IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 March 2024, accepted 25 April 2024, date of publication 1 May 2024, date of current version 9 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3395995

RESEARCH ARTICLE

Diagnosis of Multiple Sclerosis by Detecting Asymmetry Within the Retina Using a Similarity-Based Neural Network

REGAN CAIN BOLTON[1], RAHELE KAFIEH[2], (Member, IEEE), FERESHTEH ASHTARI[3], AND AMIR ATAPOUR-ABARGHOUEI[1], (Member, IEEE)

[1]Department of Computer Science, Durham University, DH1 3LE Durham, U.K.
[2]Department of Engineering, Durham University, DH1 3LE Durham, U.K.
[3]Isfahan Neurosciences Research Center, Isfahan University of Medical Sciences, Isfahan 8174673461, Iran

Corresponding author: Amir Atapour-Abarghouei (amir.atapour-abarghouei@durham.ac.uk)

ABSTRACT Multiple sclerosis (MS) is a chronic neurological disorder that targets the central nervous system, causing demyelination and neural disruption, which can include retinal nerve damage leading to visual disturbances. The purpose of this study is to demonstrate the capability to automatically diagnose MS by detecting asymmetry within the retina, using a similarity-based neural network, trained on optical coherence tomography images. This work aims to investigate the feasibility of a learning-based system accurately detecting the presence of MS, based on information from pairs of left and right retina images. We also justify the effectiveness of a Siamese Neural Network for our task and present its strengths through experimental evaluation of the approach. We train a Siamese neural network to detect MS and assess its performance using a test dataset from the same distribution as well as an out-of-distribution dataset, which simulates an external dataset captured under different environmental conditions. Our experimental results demonstrate that a Siamese neural network can attain accuracy levels of up to 0.932 using both an in-distribution test dataset and a simulated external dataset. Our model can detect MS more accurately than standard neural network architectures, demonstrating its feasibility in medical applications for the early, cost-effective detection of MS.

INDEX TERMS Optical coherence tomography, Siamese neural network, asymmetry, classification.

I. INTRODUCTION

Deep learning and its applications have received significant attention over the past few years. As an increasingly active research area, it has already made significant contributions to health and medicine. One open task in this field is the early diagnosis and detection of autoimmune diseases such as multiple sclerosis (MS). One primary difficulty is the need for doctors with specific expertise to screen for this disease as MS diagnosis is greatly influenced by clinical judgement [1]. Furthermore, it is often difficult for clinicians to assess MS early, as there is no typical pattern of early symptoms that applies to everyone [2]. Traditionally, neurologists will use an MRI scan to diagnose MS; however, this can be expensive [3] and invasive for patients due to the noisy, claustrophobic nature of the scan. Clinicians sometimes perform a lumbar puncture under local anaesthetic to provide extra information, though evidently, this can be uncomfortable for patients [4]. In this study, we address the difficulties and cost of early MS detection by leveraging recent advances in deep learning. Our work targets a research gap in the literature as there is little to no research using Siamese Neural Networks trained on OCT data for MS diagnosis, based on eye symmetry and similarity. We provide an investigation into how learning-based methods can be used to advance the screening processes currently in place, replacing traditional methods, using a similarity-based approach.

Overall, there is a need for development in the diagnostic work for MS, so that we can improve screening for this disease [5]. One approach to detecting optic neuritis (ON)

The associate editor coordinating the review of this manuscript and approving it for publication was Zhan-Li Sun.

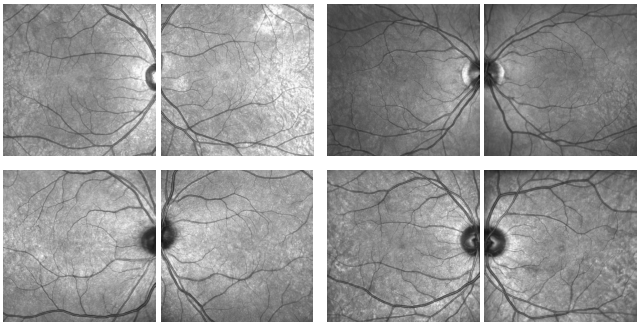

FIGURE 1. Examples of images from the left and right eyes. Images on top are of patients with MS and on the bottom without.

is by measuring the symmetry between cellular layers of the eye, obtained through an OCT scan. This is practical as our bodies are inherently symmetrical with paired organs and features such as hands, legs, lungs, etc., giving an approximate-bilateral symmetrical look (<99%), with minor differences between left and right samples [6], [7]. By using one eye as proxy for the other, clinicians can measure the degree of retinal symmetry as several statistical measures are determined to be lower in eyes with pathology, as indicated in past research [5], [8], [9], [10], [11]. A computing-based approach is deemed advantageous due to the inherent difficulty humans encounter when attempting to identify symmetry within skewed, misaligned, or poor-quality images. However, existing detection methods are not effectively leveraging retinal imagery due to their inability to accurately identify asymmetry. A Neural Network (NN) has the ability to recognise patterns where human perception falls short. This enables a more precise analysis of OCT images, as opposed to solely using human detection for MS [5]. Our methodology adheres to established medical theory; specifically, we focus on retinal symmetry as opposed to similarity (examples of data seen in Figure 1). This strategic approach ensures that our model is capable of accurately emulating a genuine diagnosis of MS following established medical practices.

In short, the primary contributions of our study can be summarised as follows:

- We propose several algorithms based on Siamese Neural Networks, which can be used for the early detection of neurological conditions such as MS (Section III-A). We eliminate poor performing models through a series of ablation and robustness experiments. (Section IV).
- Our experimental analyses demonstrate that our model is capable of outperforming standard neural network architecture, attaining accuracy levels of up to 0.932 (Section IV).
- Our analyses include measuring how well our model can identify symmetry and similarity between the eyes to follow clinical practices and provide additional explainability (Section IV-C).
- To demonstrate the robustness and adaptability of our model, we employ a simulated external test set that accurately reflects the diverse lighting and environmental conditions encountered in different Optical Coherence Tomography (OCT) machinery. This approach validates our model's effectiveness and underscores its practical applicability in real-world scenarios across various OCT systems. (Section IV-F).

To enable better reproducibility of our approach and to ensure wide-spread dissemination of our study, the source code is publicly available.[1]

II. RELATED WORK

We consider related work within three distinct areas, focusing on conventional methods for MS detection based on statistical features (Section II-A), MS detection through learning-based methods (Section II-B), and Siamese Neural Networks (Section II-C) and their applications (Section II-D).

A. STATISTICAL MS DETECTION METHODS

In medicine, the diagnosis of MS can be determined by investigating the textural information within the retinal nerve fibre layer [12]. One way of measuring this is through Dissemination in Space (DIS) and Dissemination in Time (DIT) [13]. The assessment of symmetry in the retina is conventionally quantified through the Inter-Eye Percentage Difference (IEPD) metric [14]. Recently, IEPD cutoffs are recommended to determine DIS and DIT [5], and have been revised in recent studies [9], [11], [15]. IEPD is usually measured through OCT scans since OCT has considerable advantages over MRI. OCT is an inexpensive, noninvasive, and accessible imaging technique that produces reliable quantitative measures for detecting pathology [16], [17], [18], [19], [20]. Research indicates that medical practitioners exhibit high precision in diagnosing patients afflicted with MS displaying both unilateral and bilateral ON. The downside is that this entails a significant amount of work due to its laborious, time-consuming, and manually intensive nature. Past research in this field demonstrates that a system to automatically define regions of interest is eagerly awaited [8].

B. DEEP LEARNING FOR MS DETECTION

One way of addressing this issue is through the use of learning-based approaches. Deep learning can significantly improve the diagnostic value of retinal OCT data, as explored in [5]. Another method to determine pathology in the eye is the Symmetry Index (SI). SI measures the degree of symmetry of the running pattern of the choridal vessels, relative to a horizontal line across a part of the eye called the fovea [10]. This has been used to detect Central Serous Chorioretinopathy (CSC) in the past. It has been shown that people with this disease have a lower SI than that of healthy eyes, thereby indicating its potential utility in the detection of other retinochoroidal diseases such as MS [10]. This reinforces the fact that ocular pathology can be detected through the use of binocular methods, using one eye as

[1] https://github.com/regybean/SiameseMSDiagnosis

a proxy for the other, as opposed to singular eye images. The most effective approaches in this area all utilise either conventional machine learning (ML) techniques or neural networks (NN) to ascertain the degree of asymmetry required to identify retinal pathology such as MS [7], [21], [22], [23], [24], [25], [26]. An approach taking advantage of Siamese Neural Networks excels at handling paired data, and thus has the potential to exceed traditional statistical measures as it considers all of the retina, rather than single measurements, such as the thickness of the retina.

C. SIAMESE NEURAL NETWORKS

A Siamese neural network (SNN) requires two inputs into two parallel convolutional neural sub-networks, which commonly share the same weights. The two networks are united by a similarity metric, and once fed through a classifier, the overall model can be used for various recognition-related tasks [27]. This architecture is well suited to the problem at hand as the two inputs correspond to the left and right eyes, permitting a binocular approach. Furthermore, the similarity metric (severity of ON) is learned implicitly, negating the need for manual threshold setting. SNNs are extremely useful for such problems, as they do not rely on a specific localisation of the pathology of interest, requiring only image level annotation and the binary comparison labels [26].

Overall, SNNs are inherently suited to this problem and have been extensively used in computer vision tasks [28], [29], [30], [31], [32], [33], [34] as well as speech and natural language processing [32]. In the person re-identification problem, for example, using a SNN has seen great success over other deep learning approaches [25], [35]. This success is accredited to the ability of the SNN to handle illumination variation, partial occlusions, and viewpoint invariance. This can be of value for the retinal images used in this study since OCT images are generally subject to imperfections due to the inherent nature of the imaging process. Notably, different OCT machines may exhibit varying levels of noise during image capture and the surrounding brightness can significantly affect the illumination of the images [36].

D. SIAMESE NEURAL NETWORKS IN MEDICINE

Within medical applications, SNNs are already being used extensively to diagnose and classify various diseases [37], [38]. Furthermore, within the existing literature [21], [22], [23], [24], SNNs have been used to ascertain symmetry within pairs of retinas, using fondus images or OCT images. These models are useful in detecting diverse forms of retinal pathology beyond MS, thereby providing further support for the feasibility of our study. One approach [21] applies an SNN to colour fondus images, to determine whether pairs of retinas can be used as a strong biometric. This study reinforces the idea that left and right retinas are approximately bilaterally symmetrical, up to a max accuracy of 89% on their network [21]. One takeaway from this work is that when weights are shared, compared to unshared, performance is observed to improve. In Section III-A, we discuss how weight sharing may effect the performance of the system, depending on whether it relies on symmetry or similarity for its analysis. Another similar study uses an SNN trained on fondus images, along with a test group, to determine similarity in pairs of eyes [23], with the SNN substantially outperforming human volunteers. This highlights the potential value of incorporating such models into clinical settings.

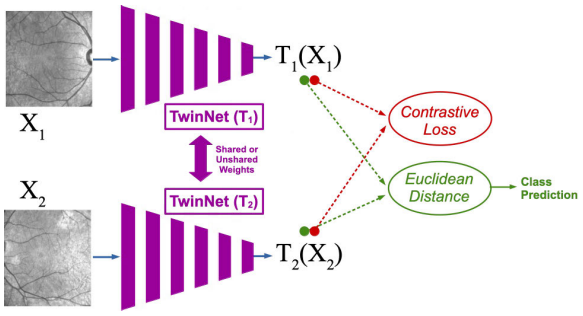

FIGURE 2. Proposed network architecture for the contrastive Siamese network, showing how input data, X_1 and X_2, are transformed into a binary class label and a loss output.

Two other studies [22], [24] use SNNs to detect Diabetic Retinopathy (DR), another retinal disease. The first study contrasts monocular and binocular approaches, giving the specificity for each as 70.7% and 82.2% respectively. This suggests that binocular methods are preferred for the detection of pathology [22]. An ML approach is considered in [24], though the feature extraction algorithms are determined to be complicated and difficult, so a deep learning approach is used instead [24]. Overall, the promising results achieved by SNNs within the area form the basis of our proposed approach towards MS diagnosis.

III. METHODOLOGY

Our methodology relies on the capability of Siamese Neural Networks to compare the similarity between the two output representations, given two images of the left and right eyes as the inputs. We propose two different ways of calculating this similarity between the two images. One method involves separately outputting the two feature representations, enabling a loss function to compute a similarity metric between them, utilising a distance function. An alternative approach involves concatenating both representations and passing them through a fully-connected layer with a single output. The key advantage is that the loss function is not required to calculate similarity, as the output can directly predict the desired class, taking advantage of an internal representation of similarity.

A. PROPOSED ARCHITECTURE

The contrastive Siamese network architecture can be seen in Figure 2. The network consists of two twin sub-networks, T_1 and T_2, each processing one of the two input samples, $\mathbf{X_1}$ and $\mathbf{X_2}$. The convolutional layers, denoted as f_{conv}, are responsible for extracting features from the input samples, while the fully-connected layers, denoted as f_{fc},

within each network generate a representation of that input sample. Mathematically, this can be described as $T_1(\mathbf{X_1}) = f_{fc}(f_{conv}(\mathbf{X_1}))$ and $T_2(\mathbf{X_2}) = f_{fc}(f_{conv}(\mathbf{X_2}))$. After processing the two input samples through their respective sub-networks, the output representations, $T_1(\mathbf{X_1})$, $T_2(\mathbf{X_2})$ and targets Y are fed into a distance or similarity metric function, L, which calculates the similarity score $L(\mathbf{Y}, T_1(\mathbf{X_1}), T_2(\mathbf{X_2}))$ between the two input samples, using a contrastive loss function. For similar pairs of input samples, the loss function tries to minimise the distance or dissimilarity between the output representations, while for dissimilar pairs, it tries to maximise this distance. The contrastive loss function trains the model such that dissimilar samples correspond to a positive case of MS.

The aggregated Siamese network architecture is illustrated in Figure 3. Once the convolutional layers f_{conv} extract the image features, the two image representations, $T_1(\mathbf{X_1})$ and $T_2(\mathbf{X_2})$, are concatenated into a single representation $(T_1(\mathbf{X_1}) \cdot T_2(\mathbf{X_2}))$ and then fed through the fully-connected layer in the network f_{fc} to give $\mathbf{Z} = f_{fc}(T_1(\mathbf{X_1}) \cdot T_2(\mathbf{X_2}))$. The fully-connected layer acts as a classifier, where linear layers reduce the dimensions of the concatenated representation down to a single metric, which determines the model's class prediction - either MS positive or MS negative. The subsequent loss function, L, can be any standard classification loss function, such as binary cross-entropy, applied as $L(\mathbf{Z}, \mathbf{Y})$. The key disparity between the two architectures lies in the approach taken to calculate similarity as each respective network has a different output.

Shared weights refer to a configuration in which both sub-networks in the SNN have identical weights. This means that the weights learned by each sub-network during training are exactly the same for each respective retina. Commonly, shared weights are appropriate for images that share similar properties, as they can learn a similar representation for each input. Instead of using shared weights, the two twin networks can be separate and independently learn representations for the left and right eye images individually. The disadvantage of using models with shared weights is that we use the same parameters for each eye, which means the model is inclined to find the same set of features in the left and right images as each eye is treated the same way. This model is likely to focus on the similarity between the two images rather than symmetry as they have no unique representations. On the other hand, the advantage of using unshared weights is that the models are free to extract their own features, for each eye, using their own individual parameters; therefore, symmetry is likely to be the primary focus rather than similarity (as is standard medical practice) [11], [15]. This is because the models will compare two different representations that are more symmetrical than similar. We analyse this further in our experiments in Section IV-C.

B. DATASET

Prior to conducting our experiments, the first challenge is to curate a dataset suitable for our neural network, as SNNs require pairs of inputs. Within the context of our problem, it is important that each data point only contains retinal imagery from a single patient.

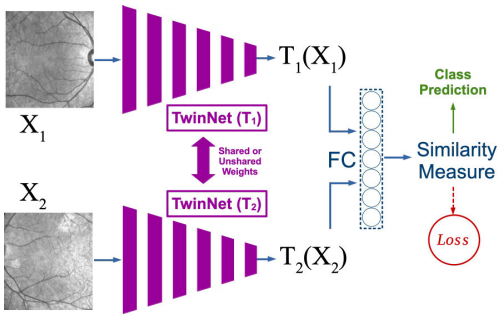

FIGURE 3. Proposed network architecture for the aggregated Siamese network, showing how input data, X_1 and X_2, are transformed into a binary class label and a loss output.

Consequently, each patient can be considered to have two data points, {Left, Right} and {Right, Left}, and one class label. In the case where the model has shared weights, the order of the images does not affect the model's interpretation of the data because symmetry is maintained, all passing through the same network. The benefit of this formulation is that we will have a larger set of training data. If a model has unshared weights, we only pass in a single data point for each patient {Left, Right} to ensure that each sub-network consistently receives the same type of input every time. For this same reason, none of our data augmentation techniques involve flipping an eye randomly, as it is imperative to preserve the individuality of left and right features. If both eyes are not mirror-symmetric, then CNNs could learn unexpected features from flipped images [6].

The OCT image dataset utilised in our study comprises 133 pairs of left and right retina, each having an associated MS positive or negative label. Due to constraints such as the necessity of acquiring both retinas and the limited availability of MS-positive retinas, the dataset is significantly small. To overcome this, we utilise data augmentation to force the model to learn more robust features and reduce overfitting.

We apply a range of data augmentation techniques, including affine transformations, such as rotation, translation, scale, and shearing with random parameters which help to simulate the nuances between different OCT machines and their calibration and will thus improve the generalisation capabilities of the model trained on augmented data. Another augmentation method we investigate is brightness and contrast shifts to account for illumination changes that may occur during the image capture process. Furthermore, the contrast change has the advantage of making blood vessels clear and more prominent in some of the images, which helps the model focus on the patterns more easily with greater emphasis on the relevant information within the images. Finally, we apply different levels of speckle and Gaussian noise to simulate the noise that occurs during image capture and increase the variability of the data for the model. Each

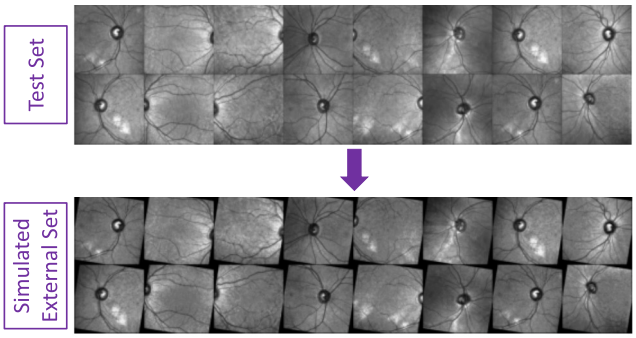

FIGURE 4. Construction of the simulated external set from the test set. Note the fixed changes in the levels of noise, brightness, contrast, and rotation.

augmentation technique is applied with a probability of 50%. We run extensive experiments to find the best augmentation techniques (Section IV-B).

While the model can be reliably tested using a classic train-test split, the generalisation capabilities of the model needs to be evaluated using external test data captured under different initial conditions. However, external data with our specific requirements, proves to be very rare in this domain. Hence, we fabricate a simulated external dataset that functions as a second test set for our experiments. This strategy enables us to measure the model's robustness to change. This simulated dataset uses the same images as the test set. However, the individual images have a different fixed level of brightness, contrast, noise, and rotation than that of the original set. The amount of transformations applied is selected from a random suitable range. We design this set in such a way that it reflects what OCT imagery may look like from other sources. Figure 4 shows how the samples are constructed, differing from our test samples.

C. TRAINING

The training process involves iteratively feeding training data to the network, and adjusting its parameters in order to minimise the loss function [39]. After constructing each model, their respective loss functions must be implemented differently as the number of outputs from each model differs.

A contrastive loss function is employed for our contrastive Siamese model, with the neighbours being pulled together and non-neighbours pushed apart [40]:

$$L(\mathbf{Y}, \vec{\mathbf{X}_1}, \vec{\mathbf{X}_2}) = (1 - \mathbf{Y})\frac{1}{2}(D_W)^2 + (\mathbf{Y})\frac{1}{2}\{max(0, m - (D_W)\}^2, \quad (1)$$

where $\vec{\mathbf{X}_1}$ and $\vec{\mathbf{X}_2}$ are representations from the input images, $\mathbf{Y}$ denotes the target, m is the minimal distance the points need to keep and D_W represents the distance between the two inputs, $D_W = \left\| \vec{\mathbf{X}_1}, \vec{\mathbf{X}_2} \right\|_2$.

For the aggregated Siamese model we used a standard binary cross-entropy loss.

$$L(\mathbf{x}, \mathbf{y}) = -mean(y_n \cdot \log x_n + (1 - y_n) \cdot \log(1 - x_n)), \quad (2)$$

where x_n is the euclidean distance between each two samples and y_n is the target for each sample.

Adam [41], RMSprop [42] and SGD [43] are used as optimisers in our experiments. We also make use of a scheduler to adjust the learning rate of the optimiser and help improve performance [44]. The different hyper-parameters are selected empirically, discussed in Section IV-A.

D. IMPLEMENTATION DETAILS

All implementation is done in PyTorch [45] and all models are tained on an NVIDIA GeForce RTX 2080 Ti GPU. Optuna [46] is used for hyper-parameter tuning, using a tree-structured parzen estimator (TPE) algorithm. Each image sample is resized to 300 × 300 and normalised. All networks fully converge after 20 epochs and in general take no longer than 2 minutes (Table 1) to fully train; therefore, we do not consider training time to be an issue. Models are evaluated by measuring the final accuracy, F_1 score, and AUC on the test set, and simulated external set.

IV. EXPERIMENTAL RESULTS

Numerous experiments are carried out to fine-tune and optimise the models for the best possible performance and all results are rigorously evaluated and discussed. Our experiments are conducted on four primary models, namely: unshared contrastive model (UC), which is our contrastive architecture in Figure 2 using separate weights for the two networks, shared contrastive model (SC) which is our contrastive model using shared weights, unshared aggregated model (UA), which is our aggregated architecture in Figure 3 without any shared weights and the shared aggregated model (SA). In order to be able to evaluate our results, we establish a baseline model that solely uses a single concatenated image of both eyes, processed through a single network. The architecture of the baseline is otherwise similar to that of our Siamese network. The dataset comprises 68 MS positive retina pairs and 65 MS negative retina pairs. We train and test using 89 and 44 retina pairs respectively (66:33 split) over 20 epochs with a batch size of 16, in all of our experiments, unless mentioned otherwise. Some of our experiments focus on improving the model by performing ablation experiments as seen in Sections IV-A, IV-B, IV-D, IV-E, in order to increase the model accuracy. Other experiments focus on improving the generalisability of the model so it is more effective when applied to different OCT imagery, captured under different conditions, subject to different levels of noise and light. These experiments can be found in Sections IV-F and IV-G.

A. HYPER-PARAMETER TUNING

Table 2 shows the result of our experiments to determine the optimal combination of hyper-parameters, utilising a sample

TABLE 1. The average runtime for a forward pass over 20 epochs. SA: Shared Aggregated; UA: Unshared Aggregated; SC: Shared Contrastive; UC: Unshared Contrastive. Showing the low complexity of the models.

Model	Average time for a single forward pass
Baseline	5.263
SA	5.369
UA	2.910
SC	5.314
UC	2.701

of 30 test cases. While a validation set is often used to tune hyper-parameters, the small number of images available in the dataset makes this very challenging. We mitigate this by testing for generalisability in later experiments (Section IV-F). The test offers a range of results, showcasing the accuracy, F_1 score, and AUC for each model, alongside their associated hyper-parameters. The baseline scores of 0.989, 0.988, 1.00 are presented, and the Siamese models perform worse with the next best model, Shared Aggregated (SA) model, achieving scores of 0.920, 0.909, 0.988 for accuracy, F_1 and AUC respectively.

This test was completed to get a general understanding of the performance we could achieve from each model. From Table 2, we can deduce that the best performing model is the baseline when the hyper-parameters are optimally fine-tuned for all models. This is due to the lack of diversity within the data as each sample is taken from the same OCT machine, from the same hospital. While the baseline appears to identify the patterns within this dataset well, it is likely to perform worse when applied to other data or augmented data (as demonstrated in Section IV-F). The next best models are the Shared Aggregated model (SA), and the Unshared Aggregated model (UA). The results clearly show that aggregated models generally outperform contrastive models. This indicates that outputting the distance rather than the image representations from the network is more suited to our problem. As contrastive models are expected to be inferior, we have omitted them from our subsequent tests.

B. DATA AUGMENTATION

This test shows the effect of using and removing different augmentations, based on the optimal values found in the previous experiment. We assess every combination of augmentations and present a range of results in Table 3. The random affine transformation performs a random rotation between $\pm$ 20 degrees, a random translation in x,y up to $\pm$ 1.5% of the images size, a random scale up to $\pm$ 10% of the images size, and a random shear up to $\pm$ 10 degrees in x,y. The brightness and contrast are randomly changed by up to $\pm$ 15%. Gaussian noise adds random noise at an amount of 0.2. Speckle noise adds noise at an amount of 0.2.

Data augmentation is not universally beneficial for all models, as the results of our experiments demonstrate that the performance of each model varies when augmented data is used. Using Table 3, we confirm our suspicion that the baseline performs poorly when trained on augmented data. In the presence of diverse data, the baseline model exhibits a worse performance, hence why the best performing metrics can be observed for the baseline where there is no data augmentation. Interestingly, UA also prefers no augmentation in terms of its performance. The spread of results we obtain from this experiment is shown to be tighter than that of the baseline, suggesting UA generally reacts better to augmentation. On the other hand, the SA model shows a significant improvement in performance with data augmentation techniques such as affine transformations, brightness & contrast adjustments, and the addition of Gaussian noise improving the model's performance.

C. SYMMETRY VS. SIMILARITY

In order to provide some explainability into the behaviour of our model, in this section, we attempt to identify whether the model is focusing on the symmetry of the eyes (as often done by medical practitioners [5]) or whether similarity between the pair of inputs dictates the output prediction. We measure how well the models can identify symmetry, by flipping one eye in the test set. This test effectively measures whether the model is focusing on symmetry as if symmetry features are learnt during the training process, then the model would be expected to perform reasonably well on a test where symmetry is enforced artificially.

It is important to explain the behaviour of the model so we can assess whether the approach is following common medical practice and measures interretinal symmetry for diagnosing MS, rather than similarity. UA achieved scores of 0.864, 0.842, 0.940, SA achieved scores of 0.784, 0.782, 0.871 and the baseline achieved scores of 0.807, 0.790, 0.942.

In a similar test, we measure the models' ability to identify similarity, where one of the images in the train and test set was flipped. This checks for whether the model is focusing on similarity. It is crucial to flip both images since not doing so would result in both symmetrical and similar features between the train and test sets, making it impossible to test which type of features the model was identifying. UA achieved scores of 0.795, 0.757, 0.940, SA achieved scores of 0.818, 0.765, 0.989 and the baseline achieved scores of 0.830, 0.800, 0.961.

Our symmetry/similarity tests identified that UA exhibited a focus on symmetry by -2.2% in accuracy - i.e. the model performs 2.2% worse in terms of accuracy in the symmetry test than in the similarity test. Conversely, SA performs with a focus on symmetry at a level of -19.3%. This suggests UA better focuses on symmetry overall as the difference between the standard dataset results is smaller. These findings are particularly relevant in medical contexts, where retinal symmetry is a critical diagnostic factor for MS. As a consequence of this, we would prefer a model that better analyses symmetry. Additionally, the similarity test demonstrates that SA places an emphasis on similarity by -15.9% whereas UA was lower at -6.8%. Larger differences in these results indicate a lesser focus on a given aspect.

TABLE 2. The best hyper-parameter tuning results. SA: Shared Aggregated; UA: Unshared Aggregated; SC: Shared Contrastive; UC: Unshared Contrastive. Demonstrating the general performance of the models.

Model	Test Accuracy	F_1 Score	AUC	Learning Rate	Loss	Optimiser
Baseline	0.989	0.988	1.00	2.61×10^{-4}	BCELoss	RMSprop
SA	0.920	0.909	0.988	3.89×10^{-5}	BCELoss	Adam
UA	0.886	0.872	0.986	4.4×10^{-4}	BCELoss	Adam
SC	0.886	0.889	0.963	3.3×10^{-4}	Contrastive	Adam
UC	0.818	0.818	0.853	1.83×10^{-5}	Contrastive	RMSprop

TABLE 3. Augmentation experiment results, demonstrating the most effective combination of data augmentation methods.

Model	Test Accuracy	F_1 Score	AUC	Affine	Contrast & Brightness	Speckle	Gaussian
Baseline	0.989	0.988	1.00	False	False	False	False
SA	0.977	0.976	0.996	True	True	False	True
UA	0.886	0.871	0.948	False	False	False	False
Baseline	0.886	0.868	0.983	False	True	True	True
SA	0.920	0.909	0.988	False	False	True	False
UA	0.841	0.811	0.963	False	True	False	False
Baseline	0.841	0.825	0.950	True	True	True	True
SA	0.864	0.846	0.973	True	False	True	False
UA	0.818	0.790	0.938	True	True	False	True

TABLE 4. A sample of layer freezing results for UA to show a performance increase.

Test Accuracy	F_1 Score	AUC	Number of Frozen Layers /64
0.932	0.923	0.975	24
0.909	0.905	0.938	34
0.909	0.900	0.983	62
0.841	0.829	0.946	10
0.727	0.714	0.865	2

TABLE 5. The results of the experiments to determine the best architecture, using optimal hyper-parameters.

Architecture	Test Accuracy	F_1 Score	AUC
ResNet50 [48]	0.955	0.952	0.988
EfficentNet v2 small [49]	0.932	0.927	0.963
DenseNet121 [50]	0.932	0.923	0.969
VGG11 [51]	0.909	0.895	0.971
AlexNet [52]	0.886	0.884	0.954
MobileNet v3 small [53]	0.773	0.792	0.832

These results confirm that SA focuses more on similarity over symmetry and UA focuses more on symmetry over similarity. These results are intuitively sensible since an unshared model can capture individual features for each eye due to them having entirely different parameters, allowing the model to focus more effectively on the symmetry between distinct features rather than the comparison of the same features in each eye. In light of this test, we work with only UA in subsequent tests to align with medical practices.

D. TRANSFER LEARNING

Another important area to explore to improve performance and robustness is the potential application of transfer learning within our training pipeline. Transfer learning makes use of pre-trained weights in model initialisation, allowing for the pre-learning of simple features such as lines and shapes. By doing so, if these simple features appear in the training data, then the model's weights are effectively pre-learned. This leads to a higher accuracy and is especially useful when only a very small amount of data is available, as is in our case. In this experiment, the weights from ImageNet1K [47] are used in our training process. Without transfer learning, UA gets scores of 0.886, 0.872, 0.986 for accuracy, F_1 Score and AUC respectively, but with the use of transfer learning, the model reaches scores as high as 0.932, 0.923, 0.975. The baseline does not improve using transfer learning, getting scores of 0.989, 0.988, 1.00 without and 0.920, 0.918, 0.973 with transfer learning.

We also experiment with freezing a number of shallow layers in the model to allow the network to learn task-specific features while retaining the useful generic features learned in the pre-training phase. Table 4 shows the results of this experiment. Using this technique, UA achieves scores of 0.932, 0.923, 0.975 for accuracy, F_1 and AUC respectively with 24 frozen layers. This improves the performance by around 5%, bringing it closer to the baseline accuracy, as the model can better focus on deeper layers which identify patterns associated with symmetry.

Interestingly, transfer learning tests demonstrate that the baseline does not improve with pre-trained weights. We can analyse that the baseline extracted different features to get to its outputs than that of the pre-trained model and therefore, its pre-trained weights are not useful. This means that the baseline is taking advantage of certain shortcuts within the existing datasets and cannot generalise to data from a different underlying distribution (a conclusion reinforced in Section IV-F).

TABLE 6. A sample of results of the baseline and UA models on the simulated external test set on random set of hyper-parameters.

Model	External Accuracy	F_1 Score	AUC	Learning Rate	Loss	Optimiser
UA	0.727	0.778	0.896	1.7×10^{-4}	DistanceContrastiveLoss	RMSprop
UA	0.750	0.792	0.925	3.3×10^{-5}	DistanceContrastiveLoss	Adam
UA	0.750	0.792	0.911	2.1×10^{-5}	DistanceContrastiveLoss	Adam
UA	0.750	0.792	0.889	1.8×10^{-4}	DistanceContrastiveLoss	Adam
UA	0.682	0.750	0.810	2.4×10^{-4}	DistanceContrastiveLoss	Adam
UA	0.750	0.792	0.925	1.0×10^{-4}	DistanceContrastiveLoss	Adam
UA	0.545	0.655	0.784	1.6×10^{-4}	DistanceContrastiveLoss	Adam
UA	0.477	0.646	0.783	2.9×10^{-4}	BCELoss	Adam
UA	0.477	0.646	0.783	2.9×10^{-4}	BCELoss	RMSprop
UA	0.477	0.646	0.783	2.9×10^{-4}	BCELoss	Adam
Baseline	0.477	0.646	0.783	2.9×10^{-4}	BCELoss	RMSprop
Baseline	0.477	0.646	0.576	6.2×10^{-5}	DistanceContrastiveLoss	RMSprop
Baseline	0.477	0.646	0.632	2.4×10^{-4}	DistanceContrastiveLoss	Adam
Baseline	0.477	0.646	0.578	9.8×10^{-5}	BCELoss	Adam
Baseline	0.477	0.646	0.576	1.1×10^{-4}	BCELoss	Adam
Baseline	0.477	0.646	0.576	2.2×10^{-4}	BCELoss	Adam
Baseline	0.534	0.226	0.576	1.5×10^{-4}	DistanceContrastiveLoss	Adam
Baseline	0.500	0.083	0.500	3.0×10^{-4}	BCELoss	Adam
Baseline	0.523	0.160	0.576	2.7×10^{-4}	BCELoss	RMSprop
Baseline	0.523	0.160	0.500	1.9×10^{-4}	BCELoss	Adam

TABLE 7. Results of our experiments on the simulated external test set with different architectures, using optimal hyper-parameters.

Architecture	External Test Accuracy	F_1 Score	AUC
UA(ResNet18)	0.932	0.933	0.986
Baseline(ResNet18)	0.455	0.625	0.705
ResNet50 [48]	0.568	0.688	0.867
EfficentNet v2 small [49]	0.777	0.808	0.913
DenseNet121 [50]	0.477	0.646	0.966
VGG11 [51]	0.795	0.816	0.957
AlexNet [52]	0.955	0.950	0.991
MobileNet v3 small [53]	0.682	0.682	0.725

E. NETWORK ARCHITECTURE

Swapping out the network architecture can have a significant impact on predictive performance as some architectures are better suited to different types of tasks. In this section, we perform extensive experiments to determine the best possible network architecture for our task. Table 5 shows a summary of the best results obtained using the different network architectures for UA. We perform hyper-parameter tuning on each pre-trained individual architecture with no augmentation, constructing 40 different test cases for each architecture. We primarily focused our investigation on smaller architectures [48], [49], [50], [51], [52], [53] as our dataset is small and a large number of learnable parameters will inevitably lead to overfitting. We can see ResNet50 achieves the highest scores of 0.955, 0.952, 0.988 for accuracy, F_1 and AUC respectively.

F. EVALUATING ROBUSTNESS

Robustness testing involves measuring how well the model adapts to alterations in the data to evaluate its generalisation capabilities. We measure this by modifying the data on which the model is trained and tested. To more accurately assess the model's ability to withstand changes, we include a simulated external test set in our experiments.

We first demonstrate that our model, has better generalisation capabilities over that of the baseline model, even in the case where hyper-parameters are completely random. Table 6 overviews how each model reacts to new simulated data. We can see that the baseline model continually fails to generalise to simulated external data, in most cases seemingly guessing at random, whereas the UA model sometimes is able to generalise. We can see that even if hyper-parameter tuning does not take place, the model still performs well and is capable of diagnosing MS compared to the baseline.

Secondly, it is important to verify the results of our experiments on the different architectures in Section IV-E, using the simulated external test set to evaluate the robustness of the architectures in the context of our problem. We use the same optimal parameters from Section IV-E and observe the simulated external test set accuracy for each architecture. We can then observe how different architectures cope with data simulated to be captured under different conditions. The results can be seen in Table 7. To measure the effectiveness of our model, we also assess the baseline simulated accuracy. We can observe AlexNet reaches a high score of 0.955, 0.950, 0.991 for accuracy, F_1 Score and AUC.

Experimenting with the different network architectures presents the challenge of selecting the best-performing architecture using our two test sets. AlexNet shows good performance on the simulated external test set but has a lower accuracy on the regular test set. This suggests it is generalising well but not generally performing well. Our experiments with other models, such as ResNet50, reveal a lower accuracy on the simulated external test set but a very high accuracy on the regular test set. This suggests that ResNet50 may have overfit to the training and in-distribution testing data, but failed to generalise well to out-of-distribution data, possibly due to its complexity. We determine that ResNet18 offers the best trade-off due to its reasonably high accuracy on both the simulated external test set and the regular test set. ResNet18 has the right number of parameters for our task and dataset and is a very well-rounded model,

TABLE 8. Results of k-fold validation on the regular and simulated external test sets for a better evaluation of model performance.

K value	Test Accuracy	External Test Accuracy	Test Accuracy	External Test Accuracy
	Baseline		UA	
2	0.789	0.489	0.910	0.752
3	0.872	0.489	0.910	0.683
5	0.857	0.490	0.940	0.699
7	0.827	0.489	0.932	0.632
9	0.888	0.496	0.932	0.671

TABLE 9. The best experimental results with tuned hyper-parameters on 50:50 split of the data to test robustness.

Model	Accuracy	F_1 Score	AUC	Learning Rate	Loss	Optimiser
UA	0.940	0.939	0.983	1.7×10^{-4}	DisContrastive	Adam
UA (External)	0.701	0.767	0.985			
Baseline	0.925	0.921	0.974	4.7×10^{-4}	BCELoss	Adam
Baseline (External)	0.492	0.660	0.759			

meaning it can perform well on out-of-distribution external data.

To test the overall capabilities of our approach, we also experiment with k-fold cross-validation. Table 8 shows the results on our regular and simulated external test sets with k-fold cross-validation, using different k values, for UA and baseline. This is done to gain a better understanding of the performance of the model over different splits of the data to help us more precisely interpret model performance.

Our k-fold validation confirms that the baseline model fails to achieve high accuracy on the simulated external test set, with less than 50% accuracy observed across all k values. In contrast, we see that UA consistently exhibits superior performance on the simulated external test set over all k values. This outcome is likely due to how the network is structured, making it ideal for diverse samples. Surprisingly, UA also outperforms the baseline model across all k values. With the changing split of data, the baseline once again continues to fall short in the face of our Siamese network, proving our model's continued robustness to change. Interestingly, our model even attains high accuracy levels when both datasets are used with lower values of k, demonstrating its ability to train on very small amounts of data, unlike the baseline.

G. 50:50 DATA SPLIT

Since the domain this study is focusing on generally suffers from data scarcity, it is important for any proposed approach to be able to perform well even when trained on a very small training dataset. In this section, we present the results of an experiment on a 50:50 spread of the test and training data after optimal performance is achieved via hyper-parameter tuning. The best results found for each model after hyper-parameter tuning can be seen in Table 9.

Interestingly, our 50:50 split test yields intriguing results, as UA surpasses the baseline in terms of accuracy on the test and simulated external test set by around 2% and 20% respectively. The UA model improves with the new split whereas the baseline performance decreased. These results, combined with the k-fold cross validation experiment, suggest that the baseline model tends to overfit and lacks generalisation capabilities. To make the baseline generalise and capture better features, we need a more diverse training dataset, which is extremely difficult and expensive to obtain.

V. CONCLUSION

Our study has investigated the effectiveness of detecting MS, through asymmetry within the retinas. We propose the use of a Siamese neural network, trained on OCT images to detect ON lesions and diagnose MS with high accuracy. Early diagnosis of MS can help alleviate symptoms and slow the progression of the autoimmune disease. With new techniques desperately needed for the diagnosis of MS [5], we have demonstrated the practicality of our model through numerous rigorous experiments. The proposed model is robust and accurate and can achieve accuracy levels of up to 0.932, even though the dataset available for training is rather small. Additionally, our approach has also been tested on a simulated external test set to evaluate its generalisation capabilities, which is highly important when training data is scarce. An effective automated MS detection model such as the proposed approach is cost-effective, only requires OCT imagery, and is very fast and efficient compared to a medical professional, being able to analyse many images per second. Additionally, such an automated diagnosis system can provide low-cost early warnings, which can significantly democratise access to essential healthcare, especially in areas where specialist services might not be readily available. Despite this, there could be challenges when applying this technique in real-world scenarios as neural networks often function in ways that are complex and difficult to interpret for human users, obscuring the understanding of their decision-making processes. This lack of transparency and explainability can be a significant issue in medical diagnostics and can face regulatory hurdles as well. Over time, other technical challenges could arise such

as an increased variability in OCT images due to factors like a patient's age or ethnicity, which can also affect the performance of the network. Additionally, the reliability of this approach also depends heavily on the quality and consistency of the images, requiring standardised imaging procedures to ensure accuracy of the overall system.

There is a vast scope for future work that could extend our research. Of course, the inclusion of a larger dataset for training would boost the performance of any automated MS detection approach. One important extension would be the use of another authentic dataset captured under different clinical and environmental conditions to evaluate the robustness of the model, rather than the simulation of one, as was used in our work. Other directions of future research could include specific region symmetry analysis within each retina as ocular pathology is more prevalent around the optic nerve and at the terminal of vascular vessels.

REFERENCES

[1] A. J. Solomon and B. G. Weinshenker, "Misdiagnosis of multiple sclerosis: Frequency, causes, effects, and prevention," *Current Neurol. Neurosci. Rep.*, vol. 13, no. 12, pp. 1–7, Dec. 2013.

[2] *Early Signs of MS*. Accessed: Aug. 10, 2023. [Online]. Available: https://www.mssociety.org.uk/about-ms/signs-and-symptoms/early-signs-of-ms

[3] G. R. Cherryman, "Cost of operating a nuclear magnetic resonance imaging system," *BMJ*, vol. 291, no. 6506, pp. 1437–1438, Nov. 1985.

[4] NHS Choices. (2022). *Diagnosis—Multiple Sclerosis*. [Online]. Available: https://www.nhs.uk/conditions/multiple-sclerosis/diagnosis/#:~:text=Diagnosing%20MS%20is%20complicated%20because,attack%20of%20MS%2Dlike%20symptoms

[5] A. Petzold, S. Chua, A. Khawaja, P. A. Keane, P. T. Khaw, C. Reisman, B. Dhillon, N. G. Strouthidis, N. J. Strouthidis, P. J. Foster, and P. J. Patel, "Retinal asymmetry in multiple sclerosis," *Brain*, vol. 144, no. 1, pp. 224–235, 2021.

[6] T. S. Kang, B. J. Kim, K. Y. Nam, S. Lee, K. Kim, W.-S. Lee, J. Kim, and Y. S. Han, "Asymmetry between right and left fundus images identified using convolutional neural networks," *Sci. Rep.*, vol. 12, no. 1, p. 1444, Jan. 2022.

[7] S. Biswas, J. Rohdin, and M. Drahansky, "Interretinal symmetry in color fundus photographs," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 1980–1983.

[8] J. Nij Bijvank, B. M. J. Uitdehaag, and A. Petzold, "Retinal inter-eye difference and atrophy progression in multiple sclerosis diagnostics," *J. Neurol., Neurosurgery Psychiatry*, vol. 93, no. 2, pp. 216–219, Feb. 2022.

[9] G. Bsteh, H. Hegen, P. Altmann, M. Auer, K. Berek, A. Zinganell, F. D. Pauli, P. Rommer, F. Deisenhammer, F. Leutmezer, and T. Berger, "Validation of inter-eye difference thresholds in optical coherence tomography for identification of optic neuritis in multiple sclerosis," *Multiple Sclerosis Rel. Disorders*, vol. 45, Oct. 2020, Art. no. 102403.

[10] H. Shiihara, S. Sonoda, H. Terasaki, N. Kakiuchi, T. Yamashita, E. Uchino, F. Murao, H. Sano, Y. Mitamura, and T. Sakamoto, "Quantitative analyses of diameter and running pattern of choroidal vessels in central serous chorioretinopathy by en face images," *Sci. Rep.*, vol. 10, no. 1, pp. 1–10, Jun. 2020.

[11] R. C. Nolan-Kenney et al., "Optimal intereye difference thresholds by optical coherence tomography in multiple sclerosis: An international study," *Ann. Neurol.*, vol. 85, no. 5, pp. 618–629, May 2019.

[12] A. Nunes, G. Silva, C. Alves, S. Batista, L. Sousa, M. Castelo-Branco, and R. Bernardes, "Textural information from the retinal nerve fibre layer in multiple sclerosis," in *Proc. IEEE 6th Portuguese Meeting Bioeng. (ENBENG)*, Feb. 2019, pp. 1–4.

[13] A. J. Thompson et al., "Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria," *Lancet Neurol.*, vol. 17, no. 2, pp. 162–173, Feb. 2017.

[14] D. Coric, L. J. Balk, B. M. J. Uitdehaag, and A. Petzold, "Diagnostic accuracy of optical coherence tomography inter-eye percentage difference for optic neuritis in multiple sclerosis," *Eur. J. Neurol.*, vol. 24, no. 12, pp. 1479–1484, Dec. 2017.

[15] R. C. Nolan, S. L. Galetta, T. C. Frohman, E. M. Frohman, P. A. Calabresi, C. Castrillo-Viguera, D. Cadavid, and L. J. Balcer, "Optimal intereye difference thresholds in retinal nerve fiber layer thickness for predicting a unilateral optic nerve lesion in multiple sclerosis," *J. Neuro-Ophthalmol.*, vol. 38, no. 4, pp. 451–458, 2018.

[16] D. Ontaneda and R. J. Fox, "Imaging as an outcome measure in multiple sclerosis," *Neurotherapeutics*, vol. 14, no. 1, pp. 24–34, Jan. 2017.

[17] S. C. Xu, R. H. Kardon, J. A. Leavitt, E. P. Flanagan, S. J. Pittock, and J. J. Chen, "Optical coherence tomography is highly sensitive in detecting prior optic neuritis," *Neurology*, vol. 92, no. 6, pp. 527–535, Feb. 2019.

[18] C. Kolluru, J. Lee, Y. Gharaibeh, H. G. Bezerra, and D. L. Wilson, "Learning with fewer images via image clustering: Application to intravascular OCT image segmentation," *IEEE Access*, vol. 9, pp. 37273–37280, 2021.

[19] J. Wang, "OCT image recognition of cardiovascular vulnerable plaque based on CNN," *IEEE Access*, vol. 8, pp. 140767–140776, 2020.

[20] H. Wei and P. Peng, "The segmentation of retinal layer and fluid in SD-OCT images using mutex dice loss based fully convolutional networks," *IEEE Access*, vol. 8, pp. 60929–60939, 2020.

[21] S. Biswas, J. Rohdin, and M. Drahansky, "Bilateral symmetry in central retinal blood vessels," in *Proc. 8th Int. Workshop Biometrics Forensics (IWBF)*, Apr. 2020, pp. 1–6.

[22] X. Zeng, H. Chen, Y. Luo, and W. Ye, "Automated diabetic retinopathy detection based on binocular siamese-like convolutional neural network," *IEEE Access*, vol. 7, pp. 30744–30753, 2019.

[23] S. Biswas, J. Rohdin, A. Biswas, and M. Drahansky, "A study of bilateral symmetry in color fundus photographs," *IEEE Access*, vol. 9, pp. 109624–109651, 2021.

[24] T. Liu, Y. Chen, H. Shen, R. Zhou, M. Zhang, T. Liu, and J. Liu, "A novel diabetic retinopathy detection approach based on deep symmetric convolutional neural network," *IEEE Access*, vol. 9, pp. 160552–160558, 2021.

[25] A. Subramaniam, M. Chatterjee, and A. Mittal, "Deep neural networks with inexact matching for person re-identification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 2675–2683.

[26] M. D. Li, K. Chang, B. Bearce, C. Y. Chang, A. J. Huang, J. P. Campbell, J. M. Brown, P. Singh, K. V. Hoebel, D. Erdoğmuş, S. Ioannidis, W. E. Palmer, M. F. Chiang, and J. Kalpathy-Cramer, "Siamese neural networks for continuous disease severity evaluation and change detection in medical imaging," *npj Digit. Med.*, vol. 3, no. 1, p. 48, Mar. 2020.

[27] J. Bromley, I. Guyon, and Y. LeCun, "Signature verification using a 'Siamese' time delay neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 6. San Mateo, CA, USA: Morgan Kaufmann, 1993, pp. 737–744.

[28] C. Fu, K. Lu, G. Zheng, J. Ye, Z. Cao, B. Li, and G. Lu, "Siamese object tracking for unmanned aerial vehicle: A review and comprehensive analysis," *Artif. Intell. Rev.*, vol. 56, no. S1, pp. 1417–1477, Oct. 2023.

[29] M. L. Battle, A. Atapour-Abarghouei, and A. S. McGough, "Siamese neural networks for skin cancer classification and new class detection using clinical and dermoscopic image datasets," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2022, pp. 4346–4355.

[30] J. Atanbori and S. Rose, "MergedNET: A simple approach for one-shot learning in Siamese networks based on similarity layers," *Neurocomputing*, vol. 509, pp. 1–10, Oct. 2022.

[31] W. Cui, W. Zhan, J. Yu, C. Sun, and Y. Zhang, "Face recognition via convolutional neural networks and Siamese neural networks," in *Proc. Int. Conf. Intell. Comput., Autom. Syst. (ICICAS)*, Dec. 2019, pp. 746–750.

[32] A. Nandy, S. Haldar, S. Banerjee, and S. Mitra, "A survey on applications of Siamese neural networks in computer vision," in *Proc. Int. Conf. for Emerg. Technol. (INCET)*, Jun. 2020, pp. 1–5.

[33] L. Zhao, Z. Shang, L. Zhao, A. Qin, and Y. Y. Tang, "Siamese dense neural network for software defect prediction with small data," *IEEE Access*, vol. 7, pp. 7663–7677, 2019.

[34] T. Jalonen, F. Laakom, M. Gabbouj, and T. Puoskari, "Visual product tracking system using Siamese neural networks," *IEEE Access*, vol. 9, pp. 76796–76805, 2021.

[35] Y. Haoran, "Face detection using the Siamese network with reconstruction supervision," in *Proc. 3rd Int. Conf. Inf. Commun. Softw. Eng. (ICICSE)*, Apr. 2023, pp. 44–50.

[36] J. M. Schmitt, S. H. Xiang, and K. M. Yung, "Speckle in optical coherence tomography," *J. Biomed. Opt.*, vol. 4, no. 1, pp. 95–105, 1999.

[37] M. S. A. Toofanee, S. Dowlut, M. Hamroun, K. Tamine, V. Petit, A. K. Duong, and D. Sauveron, "DFU-SIAM a novel diabetic foot ulcer classification with deep learning," *IEEE Access*, vol. 11, pp. 98315–98332, 2023.
[38] A. Vashishtha, A. K. Acharya, and S. Swain, "Automatically detection of multi-class Alzheimer disease using deep Siamese convolutional neural network," in *Proc. Int. Conf. Advancements Smart, Secure Intell. Comput. (ASSIC)*, Nov. 2022, pp. 1–7.
[39] S. Linnainmaa, "The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors," Ph.D. dissertation, Dept. Comput. Sci., Univ. Helsinki, Helsinki, Finland, 1970.
[40] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2006, pp. 1735–1742.
[41] Y. LeCun and F. J. Huang, "Loss functions for discriminative training of energy-based models," in *Proc. Int. Workshop Artif. Intell. Statist.*, vol. 5, 2005, pp. 206–213.
[42] G. Hinton, N. Srivastava, and K. Swersky, "Neural networks for machine learning lecture 6A overview of mini-batch gradient descent," *Cited On*, vol. 14, no. 8, p. 2, 2012.
[43] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, no. 3, pp. 400–407, Sep. 1951.
[44] R. Ge, S. M. Kakade, R. Kidambi, and P. Netrapalli, "The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 14977–14988.
[45] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 8024–8035.
[46] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 2623–2631.
[47] F.-F. Li. (2006). *ImageNet*. [Online]. Available: https://www.image-net.org/
[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[49] M. Tan and Q. V. Le, "EfficientNetv2: Smaller models and faster training," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10096–10106.
[50] Y. Fu, J. Wu, Y. Hu, M. Xing, and L. Xie, "DESNet: A multi-channel network for simultaneous speech dereverberation, enhancement and separation," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Jan. 2021, pp. 857–864.
[51] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[52] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.
[53] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1314–1324.

REGAN CAIN BOLTON received the bachelor's degree in computer science from the University of Durham. He is currently an Aspiring Data Scientist.

RAHELE KAFIEH (Member, IEEE) is currently an Assistant Professor in bioengineering with Durham University. She is also an EPSRC Women in Engineering Ambassador. Her research interests include the application of artificial intelligence (AI) in biomedical image and signal processing, in collaboration with clinicians and NHS. With over 13 years of experience in health-related projects and medical data processing using AI, she has expertise in localizing anatomical features in medical images, classifying medical data for disease diagnosis, and extracting numerical features from medical signals and images using deep learning models. She has published over 120 research articles in the field. Furthermore, she has successfully secured multiple grants, including the UKRI MRC Ageing Research Development Award, Switzerland Research Seed Money, Einstein Forum Grant, National Institute for Medical Research Development Grant, and TÜBİTAK Grant.

FERESHTEH ASHTARI received the master's degree from the Attaw MS Center, Canada. She is currently a Professor in neurology with Isfahan University of Medical Sciences. She is also the Head of the Kashani MS Center.

AMIR ATAPOUR-ABARGHOUEI (Member, IEEE) received the B.Sc. and M.Sc. degrees, in 2008 and 2010, respectively, and the Ph.D. degree from Durham University, U.K. He is currently an Assistant Professor with the Department of Computer Science, Durham University. After that, he joined Newcastle University as a Postdoctoral Research Associate and subsequently a Lecturer in computer science with the School of Computing Science, Newcastle University. His work includes generalized high-impact GANomaly anomaly detection approach, which is now part of the Intel OpenVINO/GETI AI product line and has been mentioned in more than 40 international patents as the underpinning approach in industrial monitoring/sensing solutions. He has co-organized the highly-prized CVPR-NAS workshop and workshops at IEEE BigData (BDA4CID and BDA4HM).