from a collection of 5678 ultrasound visuals. The associated Recall, Precision and F1 metrics were all recorded at an impressive 0.98. On the other hand, study [12] implemented an automated technique named SPRNet to detectionect the fetal brain, heart,face, abdomen, as well as the facial coronal view during prenatal screening. Taking inspiration from DenseNet, SPRNet was trained on both fetal and placenta ultrasound images using a partial transfer learning approach based on data. Notably, this model achieved exceptional performance with accuracy scoring 0.99 along with recall at 0.96; specification reaching 0.99; and F1 measuring 0.95.

In study [17], the researchers introduced a unique differential convolutional neural network (differential-CNN). This specially designed network seamlessly distinguishes between standard and non-standard brain planes in fetuses. By using differential operators, the model extracts enhanced differential features from the base CNN's feature maps. This enhancement improves the identification capacity without adding extra computational demands. To test this approach thoroughly, they conducted experiments on a dataset of 30,000 2D ultrasound images, which included 155 fetuses aged between 16 and 34 weeks. The results showed impressive scores for Precision (0.93), F1 (0.93), Accuracy (0.93) and Recall (0.92).

Research [19] utilized a generative adversarial network (GAN) to enhance the categorization abilities of the fetal brain using ResNet. Their approach was tested on 2249 images and achieved an AUC of 0.86, with Accuracy and F1 scores of 0.81 and 0.80 respectively. In another study [16] focused on inter-device classification for standard anatomical views including the fetal heart, abdomen, and mouth. They employed enhanced feature alignment techniques to identify both unique and consistent features across different domains. The results showed average scores of 0.77 for both F1 and Recall, and 0.78 Precision.

2.3 The Inefficiency of CNNs and Transformers with Long Sequences

When we look at how well transformer models perform on sequential data, what stands out is that their attention mechanisms excel at picking out useful pieces of information and concentrating on them such that not every item has to be processed in sequence. This property makes them more efficient than convolutional

neural networks (CNNs) but only up to a point. Because of this, transformers fail to cope with very long sequences and that can mean much longer training times and increased computational requirements which may affect result accuracy eventually. The main problem lies with the inability of the transformer to effectively compress context into relevant information.

On the other hand, RNNs are designed for sequential processing hence should theoretically work better with extended data sets. However, this processing method increases computation while exposing them to risks associated with vanishing or exploding gradients. Also, RNNs tend to struggle in preserving input data for long periods thereby becoming ineffective in managing tasks involving extended memory spans at a later stage where initial inputs are concerned.

2.4 Introduction of Mamba's Efficient Sequence Modeling

Mamba is a new type of sequence model that combines the best features of transformers and CNNs [26]. It treats long sequences more effectively than other models by segmentationmenting them and introducing key-value attention mechanism, which attends only to important parts. Thus it breaks free from sequential nature of CNN in favor of dividing them into shorter pieces that can better preserve long-term dependencies.

By employing adaptive computing along with dynamic length adjustment techniques, Mamba demonstrates accelerated training on longer inputs as well as improved performance over such records. The design relies heavily on memory optimization for GPUs while dealing with large data sets; this ensures processing power remains efficient throughout different stages. When compared against other models in terms of perplexity and accuracy measurements especially when working with very long chains, Mamba always outperforms them all indicating robustness towards context choice during computation.

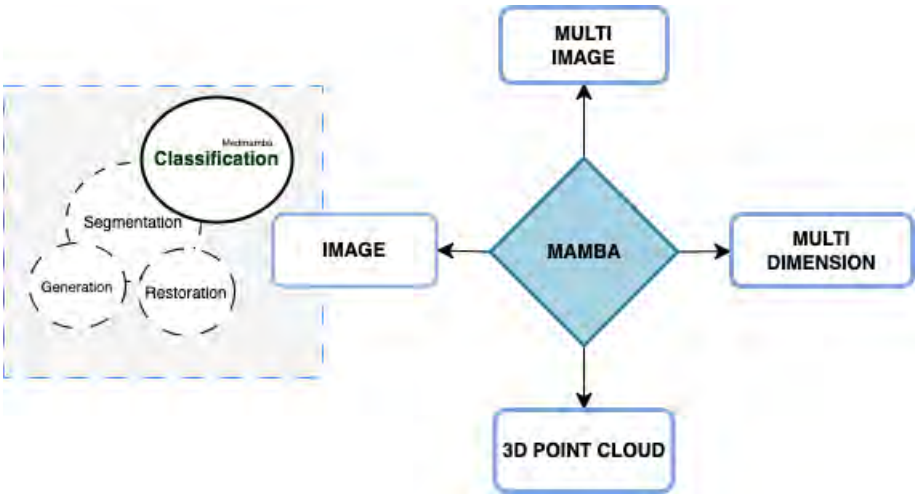

Figure 2.1: Implementations of Mamba across diverse vision sectors.

Table 2.1: Mamba based Image Classification Methods

Image Classification		
Methods	Dataset	Experiment
FER-YOLO-Mamba [34]	RAF-DB, SFEW	Facial Expression classification
RES-VMAMBA [28]	Food Dataset: CNFOOD-241	Food classification
Mamba-ND [33]	Natural Images	2D Natural Images classification
nnMamba [31]	6 3D Biomedical Image Dataset	3D Image segmentation, classification, detection
MambaMIL [38]	Whole Slide Images	Cancer Subtyping/Survival Prediction
RSMamba [29]	Remote Sensing Images	Remote Sensing Images classification
MamMIL [30]	Whole Slide Images	Cancer Subtyping
CMViM [36]	3D Medical Images (MRI & PET)	3D Medical Image classification
MedMamba [40]	2D Medical Images	2D Medical Image classification
Spectral-Spatial Mamba [32]	HSI Dataset	Hyperspectral Image classification
SpectralMamba [39]	HS Dataset	2D Medical Image classification
HSIMamba [37]	Houston 2013, Indian Pines, Pavia University	Hyperspectral Image classification
S^2Mamba [35]	Houston 2013, Indian Pines, Pavia University	Hyperspectral Image classification

2.5 MedMamba: Bridging CNNs and Transformers for Advanced Medical Image Classification

In the recent research [26], SSM-based algorithm Mamba has been introduced that handles these issues well by considering long-range interactions with a linear computational complexity. Based on this study, MedMamba has been introduced which is the first modification of the Vision Mamba network design and is customized for medical image classification [40]. The special Conv-SSM architecture of MedMamba combines the advantages of convolutional layers that allow for local feature extraction and SSM, which can capture long-range dependency. The hybrid approach described in this paper allows to perform comprehensive modeling of medical images covering different imaging modalities. The findings from the experiments in this paper illustrate MedMamba's strength in lesion detection within diversified scopes of medical imaging. Moreover, this research marks the foundation for a new era of SSM-based AI algorithms and systems, paving the way towards resourceful and cost-effective AI solutions in the healthcare sector.

Chapter 3

Background Study

3.1 Overview of the Models

3.1.1 VGG19

The VGGNet, also known as VGG, marks a significant turning point in the design and depth of convolutional neural networks (CNN). Developed by the Visual Geometry Group, this architecture stands out for its remarkable depth. The widely recognized VGG-16 and VGG-19 models have 16 and 19 convolutional layers respectively. It's worth noting that the latter model, VGG-19, boasts an additional three convolutional layers compared to its counterpart. VGG's groundbreaking architecture has revolutionized object recognition methodologies, surpassing established benchmarks across various tasks beyond its original implementation on ImageNet. Its continued relevance is evident through its prominent role in contemporary image recognition research.

First of all, images pass through multiple convolutional layers which have 3x3 filters. After that, spatial dimensions are reduced by applying max pooling. As the image moves through these layers, it extracts complex features more and more. When all convolutional layers are over, the data is arranged in a row and passed through fully connected layers. Finally, an image is passed through a softmax activation function for classification. Significantly, it makes extensive use of ReLU (Rectified Linear Unit) activation, which enables it to capture non-linear patterns by outputting input directly if positive and producing zero when otherwise.

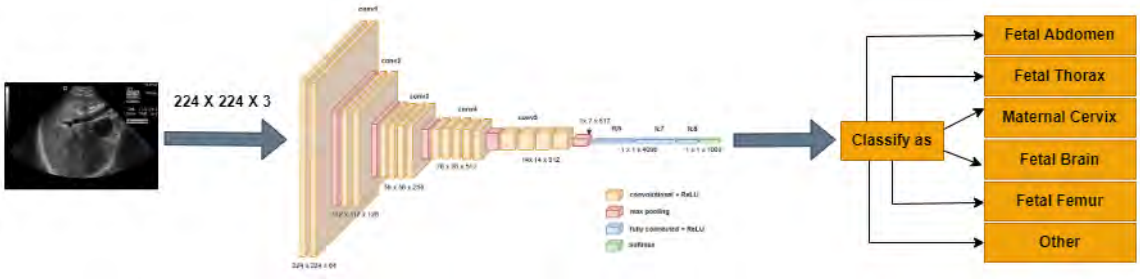

Figure 3.1: VGG19 Working Procedure

3.1.2 SENet (Squeeze-and-Excitation Network)

The first Squeeze and Excitation Network was proposed in 2018 by Hu et al. [11] for Convolutional Neural Networks. The novel method looks for better channel relationships within the process with a channel-wise attention mechanism. The main idea is to give channels adjustable weights that amplify important features while reducing insignificant ones. The Squeeze and Excitation block forms the basis of this method commonly known as SE-block. It consists of three key operations:

i. Squeeze

ii. Excitation

iii. Scaling

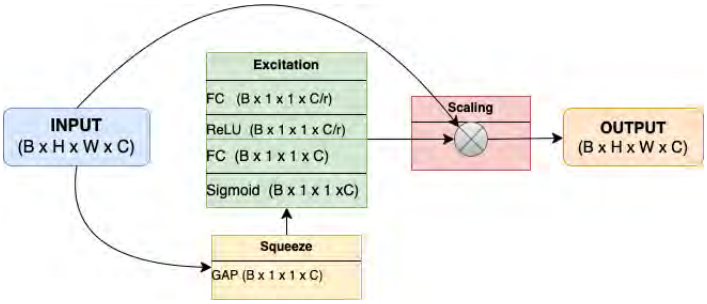

Figure 3.2: SENet Network

It is at strategic points in the network that the squeeze-and-excitations blocks are used. These block compress spatial information to produce a channel descriptor which then fine-tunes the channels using and recalibrates or excites them. As this goes through its recalibrated layers, it improves at picking up distinctive features. After going through all of the convolutional layers, the data is flattened before being fed into fully connected ones. Eventually, an image gets classified with the help of a SoftMax activation function. The dynamic ability of SENet to adjust feature responses per-channel is one of its major attributes. This means that for any input, only relevant channels will be attended by the network.

3.1.3 Swin Transformer

An innovative architecture known as Swin Transformers was introduced in 2021 [18]. Instead of processing images in patches like conventional methods, Swin Transformers splits them up into non-overlapping windows that shift position. The speed and scalability for handling huge amounts of data has been enhanced significantly by this invention. Displaced windows are the building blocks in Swin Transformers' design. The inherent quadratic complexity problem that conventional transformers encounter when working with high-resolution images is successfully addressed by implementing this well-ordered and hierarchical approach. Swin Transformers, with this automatic ability to adapt to different image resolutions in this manner are very much valuable for images of varying dataset sizes.

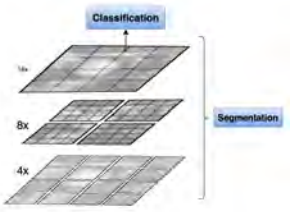

Figure 3.3: (a) Swin Transformer

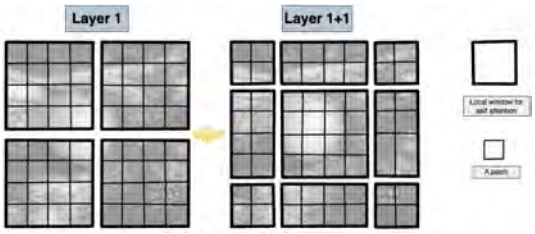

Figure 3.4: (b) Window Shift

Partitioning the input image into several non-overlapping patches of the same fixed size is the first step in the architecture of Swin Transformers. Following this, these patches undergo some changes that turn them into vectors used as initial inputs for subsequent transformer layers. In each layer, the Swin model has local self-attention, where each location attends to only a few other positions within a small window. Through different layers, these windows are adjusted so as to have a better understanding of the whole image. In deeper levels, smaller patches are combined together and therefore they form larger ones creating hierarchical representation. Local-to-global technique is how Swin achieves efficient computation. Eventually after this process; global average pooling is performed and then classifier head is put in place to obtain the final result for example image classification tasks. Layer normalization ensures stability during training stage in deep transformer systems.

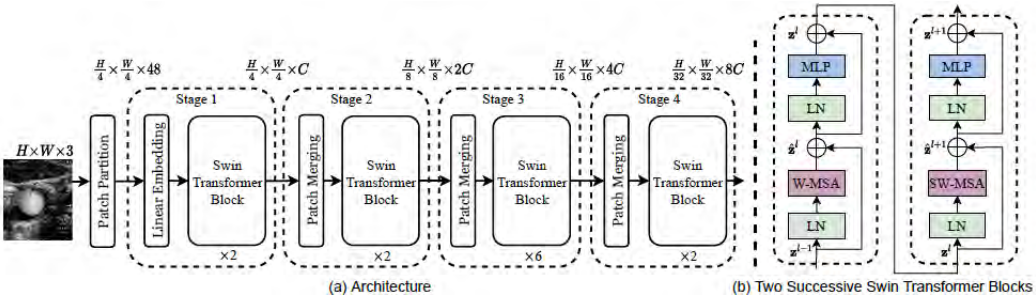

Figure 3.5: Swin Transformer Working Procedure

3.1.4 MedMamba

For classifying medical images like X-rays and scans, MedMamba is a deep learning model made especially for this purpose. It incorporates some distinctive alternatives to the traditional convolutional neural networks (CNNs) that make it an ideal architecture for medical imaging. Key features are Spatial Attention Modules which highlight the most important part of an image in order to capture important anatomical details, and Multi-Scale Fusion which brings together information from different resolutions of images so as to improve on details and context recognition. Extensively trained on datasets such as chest X-rays and mammographies, its performance has been superior to those of other state-of-the-art models thereby indicating that it can significantly enhance health care through computer aided diagnosis. Thus, this is a useful tool in improving the analysis of medical images which may thus, result in better diagnostic decisions, treatment plans and general life directions of patients.

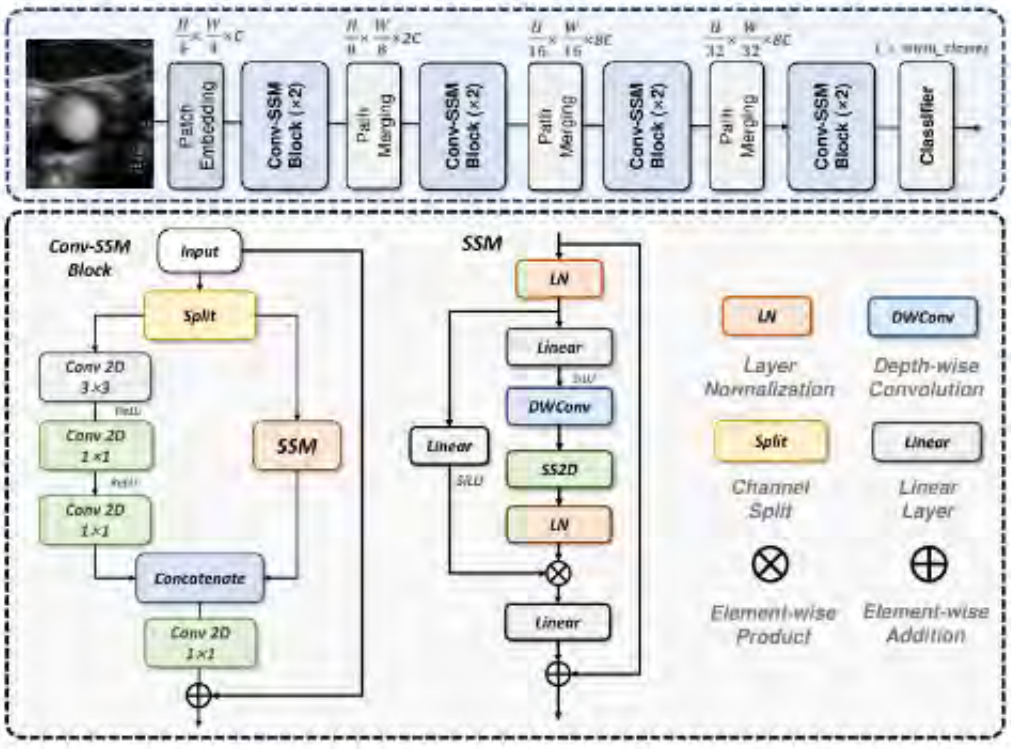

Figure 3.6: Overview of MedMamba Architecture

The Medmamba architecture consists of three main components: (1) a patch embedding layer (2) SS-Conv-SSM block and (3) patch merging layer [40]. Firstly, the input images with $H \times W \times 3$ dimensions undergoes a patch embedding layer. This stage divides the image into the size of 4×4 smaller, non-overlapping patches. This division reduces the dimensionality to $[\frac{H}{4} \times \frac{W}{4} \times C]$ where C is the number of channels, typically equals 96. Before these patches are utilized further, they undergo normalization through a Layer Normalization technique. The core structure of MedMamba is built from four main stages. Notably, after the first three stages, the system applies patch merging layers to condense the feature map's height and width while expanding the channel capacity. The stages are configured with $[2, 2, 4, 2]$ SS-Conv-SSM blocks, each adjusting the channel configuration to $[C, 2C, 4C, 8C]$ to enhance processing capability.

3.2 Dataset Description

For our research we have used a maternal fetal untrasound image dataset, developed by Burgos-Artizzu et al [14]. This dataset is widely recognized as the most comprehensive collection of ultrasound images from singleton pregnancies available today. The dataset comprises over 12,400 images sourced from 1,792 patients, providing a diverse and comprehensive set for research purposes.

The dataset strictly follows the clinical US screening guidelines established by a scientific committee [14], ensuring consistency and minimizing both inter-observer and intra-observer variations. Noteworthy clinicians have provided annotations for each image.

The images fall into six main categories. Four of these categories correspond to the most frequently examined fetal anatomical regions: Abdomen, Brain, Femur, and Thorax. Additionally, there's a distinct category for cervical images of the mother, vital for screenings related to premature birth. The last group includes less commonly captured anatomical views. Within the category designated for fetal brain images, there are more specific subdivisions, focused on three major planes of the fetal brain: Trans-thalamic, Trans-cerebellum, and Trans-ventricular. The dataset also includes an "Other" category.

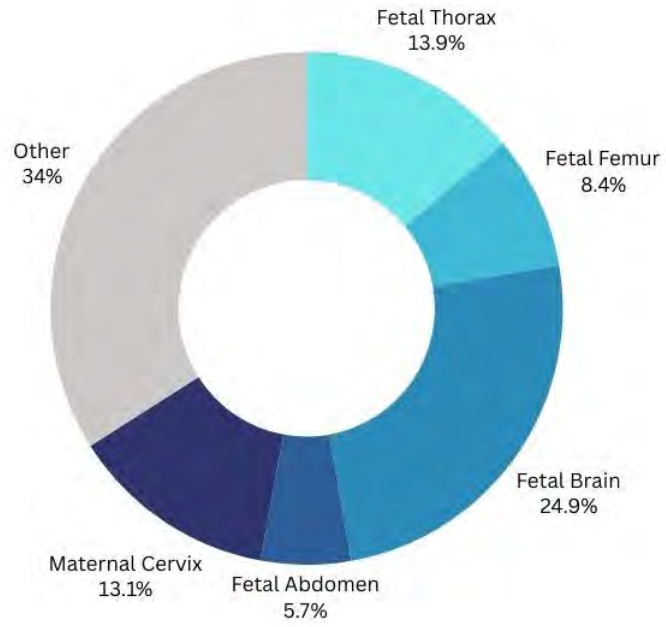

Figure 3.7: Distribution of the classes in the Dataset

3.3 Dataset Implementation

3.3.1 Data Extraction and Partitioning

The first step in our preprocessing pipeline involved extracting the image paths and associated labels from a meticulously curated CSV file. To ensure the integrity of the dataset, entries with missing image path were promptly removed. The categorical labels were then converted into numerical format to facilitate the machine learning process. The dataset was divided using stratified random sampling, allocating 80% for training and 20% for testing. A subset of the training data was further used for validation.

Table 3.1: Dataset Split in Neural Network Models

Category	Train Data	Test Data	Validated Data	Total
Fetal Abdomen	455	155	101	711
Fetal Thorax	1122	319	277	1718
Fetal Femur	662	201	177	1040
Fetal Brain	1973	629	490	3092
Maternal Cervix	1024	341	261	1626
Other	2700	835	678	4213
Total	7936	2480	1984	12400

For the SSM-based model MedMamba, we have divided the data into a 60:20:20 ratio for train, test and validation purpose.

Table 3.2: Dataset Split in MedMamba

Category	Train Data	Test Data	Validated Data	Total
Fetal Abdomen	352	140	219	711
Fetal Thorax	1053	322	343	1718
Fetal Femur	526	227	287	1040
Fetal Brain	1625	679	788	3092
Maternal Cervix	1286	157	183	1626
Other	2598	955	660	4213
Total	7440	2480	2480	12400

3.3.2 Model-Specific Preprocessing

Leveraging the PyTorch framework, SENet required the images to be resized, center-cropped, and transformed into tensor format. Image normalization was applied using predefined mean and standard deviation values compatible with ImageNet.

Unique to SWIN was its dual compatibility with both PyTorch and TensorFlow frameworks. It utilized torchvision's transforms module for image transformation and featured a function to convert PyTorch DataLoaders into TensorFlow datasets.

In contrast to other models, VGG19 employed preprocessing steps such as rescaling, rotations, shifts, and flips. The final step involved creating generator objects for image batch generation.

As data preprocessing steps for MedMamba, we have utilized image resizing, horizontal flip and converting the images into tensor format. After that image normalization has been applied using standard deviation and mean value.

Each of these preprocessing pipelines was fine-tuned to cater to the specific needs and characteristics of the corresponding deep learning architecture, thereby optimizing the dataset for the highest possible performance metrics.

Chapter 4

Methodology

The proposed MedMambaSE model builds upon the original MedMamba architecture by incorporating a Squeeze-and-Excitation (SE) block. It includes three key modules: collecting the dataset (utilizing 2D Ultrasound Fetal Image dataset [14]), establishing model specific preprocessing, creating training and testing datasets to educate the customized MedMambaSE architecture. The top level architecture of MedMambaSE has been represented in Fig 4.2.

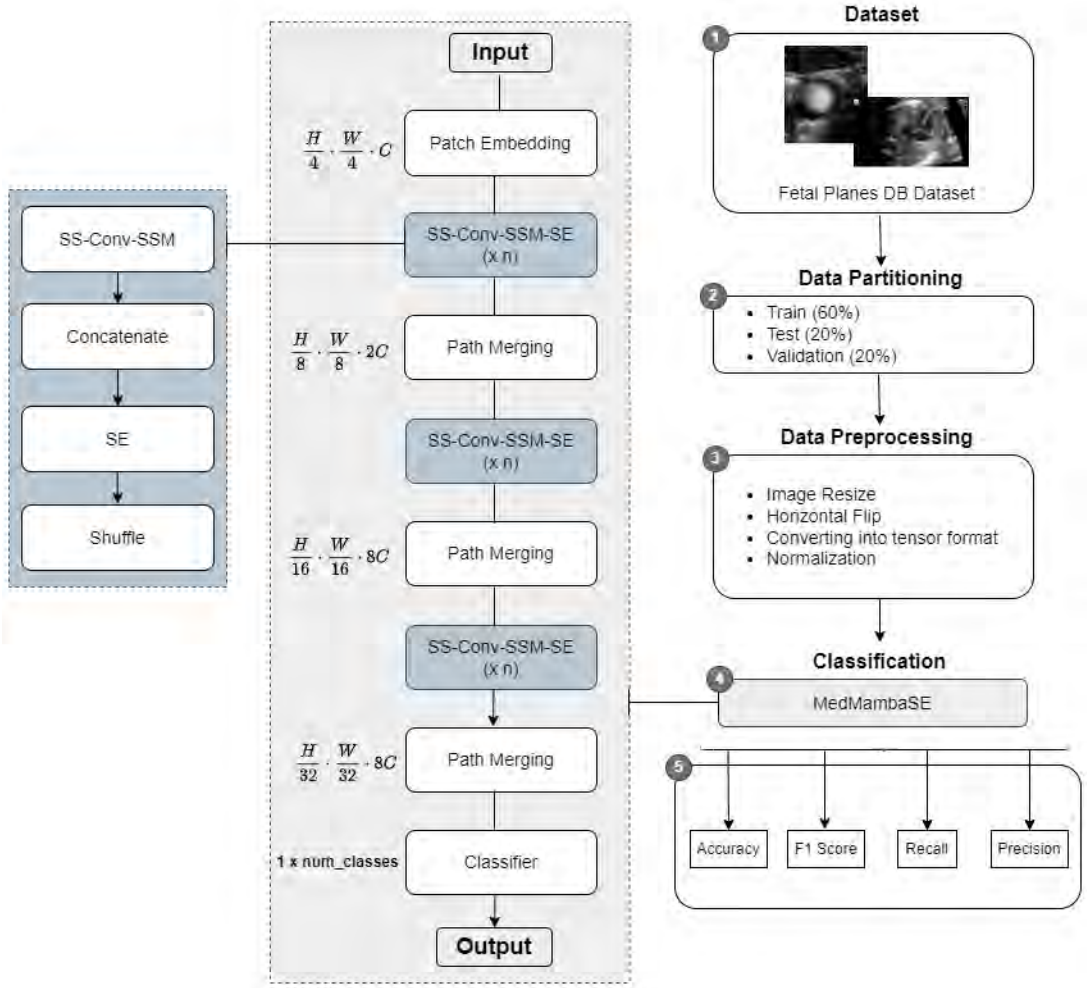

Figure 4.1: Top Level Overview of MedMambaSE

In the proposed MedMambaSE scheme, the Fetal Planes DB dataset is used. Firstly, the dataset has been partitioned into training (60%), testing (20%), and validation (20%) sets. Next, during the data preprocessing step, images resize, horizontal flip, conversion into tensor format and normalization have been used. Therefore, the classification step employs the MedMambaSE model. Various metrics such as accuracy, F1 score, recall, and precision have been used to evaluate the model's performance.

4.1 Dataset Patitioning

For our proposed model (MedMambaSE), we have divided the data into a 60:20:20 ratio and distributed them into separate train, test, and validation folders. Each of these folders contain six subfolders corresponding to our six classes, which hold the relevant images.

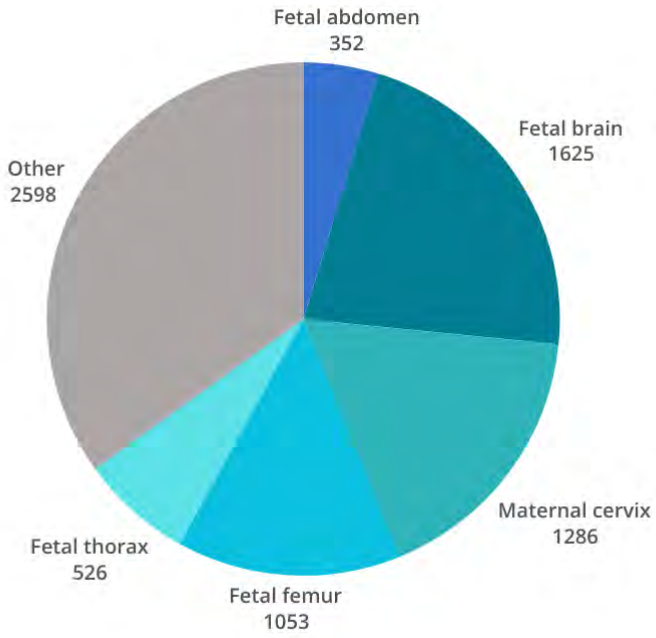

Figure 4.2: Train Data Split in MedMambaSE

4.1.1 Data Preprocessing

As part of our data preparation process, we have employed a series of image transformations so that our proposed model receives input that is both standardized and augmented. These transformations firstly include resizing all the images to $128 \times 128 \times 3$. Additionally, we have applied a random horizontal flip to the training images. Moreover, the images were then converted into tensor format. This is important because it allows us to work with multidimensional arrays in PyTorch-based neural networks where computations involve multi-dimensional arrays required by

neural networks for processing different layers. Subsequently, we have normalized the pixel values of these tensor images to have a mean and standard deviation of 0.5 across all color channels. This normalization serves the purpose of scaling the pixel values to a standard range.

4.2 Architectural Overview

4.2.1 Background of MedMambaSE

Modern SSM-based models like the Structured State Space Sequence Model (S4) and Mamba involve using classic systems. They take simple one-dimensional input functions or sequences called $w(t)$, and pass them through intermediate states $u(t)$ to produce an output, $v(t)$ [40]. These steps can be described using simple linear ODE (Ordinary Differential Equations):

$$\text{State Equation:} \qquad u'(t) = Xu(t) + Yw(t) \tag{4.1}$$

$$\text{Observation Equation:} \qquad v(t) = Zu(t) \tag{4.2}$$

Here, the state equation describes how the internal state of the system evolves with time. The observation equation relates that internal system of the state to the observations that are made. To simply put, X, Y and Z are learning parameters that can change. The $u(t)$ is the implicit state and $w(t)$ is our input.

To adapt their continuous systems for deep learning, S4 and Mamba models add a time parameter Δ and change parameters X and Y to $\overline{X}$ and $\overline{Y}$ respectively using a certain technique which commonly employs zero-order hold (ZOH) method [40]. It can be defined through the following equation:

$$\overline{X} = \exp(\Delta X) \tag{4.3}$$

$$\overline{Y} = (\Delta X)^{-1}(\exp(\Delta X) - I) \cdot \Delta Y \tag{4.4}$$

After these modifications, the SSM-based models can work in two main modes – stepwise linear mode or full convolutional mode described by equations:

$$u'(t) = \overline{X}u(t) + \overline{Y}w(t) \tag{4.5}$$

$$v(t) = Zu(t) \tag{4.6}$$

$$K = (Z\overline{Y}, C\overline{XY}, \ldots, Z\overline{X}^{L-1}\overline{Y}) \tag{4.7}$$

$$v = w * \overline{K} \tag{4.8}$$

$\overline{K}$ is a structured convolutional kernel, and L is the length of the input sequence x.

2D-selective-scan (SS2D)

2D-selective scan (SS2D) is the core part of the SS-Conv-SSM Block. SS2D is made up of three components: scan expansion, S6 block, and scan merging:

Firstly, the scan expansion expands input image in four primary directions: top, bottom, left, and right. It enables the network to capture rich pattern information under various possible directions, so no significant information is missed due to the input image's initial orientation.

Afterward, the S6 block processes the extended sequences. The S6 block is a more refined version of Mamba's S4 structured in a way that it is capable of extracting every part of the information from the input sequences. It does this work by scanning the stretched sequences. It adopts an intelligent mechanism through which it tunes the parameters of the known Structured State Space Model in tune with the input data of the SSM. This seems to keep all the essential information gracefully by removing all the irrelevant pieces of information.

Lastly, the scan merging process merges all the processed sequences by the S block from every direction. This is an essential process in returning the output to the original size, thus allowing the information gathered from the various directions to be effectively combined. The information has been collected through each directional scan merged; in this way, the model has a whole picture of what the input data is about.

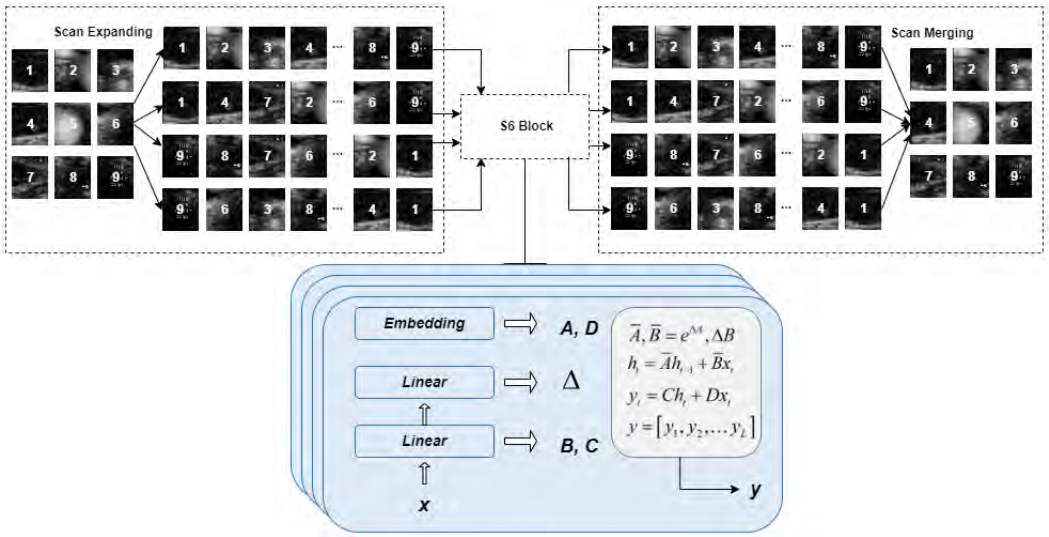

Figure 4.3: Visualization of the 2D-Selective-Scan (SS2D) process

SS-Conv-SSM-SE Block

Core Module of MedMambaSE is SS-Conv-SSM-SE Block. This unit uses a dual branch system without any complicated methods. To begin with, it splits its input into two equal-sized sub-inputs through a channel splitting operation. Then each of these sub-inputs is fed into separate branches; namely Conv-branch and SSM-branch.

In the Conv-branch, simple convolutional layers are employed for processing local features in the input. The convolutional branch specifically uses Batch Normalization (BN) and the ReLU activation function for improved performance. On the

other hand, SSM-branch starts by layer normalizing the input where it should be noted that after normalization, the input is split into two parts. The first part of SSM involves passing linearly through SiLU activation function while on the second part, after another linear layer, depthwise separable convolution and SiLU activation function have been applied to it. Then, the 2D Selective Scan module (SS2D) is used for better feature extraction. These processes are followed by normalizing features using Layer Normalization which are element-wise combined with outputs from the first part thereby merging two streams, and lastly, a linear layer blends these features together to give the final output from the SSM-branch.

The outputs of the Conv-branch and the SSM-branch are then concatenated. From this global concatenation, the features are submitted to the Squeeze-and-Excitation block for dynamic channel-wise recalibration of the concatenated features. This is accomplished through a squeeze operation, which employs the global pooling of the features into one channel descriptor, followed by excitation, which scales the original features. This ensures that important features are highlighted, while those that are less relevant are suppressed.

These recalibrated features are forwarded to a Shuffle block. The block reshapes the channels further to mix the feature representations, whereby the model learns intricate patterns of information by making mixed and rich interactions from different feature maps. This completes the process of obtaining a more robust and generalized feature representation. The output from the Shuffle block is finally passed forward to other layers for further processing or final prediction.

4.2.2 Detailed Structure

The proposed model aims to refine the feature recalibration capabilities of the network, thereby improving the sensitivity and accuracy of the model in detecting abnormalities in ultrasound images. The SE block dynamically scales the feature channels to emphasize important features while suppressing less useful ones.

This architecture strategically positions the SE block to process the output from the Convolutional Self-Attention module before further passing it to the remaining layers. This is specifically implemented in the `ConvSSM` class, which is a convolutional block with designed self-attention mechanisms, specialized to deal with ultrasound image analysis. Below is a more elaborated description of how the SE block was embedded and works in this architecture.

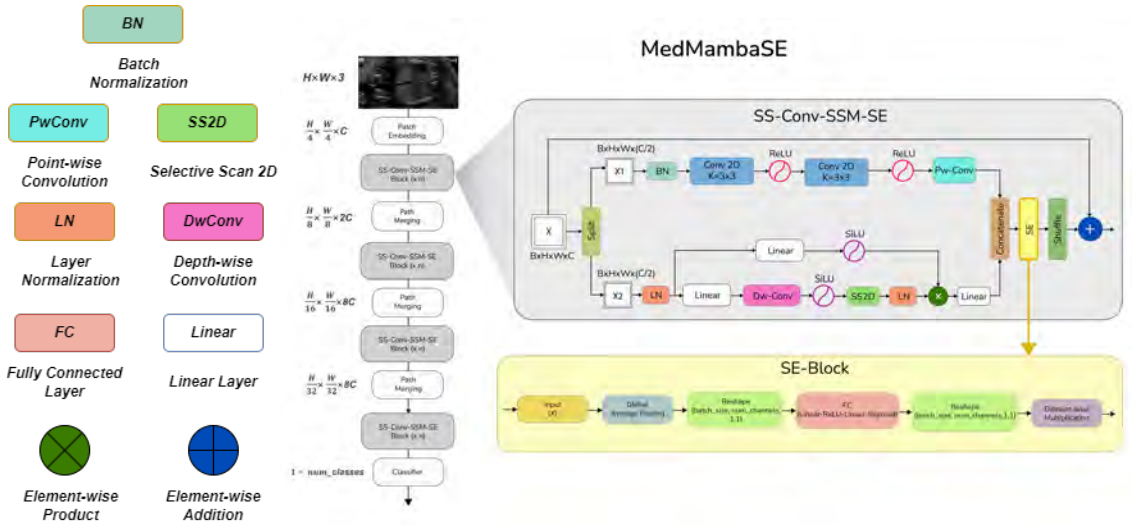

Figure 4.4: Overall Architecture of MedMambaSE

Input Splitting

The input feature map is divided into two halves. The first one is processed directly using several convolutional operations (three 3x3 convolutions followed by a 1x1 convolution) that produce many intermediate features. The other half is made to go through a self-attention mechanism.

Self-Attention Processing

The self-attention mechanism, now the `SS2D` class, operates over the second part of the input. This part of the model captures long dependencies of the feature map; it portrays the general surrounding context within which the image features fall.

Feature Concatenation

Further down the two streams, the processed features from both the convolutional path and the self-attention path are concatenated. This concatenation allows the model to integrate the local feature interactions learned by the convolutions with the global contextual understanding of the self-attention mechanism.

Squeeze-and-Excitation

The concatenated features are then passed through the SE block. The features are then passed through the squeezing operation, where the global spatial information gets compressed to a channel descriptor through employing global average pooling. The descriptor thus obtained is then passed through two fully connected layers, called the excitation operation, where it learns to recalibrate the feature responses in a channel-wise fashion by emphasizing informative features and suppressing less useful ones.

Feature Rescaling

The output of the SE block is a rescaled collection of channel weights, which are then used to rescale the original concatenated features. Rescaling refines the feature representation to make the network more sensitive to the important features, yet less sensitive to the noises and other less important information.

Output Projection

The recalibrated features are then finally projected back to the desired channel dimensionality for further processing or for the purposes of making the final predictions.
In comparison to the standard MedMamba model, the inclusion of the Squeeze-and-Excitation block into the MedMambaSE model has the following advantages:

- **Improved Feature Representation:** Channel-wise recalibration of the feature maps enables MedMambaSE to focus more on the important features critical to classification, making it more robust and reliable in the inference, while it focuses less on irrelevant or misleading features.
- **Enhanced Model Sensitivity and Specificity:** Dynamic recalibration will make the model more sensitive to subtle abnormalities in ultrasound images, which are often paramount for early diagnosis. This sensitivity, combined with the ability to suppress less relevant features, also enhances the specificity of the model.
- **Adaption to Different Imaging Conditions:** MedMambaSE can be adapted to large variations in imaging conditions and noise levels, made possible by the SE block's dynamic change of importance of features with respect to the specific content of an image.

In general, MedMambaSE is significantly better than the MedMamba architecture since it has a higher potential to address the complexities of medical image analysis, more particularly in the prenatal ultrasound imaging domain. This not only increases classification accuracy but also makes the system more reliable and trustworthy for clinical applications.

Chapter 5

Implementation & Result analysis

For our proposed model (MedMambaSE) we have employed Adam optimizer with learning rate = 0.0001, weight decay = 1e-4, $\beta 1 = 0.9$ and $\beta 2 = 0.999$. We have also used Cross Entropy Loss so that the model parameters are optimized. We have utilized the Pytorch framework. We have trained the dataset for 20 epochs with batch size of 16 due to the longer training duration and resource limitation. Moreover, we have not used any data augmentation techniques and pre trained weights to evaluate the performance from the original model. The training has been conducted on a computer with Windows Operating System and an NVIDIA GeForce RTX 4090 GPU.

Table 5.1: Summary of Experimental Setup for MedMambaSE

Experimental Setup	
Operating System	Windows 11
GPU Accelerators	NVIDIA® GeForce RTX 4090
CPU	13th Gen Intel(R) Core(TM) i9-13900K
DL Framework	PyTorch
Optimizer	Adam (learning rate = 0.0001, weight decay = 1e-4, $\beta_1 = 0.9$, $\beta_2 = 0.999$)
Loss Function	Cross Entropy Loss
Epochs	20
Batch Size	16
Image Size	128
Additional Techniques	No data augmentation No pre-trained weights used

5.1 Performance Evaluation Metrics

Here we have used precision, recall and f1 as our evaluation metrics. The metrics false positives (FP), true positives (TP), false negatives (FN), and true negatives (TN) have been used to calculate these measures. TN and TP refer to the number of positive and negative samples correctly classified by the model. On the other hand, FN and FP refer to the number of positive and negative samples that were incorrectly classified.

Precision measures the accuracy of the positive predictions made by the model. It is defined as the ratio of true positive results to the total number of positive results predicted by the model. The formula for precision is:

$$\begin{aligned} \text{Precision} &= \frac{\text{Number of True Positives}}{\text{Total Number of False positives}} \\ &= \frac{TP}{TP + FP} \end{aligned} \tag{5.1}$$

A higher precision score indicates that the model returns more relevant results.

Recall, also known as sensitivity, measures the model's ability to identify all relevant instances within a dataset. It is calculated as the ratio of true positive results to the sum of true positives and false negatives:

$$\begin{aligned} \text{Recall} &= \frac{\text{Number of True Positives}}{\text{Total Number of Predicted positives}} \\ &= \frac{TP}{TP + FN} \end{aligned} \tag{5.2}$$

High recall indicates that the model captures a large proportion of positive samples.

F1 Score is the harmonic mean of precision and recall and serves as a single metric to balance both the precision and recall of a model. It is particularly useful when the classes are imbalanced. The F1 score is calculated as:

$$\begin{aligned} F1 &= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \\ &= \frac{2 \times TP}{TP + FP + FN} \end{aligned} \tag{5.3}$$

An F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

5.2 Overview of Existing Model Results

5.2.1 Efficacy Analysis

Neural Network Models

The implemented techniques are built upon three distinct neural network architectures: SENet, VGG19, and Swin Transformer. Each of these models was chosen for its unique capabilities and performance characteristics. When evaluated individually, SENet emerged as the highest performing model, achieving an accuracy rate of 94.88%. It is closely followed by VGG19, which achieved 90.44% accuracy, while Swin Transformer lagged behind with a 75.36% accuracy rate.

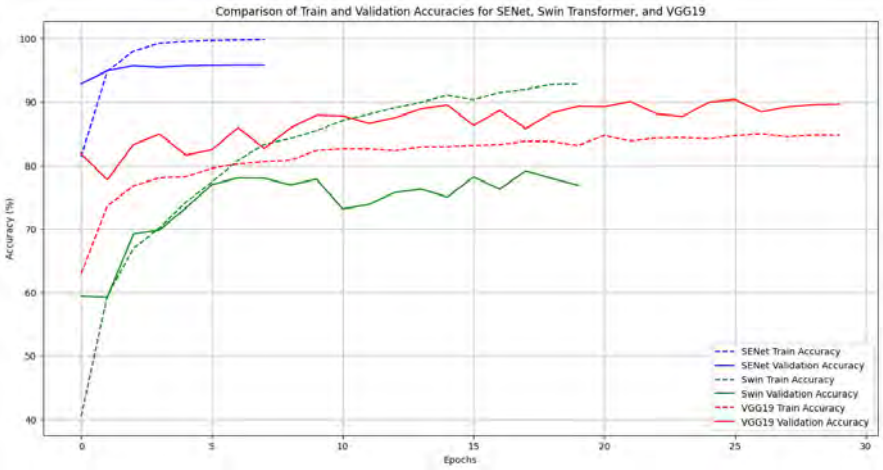

Figure 5.1: Model Accuracy Comparison

SSM-Based Model: MedMamba

Looking at the figure 5.2, we observe a consistent rise in training accuracy over epochs. Initially, the validation accuracy was fluctuating, however over time, it also demonstrated a steady upward trend. The difference between these two accuracies is quite reasonable which shows that there is no overfitting of the model. According to this graph, employing additional epochs could potentially further enhance accuracy.

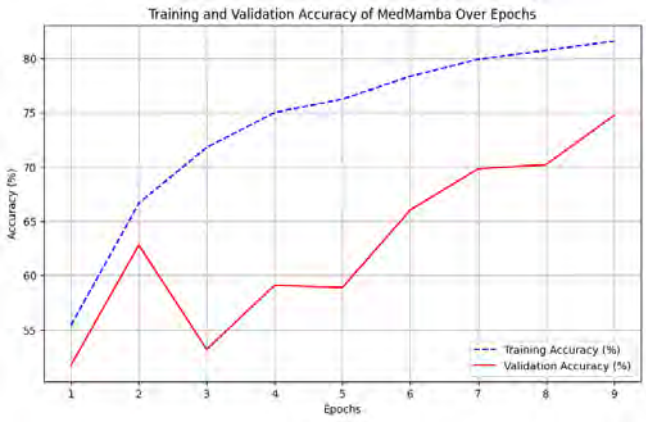

Figure 5.2: MedMamba Training and Validation Accuracy Over Epochs

5.3 Comparative Performance Analysis

In terms of global accuracy, the SENet algorithm outperformed its counterparts. Intriguingly, the algorithm displayed a balanced relationship between its precision, F1-score, and overall accuracy, indicating its adeptness at both identifying and categorizing ultrasound images. Specifically for SENet, the precision stood at 94.92%, with a recall of 94.88% and an F1-Score of 94.88%. These attributes position SENet as a highly suitable option for computational tasks that necessitate elevated levels of sensitivity and specificity.

Contrastingly, VGG19, despite lagging behind SENet in terms of overall accuracy, exhibited a commendable equilibrium across key performance indicators such as precision, recall, and F1-score. For VGG19, the precision was 90.04%, recall was 90.00%, and the F1-Score amounted to 90.01%. This observation suggests that VGG19 may be more apt for contexts that demand a balanced distribution of Type I (False Positive) and Type II (False Negative) errors.

On the other end of the accuracy spectrum, the Swin Transformer reported an overall precision of 76.43%, a recall of 75.08%, and an F1-Score of 75.65%. This finding implies that Swin Transformer may be optimally configured for computational tasks where a correct classification among the top-ranking predictions is deemed sufficient.

SENet and VGG19 have higher performance metrics than MedMamba, which only has 77.99% precision, 79.06% recall and 76.77% F1 Score. However, these performance metrics still outperform Swin Transformer. Moreover, these results are quite satisfactory given that the model was trained for only 10 epochs. This suggests that with further training, MedMamba could potentially achieve even better performance metrics.

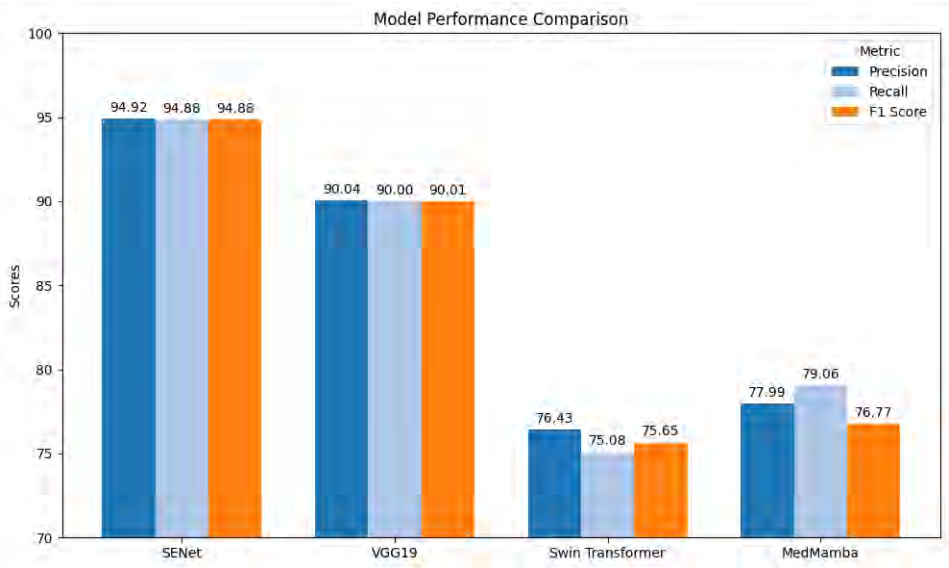

Figure 5.3: Precision, Recall, F1 bar chart comparison of four models

5.4 Overview of our Proposed Model Result: MedMambaSE

The MedMamba model's performance was evaluated over 20 epochs with early stopping. These periods exposed a trend of development in training loss, training accuracy as well as validation accuracy. At first, the training accuracy was 55.39% and the validation accuracy was 51.37%. Nevertheless, it grew to be 90.6% for training and 87.21% for validation by the end of last epoch. Moreover, the training loss significantly showcases a consistent decrease over per epoch which indicates that it is effectively minimizing the error over time.

However, the accuracies have just touched the 90%. This is majorly because the number of epochs used for training is only 20 which is relatively short for most complex models such as MedMambaSE to learn enough features needed for good performance; hence they require more time to train if better results are expected from them in terms of performance improvement during learning stage. If trained further these additional learns will enable this system recognize features correctly thus achieving higher rates.

Figure 5.4: Training Loss, Training and Validation Accuracy of MedMambaSE

Table 5.2: Training, Test and Validation Accuracy of Implemented Models

Model	**Training Accuracy (%)**	**Test Accuracy (%)**	**Validation Accuracy (%)**
VGG19	84.77	90.44	89.67
SENet	97.81	94.96	95.46
Swin Transformer	88.61	73.95	75.5
MedMamba	87.56	81.2	83.49
MedMambaSE	90.6	84.23	87.21

5.4.1 Assessment of Classification Performance and Predictive Accuracy of MedMambaSE

The confusion matrix in Figure 5.5 provides valuable insights into classification across six classes. For instance, the model showcases higher performance in classifying certain classes such as "Fetal Brain" and "Maternal Cervix". These classes show a high number of true positives which indicates that the model can effectively recognize and classify these images. This high performance is likely due to the good number of images in the training dataset which allowed the model to learn detailed and distinctive features necessary for the prediction.

However, classes that have fewer images such as "Fetal Abdomen" and "Fetal Femur", has a higher rate of false negatives. This issue seems to occur due to the imbalanced distribution of images. Classes with fewer training data don't provide the model with enough examples to learn from. Eventually it results in poor generalization capabilities to unseen data. Hence, it is evident in the model's tendency to confuse these less represented classes with "Other" category which is more frequently represented.

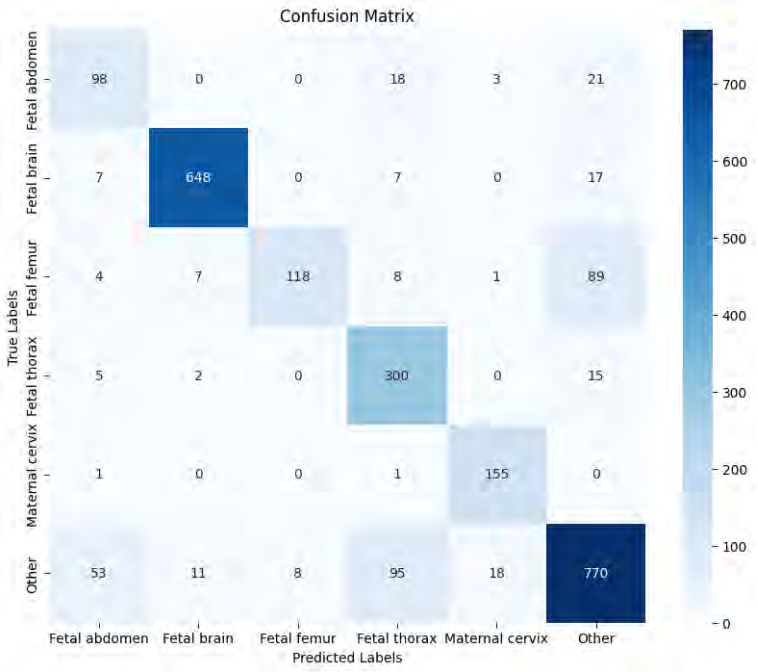

Figure 5.5: Confusion Matrix of MedMambaSE across six classes

Figure 5.6: Correctly Classified Images

Figure 5.7: Incorrectly Classified Images

5.4.2 Weaknesses in Predictive Accuracy

Although MedMambaSE shows a promising result in certain classes, the imbalance in the training data causes a challenge to its overall effectiveness. For instance, from the Figure 5.8 it is seen that, "Fetal Abdomen" has low precision which suggests that it often incorrectly predicts the class. Moreover, the "Fetal Femur" class has a low recall which indicates that the model misses a substantial number of actual fetal femur cases. Moreover, there is moderate score in the "Maternal Cervix". On the contrary, if we look at "Fetal Thorax" and "Fetal Brain", these classes show excellent performance which is likely due to the distinct and well represented features in the dataset. Hence, to address this imbalance we can try several data balancing strategies such as data augmentation, under-sampling, over-sampling or class weighting. However, we wanted to evaluate our proposed model performance on the actual dataset.

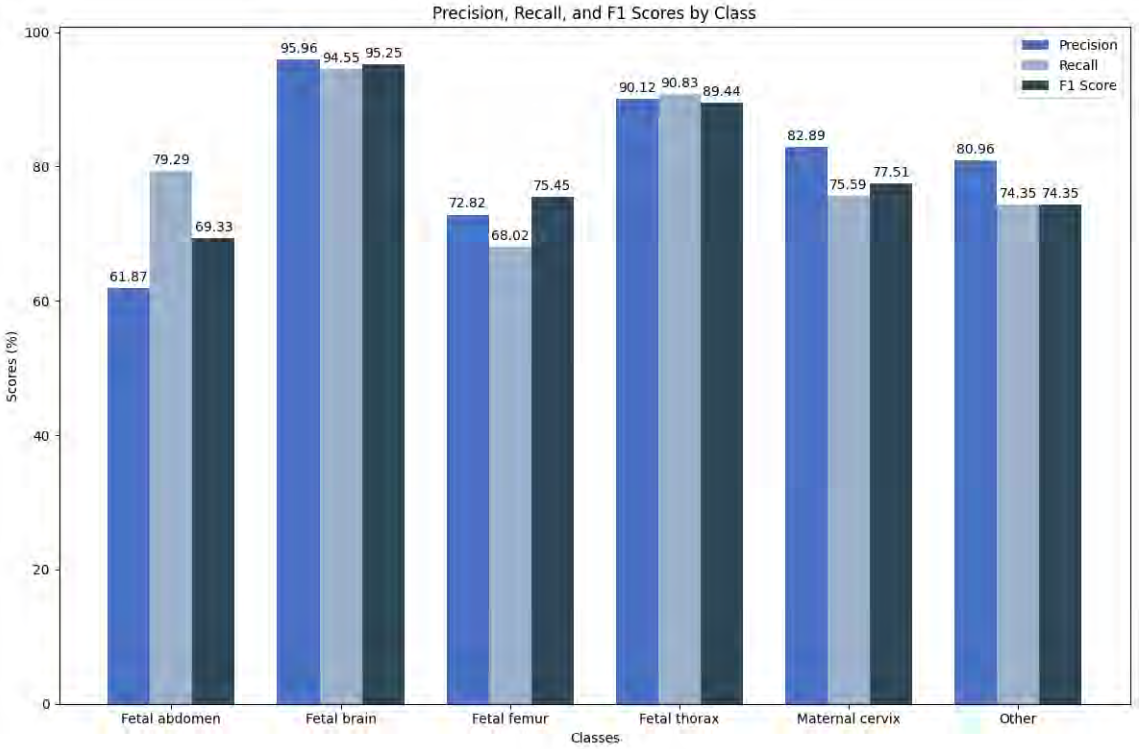

Figure 5.8: Performance Evaluation Metrics of MedMambaSE across six classes

5.4.3 Comparative Analysis of Feature Recalibration

Squeeze-and-Excitation (SE) blocks are shown to be a critical improvement in hybrid models that advance the performance of both convolutional neural networks (CNNs) and self-attention mechanisms for ultrasound image analysis. In order to illustrate the effectiveness of feature recalibration, this section compares the MedMamba model with the MedMambaSE model. To empirically demonstrate the impact of SE blocks, we extracted and visualized the feature maps from both MedMamba and MedMambaSE models 5.9.

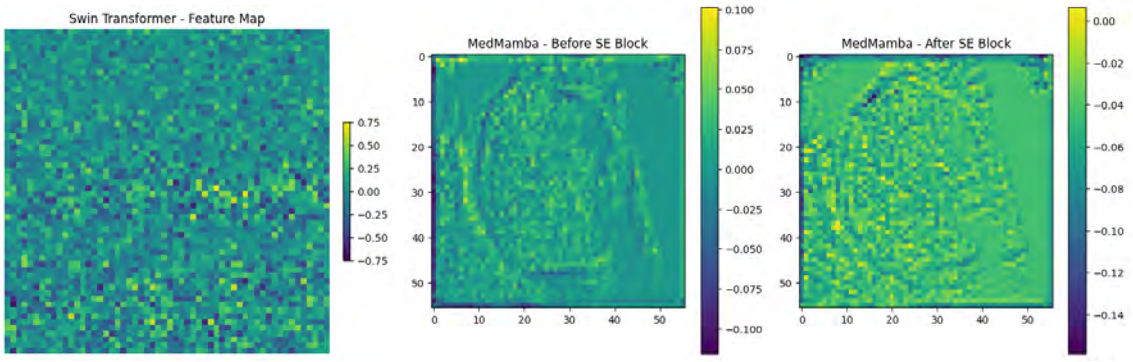

Figure 5.9: Comparison of Feature Maps Before and After SE Block Recalibration in MedMamba and Swin Transformer

MedMamba is a hybrid model incorporating convolutional neural networks (CNNs) and self-attention mechanisms designed for medical image analysis. The MedMamba model has the following observations:

- The feature map visualization for MedMamba, shown in the left panel, exhibits a broad distribution of activation values ranging from -0.10 to 0.10.
- The spread of values indicates that the model treats many features similarly, without distinguishing their relative importance. This uniform treatment can lead to the inclusion of irrelevant or redundant features, potentially hindering model performance.

On the other hand, MedMambaSE is an enhanced version of MedMamba, where SE blocks are integrated to recalibrate feature maps dynamically, improving the model's ability to focus on salient features. The MedMambaSE model has the following observation:

- The feature map visualization for MedMambaSE, depicted in the right panel, reveals a more concentrated range of activation values from -0.14 to 0.00.
- SE blocks recalibrate the feature maps by emphasizing more informative features while suppressing less relevant ones. This recalibration is evident in the refined distribution of feature activations, where crucial patterns are enhanced.

The SE blocks contribute to enhanced feature Representation. By recalibrating the features, the model can focus on more relevant patterns, improving its ability to generalize from the training data to unseen data. Moreover, SE blocks help in

mitigating the influence of noisy or redundant features, thereby enhancing the robustness of the model.

Superior feature recalibration capabilities are shown by the MedMambaSE architecture, which is the product of integrating SE blocks into the MedMamba model. The advantages of this strategy are well shown by the empirical display of feature maps before and after the SE block. The model's enhanced focus and performance are shown by the recalibrated features of MedMambaSE, which more effectively highlight noteworthy patterns.

Looking at the feature map for the Swin Transformer, it is observed that the feature map is almost entirely deep green, with an activation value of between -0.75 and 0.75. This means that the focus spreads over all the features without emphasizing the relevant ones most while underemphasizing those less relevant. This highlights the need for the SE block in that it attends to the most relevant features to ensure critical patterns are put forth while the less relevant information is suppressed.

Due to the higher computational requirements, the MedMambaSE model was only trained for 20 epochs; still, it has promise. It appears from the feature recalibration that the accuracy of the MedMambaSE model may be even greater with longer training. Based on the same dataset, the SENet model performed remarkably well, achieving 97% accuracy, which supports this notion. So, MedMambaSE's potential for better performance in ultrasound image analysis is well supported by the feature recalibration proof.

5.4.4 Training Time Analysis of MedMambaSE

Gradient Accumulation and Training Time

In deep learning, the training process can be computationally intensive, especially when dealing with large datasets and complex models. One technique used to manage memory consumption and improve training efficiency is gradient accumulation. Gradient accumulation works by splitting a batch of training data into several smaller sub-batches (or micro-batches). Instead of updating the model parameters after processing each sub-batch, the gradients are accumulated over multiple sub-batches. After processing the specified number of sub-batches, the accumulated gradients are used to update the model parameters. This process helps in training with larger effective batch sizes without requiring a large amount of memory, as the model parameters are updated less frequently.

While gradient accumulation helps manage memory consumption and allows training with larger batch sizes, it can also significantly increase the overall training time. The primary reasons for this increase in training time include:

- **Increased Number of Forward and Backward Passes:** Since the model processes multiple sub-batches before updating the parameters, the number of forward and backward passes through the network increases. Each pass involves complex computations that contribute to the overall training time.

- **Synchronization Overhead:** After processing each sub-batch, gradients need to be synchronized and accumulated. This synchronization step introduces additional overhead, especially when training on multiple GPUs or distributed systems.
- **Delayed Parameter Updates:** The delay in updating model parameters means that the model takes longer to converge. More iterations are required to achieve the same level of training progress compared to more frequent updates.

In our MedMambaSE model, the use of gradient accumulation is essential due to the high-dimensional nature of the input ultrasound images and the complexity of the network architecture. By accumulating gradients over multiple sub-batches, we can effectively manage the memory requirements. However, this comes at the cost of increased training time, which is a trade-off necessary for achieving efficient and stable training.

```
using cuda:0 device.
Using 8 dataloader workers every process
using 7440 images for training, 2480 images for validation.
  0%|          | 0/465 [00:00<?, ?it/s]Step 0:
  Data transfer: 0.000000 seconds
  Zero grad: 0.000000 seconds
  Forward pass: 0.975357 seconds
  Loss computation: 0.006183 seconds
  Backward pass: 35.738770 seconds
    Backward data handling: 35.738770 seconds
      Grad accumulation: 35.732317 seconds
      Parameter update: 0.006453 seconds
    Backward calculation: 0.000000 seconds
    Backward communication: 0.000000 seconds
  Optimizer step: 0.023875 seconds
  Total step time: 36.747204 seconds
```

Figure 5.10: Impact of Gradient Accumulation on Training Time

5.5 Challenges and Issues

5.5.1 GPU Utilization

Despite having access to a high-performance NVIDIA RTX 4090 GPU, initial tests showed that TensorFlow was not fully utilizing its capabilities. This became a bottleneck, especially for computationally expensive models like MedMamba and MedMambaSE. Tweaking the GPU settings and modifying the code to enable full GPU utilization was a vital step in the project.

5.5.2 Training Duration and Resource Limitations

One of the main challenges we have faced was the extensive training duration per epoch. Completing 20 epochs with early stopping took nearly 5 days. Moreover, lower-spec GPUs were inadequate for the model implementation. Hence, we needed high end GPUs like RTX-4090. As a result, our testing and experimentation were further constrained by limited access to these high-resource GPUs.

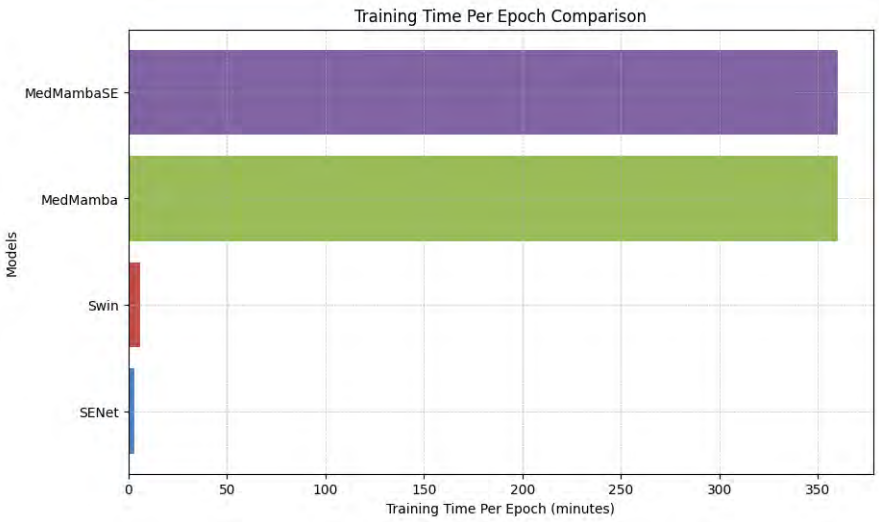

Figure 5.11: Training Time Comparison

5.5.3 Memory Exhaustion

ResourceExhaustedErrors mainly arose when running the model, due to its depth and the large number of parameters. A number of techniques, such as, reducing the batch size and image size, were employed, as a workaround for this.

5.5.4 Module Import Errors

Numerous challenges were encountered during the initial setup phase, especially, concerning module import errors. Several issues surfaced due to missing python packages and incorrect environment settings, which affected data manipulation and model training. The resolution of this issues was pivotal for safeguarding the integrity of the research and verifying that the chosen machine learning libraries functioned as anticipated.

Chapter 6

Conclusion

Development of the MedMambaSE model is a considerable step forward in the applications of deep learning for prenatal ultrasound diagnostics. It is possible to show that integration of the Squeeze-and-Excitation (SE) block into the MedMamba architecture brings massive improvement in the ability of the model for dynamic recalibration, as well as in the emphasis of important features at the cost of pressing down those which are less informative, hence lifting both sensitivity and specificity of ultrasound image classification.

The initial implementation of MedMambaSE has shown encouraging results, with increased accuracy and timely diagnostic results, as is important in prenatal care. Nevertheless, there is always a continuous scope for its betterment and optimization in full utilization of such potential in clinical implementations. This will set the foundation for further development, which will have a quick training duration of the model with advanced computational strategies and the challenge of database imbalance, which remains an issue to provide the robustness and reliability of the model in different clinical settings and patient demographics. In further epochs of training and the testing of the developed model over increased datasets, the performance will be duly tuned to handle diverse real-life scenarios.

These advances will make the MedMambaSE model a much more potent tool in the area of medical imaging, thereby contributing much to furthering prenatal healthcare. This development will be continuous, meaning that the model will not only service the demands of medical diagnostics but also be helpful to medical personnel in the context of providing information that is reliable, accurate, and timely for diagnoses.

In addition, our future research will focus on a few critical improvements to the MedMambaSE model in order to maximize its use in prenatal ultrasound diagnosis. To begin with, the primary goal is to reduce the training period through the employment of modern optimization strategies that have a lower computational overhead without any decrease in performance. Secondly, the imbalance in dataset will be fixed as a major task. In addition, more advanced forms of data augmentation and synthetic data creation will be employed for better accuracy of the model concerning various kinds of data. Therefore, it would involve longer training epochs on bigger datasets including but not limited to other populations and imaging conditions. The

purpose of this research is to improve MedMambaSE's diagnostic capabilities so that it is able to consistently perform well in different clinical situations and significantly contribute to medical imaging technology advancement.

Bibliography

[1] N. Siauve and A. Coulon, "Abdominal imaging," *Journal de Radiologie*, vol. 86, no. 7-8, pp. 833–838, 2005.

[2] X. Liu, P. Annangi, M. Gupta, *et al.*, "Learning-based scan plane identification from fetal head ultrasound images," in *Medical Imaging 2012: Ultrasonic Imaging, Tomography, and Therapy*, SPIE, vol. 8320, 2012, pp. 90–95.

[3] B. Lei, L. Zhuo, S. Chen, S. Li, D. Ni, and T. Wang, "Automatic recognition of fetal standard plane in ultrasound image," in *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, IEEE, 2014, pp. 85–88.

[4] B. Lei, E.-L. Tan, S. Chen, *et al.*, "Automatic recognition of fetal facial standard plane in ultrasound image via fisher vector," *PloS one*, vol. 10, no. 5, e0121838, 2015.

[5] C. Baumgartner, K. Kamnitsas, J. Matthew, *et al.*, "Medical image computing and computer-assisted intervention–miccai 2016," 2016.

[6] C. P. Bridge, C. Ioannou, and J. A. Noble, "Automated annotation and quantitative description of ultrasound videos of the fetal heart," *Medical image analysis*, vol. 36, pp. 147–161, 2017.

[7] S. Fekri-Ershad and F. Tajeripour, "Impulse-noise resistant color-texture classification approach using hybrid color local binary patterns and kullback–leibler divergence," *The Computer Journal*, vol. 60, no. 11, pp. 1633–1648, 2017.

[8] X. Zhu, H.-I. Suk, L. Wang, S.-W. Lee, D. Shen, A. D. N. Initiative, *et al.*, "A novel relational regularization feature selection method for joint regression and classification in ad diagnosis," *Medical image analysis*, vol. 38, pp. 205–214, 2017.

[9] P. Kong, D. Ni, S. Chen, S. Li, T. Wang, and B. Lei, "Automatic and efficient standard plane recognition in fetal ultrasound images via multi-scale dense networks," in *Data Driven Treatment Response Assessment and Preterm, Perinatal, and Paediatric Image Analysis*, A. Melbourne, R. Licandro, M. DiFranco, *et al.*, Eds., Cham: Springer International Publishing, 2018, pp. 160–168, ISBN: 978-3-030-00807-9.

[10] Z. Yu, E.-L. Tan, D. Ni, *et al.*, "A deep convolutional neural network-based framework for automatic fetal facial standard plane recognition," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 3, pp. 874–885, May 2018. DOI: 10.1109/jbhi.2017.2705031. [Online]. Available: https://doi.org/10.1109/jbhi.2017.2705031.

[11] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, *Squeeze-and-excitation networks*, 2019. arXiv: 1709.01507 [cs.CV].

[12] J. Liang, R. Huang, P. Kong, S. Li, T. Wang, and B. Lei, "Sprnet: Automatic fetal standard plane recognition network for ultrasound images," in *Smart Ultrasound Imaging and Perinatal, Preterm and Paediatric Image Analysis: First International Workshop, SUSI 2019, and 4th International Workshop, PIPPI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13 and 17, 2019, Proceedings*, Shenzhen, China: Springer-Verlag, 2019, pp. 38–46, ISBN: 978-3-030-32874-0. DOI: 10.1007/978-3-030-32875-7_5. [Online]. Available: https://doi.org/10.1007/978-3-030-32875-7_5.

[13] P. Sridar, A. Kumar, A. Quinton, R. Nanan, J. Kim, and R. Krishnakumar, "Decision fusion-based fetal ultrasound image plane classification using convolutional neural networks," *Ultrasound in Medicine &amp Biology*, vol. 45, no. 5, pp. 1259–1273, May 2019. DOI: 10.1016/j.ultrasmedbio.2018.11.016. [Online]. Available: https://doi.org/10.1016/j.ultrasmedbio.2018.11.016.

[14] X. P. Burgos-Artizzu, D. Coronado-Gutiérrez, B. Valenzuela-Alcaraz, *et al.*, "Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes," *Scientific Reports*, vol. 10, no. 1, p. 10 200, 2020, ISSN: 2045-2322. DOI: 10.1038/s41598-020-67076-5. [Online]. Available: https://doi.org/10.1038/s41598-020-67076-5.

[15] S. Fekri-Ershad, "Bark texture classification using improved local ternary patterns and multilayer neural network," *Expert Systems with Applications*, vol. 158, p. 113 509, 2020.

[16] Q. Meng, D. Rueckert, and B. Kainz, *Unsupervised cross-domain image classification by distance metric guided feature alignment*, 2020. arXiv: 2008.08433 [cs.LG].

[17] R. Qu, G. Xu, C. Ding, W. Jia, and M. Sun, "Standard plane identification in fetal brain ultrasound scans using a differential convolutional neural network," *IEEE Access*, vol. 8, pp. 83 821–83 830, 2020. DOI: 10.1109/access.2020.2991845. [Online]. Available: https://doi.org/10.1109/access.2020.2991845.

[18] Z. Liu, Y. Lin, Y. Cao, *et al.*, *Swin transformer: Hierarchical vision transformer using shifted windows*, 2021. arXiv: 2103.14030 [cs.CV].

[19] A. Montero, E. Bonet-Carne, and X. P. Burgos-Artizzu, "Generative adversarial networks to improve fetal brain fine-grained plane classification," *Sensors*, vol. 21, no. 23, p. 7975, Nov. 2021. DOI: 10.3390/s21237975. [Online]. Available: https://doi.org/10.3390/s21237975.

[20] X. Yang, Y. Huang, R. Huang, *et al.*, "Searching collaborative agents for multi-plane localization in 3d ultrasound," *Medical Image Analysis*, vol. 72, p. 102 119, Aug. 2021. DOI: 10.1016/j.media.2021.102119. [Online]. Available: https://doi.org/10.1016/j.media.2021.102119.

[21] B. Zhang, H. Liu, H. Luo, and K. Li, "Automatic quality assessment for 2d fetal sonographic standard plane based on multitask learning," *Medicine*, vol. 100, no. 4, e24427, Jan. 2021. DOI: 10.1097/md.0000000000024427. [Online]. Available: https://doi.org/10.1097/md.0000000000024427.

[22] S. Belciug, "Learning deep neural networks' architectures using differential evolution. case study: Medical imaging processing," *Computers in Biology and Medicine*, vol. 146, p. 105 623, Jul. 2022. DOI: 10.1016/j.compbiomed.2022.105623. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2022.105623.

[23] T. B. Krishna and P. Kokil, "Automated detection of common maternal fetal ultrasound planes using deep feature fusion," in *2022 IEEE 19th India Council International Conference (INDICON)*, 2022, pp. 1–5. DOI: 10.1109/INDICON56171.2022.10039879.

[24] M. Micucci and A. Iula, "Recent advances in machine learning applied to ultrasound imaging," *Electronics*, vol. 11, no. 11, 2022, ISSN: 2079-9292. DOI: 10.3390/electronics11111800. [Online]. Available: https://www.mdpi.com/2079-9292/11/11/1800.

[25] H. Ghabri, M. S. Alqahtani, S. B. Othman, *et al.*, "Transfer learning for accurate fetal organ classification from ultrasound images: A potential tool for maternal healthcare providers," May 2023. DOI: 10.21203/rs.3.rs-2856603/v1. [Online]. Available: https://doi.org/10.21203/rs.3.rs-2856603/v1.

[26] A. Gu and T. Dao, *Mamba: Linear-time sequence modeling with selective state spaces*, 2023. arXiv: 2312.00752 [cs.LG].

[27] T. B. Krishna and P. Kokil, "Automated classification of common maternal fetal ultrasound planes using multi-layer perceptron with deep feature integration," *Biomedical Signal Processing and Control*, vol. 86, p. 105 283, 2023, ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2023.105283. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809423007164.

[28] C.-S. Chen, G.-Y. Chen, D. Zhou, D. Jiang, and D.-S. Chen, *Res-vmamba: Fine-grained food category visual classification using selective state space models with deep residual learning*, 2024. arXiv: 2402.15761 [cs.CV].

[29] K. Chen, B. Chen, C. Liu, W. Li, Z. Zou, and Z. Shi, *Rsmamba: Remote sensing image classification with state space model*, 2024. arXiv: 2403.19654 [cs.CV].

[30] Z. Fang, Y. Wang, Z. Wang, J. Zhang, X. Ji, and Y. Zhang, *Mammil: Multiple instance learning for whole slide images with state space models*, 2024. arXiv: 2403.05160 [cs.CV].

[31] H. Gong, L. Kang, Y. Wang, X. Wan, and H. Li, *Nnmamba: 3d biomedical image segmentation, classification and landmark detection with state space model*, 2024. arXiv: 2402.03526 [cs.CV].

[32] L. Huang, Y. Chen, and X. He, *Spectral-spatial mamba for hyperspectral image classification*, 2024. arXiv: 2404.18401 [cs.CV].

[33] S. Li, H. Singh, and A. Grover, *Mamba-nd: Selective state space modeling for multi-dimensional data*, 2024. arXiv: 2402.05892 [cs.CV].

[34] H. Ma, S. Lei, T. Celik, and H.-C. Li, *Fer-yolo-mamba: Facial expression detection and classification based on selective state space*, 2024. arXiv: 2405.01828 [cs.CV].

[35] G. Wang, X. Zhang, Z. Peng, T. Zhang, X. Jia, and L. Jiao, *S^2mamba: A spatial-spectral state space model for hyperspectral image classification*, 2024. arXiv: 2404.18213 [cs.CV].

[36] G. Yang, K. Du, Z. Yang, Y. Du, Y. Zheng, and S. Wang, *Cmvim: Contrastive masked vim autoencoder for 3d multi-modal representation learning for ad classification*, 2024. arXiv: 2403.16520 [cs.CV].

[37] J. X. Yang, J. Zhou, J. Wang, H. Tian, and A. W. C. Liew, *Hsimamba: Hyperpsectral imaging efficient feature learning with bidirectional state space for classification*, 2024. arXiv: 2404.00272 [cs.CV].

[38] S. Yang, Y. Wang, and H. Chen, *Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology*, 2024. arXiv: 2403.06800 [cs.CV].

[39] J. Yao, D. Hong, C. Li, and J. Chanussot, *Spectralmamba: Efficient mamba for hyperspectral image classification*, 2024. arXiv: 2404.08489 [cs.CV].

[40] Y. Yue and Z. Li, *Medmamba: Vision mamba for medical image classification*, 2024. arXiv: 2403.03849 [eess.IV].

[41] *An image classification deep-learning algorithm for shrapnel detection from ultrasound images — Scientific Reports — rdcu.be*, https://rdcu.be/dmewE, [Accessed 15-09-2023].