ENAS-B: Combining ENAS with Bayesian Optimisation for Automatic Design of Optimal CNN Architectures for Breast Lesion Classification from Ultrasound Images

Mohammed Ahmed [a], Hongbo Du [a], Alaa AlZoubi [b]

[a]*School of Computing, The University of Buckingham, Buckingham, United Kingdom*
[b]*School of Computing and Engineering, University of Derby, Derby, United Kingdom*
mhamdit@gmail.com [a], *hongbo.du@buckingham.ac.uk* [b], *a.alzoubi@derby.ac.uk* [c]

1. Abstract

Efficient Neural Architecture Search (ENAS) is a recent development in searching for optimal cell structures for Convolutional Neural Network (CNN) design. It has been successfully used in various applications including ultrasound image classification for breast lesions. However, the existing ENAS approach only optimises cell structures rather than the whole CNN architecture nor its trainable hyperparameters. This paper presents a novel framework for automatic design of CNN architectures by combining strengths of ENAS and Bayesian Optimisation in two folds. Firstly, we use ENAS to search for optimal normal and reduction cells. Secondly, with the optimal cells and a suitable hyperparameter search space, we adopt Bayesian Optimisation to find the optimal depth of the network and optimal configuration of the trainable hyperparameters. To test the validity of the proposed framework, a dataset of 1,522 breast lesion ultrasound images is used for the searching and modelling. We then evaluate the robustness of the proposed approach by testing the optimized CNN model on three external datasets consisting of 727 benign and 506 malignant lesion images. We further compare the CNN model with the default ENAS-based CNN model, and then with CNN models based on the state-of-the-art architectures. The results (error rate of no more than 20.6% on internal tests and 17.3% on average of external tests) showed that the proposed framework generates robust and light CNN models.

Keywords: Breast Lesion Classification, Convolutional Neural Networks, Efficient Neural Architecture Search, Bayesian Optimisation, Ultrasound Image Classification.

2. Introduction

Breast cancer is one of the most common cancer types [1]. It is the second deadliest cancer for woman [2]Previous studies show that early detection of breast cancers followed by appropriate treatment is responsible for 38% reduction in mortality rate from 1989 to 2018 [1]. Ultrasound (US) imaging has the benefits of being safe and less costly than other imaging modalities such as Magnetic Resonance Imaging (MRI), and hence widely used in breast cancer diagnosis. The clinical needs as well as technological advances in deep learning have motivated us to develop a new automated recognition approach for classifying breast lesions into benign or malignant types.

In recent years, Computer-Aided Diagnosis (CAD) systems have been applied to medical image analysis including classifying ultrasound images of breast lesions [3]. At the same time, deep learning Convolutional Neural Network (CNN) has shown great success in natural image classification. Many existing CNN architectures such as VGG net [4] and GoogLeNet [5] were designed. Because of model complexity and shortage of annotated medical images, most existing research focuses on customising the existing CNN architectures to the medical images via transfer learning [3]. However, such customised CNN models are still inherently large and complex with an increased risk of model overfitting. Attempts have also been made to design CNN architectures specifically for breast lesion classification from US images. An architecture (CNN3) of three convolutional layers followed by Batch normalisation, Relu and MaxPooling was proposed [6]. Another architecture (CNN4) of four convolutional layers with filters of different sizes and numbers was also reported [7]. More recently, the Fus2Net [8] architecture consists of three convolutional layers followed by two consecutive modules each of which consists of several convolutional layers using filters of different sizes. Despite all the efforts already made in building and customising CNN architectures for breast lesion image classification via manual designs of the layers and hyperparameters, the need for accurate, robust, and light CNN models remains constant.

CNN architecture design involves setting many hyperparameters. Manually obtaining the optimal settings for them is challenging and time-consuming [9]. Therefore, the interest in automatic search for optimal CNN architectures is increasing. Several approaches, such as Generic Algorithms (GA), Reinforcement Learning (RL) and Bayesian Optimization (BO), have been developed [10]. Neural architecture search (NAS) is a RL-based framework [9], but it is computationally expensive because the number of architectural options to explore grows exponentially. Efficient Neural Architecture Search (ENAS) overcomes this limitation through weight sharing during the search phase [11]. In ENAS, a single CNN network known as Supernet with all operations within a search space is trained, and the generated CNNs share trained weights of the Supernet. Two types of search space can be used by the RNN controller within the ENAS framework: the macro space where the controller searches for an entire network or the micro space where the controller generates cells containing operations and connections between them. Evidence shows that the micro search space is more efficient [11].

Automatic search of CNN architectures has been attempted for medical images recently. A hybrid NAS framework for classifying and segmenting thyroid cancer from ultrasound images was proposed in [12]. ENAS with micro search space was adopted for breast lesion classification from US images [13]. The generalisation gap of ENAS models was further investigated [14]. Nevertheless, the ENAS approach has its own limitations. First, the number of blocks of cells is still determined manually through trials. Secondly, trainable hyperparameters critical for designing effective and efficient CNN architectures are manually set by trials.

This paper addresses these limitations by adopting Bayesian Optimization for optimizing the number of blocks of ENAS cells and trainable hyperparameters. Bayesian Optimization, as an efficient method for optimizing noisy and expensive functions, provides a better approach than other optimizers to model uncertainty and allow exploration and exploitation to be automatically balanced during the search [10]. The paper therefore proposes a novel automatic “end-to-end” CNN design framework by combining ENAS cells with Bayesian Optimisation search. To evaluate this framework, the optimised classification model is tested on images captured by US machines of different makes and from different medical centres in different countries. A further comparison is made between our model and state-of-the-art models based on hand-crafted architectures.

3. Materials and Methods

3.1. Data Collection and Preparation

In this study, five datasets of US images of breast lesions were used. Four were collected by our sponsor from three hospitals in Shanghai China including Pudong New Area People's Hospital, No.6 Hospital and No.10 Hospital after ethical approvals by the hospitals. The ground-truth for each image (benignity or malignancy) is based on pathology reports. Experienced radiologists from the hospitals manually cropped the region of interest (RoI) for each US image in every dataset. A RoI bounding box image was generated and used as the input image. The fifth is a public domain dataset (BUSI) collected from a hospital in Egypt with associated class labels and cropped lesion areas [15]. All images were captured using US machines of different makes (Siemens, Toshiba, GE, Philips and LOGIQ E9). This research was granted ethics approval by the Research and Ethics Committees of University of Buckingham. The datasets are split into two collections:

1) *Modelling Dataset:* Two of the four datasets from two of the three hospitals in China respectively containing 1,102 images (726 benign and 376 malignant) and 420 images (278 benign and 142 malignant) were merged into a dataset of 1,522 images. This set is used for developing ENAS-B.
2) *External Test Sets:* The BUSI dataset (External A) consists of 565 images (355 benign and 210 malignant). The other two datasets (External B and External C) from two of the three hospitals in China respectively consist of 500 images (300 benign and 200 malignant) and 168 images (72 benign and 96 malignant). The three datasets were separately used for testing purposes. Figure 1 shows some examples of US images from the datasets.

A new dataset of ultrasound images of breast lesions just became available [16]. It consists of 109 images of benign and 123 images of malignant lesions all of which have been confirmed by histopathologic results including fine needle aspiration, core needle, or open biopsies. After removing the images with artefacts, 207 images (95 benign and 112 malignant) were used as another external test set (External D).

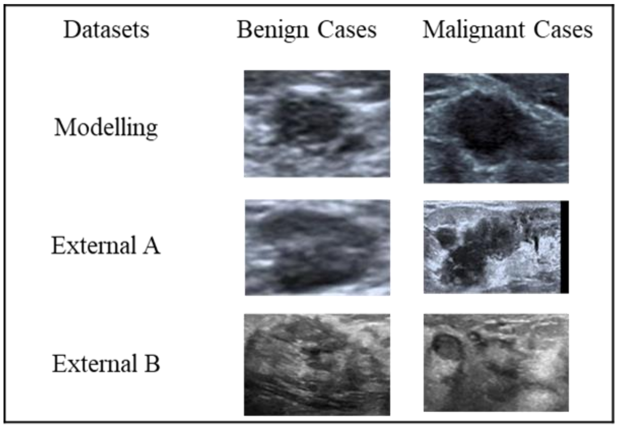

Figure 1:Samples of RoI images for modelling and external test sets

3.2. Bayesian Optimisation for ENAS-based Architecture Design

The proposed framework is shown in Figure 2. It consists of three main phases. Phase I is a general preparation of US images including RoI (i.e. the lesion region) cropping, image resizing, and increasing the number of training examples. Phase II is intended to obtain an optimized backbone deep CNN (DCNN) architecture and a set of optimized trainable hyperparameters. Phase III finally uses the optimized architecture and hyperparameters to train a classification model.

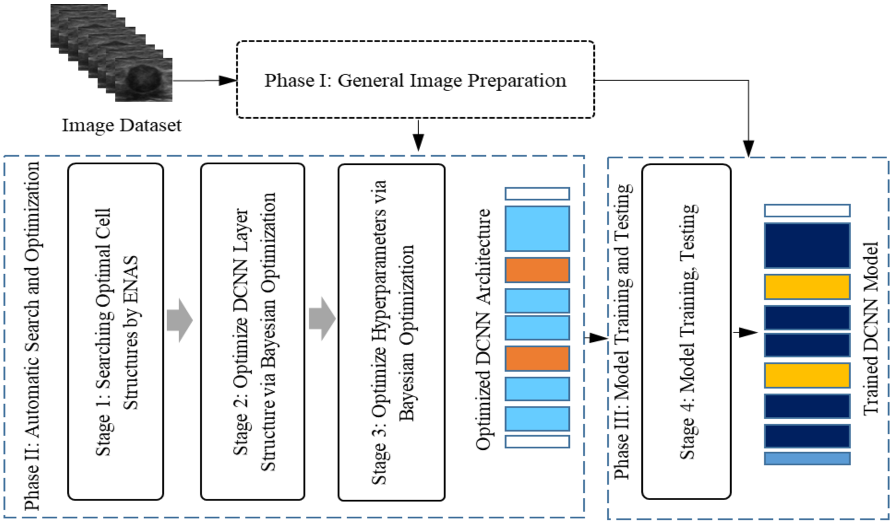
Image Dataset
Phase I: General Image Preparation
Phase II: Automatic Search and Optimization
Stage 1: Searching Optimal Cell Structures by ENAS
Stage 2: Optimize DCNN Layer Structure via Bayesian Optimization
Stage 3: Optimize Hyperparameters via Bayesian Optimization
Optimized DCNN Architecture
Phase III: Model Training and Testing
Stage 4: Model Training, Testing
Trained DCNN Model

Figure 2:The proposed framework for automatic CNN model designs for breast lesion classification from US images

3.2.1. Image Preparation

The RoI image of lesion was cropped from the whole US image for accurate recognition. A free-hand cropping tool reported in [17] was used by the radiologists to identify, collect and store the coordinates of the pixel points on the border of a lesion. A rectangular bounding box was then generated for each lesion by fitting the border points into a minimum area rectangle. The tumour microenvironment (TME) is the cellular environment in which a tumour exists, and it includes various components such as immune cells, blood vessels, fibroblasts, and extracellular matrix [18]. The TME plays a crucial role in cancer progression and can have a significant impact on disease management and diagnosis [19]. Therefore, the accurate assessment of the tumour microenvironment (TME) within breast cancer plays a pivotal role in disease management when utilizing ultrasound images. Selecting RoI is a crucial factor in the development of machine learning models for breast cancer classification from ultrasound images [20]. Furthermore, integrating information about the TME into these models by RoI margin enhances model performance [21].

Furthermore, using a small margin of background around the lesion can provide contextual and spatial information that can aid in the lesion classification task and mitigate the effects of image cropping. Based on the work in [17], [22]margin of 8% of the lesion width and height was then added for the final cropped RoI image. To satisfy the training requirements of our proposed framework, the cropped RoI images were resized to 100×100 pixels.

Searching and training a complex DCNN also requires large datasets. One way to meet the requirement is to enlarge the training set through data augmentation. Two augmentation methods reported in [17] were adopted. The geometric methods use both image mirroring and rotation (90, 180 and 270 degrees), and the singular value decomposition (SVD) method respectively takes 45%, 35% and 25% ratios of the selected top singular values. The methods generated seven additional images from one RoI image.

3.2.2. Automatic Search and Optimization

Phase II of our framework consists of three stages as shown in Figure 2. At Stage 1, the ENAS method is used to search for the optimal internal structures of normal and reduction cells. At Stage 2, the optimized cells are stacked in a process controlled by the Bayesian Optimization algorithm, creating a sequential layer structure of the cells for the whole network. At Stage 3, Bayesian Optimization is again employed to optimize trainable hyperparameters within the optimized network structure, creating the final optimized DCNN architecture for modelling.

Optimal Cells Search Using ENAS. The ENAS micro approach consists of two stages [11]. The first stage searches for an optimal pair of Normal (N) and Reduction (R) cells in a pre-defined architecture (i.e. Supernet) based on validation accuracy. The Supernet consists of a 3×3 standard convolution layer named *stem conv* and 7 cells (1N, 1R, 1N, 1R, 3N). The default search operations of ENAS [11] were provided to the ENAS controller. We set the mini batch size to 8 and all other hyperparameters as ENAS default [10]. The RNN controller is trained for 150 epochs and each epoch generated 10 pairs of N and R cells. In the searching stage, the Modelling Dataset (see Section 3.1) was used under a single split policy (see Section 4). Figure 3 shows the searched optimal cell structures from ENAS based on the modelling dataset.

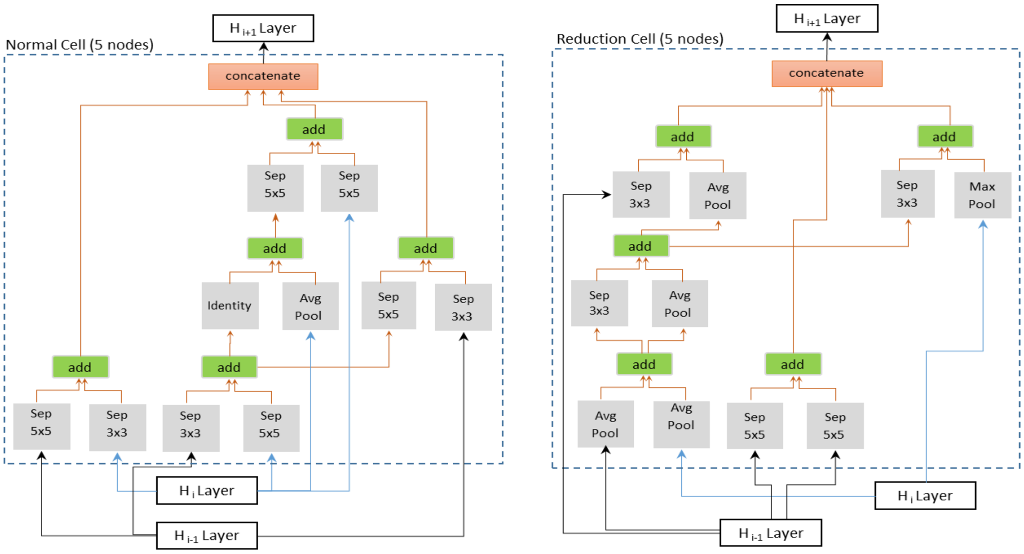

Figure 3:Example of optimal cell structure (Normal and Reduction cells) generated from our data set

ENAS-B. The proposed ENAS-B search involves three key elements: a backbone architecture, a search space and a search strategy. First, we define a CNN backbone architecture with optimisable structural hyperparameters and their search spaces. Second, we perform automatic architecture search (the first optimisation stage) using Bayesian optimisation to identify the optimal number of normal cells (or the depth of the architecture) that results in a new architecture called ENAS-B-1. Finally, we use ENAS-B-1 as a backbone architecture, define optimisable training hyperparameters and their search space, and perform automatic architecture search using Bayesian optimisation to optimize training hyperparameters. This second optimisation stage results in ENAS-B. It is worth noting that the first and the second stages use the same Bayesian optimisation algorithm but with different inputs.

Backbone Architecture. We define the backbone architecture (B_A) as follows. A stem convolution layer, i.e. a convolutional layer with 108 filters of size 3×3 stride 1 followed by ReLU and batch normalisation, is included immediately after the input layer. The architecture then contains several blocks. Each block consists of one or more normal cells and one reduction cell. The output of the final layer (final normal cell) is followed by Global Average Pooling (GAP) for reducing the feature map dimensionality. The final layer consists of two nodes for the two classes followed by SoftMax for classification. Since the reduction cells are used as a pooling layer to reduce the feature map size by half, to control the output size, the backbone architecture in this study has two reduction cells determined by the input image size and the intention to avoid input vanishing. Figure 4 shows the proposed backbone architecture.

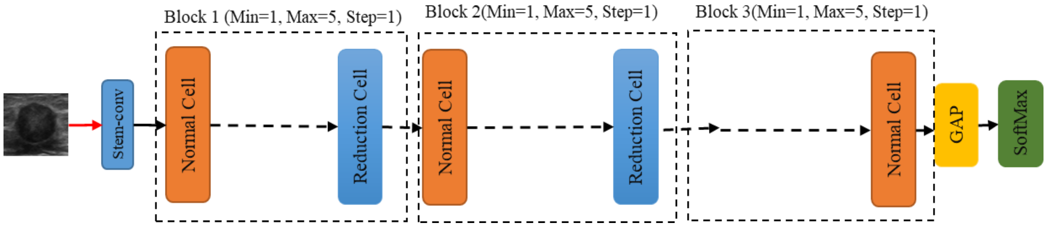
Block 1 (Min=1, Max=5, Step=1)
Block 2(Min=1, Max=5, Step=1)
Block 3(Min=1, Max=5, Step=1)
Stem-conv
Normal Cell
Reduction Cell
Normal Cell
Reduction Cell
Normal Cell
GAP
SoftMax

Figure 4: Backbone architecture (BA) for Bayesian optimization search

Structural Hyperparameter Search Space. The structural search by Bayesian Optimization aims at utilizing the optimized normal and reduction cells within the backbone architecture (B_A). In fact, Bayesian Optimiser searches for the optimal number of normal cells in each of three blocks (Block 1 (d_1), Block 2 (d_2), Block 3 (d_3)) in Figure 4. Thus, the structural search space is the number of normal cells per block d_i. The search range for d_i is therefore defined as Min=1, Max= 5 and step=1. Given this setting, the deepest architecture may have 15 normal and 2 reduction cells, while the shallowest architecture 3 normal and 2 reduction cells. The full details of the Bayesian Optimization algorithm are presented in the following Search Strategy section.

Trainable Hyperparameters Search Space. A suitable search space of trainable hyperparameters is needed as the input for Bayesian optimizer to build the optimal CNN architecture. In this paper, the search space ($L_r, Opz, L_f, W_i, Drp, L_n, L_{2r}$) is composed by Learning rate L_r, Optimization Opz, Loss function L_f, Weight Initialization W_i, Dropout Rate Drp, Layer Normalization L_n, and regularization

L_{2r}. Based on the literature [12] and our knowledge in deep learning architecture design, the following values and ranges of hyperparameters for the search space are carefully defined: L_r: [0.00001, 0.001]; Opz: (Adam, SGD, RMSprop); [12]L_f: (Sparse Categorical Cross-Entropy (SCCE), Binary Cross-Entropy (BCE)); W_i: (He normal, Glorat normal); Drp: [0%, 90%]; L_n: (Batch Normalization, Group normalization (4)); and L_{2r}: [0.00001, 0.001].

Search Strategy: The Bayesian Optimisation is conducted in two sequential stages. Given B_A and our definition of the structural search space, Bayesian Optimiser first searches for the optimal number of N cells in each block. The Bayesian Optimisation algorithm consists of six steps. Step 1, a hyperparameter setting S_s is defined as one set of possible values of optimisable structural hyperparameters (d_1, d_2, d_3). Therefore, it is defined as $S_s = \{S_{s1}, \dots, S_{si}, \dots, S_{sj}\}$ where S_{si} is the value of the optimizable parameter i in the hyperparameter setting S_s and j is number of hyperparameters that are being optimised ($j = 3$ in the first search stage). Step 2, we define an objective function $f(S_s)$ as the validation accuracy (the model accuracy on the test set when modelling the backbone architecture with hyperparameter setting S_s) that is maximised at each iteration. Step 3, Bayesian optimiser randomly selects t number of hyperparameter settings known as the initial seed points that the Bayesian optimiser examines before starting the search process. We set $t = 3$ as illustrated in our experiment. Using three initial hyperparameter settings, Bayesian optimiser models the backbone architecture to calculate the objective function $f(S_s)$. In step 4, Bayesian optimiser builds the surrogate model $G(S_s)$ which is based on Gaussian Process Regression. Given the initialisation of the $G(S_s)$, Bayesian optimiser uses Expected Improvement as acquisition function to select the next hyperparameter setting in Step 5. Where the next hyperparameter setting $f(S_{si})$ with the highest expected improvement over the current best observed setting of the objective function is selected using the Expected Improvement as follow:

$$A_q(S_s) = E(max(G(S_s) - f'(S_s), 0)$$

where $f'(S_s)$ is the best observed point of the objective function and $G(S_s)$ is the posterior distribution of the surrogate model. At each iteration, the hyperparameter setting S_s that maximises the $A_q(S_s)$ is selected as the next setting for evaluation. The surrogate model is updated with the newly evaluated hyperparameter setting after each iteration. The search process is repeated for 30 iterations. The number of iterations were defined empirically in this study. Finally, in Step 6, the architecture that provides the highest classification accuracy from Bayesian Optimisation search was selected and named as ENAS-B-1. Figure 5 shows ENAS-B-1 architecture.

In the second stage of the optimisation, given the architecture of ENAS-B-1 as a backbone and our definition of the trainable search space, the Bayesian Optimiser algorithm searches for the optimal trainable hyperparameter setting. In particular, the same steps 1 to 6 are used, but $S_s = (L_r, Opz, L_f, W_i, Drp, L_n, L_{2r})$ and $j = 7$. This stage of the search results in an optimal CNN architecture ENAS-B. Section 4 will provide the details on the optimized trainable hyperparameters.

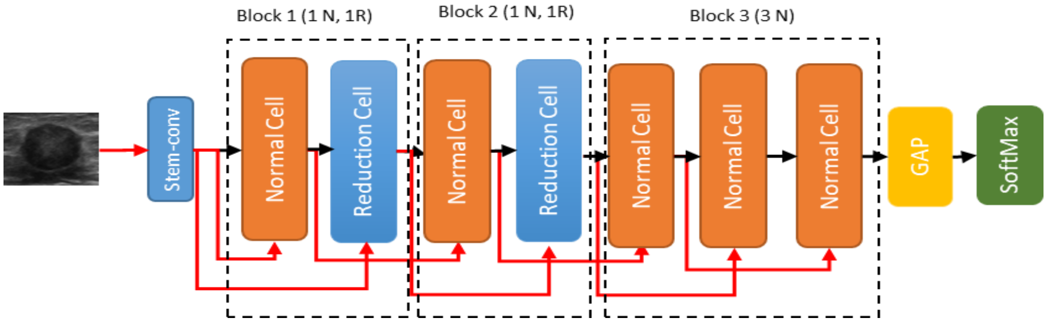

Figure 5: Fixed Backbone Architecture (ENAS-B-1) for trainable Hyperparameters Search stage

An interesting alternative of the two-stage search as described earlier is a combined search strategy where optimal combinations of the number of normal cells and the trainable hyperparameters are searched using the Bayesian Optimiser. We further explore this alternative search strategy and compare the searched architecture with ENAS-B. Further details of the architecture obtained will be shown in Section 4.

3.3. Experiment Setups

Experiments have been conducted to find the optimal CNN architectures and evaluate the classification performance of ENAS-B models. All experiments were run on a station with Intel Xeon(R) W-2102 CPU@2.90GHz with 16.0GB RAM. The Modelling Dataset (Section 3.1) was used for searching for the optimal cells in ENAS, searching for the optimal number of normal cells and trainable hyperparameters using Bayesian Optimisation, and finally training the ENAS-B model from scratch. A 5-fold stratified cross validation protocol was followed. At each iteration, the modelling data was split into 20% for testing (Internal test) and 80% for training. The training part was further split into 10% for validation and 90% for training. One split out of the five was used for the optimization (Section

3.2.2). To determine the classification error rates, all 5 folds were used. The imbalance ratio between benignity and malignancy (1.92:1) was upheld in the modelling and searching stages based on the findings in [14]. All images were pre-processed, and the training set enlarged using the pre-processing and data augmentation methods as described in 3.2.1.

Optimising ENAS-B Architecture: After ENAS generates the optimal cells, the Bayesian optimisation algorithm searched for 30 networks each of which was trained from scratch on the Modelling dataset with 50 epochs during the search. The primary criterion for selecting the optimal among the generated architectures is the validation accuracy, but the architecture complexity in terms of the number of weight parameters within the model is also considered. As a result, the optimal architecture has the block configuration of (1N, 1R, 1N, 1R, 1N) as depicted in Figure 5. Then, the search for trainable hyperparameters of the architecture is conducted under the following setting. The number of trials (sample model) is 30. Each model is trained on the unbalanced dataset for 50 epochs with batch size 8. The maximum batch size was constrained by the available computational power and the number of epochs was determined experimentally. The final ENAS-B architecture has the following hyperparameters: Learning Rate = 0.0001; Optimization function = SGD; Loss function = SCCE; Weight initialization = He Normal; Dropout rate = 0.3; Normalization Layer = Group Normalization; and L2 Regularization = 0.00036.

For the optimized architecture using the combined search strategy (ENAS-B Combined), the number of trial network models is set to 40. Each model is trained on the unbalanced dataset for 50 epochs with batch size 8. The final optimal architecture has the block configuration of <5N, 1R, 1N, 1R, 4N> with Learning Rate = 0.0001; Optimization function = Adam; Loss function = BCE; Weight initialization = He Normal; Dropout rate = 0; Normalization Layer = Batch Normalization; and L2 Regularization = 0.00042.

Model performance is measured by Sensitivity, Specificity, Accuracy and F1-score. Sensitivity refers to the proportion of known malignant test examples being classified as malignant, whereas Specificity refers to the proportion of known benign test examples being classified as benign. Accuracy refers to the proportion of correctly predicted test examples out of the total, and F1-score is the harmonic mean of Accuracy and Sensitivity.

4. Results

Breast Lesion Classification. The optimized ENAS-B is then trained from scratch on the unbalance dataset. All the data augmentation methods as mentioned are used to expand the training set. The number of epochs for training the EBAS-B models is set to 50. Figure 6 shows the loss-accuracy of ENAS-B training and validation.

Comparison with State-of-Art Purposely Built CNNs. We first compared ENAS-B with three existing state-of-art networks manually designed specifically for classifying breast lesions in US images, i.e. CNN3 [6], CNN4 [7], and Fus2Net [8]. Each CNN was trained and tested on the Modelling Dataset under the same cross validation protocol with the same folds as used for ENAS-B. As shown in Table 1, ENAS-B model outperforms all three network models by a large margin with higher overall accuracy of 4.5%, 18.8% and 13.3% respectively. ENAS-B also outperforms CNN3, CNN4 and Fus2Net by at least 6.6% when tested on the external sets A, B and C. With the new dataset (External D) [16], the results show that ENAS-B still achieved the highest overall accuracy of 67.4% (specificity 45.3% and sensitivity 89.5%) while CNN3, CNN4 and Fus2Net achieved overall accuracies of 61.5% (specificity 87.4% and sensitivity 35.7%), 53.7.5% (specificity 56.4% and sensitivity 50.9%), and 60.5% (specificity: 40% and sensitivity: 81.1%) respectively.

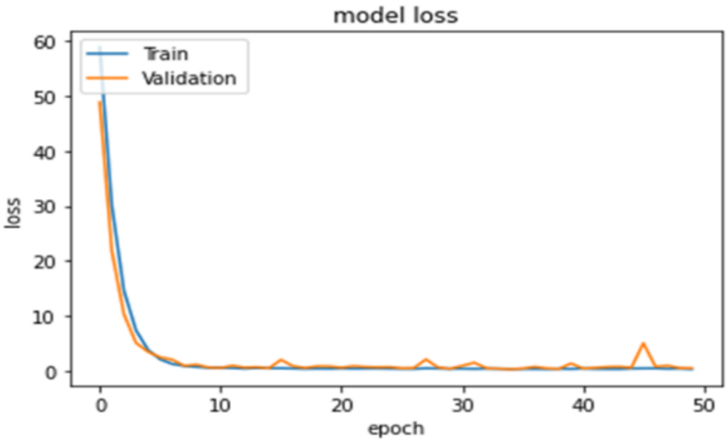

Figure 6: Training and Validation loss-accuracy of ENAS-B model

The classification performance of the ENAS-B models is presented in Table 1. With a 5-fold cross validation, ENAS-B achieved an average overall accuracy of 79.4% (88.2% specificity and 70.5% sensitivity). Further, we tested all 5 ENAS-

B models on the four external datasets (External A, B, C and D). ENAS-B generalizes well on the unseen data and achieved average accuracies (average of 5 models) of 80.4%, 89.7%, 78.0%, 67.4% on External A, B, C and D respectively.

Table 1: ENAS-B Performance and Comparison against Other State-of-the-art Breast Lesion CNNs

Models	Test Sets	Specificity	Sensitivity	ACC	F1	# Parameters
CNN3 [6]	Internal	88.7 ±6	61.1 ±13	74.9 ±5	66.0 ±8	619,202
	External A	78.8 ±4	68.1 ±12	73.5 ±4	66.2 ±7	
	External B	78.9 ±11	86.4 ±7	82.6 ±3	79.4 ±3	
	External C	73.6 ±10	71.0 ±12	72.3 ±1	73.9 ±5	
	External D	87.4 ±6	35.7 ±15	61.5 ±5	46.9 ±14	
CNN4 [7]	Internal	91.3 ±13	29.9 ±34	60.6 ±11	30.8 ±26	628,418
	External A	88.2 ±11	39.8 ±33	64.0 ±11	40.9 ±28	
	External B	89.3 ±18	39.4 ±34	64.4 ±9	42.8 ±25	
	External C	79.7 ±29	36.9 ±31	58.3 ±6	41.4 ±25	
	External D	56.4 ±45	50.9 ±44	53.7 ±6	41.7 ±31	
Fus2Net [8]	Internal	83.0 ±15	49.2 ±38	66.1 ±12	44.0 ±29	889,714
	External A	63.1 ±34	56.2 ±40	59.7 ±10	46.5 ±28	
	External B	84.9 ±14	64.3 ±42	74.6 ±15	59.1 ±33	
	External C	68.9 ±27	52.7 ±41	60.8 ±9	48.1 ±35	
	External D	40.0 ±19	81.1 ±22	60.5 ±5	68.3 ±11	
ENAS17 [11]	Internal	86.4 ±1.1	81.1 ±3.4	83.8±1.4	78.2 ±1.7	3,927,636
	External A	90.0 ±2.8	63.0 ±3.5	76.5±0.5	70.0 ±1	
	External B	84.9 ±4.3	89.3 ±3.4	87.1±1.4	84.3 ±1.7	
	External C	74.4 ±4	73.8 ±4.7	74.1±1.7	76.4 ±2.4	
	External D	55.2 ±9	81.9 ±6	68.6 ±2	74.8 ±1	
ENAS-B	Internal	88.2 ±2	70.5 ±4	79.4 ±1	73.0 ±2	1,053,398
	External A	88.8 ±4	72.0 ±8	80.4 ±2	75.2 ±2	
	External B	84.3 ±4	95.0 ±3	89.7 ±1	87.0 ±1	
	External C	75.6 ±8	80.4 ±3	78.0 ±3	81.0 ±1	
	External D	45.3 ±19	89.5 ±6	67.4 ±6	76.0 ±2	

We then compared ENAS-B with the original ENAS [11]. Based on our earlier findings as reported in [13], we chose ENAS17 for the comparison. Using the optimal cells as shown in Figure 3, ENAS17 architecture consists of 15 Normal cells (N) and two Reduction cells (R) in a configuration of (5N, R, 5N, R, 5N) and trained on the Modelling dataset under the same 5-fold cross validation protocol.

Although the ENAS17 models achieved higher accuracy in internal test, ENAS-B generalized better and achieved higher overall accuracy than ENAS17 on the external datasets except External D where ENAS17 has a marginally better overall accuracy. On the other hand, the number of weight parameters of ENAS-B is about 3.73 times fewer than that of ENAS17.

The performance of ENAS-B demonstrates the effectiveness of our approach in optimising the number of layers and trainable hyperparameters for accurate and robust networks. To confirm whether the differences in the model accuracies on external datasets are statistically significant, a paired sample t-test upon the ENAS-B model and each of CNN3, CNN4 and Fus2net were separately conducted, and the t-statistics and p-values were calculated. The p-values for ENAS-B vs CNN3, ENAS-B vs CNN4 and ENAS-B vs Fus2Net are respectively 0.000487, 0.001484 and 0.016456, all well below the general threshold of p = 0.05. Therefore, the ENAS-B model significantly outperforms the other manually designed CNN models.

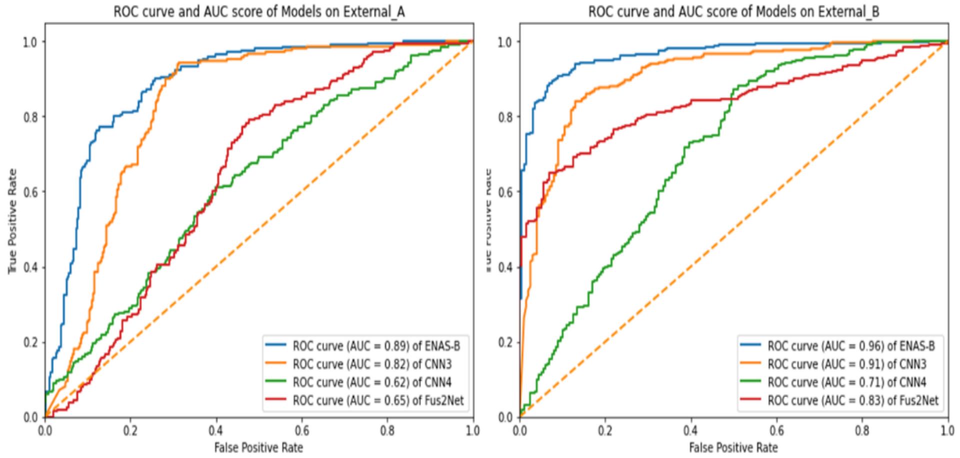

Figure 7: Presents ROC curve and AUC score of ENAS-B, CNN3, CNN4 and Fus2Net on External_A and External_B.

To further explore the predictive power of ENAS-B on external datasets, the Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) were calculated on the External datasets A and B because both datasets have more than 500 images collected from two different sources (see Section 3.1). For calculating the ROC curve different thresholds were used between 0 to 10. Figure 7 shows the ROC curves and AUC scores of ENAS-B (0.89 on External A and 0.96 on External B). AUC score in general demonstrates how well a classifier can discriminate between classes. The AUC scores have demonstrated that the capability of ENAS-B in distinguishing malignant lesions is better than CNN3, CNN4 and Fus2Net on both External A and B respectively. Moreover, we calculated the Delong test for ENAS-B against CNN3, CNN4 and Fus2Net on both External A and B. The DeLong value of ENAS-B vs CNN3, ENAS-B vs CNN4 and ENAS-B vs Fus2Net are respectively 0.3, 2.4 and 1.2 on External A, and 2.9, 4.9 and 3.4 on External-B. The DeLong test values all greater than zero indicate that ENAS-B different from those purposely designed CNN models [23] The larger the DeLong test statistic is, the stronger the evidence supporting the difference in AUC values between the two models. Since the DeLong test statistics of ENAS-B against the other models are mostly greater than 1, it implies that the ENAS-B model has a significantly higher AUC compared to the other models. Although, the DeLong value of 0.3 for ENAS-B vs CNN3 on External-A is less than 1 and closer to zero, the p-value for ENAS-B vs CNN3 is 0.000487. All the results indicate that ENAS-B has a significantly better performance than the other purposely designed CNN models.

Comparison with State-of-the-art Generic CNNs. It is also interesting to know how ENAS-B models compare with known CNN architectures originally designed for ImageNet. We selected some well-known architectures (VGG16[4], ResNet50 [24], InceptionV3 [25], MobileNet V2 [26], DenseNet [27], EfficientNetB0 [28], NasNet Mobile [29] and XceptionNet [30]) and then customized them for breast lesion recognition from US images. Both training the architectures from scratch and training them with transfer learning (TL) have been attempted. The number of epochs was set to 50 for the former, and 25 for the latter. The batch size was set as 16 for all the models in both situations. For fairness of the comparison, all the network models were trained on the Modelling Dataset under the same setting as for the ENAS-B models. Table 2 shows that ENAS-B achieved the highest overall accuracy on the internal tests except XceptionNet TL (with a small margin), and the highest average overall accuracy on the external tests.

Comparison with ENAS-B Combined. We further compare ENAS-B with ENAS-B Combined. Figure 8 summarizes the performance of the 12-layer ENAS-B Combined models on the internal test data and three external test datasets (without External D). Although the ENAS-B Combined models still perform better than all other purposely built CNNs, the performance is worse than that of ENAS-B for both internal and external tests.

Table 2: Comparison results of state-of-the-art and ENAS-B to classify US images of breast lesions (SP: Specificity; ST: Sensitivity; ACC: Overall Accuracy; F1: F1-Score): internal average vs external average

Network Models	**Test sets**	**CNN models from scratch**				**CNN models with TL**			
		SP	**ST**	**ACC**	**F1**	**SP**	**ST**	**ACC**	**F1**
VGG16	Internal	**100**	0	50	N/A	**100**	0	50	N/A
	External	**100**	0	50	N/A	**100**	0	50	N/A
Resnet50	Internal	74.5	50.4	62.5	57.4	87.1	48	67.6	54.5
	External	65.4	57.6	61.5	60.1	80.2	52.9	66.6	57.6
InceptionV3	Internal	84.2	48.9	66.6	53.6	94.7	32.2	63.4	40.2
	External	76.3	60.7	68.5	63.0	90.2	47.4	68.8	55.6
MobileNet V2	Internal	19.9	80.2	50	41	77.9	77.5	77.7	72.2
	External	20.0	80.1	50.0	49.3	67.4	87.7	77.6	78.1
DenseNet	Internal	83.5	52.4	68	58	94.2	42.8	68.5	51.1
	External	80.6	67.9	74.2	70.7	87.1	50.0	68.5	53.4
EfficientNet B0	Internal	82.9	58.4	70.7	63.1	87.3	65.5	76.4	68.3
	External	73.5	68.9	71.2	69.8	80.6	78.9	79.7	77.6
NasNetMobile	Internal	50.2	91.1	70.7	64.9	73.3	36.9	55.1	24.7
	External	34.1	**95.9**	65.0	69.9	70.1	37.7	53.9	27.2
XceptionNet	Internal	88.3	57.3	72.8	63.7	87.3	73.8	**80.6**	**74.4**
	External	82.3	67.4	74.9	71.2	77.5	**84.8**	81.2	79.8
ENAS-B	Internal	**88.2**	**70.5**	**79.4**	**73**	-	-	-	-
	External	**82.9**	**82.5**	**82.7**	**81.1**	-	-	-	-

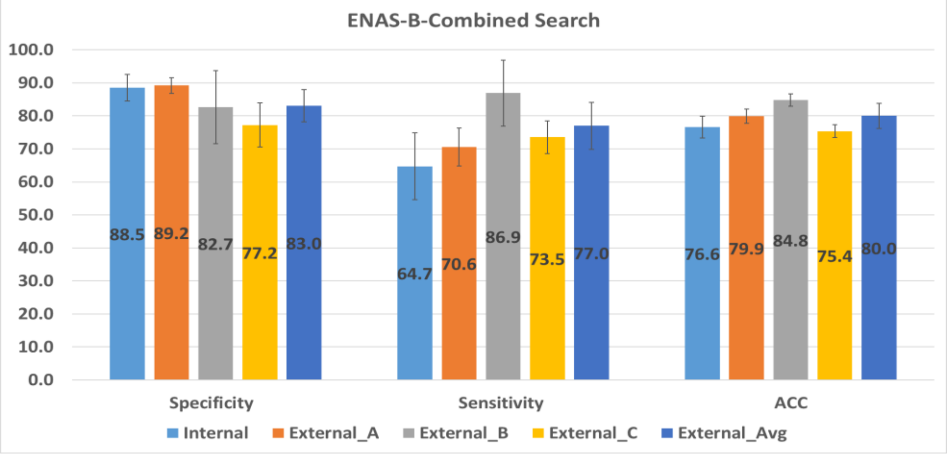

Figure 8: Performance of ENAS-B-Combined search on internal and external test sets

ENAS-B for Thyroid Cancer Classification. A pilot study was conducted by searching for an optimal ENAS-B architecture using breast lesion US images, and then train an ENAS-B model for thyroid nodule classification in two different scenarios using the same data augmentation methods to enlarge the training sets under a 5-fold cross validation evaluation framework. In the first scenario, a balance dataset of 500 ultrasound images (250 Benign and 250 Malignant) was used with the result showing that the ENAS-B architecture achieved the average overall accuracy of 73.6% (specificity: 54% and sensitivity: 93.2%) in classifying thyroid nodules. In the second scenario, following our approach of using unbalance classes, the ENAS-B models were trained on an unbalanced thyroid dataset (480 Benign and 250 Malignant) with ratio (1.92:1) with the results showing that the ENAS-B models achieved the average overall accuracy of 67.9% (specificity: 67.8% and sensitivity: 68%). In both scenarios, the specificity is close to random guess whereas sensitivity has substantial lifts. However, the potentials of transfer learning aspects of ENAS-B still require further research.

5. Discussions

The comparisons have revealed several advantages of ENAS-B over the existing approaches. First, the ENAS-B models outperform all exiting handcrafted networks purposely built for breast lesion classification from US images (Table 1). Second, the ENAS-B models in general maintain a smaller difference between sensitivity and specificity with more balanced performance on both classes (Tables 1 and 2). Furthermore, the ENAS-B models have much smaller number of weight parameters in comparison with ENAS17 (Table 1) and other known generic architectures as reported in the literature. Although the purposely build networks tend to be slimer, they underperform on bother internal and external tests (Table 1). ENAS-B also has its own limitations. Like all automatic search methods, ENAS-B requires resources to conduct searching and then training. It is worth noting that our two-stage approach for optimization has already reduced the demand on resources comparing to the combined search. Due to resource constraint, ENAS-B purposely controls the sizes of the search space by defining a backbone architecture framework, which might influence the best optimal outcomes.

The results of the comparison between the two search strategies (two-stage vs combined) show that the two-stage ENAS-B outperforms ENAS-B Combined by a margin of 3% on overall accuracy in the internal tests and nearly 5.5% better average sensitivity in the external tests while the average specificity remains marginally the same (Table 1 and Figure 8). Although the ENAS-B search principle is consistent with the ENAS's two-step principle [11], such finding is still surprising because a combined search space offers more hyperparameter combinations and hence should increase the possibility for finding the global optimum. It is possible that reaching an optimal CNN may require more iterations and hence prolong the overall time for searching.

Radiomics refers to the high-throughput analysis of quantitative image features for improving diagnostic accuracy in a clinical decision support system[31] For the rigor of studies and clinical relevance, a radiomics quality score (RQS) system was recently introduced in a landmark article. Although this paper is not a direct clinic-based study, it is desirable to evaluate the quality aspect of our study using the RQS score. After discarding 6 key components irrelevant to our study (key elements 4, 6, 7 11, 14 and 15), our study scored 16 out of 23 points on the remaining 10 key components. Although we have no direct control over the image acquisition and collection due to the data collection protocol agreed with the sponsor, US images from scanners of different makes and models were purposely collected and all lesions were cropped by experienced radiologists from their medical centres (see Section 3.1). Our deep learning approach follows an end-to-end workflow instead of examining each stage of image processing separately. The embedded convolutional operators optimally placed in a CNN architecture extract features at different levels of data abstraction. The ENAS reduction cells and the GAP layer are used for feature reduction. The performance of ENAS-B has been evaluated through internal and external tests, and various discrimination and calibration statistics have been used (Section 4). Although not all of our datasets are publicly accessible, one external test dataset (BUSI) is available from open sources. The ENAS methods for cell search are based on python codes in [32] and the Bayesian Optimization adapts the program codes in KerasTuner [33]. The radiomics analysis has also revealed the need for bringing our study closer to clinical practice. We therefore plan to conduct prospective tests in a clinical setting at the next phase of our investigation.

6. Conclusion and Future Work

This paper presented a novel framework for automatically searching CNN architectures for breast lesion classification from US images. We combined ENAS cell search with Bayesian Optimisation of network layers and trainable hyperparameters. The proposed framework yields efficient, shallow and robust CNN models that outperformed the state-of-the-art CNN models developed for the same purpose. The results show that cell structures, network depth and trainable hyperparameters are all important parameters to be optimised. Another finding is the importance of the search strategy. Evidence has shown that the two-stage approach (ENAS-B) allows the Bayesian optimiser to narrow the search and provide a robust CNN model. In the future, we plan to expand the search space by including other hyperparameters such as the number of filters, RoI margin size and the connectivity between cells. In addition, we plan to automatically optimize the depth and trainable hyper-parameters of existing CNNs such as ResNets, GoogleNet and MobileNet by using their blocks as the search spaces. We will further compare the classification accuracies of the ENAS-B model with expert radiologists on datasets from different sources. Finally, we plan to evaluate the performance of ENAS-B with pre-processed images to ensure that they have consistent characteristics.

Acknowledgement

This research work is sponsored by TenD AI Medical Technology.

7. References

[1] H. Sung *et al.*, ‘Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries’, *CA Cancer J Clin*, vol. 71, no. 3, pp. 209–249, 2021, doi: 10.3322/caac.21660.

[2] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, ‘Cancer statistics, 2023’, *CA Cancer J Clin*, vol. 73, no. 1, pp. 17–48, Jan. 2023, doi: 10.3322/caac.21763.

[3] Liu, S., Wang, Y., Yang, X., Lei, B., Liu, L., Li, S.X., Ni, D. and Wang., ‘Deep Learning in Medical Ultrasound Analysis: A Review’, Engineering, no. 157-22, 2019, doi: 10.1016/j.eng.2018.11.020.

[4] K. Simonyan and A. Zisserman, ‘Very Deep Convolutional Networks for Large-Scale Image Recognition’, pp. 1–14, 2014, [Online]. Available: http://arxiv.org/abs/1409.1556

[5] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., ‘Going deeper with convolutions’, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 07-12-June, pp. 1–9, 2014, doi: 10.1109/CVPR.2015.7298594.

[6] T. Xiao, L. Liu, K. Li, W. Qin, S. Yu, and Z. Li, ‘Comparison of Transferred Deep Neural Networks in Ultrasonic Breast Masses Discrimination’, *Biomed Res Int*, vol. 2018, pp. 1–9, 2018, doi: 10.1155/2018/4605191.

[7] B. Zeimarani, M. G. F. Costa, N. Z. Nurani, S. R. Bianco, W. C. De Albuquerque Pereira, and C. F. F. C. Filho, ‘Breast Lesion Classification in Ultrasound Images Using Deep Convolutional Neural Network’, *IEEE Access*, vol. 8, pp. 133349–133359, 2020, doi: 10.1109/ACCESS.2020.3010863.

[8] Ma, H., Tian, R., Li, H., Sun, H., Lu, G., Liu, R. and Wang,., ‘Fus2Net: a novel Convolutional Neural Network for classification of benign and malignant breast tumor in ultrasound images’, Biomed Eng Online, vol. 20, no. 1, pp. 1–15, 2021, doi: 10.1186/s12938-021-00950-z.

[9] B. Zoph and Q. V. Le, ‘Neural Architecture Search with Reinforcement Learning’, pp. 1–16, 2017, [Online]. Available: http://arxiv.org/abs/1611.01578

[10] Radhakrishnan, R. and Alzoubi, A., 2022. Vehicle Pair Activity Classification using QTC and Long Short Term Memory Neural Network. In VISIGRAPP (5: VISAPP) (pp. 236-247).

[11] H. Pham, M. Y. Guan, B. Zoph, Q. V. Le, and J. Dean, ‘Efficient Neural Architecture Search via Parameter Sharing’, 2018, [Online]. Available: http://arxiv.org/abs/1802.03268

[12] Qian, J., Li, R., Yang, X., Huang, Y., Luo, M., Lin, Z., Hong, W., Huang, R., Fan, H., Ni, D. and Cheng, J., ‘HASA: Hybrid Architecture Search with Aggregation Strategy for Echinococcosis Classification and Ovary Segmentation in Ultrasound Images’, vol. 00, pp. 1–17, 2022.

[13] M. Ahmed, H. Du, and A. AlZoubi, ‘An ENAS Based Approach for Constructing Deep Learning Models for Breast Cancer Recognition from Ultrasound Images’, *ArXiv*, 2020.

[14] M. Ahmed, H. Du, and A. AlZoubi, ‘Improving generalization of enasbased cnn models for breast lesion classification from ultrasound images’, in Annual Conference on Medical Image Understanding and Analysis, 2021.

[15] W. Al-dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, ‘Dataset of breast ultrasound images’, *Data Brief*, vol. 28, p. 104863, 2020, doi: 10.1016/j.dib.2019.104863.

[16] A. Abbasian Ardakani, A. Mohammadi, M. Mirza-Aghazadeh-Attari, and U. R. Acharya, ‘An open-access breast lesion ultrasound image database: Applicable in artificial intelligence studies’, *Comput Biol Med*, vol. 152, p. 106438, Jan. 2023, doi: 10.1016/j.compbiomed.2022.106438.

[17] Zhu, Y.C., AlZoubi, A., Jassim, S., Jiang, Q., Zhang, Y., Wang, Y.B., Ye, X.D. and Hongbo, D.U., ‘A generic deep learning framework to classify thyroid and breast lesions in ultrasound images’, Ultrasonics, vol. 110, p. 106300, 2021, doi: 10.1016/j.ultras.2020.106300.

[18] Khalili, N., Kazerooni, A.F., Familiar, A., Haldar, D., Kraya, A., Foster, J., Koptyra, M., Storm, P.B., Resnick, A.C. and Nabavizadeh, A., 2023. Radiomics for characterization of the glioma immune microenvironment. NPJ Precision Oncology, 7(1), p.59.

[19] L. J. Rich and M. Seshadri, ‘Photoacoustic imaging of vascular hemodynamics: Validation with blood oxygenation level-dependent MR imaging’, *Radiology*, vol. 275, no. 1, pp. 110–118, Apr. 2015, doi: 10.1148/radiol.14140654.

[20] X. Han, W. Cao, L. Wu, and C. Liang, ‘Radiomics Assessment of the Tumor Immune Microenvironment to Predict Outcomes in Breast Cancer’, *Front Immunol*, vol. 12, Jan. 2022, doi: 10.3389/fimmu.2021.773581.

[21] Mohammadi, A., Mirza-Aghazadeh-Attari, M., Faeghi, F., Homayoun, H., Abolghasemi, J., Vogl, T.J., Bureau, N.J., Bakhshandeh, M., Acharya, R.U. and Abbasian Ardakani, A., 2022. Tumor microenvironment, radiology, and artificial intelligence: Should we consider tumor periphery?. Journal of Ultrasound in Medicine, 41(12), pp.3079-3090.

[22] T. Hassan, A. Alzoubi, H. Du, and S. Jassim, ‘Towards optimal cropping: breast and liver tumor classification using ultrasound images’, in *Multimodal Image Exploitation and Learning 2021*, S. S. Agaian, S. A. Jassim, S. P. DelMarco, and V. K. Asari, Eds., SPIE, Apr. 2021, p. 15. doi: 10.1117/12.2589038.

[23] E. R. , D. M. D. and D. L. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, ‘Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach’, *Biometrics*, vol. 44, no. 3, p. 837, Sep. 1988, doi: 10.2307/2531595.

[24] K. He, X. Zhang, S. Ren, and J. Sun, ‘Deep residual learning for image recognition’, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, pp. 770–778, 2015, doi: 10.1109/CVPR.2016.90.

[25] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, ‘Rethinking the Inception Architecture for Computer Vision’, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, pp. 2818–2826, 2016, doi: 10.1109/CVPR.2016.308.

[26] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, ‘MobileNetV2: Inverted Residuals and Linear Bottlenecks’, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018, doi: 10.1109/CVPR.2018.00474.

[27] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, ‘Densely connected convolutional networks’, *Proceedings - 30th IEEE*

Conference on Computer Vision and Pattern Recognition, CVPR 2017, vol. 2017-Janua, pp. 2261–2269, 2017, doi: 10.1109/CVPR.2017.243.

[28] M. Tan and Q. V. Le, 'EfficientNet: Rethinking model scaling for convolutional neural networks', *36th International Conference on Machine Learning, ICML 2019*, vol. 2019-June, pp. 10691–10700, 2019.

[29] B. Zoph and Q. Le, 'Learning transferable architectures for scalable image recognition', *Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 8697–8710, 2018.

[30] F. Chollet, 'Xception: Deep learning with depthwise separable convolutions', *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 1800–1807, 2017, doi: 10.1109/CVPR.2017.195.

[31] L, P., Leijenaar, R.T., Deist, T.M., P, J., De Jong, E.E., Van Tim, J., S, S., Larue, R.T., Even, A.J., A. and Van Wijk, Y., 'Radiomics: the bridge between medical imaging and personalized medicine', Nat Rev Clin Oncol, vol. 14, no. 12, pp. 749–762, Dec. 2017, doi: 10.1038/nrclinonc.2017.141.

[32] Hieu Pham*, Melody Y, Guan*, Barret Zoph, Quoc V. Le, and Jeff Dean, 'Efficient Neural Architecture Search via Parameter Sharing-github', *Github,* 2019.

[33] O'Malley, Tom, B., Elie, Long, James, Chollet, Fran, Jin, Haifeng, Invernizzi, Luca., 'KerasTuner', https://github.com/keras-team/keras-tuner, 2019.