Received: April 22, 2024. Revised: June 6, 2024.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

MSAG_ENET Based Medical Image Augmentation and Classification of 2D-US Fetal Brain Anomalies

D. Vetriselvi[1] R. Thenmozhi[1]*

[1]*Department of Computing Technologies, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chennai, India*
* Corresponding author's Email: thenmozr@srmist.edu.in

Abstract: Medical imaging is essential to modern life support. Different approaches and standards exist for detecting life-threatening disorders. However, manually detecting anomalies takes time and is prone to error. Additionally, it demands great expertise and competence. Deep learning in medical imaging improves accuracy and provides several benefits. Privacy concerns limit access to medical data, especially images. This problem can be solved with data augmentation. Image quality determines prediction accuracy. Preprocessing images improves quality. This article addresses data augmentation and image enhancement with a multiscale self-attention generator. Foundational network is Inception. Elastic_net used for selection and regularisation of features. The anisotropic diffusion filter used for reducing speckle noise in the model. The ZONODO foetal planes dataset is used as input. The proposed model produces better results for classification as accuracy is 98.7, precision is 97.5, recall is 97.8 and F1-Score is 97.5 when compared with existing methods.

Keywords: Image augmentation, Multiscale self-attention generator, Convolutional neural network, GoogleLeNet, Elastic_net, Anisotropic diffusion filter.

1. Introduction

Prenatal Neurosonography, sometimes called foetal neuroimaging, is a medical imaging procedure used to monitor the baby's brain and central nervous system (CNS) throughout pregnancy. It diagnoses problems in the foetus' neurological development using sophisticated prenatal ultrasonography. Once the foetus' brain and CNS are fully developed in the second and third trimesters, Neurosonography is done regularly. Primary goals of prenatal Neurosonography are detection of abnormalities. Neurosonography can detect brain and CNS disorders in development. Agenesis of the corpus callosum, neural tube anomalies, brain cysts, ventricular abnormalities, and other neurological disorders are examples. Medical imaging lets doctors track the unborn brain's progress during pregnancy. Disparities from expected developmental milestones might be assessed. Prenatal Neurosonography can help doctors diagnose neurological problems, choose the best treatment, and create foetal and postpartum care plans. Prenatal Neurosonography uses sound waves instead of ionising radiation, making it safe and non-invasive. Professional sonographers or maternal-fetal medicine specialists execute the procedure in prenatal ultrasound centres. Preprocessing prenatal Neurosonography images is essential for assessing and interpreting ultrasound results. Pre-processing improves image quality, reduces noise, and standardises data for analysis, visualisation, and diagnosis. Ultrasound image preprocessing encompasses a range of approaches designed to improve the quality, precision, and comprehensibility of ultrasound pictures.

Image augmentation is a widely employed technique in computer vision and deep learning that artificially enhances the variety of a training dataset. The procedure entails implementing diverse alterations to the initial images, resulting in novel iterations that are somewhat modified. The augmented dataset is subsequently employed to train

machine learning models, hence improving their capacity to effectively generalise to unfamiliar data. Image augmentation is especially beneficial when the underlying dataset is constrained or lacks variety.

The structure of this article is as follows. In Section 2, we present substantial works that are very pertinent to our contribution. Section 3 provides a detailed description of the preferred approach or procedure. Section 4 offers a thorough explanation of the database and experimental setup, followed by a distinct presentation and discussion of the results. The fifth and last component of this paper functions as a conclusion and underscores the necessity for future research.

2. Literature survey

The widespread use of Convolutional Neural Networks (CNNs) has demonstrated their effectiveness in addressing large-scale picture classification problems [1]. Additionally, CNN models have shown promise in tackling Fine-Grained Visual Categorization (FGVC) tasks by leveraging their ability to identify subtle local features.

Fu et al. [2] acquaint with WS-DAN, a deep network designed to handle fine-grained visual categorization (FGVC) using weak supervision. The network utilises attention mechanisms to enhance local features and direct the augmentation process. FGVC was a prevalent issue in medical imaging because to the spatial resemblance between infections. Qin et al. [3] introduced a detailed categorization Convolutional Neural Network (CNN) to classify various forms of lung cancer in PET and CT images.

In visual tasks, attention typically refers to a scalar matrix that represents the relative importance and internal relevance of local features [4]. This nonuniform representation was generated by specially developed modules [5]. Studies have indicated that focusing attention on CNNs designed for categorization can offer a straightforward method for localising target objects, aiding in the identification of visual characteristics through local representation. Gondal et al. [6] demonstrated a strategy that utilises attention mechanism to improve the localization and detection of Diabetic Retinopathy (DR). Zhang et al. [7] improved the focus of a deep model by training self-attention blocks for skin lesion categorization and outperformed the baseline models.

Zhun Zhong et al. [8] used Random erasing in data augmentation during training, where random sections of a picture are erased or masked away. This process replicates occlusions or areas of the image

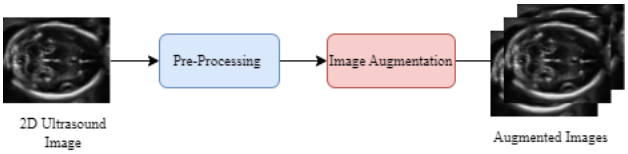

Figure. 1 Structure of Proposed Model

that are missing, so compelling the model to acquire more resilient characteristics. Humza Naveed et al. [9] utilises a convex combination of images and labels to prompt the model to acquire more universally applicable characteristics by exploring the areas of the input space that lie between different classes. Sangdoo Yun et al. [10] entails the extraction and relocation of certain areas from images, together with their related labels, to generate novel training instances. Ekin D.Cubak et al [11] utilises reinforcement learning to automate the selection of augmentation policies. Auto Augment automates the process of finding optimal augmentation policies that enhance the model's performance on a validation set, eliminating the need for manual design. Ryo Takaharhi et al. [12] introduced various types of transformations are randomly performed to the input images during the training process. Xu Zhang et al. [13] used the style of one image, commonly known as the style image, is often applied to another image, known as the content image. All the conventional methods may lead to missing of fine details which tends to misclassification.

3. Proposed methodology

The proposed model is based on the deep learning technology called convolutional neural network. There are several variations in the convolutional neural network. Among them the multiscale self-attention generator is used for the purpose of image augmentation. The multi-scale self-attention generator can be used in several scale. And various attention mechanisms like spatial attention and channel attention can be used.

The overall flow of proposed model shown in fig. 1. The proposed model includes two phases. They are image pre-processing and image augmentation. Attention mechanisms enhance the capability of models to concentrate on segments of input data, enabling them to selectively analyse specific regions or characteristics when making predictions. Within the framework of Convolutional Neural Networks (CNNs), attention mechanisms are frequently utilised to augment the model's capacity to extract pertinent information from various spatial positions within an image. In ultrasound scan images, the speckle noises are prominently present. To reduce the speckle noise the anisotropic diffusion filter is used.

3.1 Image pre-Processing

Image preprocessing is an essential and vital stage in the process of constructing computer vision models. Image processing encompasses a range of procedures aimed at improving image quality, extracting significant characteristics, and preparing data for training purposes. Prior to applying augmentation to ultrasound pictures, it is crucial to carry out preprocessing to improve image quality, eliminate artefacts, and prepare the data for efficient augmentation. In the proposed model we have adopted Z-Score Normalization which is followed with Contrast Limited Adaptive Histogram Equalization. Then to mitigate speckle noise while maintaining crucial structures Speckle Reducing Anisotropic Diffusion was incorporated.

3.1.1. Z-Score normalization

It involves subtracting the mean and dividing by the standard deviation. The outcome of this operation yields a distribution characterised by a mean value of 0 and a standard deviation value of 1. Z-score normalisation is a commonly employed technique in a variety of machine learning and statistical applications. Normalise pixel values by subtracting the average and dividing by the standard deviation [14]. The equation for z-score normalisation is provided by

$$Z - Score = \frac{Original\ Value - Mean\ of\ Feature(\mu)}{Standard\ Deviation\ of\ \ Feature(\sigma)} \quad (1)$$

Step 1: Determine the mean(μ) and the standard deviation(σ) of the feature.

Step 2: Utilise the $Z - Score$ calculation as per the equation (1). Normalisation refers to the process of organising data in a database to eliminate redundancy and improve efficiency. Calculate the value of Z for each given value of X in the feature using the specified formula.

3.1.2. Contrast limited adaptive histogram equalization

CLAHE is utilised to augment the regional disparity of an image. The steps followed in CLAHE is shown in Fig.2 below.

Step 1: CLAHE performs operations on specific parts of the image instead of the entire image. The process involves partitioning the image into distinct and non-overlapping tiles or sub-images.

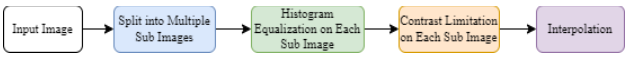

Figure. 2 Steps in CLAHE

Step 2: Histogram Equalisation is individually performed to the histogram of each tile. This guarantees that the intensity distribution within each specific zone is amplified.

Step 3: Bad contrast regions are inhibited to prevent noise amplification. Restricting the cumulative histogram to a threshold does this.

Step 4: Histogram equalisation can cause discontinuities at tile boundaries. Interpolation blends these borders, creating a smooth transition between tiles.

$$T(x,y) = \frac{(L-1)}{\max(C,H(x,y))} \sum_{i=0}^{I(x,y)} h(i) \quad (2)$$

Where, $T(x,y)$ in equation (2) is the transformed pixel value at the coordinates (x,y). L denotes the total number of intensity levels, which is typically 256 for 8-bit pictures. C is the contrast limit parameter. $H(x,y)$ is the cumulative histogram of the pixel intensities in the tile located at (x,y). $I(x,y)$ denotes the pixel intensity at the specific location (x,y), while $h(i)$represents the original histogram of the pixel intensities [15].

3.1.3. Speckle reduction with multiscale attention based anisotropic diffusion filter

Anisotropic diffusion is a method employed in image processing and computer vision to decrease image noise while preserving edges and other important aspects of the image [16][17]. Anisotropic diffusion involves iteratively smoothing an image using a diffusion technique that can vary across the image while preserving the underlying structures and edges.

$$\frac{\partial I}{\partial t} = \nabla.(c(x,y,t).\nabla I \quad (3)$$

The image is represented by the variable I in equation (3) and t represents time. The symbol ∇ symbolises the gradient operator. $c(x,y,t)$ is the diffusion coefficient at position x,y and time t, that regulates the diffusion rate at each location in the image. This coefficient is often determined by the local image gradient.

$$\frac{\partial I}{\partial t} = div(x,y,t)\nabla I = \nabla c.\nabla I = c(x,y,t)\Delta I \quad (4)$$

In the above equation (4), ∇ denotes the Laplacian. div represents the divergence function. For t >0, the output image id $I(t)$) where larger t produces blurred images. The diffusion coefficient can be

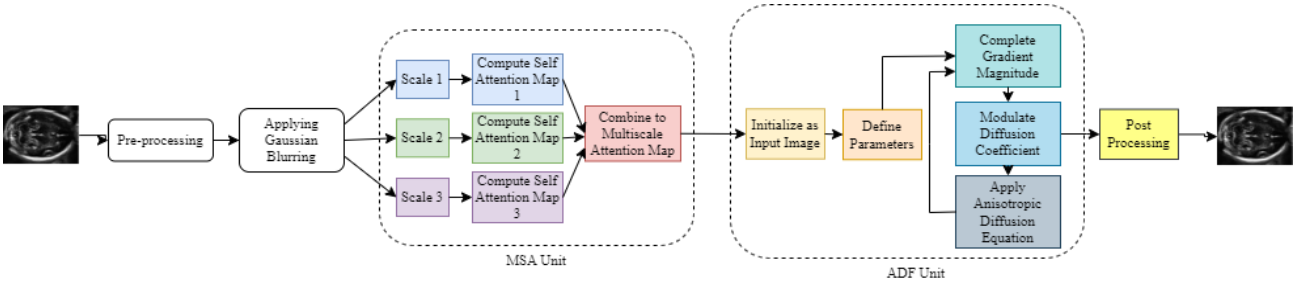

Figure. 3 MSA-ADF: Multi scale Self Attention based Anisotropic Diffusion Filte

$$c\left(||\nabla I||\right) = e^{-\left(\frac{||\nabla I||}{k}\right)^2} \quad (5)$$

$$c\left(||\nabla I||\right) = \frac{1}{1+\left(\frac{||\nabla I||}{k}\right)^2} \quad (6)$$

In equation (5) and (6), k is a constant which controls the sensitivity to the edges. The value of k is chosen as a function of the noise of the image. The anisotropic diffusion filter can be improved by using a multiscale self-attention mechanism. The processes allow the model to focus on various parts of the input at different scales, resulting in more resilient and contextually detailed representations. Combining multiscale self-attention with anisotropic diffusion filtering can provide numerous advantages. It offers adaptive contextual smoothing, enhanced edge preservation, improves robustness and in overall improves the efficiency. The steps involved in Multi scale Self Attention based Anisotropic Diffusion Filter for reducing the speckle noise is shown in Figure 3.

The first step is preprocessing, in which normalization is done for ensuring that the pixel values fall within a suitable range ([0,1]). The pre-processed image is then entered into the Multiscale Self Attention Unit (MSAU) which is responsible for generating multiscale attention map. To create multiscale representation of a given input image, multiple scales are defined. This is done by applying Gaussian blurring at different levels. Then by using transformer architecture at each scale, self-attention maps are computed. The attention maps from different scales are then combined to form the multiscale attention map. To calculate self-attention at a specific scale s and position i,

$$Atten(Qi, K, V) = softmax\left(\frac{Q_i K^T}{\sqrt{d_k}}\right).v \quad (7)$$

In equation (7), Qi is the query vector at position i, K is the matrix containing all key vectors in the sequence, V is the matrix containing all value vectors in the sequence, and d_k is the dimensionality of the key vectors. In the Anisotropic Diffusion Filter Unit (ADFU), for the current input image the gradient magnitude is computed. The generated multiscale attention map is used to modulate the diffusion coefficient function which is smoothing out of noise by focusing on relevant features. Then the anisotropic diffusion equation is applied to update the output image. It considers the modulated diffusion coefficient and gradient magnitude. Then in the post processing contrast enhancement is done.

3.2 Image augmentation

Image augmentation with an attention generator entail integrating attention mechanisms into the augmentation process, enabling selective application of attention to certain regions of an image. Utilised an attention generator to generate attention masks that specifically emphasise parts or characteristics of the image. The attention masks can be constructed according to certain criteria or traits that we wish to highlight during augmentation. Performed element-wise multiplication between the attention masks and the original images. This procedure assigns greater significance to the locations that are emphasised by the attention masks during the succeeding augmentation changes.

Applied attention-based image augmentation techniques, such as attention mix-up, random erasing with attention and attention drop-out to the transformed images with attention applied. The attention-guided alterations guarantee the preservation or transformation of certain regions in a precise manner. Combined the enhanced photos with the original images to generate a varied dataset that preserves the significant characteristics emphasised by the attention masks. Iterated through each image in the training dataset and consistently apply attention-guided augmentation.

The overall structure of the proposed Multiscale Self Attention generator is shown in Fig. 4. The pre-processed image is sent to the convolutional feature extractor for generating feature map.

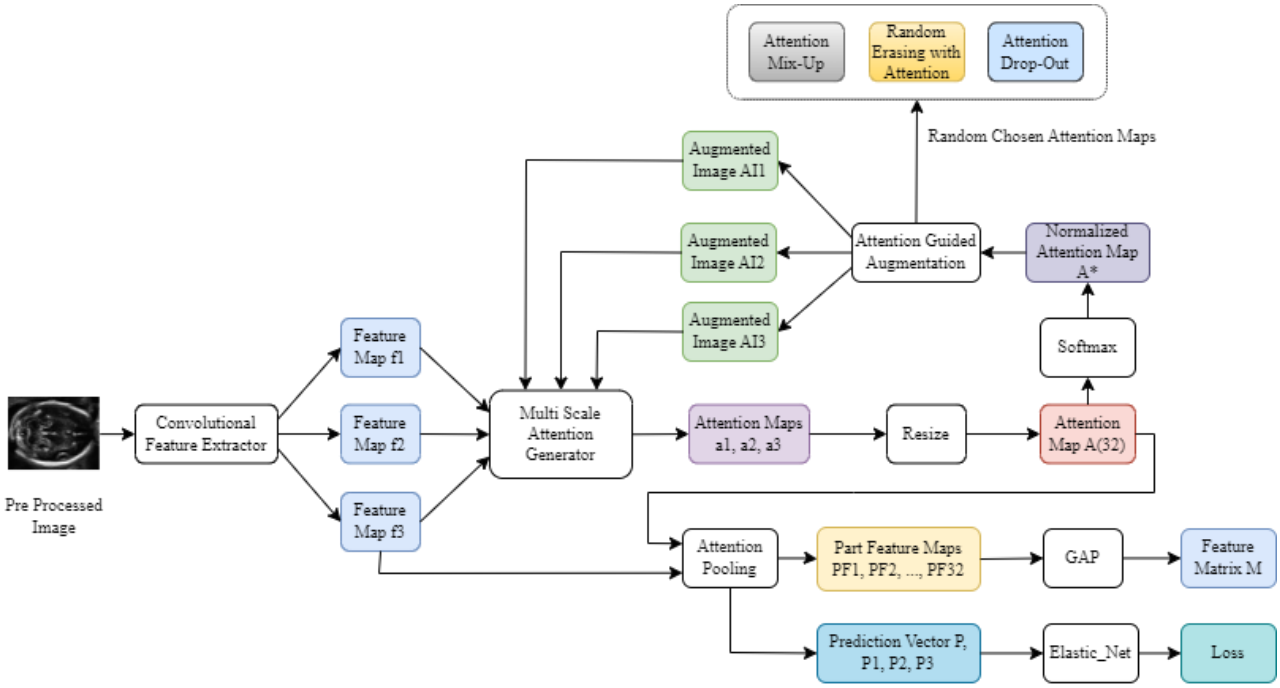

Figure. 4 Image Augmentation with Multi Scale Self Attention Generator

3.2.1. Convolutional feature extractor

In our proposed model we have used the Inception that is GoogleLeNet is used as back bone network for extracting features and generating the feature maps in different sizes in the local inception blocks. The feature maps f1, f2 and f3 are selected with size 512*28*28, 1024*14*14 and 2048*7*7 respectively.

3.2.2. Multi-Scale self -attention generator

The self-attention mechanism enables every element in the input sequence to consider all other elements while producing its output. It allocates distinct attention weights to various elements, depending on their pertinence to the present element. To gather information at various levels, the model integrates numerous self-attention processes that function at different resolutions or levels of abstraction. This may entail analysing the input sequence at different degrees of detail or employing varying window widths for attention. In Multi-Scale Self -Attention generator, based on the feature maps attention maps a1, a2, a3 are generated. Feature map f1 is convolved with 1*1 convolution operation to generate a1 then it is down sampled to 7*7. Similarly feature maps f2 and f3 are convolved with 1*1

convolution operation and down sampled. Totally 32 attention maps are generated in this way with dimension 1*7*7.

Multi-scale self-attention mechanism employing multi-head attention, a widely utilised component in transformer-based designs. The input features will be represented as X, and we will assume that we have N attention heads. Initially, input features are transformed into query (Q), key (K), and value (V) spaces by employing trainable weight matrices W_Q, W_K, and W_V, respectively as given in (8), (9) and (10).

$$Q = XW_Q \tag{8}$$

$$K = XW_K \tag{9}$$

$$V = XW_V \tag{10}$$

To compute the attention scores (α_i) for each attention head i (where i ranges from 1 to N), query and key matrices were utilized.

$$\alpha_i = softmax\left(\frac{QW_{Qi}^T(KW_{Ki})^T}{\sqrt{d_k}}\right) \tag{11}$$

The expression (11) represents the softmax function applied to a set of variables.

The value of α is determined by applying the softmax function to the variable d_k. Here, W_{Qi}^T and W_{Ki} are weight matrices that can be learned. They are used for the query and key projections in the ith attention head. Additionally, d_k represents the dimensionality of the key vectors. The attention

scores are utilised to compute a weighted sum of the value vectors (V).

$$Head_i = \alpha_i V \quad (12)$$

The term "Head" in equation (12) refers to the variable "α_i" multiplied by the variable " V ". The results of all attention heads are combined and then converted using a weight matrix that can be learned, denoted as W_O as in equation (13) to calculate $Concatenated$ and multiplied with W_O to produce $Output$ as given in equation (14).

$$Concatenated = Concat(Head_1, Head_2, \dots, Head_N) \quad (13)$$

$$Output = ConcatenatedW_O \quad (14)$$

3.2.3. Attention pooling

By multiplying feature map f3 and attention map A attention pooling is happened. From attention pooling the part feature maps PF1, PF2….PF32 are generated. Totally 32 number of part feature maps with dimension 2048*7*7 are generated. Global Average pooling is applied on them to generate the feature matrix M with size 65530*1*1. Global Average Pooling (GAP) is a method employed in convolutional neural networks (CNNs) and deep learning architectures. This is a form of spatial pooling that reduces the spatial dimensions of each feature map to a single value by computing the average of all values in the feature map. Global average pooling differs from conventional pooling layers, such as max pooling or average pooling with a fixed-size window, by calculating a single value for each feature map. This approach yields a fixed-size output, irrespective of the input size.

$$Y_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{ijc} \quad (15)$$

In equation (15) when provided with a feature map X that has dimensions $H \times W \times \mathrm{C}$, where H represents the height, W represents the width, and C is the number of channels, the Global Average Pooling procedure generates an output vector Y_c.

3.2.4. Attention guided augmentation

Attention techniques are used to highlight important parts of an image. These algorithms can give different image parts weights or attention scores. Attention-guided augmentation uses attention maps from the attention mechanism to guide augmentation. Augmentation does not randomly change the image. It targets parts sensed by the attention mechanism. The attention system detects regions of interest, which are augmented with rotation, scaling, cropping, or flipping. This ensures the enriched data retains valuable features highlighted by the attention process. The primary activity and direct augmentation attention technique are often trained together. The model learns to identify job-relevant areas and modifies the augmentation procedure accordingly. Both unaltered and changed input data are fed to the model during training. Attention-guided augmentation prioritises informative locations, improving model robustness and generalisation. In the proposed model, the attention map A are normalized by applying softmax to generate A* and based on A* the attention guided augmentation is done. Three different attention-based augmentation are used to generate the augmented images AI1, AI2, AI. They are attention mix-up, random erasing with attention, and attention dropout.

Attention Mix-Up: Mix-up is a data augmentation approach used in training, where input samples and their related labels are blended in pairs. By amalgamating knowledge from many examples, it stimulates the model to acquire more resilient and comprehensive features. Please direct your focus and concentration towards this matter. Mix-up expands upon this concept by integrating attention mechanisms, enabling the model to concentrate on regions while doing the Mix-up procedure.

Step 1: selection of sample pairs - Stochastically choose pairs of input samples and their matching labels from the training dataset.

Step 2: Mix-up - Utilise linear interpolation to merge the chosen pairs:

$$Com_{Input} = \lambda \times Input_1 + (1 - \lambda) \times Input_2 \quad (16)$$

This equation (16) reflects the aggregate input produced from attention mix-up. In this context, λ is a parameter that governs the proportion in which the two inputs, referred to as $\mathrm{Input_1}$ and $\mathrm{Input_2}$, are mixed. The initial input ($Input_1$) is multiplied by the scalar λ, whereas the second input ($Input_2$) is multiplied by the scalar $(1 - \lambda)$. The scaled inputs are aggregated to create the combined input.

$$Com_{Label} = \lambda \times Label_1 + (1 - \lambda) \times Label_2 \quad (17)$$

Here, λ is a randomly selected value obtained from a Beta distribution. This equation (17) depicts the aggregate label produced by attention mix-up. Like previous instances, the parameter λ determines the proportion of $Label_1$ and $Label_2$ in the mixture. The initial label, denoted as $Label_1$, is multiplied by the scalar λ. Similarly, the second label, denoted as $Label_2$, is multiplied by the scalar $(1-\lambda)$. The resulting scaled labels are then joined together to create the combined label.

Step 3: Attention Mechanism - Integrate attention mechanisms to calculate attention weights for every spatial location or feature in the combined input.

$$Att_{Wt} = Att(Com_{Input}) \quad (18)$$

In equation (18), Att_{Wt} denotes the attention weights allocated to the merged input. The weights represent the significance or pertinence of various components of the combined input within the specific task's context. $Att(Com_{Input})$ signifies the utilisation of an attention mechanism (Att) on the merged input (Com_{Input}). An attention mechanism is a computational function or component of a neural network that calculates attention weights based on the input.

The attention mechanism can be implemented using convolutional layers, self-attention mechanisms (such as Transformer attention), or any other appropriate attention mechanism for the given task.

Step 4: Weighted summation using attention. Utilise the attention weights to combine the combined input and produce the ultimate attention-combined input:

$$Att_{Com_{Input}} = Att_{wt} \times Com_{Input} \quad (19)$$

The attended representation in equation (19) of the combined input is $Att_{Com_{Input}}$. The result of applying attention weights to the combined input focuses on relevant or important areas or features. As described above, Att_{wt} are the combined input attention weights. These weights show the relative importance of input components. The mix-up operation yields Com_{Input}. It contains data from two inputs mixed in some ratio. The attention combined input is equal to the attention weights multiplied by the mixed input.

Step 5: Training - Train the model using the attention-combined input and mixed label as if it were a standard training case.

Random Erasing with Attention: Random Erasing randomly selects a rectangular region in the input image and replaces its pixels with random values. It simulates occlusion or dropout, making the model more robust. Attention mechanisms direct the occlusion process to picture locations, improving Random Erasing. The following structure is for attention-based Random Erasing:

Step1: Choose Attention Regions: Utilize an attention mechanism to detect areas of significance in the input image. An attention mechanism can be achieved using convolutional layers, self-attention mechanisms, or any other appropriate attention mechanism.

Step 2: Random Erasing with Attention: Choose a rectangular area at random from the attention zones that have been detected.

Remove or substitute the pixels included in that rectangular area with random values.

Step 3: Model Training: Utilize the randomly erased image for training the model as if it were a typical training example.

Attention Dropout: Attention Dropout modify the attention process to include dropout in attention weight calculations. Dropout is a common neural network regularisation method that reduces overfitting. A portion of input units is randomly zeroed during training. Attention Dropout reduces specific attention weights in attention mechanisms, forcing the model to depend on different input sequence segments. This reduces the attention mechanism's pattern dependence, improving generalisation. Integration of Attention Dropout is outlined below.

Step 1: Attention Mechanism - Calculate attention weights using a conventional attention mechanism.

Step 2: Implement Dropout - Implement the dropout technique on the calculated attention weights. Within each training iteration, a subset of the attention weights is assigned a value of zero.

Step 3: Computation of Weighted Sum - Utilize the adjusted attention weights to calculate the weighted sum of the input sequence, like a conventional attention mechanism.

Step 4: Training - Incorporate attention dropout into the attention mechanism during model training.

3.2.5. Elastic_net

The augmented images AI1, AI2, AI3 are then fed to multi-scale self-attention generator where the prediction vectors P, p1, p2 and p3 are generated. Elastic_net is used here to calculate the loss. Elastic Net is a regularisation method that incorporates both L1 (Lasso) and L2 (Ridge) regularisation penalties into a linear regression model. Its purpose is to

overcome some constraints of L1 and L2 regularisation when employed separately. The regularisation term in Elastic Net is a linear mixture of the L1 and L2 penalties, and it is determined by two hyperparameters: α and λ.

The Elastic Net regularization term is expressed as follows:

$$Elas - Net_{Penalty} = \alpha \sum_{i=1}^{n} |w_i| + (1-\alpha) \sum_{i=1}^{n} w_i \quad (20)$$

The equation (20) comprises two terms, the term $\alpha \sum_{i=1}^{n} |w_i|$ represents the L1 regularisation term. The sum of the absolute values of all model parameters is calculated, with each value scaled by the alpha parameter.

The term $(1-\alpha) \sum_{i=1}^{n} w_i$ denotes the L2 regularisation term. The expression calculates the total of the squares of all model parameters, which are then multiplied by $(1-\alpha)$. n denotes the aggregate count of model parameters or weights.

4. Results and discussion

The integration of a Multiscale Self Attention Generator has significantly enhanced the image augmentation procedure. The generator effectively emphasised significant structures and minimised potential distortions caused during augmentation by adaptively focusing on crucial features at various scales. The proposed method has demonstrated enhanced efficacy in detecting foetal brain regions in ultrasound pictures. The implementation of advanced image augmentation techniques has resulted in improved identification capabilities, ensuring greater reliability and accuracy in demanding scenarios involving fluctuations in foetal location and imaging settings.

The Multiscale Self Attention Generator has been useful in maintaining spatial context during the augmentation process. Preserving the anatomical links of structures is crucial in medical imaging to ensure precise diagnosis. The augmented images, produced using the Multiscale Self Attention Generator, have exhibited enhanced resilience to frequent variations found in ultrasound imaging, including speckle noise, variations in probe orientation, and changes in foetal position.

4.1 Dataset description

For foetal brain imaging, we utilise the foetal planes ZONODO Dataset [18], a comprehensive collection of maternal-fetal screening ultrasound pictures. These images were obtained from multiple operators and ultrasound equipment at two distinct institutions. A professional maternal foetal specialist personally labelled all the images. Totally six categories of images are taken for analysis namely mother's cervix, four foetal anatomical planes Abdomen, Brain, Femur, and Thorax and general category. Foetal brain scans are categorised into three main planes: Trans-thalamic, Trans-cerebellum, and Trans-ventricular. These planes are used to assess the ability to classify tiny details with high precision. In this suggested investigation, we employed a subset of data to synthesis.

The Convolutional Neural Network (CNN) was trained using stochastic gradient descent optimization, with the weights updated based on the calculated loss function in training. The optimal parameter values for the multiscale self-attention CNN were identified through iterative experimentation.

4.2 Evaluation metrics

When assessing the augmentation of ultrasound images, a full evaluation is typically conducted using a combination of quantitative and qualitative indicators. Quantitative metrics provide numerical measurements, but qualitative assessments entail visual inspections and expert judgements.

4.2.1. Peak signal to noise ratio (PSNR)

Peak Signal-to-Noise Ratio (PSNR) is a commonly employed metric for assessing the quality of an image. It measures the degree of similarity between an unaltered image and a distorted or compressed version of that image. The $PSNR$ is calculated by determining the mean squared error (MSE) between the original and distorted image as in (21).

$$PSNR = 10 \log_{10}\left(\frac{max^2}{MSE}\right) \quad (21)$$

The maximum pixel value of the image is denoted by max, which is typically 255 for grayscale images with 8 bits per pixel. The mean squared error (MSE) is the arithmetic mean of the squared differences between corresponding pixels in the original and distorted pictures.

4.2.2. Accuracy

$Accuracy$ is calculated based on the ratio of correctly predicted instances $No. Of\ True\ Prediction$ to the total number of instances $Total\ Prediction$. It provides a thorough

evaluation of the model's effectiveness that shown in equation (22).

$$Accuracy = \frac{No.Of\ True\ Prediction}{Total\ Prediction} \quad (22)$$

4.2.3. Precision

$Precision$ is the ratio achieved by dividing the number of correctly predicted positive observations $True\ Postives$ by the total number of predicted positives $true\ Postives{+}False\ Postives$ as per equation (23). The statement measures the level of accuracy of the positive predictions.

$$Precision = \frac{True\ Postives}{True\ Postives+False\ Postives} \quad (23)$$

4.2.4. Recall

$Recall$ refers to the ratio of correctly predicted positive observations $True\ Postives$ to the total number of actual positive observations $True\ Postives + False\ Negatives$ given in equation (24). It is used to describe the model's ability to appropriately represent all relevant cases.

$$Recall = \frac{True\ Postives}{True\ Postives+False\ Negatives} \quad (24)$$

4.2.5. F1-Score

The $F1-Score$ is computed by taking the reciprocal of the average of the reciprocals of $Precision$ and $Recall$. It provides a balanced blend of precision and comprehensiveness, making it particularly useful in scenarios where there is an uneven distribution of categories given in Eq. (25).

$$F1-Score = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (25)$$

4.3 Comparison methods

Proposed approach: Our study employs a Multi Scale Self Attention Convolutional Neural Network (MSAG-ENET) to implement diverse modifications on the initial image. This facilitates the creation of an extensive collection of images including different levels of distortion and noise. This collection is then utilised to train deep learning models that exhibit greater resilience to fluctuations in the input data.

Table 1. Comparison of Performance

	Accuracy	Precision	Recall	F1-Score
MSAG_ENET	98.8	97.4	97.7	97.5
Method A	95.2	94.2	96.3	95.2
Method B	97.2	91.8	93	92.4
Method C	90.4	95.1	90.8	92.9
Method D	91.8	91.6	96.1	93.8
Method E	96.3	89.7	91.8	90.7
Method F	97.6	87.5	93.2	90.3

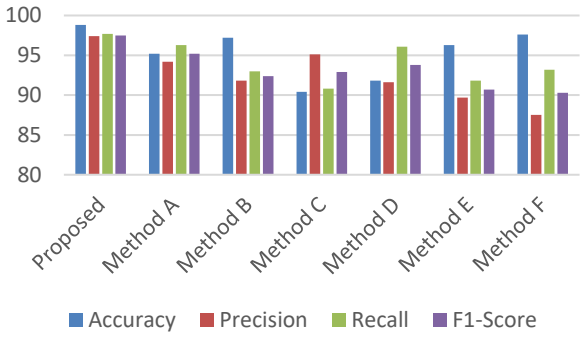

Figure. 5 Comparative Results of Proposed MSAG-ENET Model with Other Models

Random Erasing – Method A [19]: Random Erasing is a data augmentation technique frequently employed to train Convolutional Neural Networks (CNNs) to improve the overall ability of the neural network to generalise and standardize the pixel values using arbitrary reference points.

Mix-up – Method B [20]: The mixup technique is used to tackle the problem of imbalance and enhance the model's resilience by blending pixel values of two distinct images from the training dataset.

CutMix – Method C [21]: The CutMix technique involves substituting the removed pixel area with an image patch acquired from adjacent images.

Auto Augment – Method D [22]: The utilisation of an easy approach named Auto Augment allows for the automatic search of improved data augmentation strategies.

Rand Augmentation – Method E [23]: Employs a grid search technique to identify the optimal augmentation strategy for a given dataset and network. This involves exploring a range of magnitudes and applying transformations in a sequential manner.

Style Transfer Augmentation – Method F [24]: Style transfer is a technique that preserves the original high-level meaning of an image while

replacing the low-level textural details with the style of a randomly selected source image.

4.4 Performance analysis

The performance of the proposed MSAG_ENET is compared with existing methods based on Accuracy, Precision, Recall and F1-Score shown in Table 1. Fig. 5 shows the comparison as a visual result.

Fig. 6 displays the accuracy analysis results derived from the training epochs. The optimal augmentation model enhances the deep learning model's training by generating fresh data samples through several image alteration techniques on the current training data.

The accuracy consistently improves as the number of epochs rises, eventually stabilising at a high level of around 97.72% after 50 epochs. The augmentation uses a multiscale self-attention CNN model, which can collect and encode intricate and profound representations of the input image [25].

Fig. 7 demonstrates the analysis of PSNR values in relation to various degrees of noise, specifically focusing on reducing speckle noise. This technique aims to diminish speckle noise while preserving the fundamental image elements in their natural state without distortion. The graph shows that the suggested CNN-based technique effectively reduces speckle noise, resulting in Peak Signal-to-Noise Ratio (PSNR) values between 42 and 15 dB across different noise levels. As noise level increases, PSNR lowers but remains within an acceptable range considered good.

A key feature of an effective speckle noise reduction method is its capacity to uphold high Peak Signal-to-Noise Ratio (PSNR) values despite substantial fluctuations in noise levels.

4.5 Comparative analysis

The accuracy values for different augmentation techniques are presented in Fig. 8. The proposed method exhibits a classification accuracy of 98.8%, which is greater than accuracy of other methods. The existing approaches have achieved accuracy rates ranging lower than the proposed model up to 97.6%. This result suggests that the proposed method has a greater capability for accurate and reliable diagnosis.

Fig. 9 depicts the analysis of precision for different approaches. The proposed strategy achieved a precision of 97.4%, exceeding the precision of the other methods being compared. The existing methodologies produce precision rates till 95.1%. The study demonstrates that the suggested approach is very successful in accurately detecting true positive

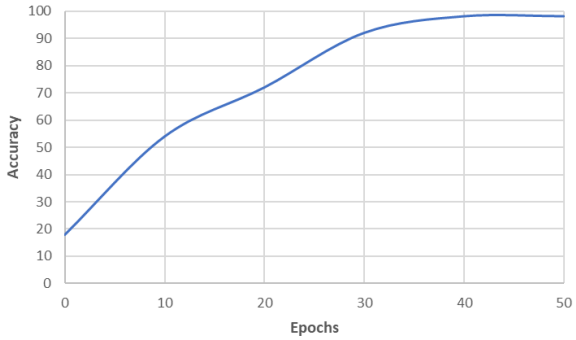

Figure. 6 Evaluation of the Proposed Augmentation Model's Accuracy

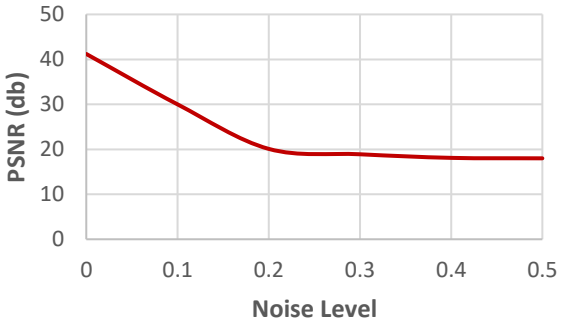

Figure. 7 PSNR Validation for Proposed Noise Reduction Model

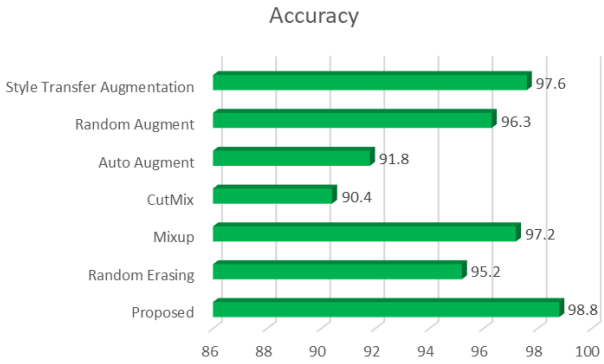

Figure. 8 Accuracy Validation

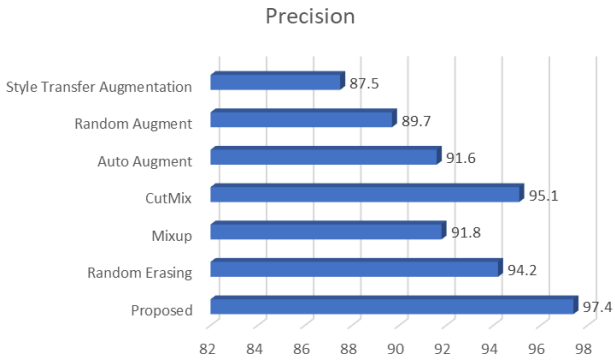

Figure. 9 Precision Validation

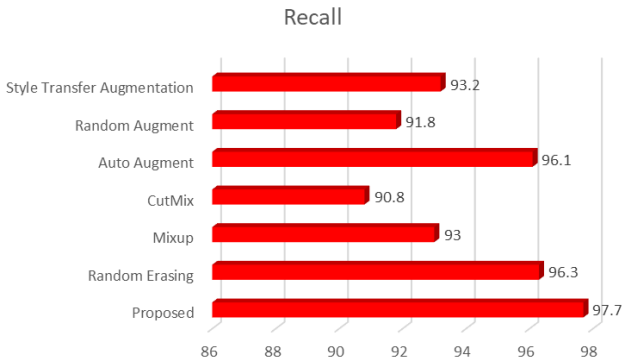

Figure. 10 Recall Validation

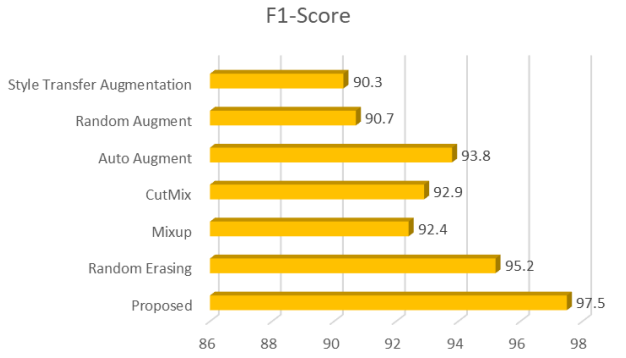

Figure. 11 F1-Score Validation

cases and reducing the number of false positive cases. This is an essential requirement in clinical applications.

The recall analysis graph described in Fig. 10 showcases the efficacy of several methodologies. The proposed method attained a recall in the rate of 97.7%, above that of the prior methodologies yield recall rates up to 96.3%. This result demonstrates that the proposed methodology has higher sensitivity in accurately identifying true positives, rendering it more appropriate for clinical applications that necessitate exact diagnosis.

The graph in Fig. 11 illustrates the F1-score analysis for various existing methodologies. The proposed method accomplished an F1-score of 97.5%, outperforming the performance of the compared strategies which are till 95.2%. The maximum F1-score attained by the projected approach demonstrates its superior capacity to optimise the balance between precision and memory, leading to more precise diagnoses.

The results of the augmentation comparison show that the suggested strategy surpasses the existing methods, highlighting its ability to enhance the resilience and adaptability of deep learning models. The method creates an expanded dataset of pictures showcasing various levels of distortion and noise. The dataset is used to build deep learning models that show enhanced resilience to changes in the input data.

Table 2. Symbols Used

Symbols	Description
μ	Mean
σ	Standard Deviation
$T(x,y)$	Transformed Pixel Value at (x,y)
L	Total Number of Intensity Levels
C	Contrast Limit Parameter
$H(x,y)$	Cumulative Histogram at (x,y)
$I(x,y)$	Pixel Intensity at(x,y)
$h(i)$	Original Histogram
I	Input Image
t	Time
$(c(x,y,t)$	Diffusion Coefficient at Position (x,y) in time t
∇	Laplacian
$div(x,y,t)$	Divergence Function
Qi	Query Vector at Position i
K	Key Vector Matrix
V	Value Vector Matrix
d_k	Dimensionality of Key Vector
α_i	Attention Score
W_{Ki}	Weight Matrix
Y_c	Output Vector
H	Height
W	Width
C	Number of Channels
$\alpha \sum_{i=1}^{n} \lvert w_i \rvert$	L1 Regularization Term
$(1 - \alpha) \sum_{i=1}^{n} w_i$	L2 Regularization Term
n	Aggregate Count of Model Parameters or Weights

5. Conclusion

The proposed method, has shown notable progress in the field of medical image processing, particularly in detecting ultrasound foetal brain structures. The application of a Multiscale Self Attention Generator for picture augmentation has demonstrated its worth as an effective improvement, enhancing the accuracy and resilience of foetal brain structure detection. Data augmentation plays a vital role in reducing the data limitation. And image quality has been enhanced through the preprocessing steps. For reducing speckle noise, the anisotropic diffusion filter used. As a net outcome the accuracy, precision, recall and F1 score are improvised to 98.8, 97.4, 97.7 and97.5 respectively. The achievement of the Multiscale Self Attention Generator in this work paves the way for additional investigation and advancement. Possible future paths may involve examining the incorporation of supplementary attention mechanisms, examining the adaptability of the methodology to various medical imaging

methods, and undertaking thorough clinical validation studies.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, Vetriselvi D and Thenmozhi R; methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation Vetriselvi D; writing—review and editing, Vetriselvi D and Thenmozhi R; visualization, Vetriselvi D and Thenmozhi R; supervision, Thenmozhi R; project administration, Thenmozhi R.

Acknowledgments

This work was not supported by any organization and funding agencies.

References

[1] M. Tripathi, "Analysis of Convolutional Neural Network Based Image Classification Techniques", *Journal of Innovative Image Processing*, Vol.3, No.2, pp.100–117, 2021.

[2] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition", In: *Proc. IEEE Conf on Computer Vision and Pattern Recognition*, pp. 4438–4446, 2017.

[3] R. Qin, Z. Wang, L. Jiang, K. Qiao, J. Hai, J. Chen, J. Xu, D. Shi, and B. Yan, "Fine-grained lung cancer classification from PET and CT images based on multidimensional attention mechanism", *Complexity*, vol. 2020, pp. 1–12, 2020

[4] S. Jetley, N. Lord, N. Lee, and P. H. Torr, "Learn to pay attention", In: *Proc. on Learning Representations*, Vancouver Canada, pp.1-14, 2018.

[5] J. Wang, Y. Bao, and Y. Wen, "Prior-attention residual learning for more discriminative COVID-19 screening in CT images", *IEEE Transaction on Med. Imaging*, Vol. 39, No. 8, pp. 2572–2583, Aug. 2020.

[6] W. M. Gondal, J. M. Kohler, R. Grzeszick, G. A. Fink, and M. Hirsch, "Weakly-supervised localization of diabetic retinopathy lesions in retinal fundus images", In: *Proc. IEEE Int. Conf. Image Processing*, pp. 2069– 2073, 2017.

[7] J. Zhang, Y. Xie, Y. Xia, and C. Shen, "Attention residual learning for skin lesion classification", *IEEE Trans. Med. Imag* , Vol. 38, No. 9, pp. 2092–2103, Sep. 2019.

[8] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation", In: *Proc. of the AAAI Conference on Artificial Intelligence*, New York, USA, pp. 13001–13008, 2020

[9] H. Naveed, S. Anwar, M. Hayat, K. Javed, and A. Mian, "Survey: Image mixing and deleting for data augmentation", *Engineering Applications of Artificial Intelligence*, Vol.131, 107791, 2024.

[10] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "Cutmix: Regularization strategy to train strong classifiers with localizable features", In: *Proc. of 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[11] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning augmentation strategies from data", In: *Proc. of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[12] R. Takahashi, T. Matsubara, and K. Uehara, "Data augmentation using random image cropping and patching for Deep CNNs", *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 2917–2931, 2020.

[13] X. Zheng, T. Chalasani, K. Ghosal, S. Lutz, and A. Smolic, "Stada: Style transfer as data augmentation", In: *Proc. of 14th International Joint Conference on Computer Vision*, Imaging and Computer Graphics Theory and Applications, Prague, Czech, 2019.

[14] S. G. Patro and K. K. Sahu, "Normalization: A preprocessing stage", *International Advanced Research Journal in Science*, Engineering and Technology,Vol.2, No.3, pp.20–22, 2015.

[15] P. Musa, F. A. Rafi, and M. Lamsani, "A Review: Contrast-Limited Adaptive Histogram Equalization (CLAHE) methods to help the application of face recognition", In: *Proc. Third International Conference on Informatics and Computing (ICIC)*, Palembang, Indonesia, 2018, pp. 1-6

[16] C. Li, Y. Wang, C. Xiao, and X. Lu, "A New Speckle Reducing Anisotropic Diffusion for Ultrasonic Speckle", *Acta Automatica Sinica, Vol.38*, pp.412-418, 2012.

[17] Y. Yongjian and S. T. Acton, "Speckle reducing anisotropic diffusion", *IEEE Transactions on Image Processing*, Vol.11, No.11, pp.1260–1270, 2002.

[18] X. P. Burgos-Artizzu, "Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes", *Sci Rep*, vol. 10, no. 1, pp. 10200, Dec. 2020.

[19] V. Mingote, A. Miguel, D. Ribas, A. Ortega, and E. Lleida, "Knowledge Distillation and Random Erasing Data Augmentation for Text-Dependent Speaker Verification", In: *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 6824-6828.

[20] Y. Ma, X. Xu, and Y. Li, "LungRN+NL: An improved adventitious lung sound classification using non-local block ResNet neural network with mixup data augmentation", In: *Proc. Interspeech, Shanghai*, China, 2020.

[21] E. Harris, A. Marcu, M. Painter, M. Niranjan, A. Prügel-Bennett, and J. Hare, "Fmix: Enhancing mixed sample data augmentation", *arXiv preprint arXiv:2002.12047*, 2020

[22] S. Yang, W. Xiao, M. Zhang, S. Guo, J. Zhao, and F. Shen, "Image data augmentation for deep learning: A survey", *arXiv preprint arXiv:2204.08610 2022*.

[23] K. Faryna, J. Van der Laak, and G. Litjens, "Automatic Data Augmentation to Improve Generalization of Deep Learning in H&E Stained Histopathology", *Computers in Biology and Medicine*, Vol.170, 2023.

[24] R. Yamashita, J. Long, S. Banda, J. Shen, and D. L. Rubin, "Learning domain-agnostic visual representation for Computational Pathology using medically-irrelevant style transfer augmentation", *IEEE Transactions on Medical Imaging*, Vol.40, No.12, pp. 3945–3954, 2021.

[25] D. Vetriselvi and R. Thenmozhi, "Advanced image processing techniques for ultrasound images using multiscale self attention CNN", *Neural Processing Letters*, Vol.55, No.9, pp. 11945–11973, 2023.