2022 IEEE 13th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) | 978-1-6654-6316-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/IEMCON56893.2022.9946550

Cataract Detection and Grading Using Ensemble Neural Networks and Transfer Learning

Renato R. Maaliw III
College of Engineering
Southern Luzon State University
Lucban, Quezon, Philippines
rmaaliw@slsu.edu.ph

Alvin S. Alon
Digital Transformation Center
Batangas State University
Batangas City, Philippines
alvin.alon@g.batstate-u.edu.ph

Ace C. Lagman
Information Technology Dept.
FEU Institute of Technology
Sampaloc, Manila, Philippines
aclagman@feutech.edu.ph

Manuel B. Garcia
Information Technology Dept.
FEU Institute of Technology
Sampaloc, Manila, Philippines
mbgarcia@feutech.edu.ph

Marmelo V. Abante
Graduate School
World Citi Colleges
Quezon City, Philippines
dmva888@gmail.com

Rodrigo C. Belleza Jr.
College of Computing Studies
Manuel S. Enverga University
Lucena City, Quezon, Philippines
rodrigo.belleza@mseuf.edu.ph

Jose B. Tan Jr.
College of Computing Studies
Manuel S. Enverga University
Lucena City, Quezon, Philippines
jose.tanjr@mseuf.edu.ph

Roselyn A. Maaño
College of Computing Studies
Manuel S. Enverga University
Lucena City, Quezon, Philippines
roselyn.maano@mseuf.edu.ph

***Abstract*—Artificial intelligence-based medical image analysis promises an efficient and reliable diagnosis in today's healthcare. Traditional approaches for cataract screening by medical practitioners often results in subjectivity due to their varying levels of knowledge and expertise. Using transfer learning, ensembles of pre-trained convolutional neural networks, and stacked long short-term memory networks, we developed a non-invasive and streamlined pipeline for automatic cataract severity classification. Empirical results show that our proposed combined models of *AlexNet*, *InceptionV3*, *Xception*, and *InceptionResNetV2* using a weighted average algorithm produces 99.20% (normal vs. cataract) and 97.76% (normal to severe) accuracies compared to standalone models. Furthermore, the ensemble model reduces classification error rates by an average of 2.17%. This study has the potential to help doctors to specify the magnitude of cataract stages with highly acceptable precision.**

Keywords—combined models, image processing, long short-term memory networks, convolutional neural networks, machine learning, transfer learning

I. Introduction

Cataracts are the primary cause of visual impairment and blindness. Its occurrence clouds and prevents light from reaching the eye's crystalline lens, resulting in decreased visual function. According to World Health Organization (WHO), over 295 million (M) individuals worldwide suffer from this condition, with 41 M having a permanent and the rest (77%) with limited sight [1]. Statisticians reveal that by 2025, it will reach 43 M instances, bolstered chiefly by third-world countries that lack professional ophthalmologists. The startling numbers demonstrate that the eye care system has not improved significantly over the last decades, and there is still an urgent need to enhance its diagnosis promptly. It is treatable through early identification, averting costly surgical operations, yet it accounts for the lion's share (33%) of blindness cases more than glaucoma and diabetic retinopathy [2].

Due to the wide range of lesions, eye tones, dimensions, structures, and positions, cataract detection using various approaches is a challenging task. Visual acuity tests, although non-invasive, present inaccuracies due to a doctor's subjective experiences [3]. Meanwhile, slit-lamp and retro-illumination microscopy can provide needed accuracy, but it discomforts photophobic patients due to their high illumination [4]. A retinal exam also creates an uneasy experience where a doctor puts eye drops to dilate a patient's pupil to find eye blockage signs. Applanation tonometry is another alternative identification process through measuring the eye's fluid pressure [5]. Many people find the procedures mentioned above intrusive, complex, and expensive. Specialists today highly recommend the low-cost, non-invasive nature of fundus images as it provides lofty retinal structural details against other forms of imaging, resulting in an accurate prognosis. Cataracts are identified and graded into five categories using fundus images based on the percentage (degree) of obscured retinal space shown in Table 1 and Fig. 1.

TABLE I. Cataract Grades [6]

Retinal space obscurities (%)	Category
< 3	Grade 0 (normal)
4 to 35	Grade 1 (mild)
36 to 55	Grade 2 (moderate)
56 to 85	Grade 3 (pronounced)
> 85	Grade 4 (severe)

Nevertheless, there are many obstacles related to fundus photographs for cataract assessments. Numerous researches relied on complex, labor-intensive, and time-consuming manual feature extraction for its evaluation and classification. Moreover, ophthalmologists have varying evaluations even with considerable knowledge and experience. These deviations are rooted in an individual's subjectivity. Thus, an automated system for identifying and grading eye abnormality is necessary.

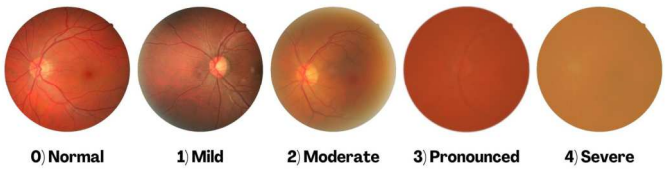

Fig. 1. Medical cataract grades based on retinal vascular obscurities.

Commission on Higher Education (CHED) – Institutional Development and Innovation Grants (IDIG)

Literature indicates considerable progress in classifying retinal illnesses [7] [8] due to multiple methods for cataract grading and detection using fundus images. Traditional methods for diagnosis relied heavily on optical coherence tomography (OCT) based on human-engineered features such as lens density, anterior surface curvature, and retinal opacity [9]. However, these features do not always correspond to clinical observations and may not be able to give an accurate diagnosis. Authors [10] obtained wavelet properties and utilized principal component analysis to minimize feature dimensions for machine learning (ML) categorization. The work of [11] used a histogram's gradient via a minimal distance classifier. Researchers [12] developed a supervised paradigm (extraction and weighting) for features such as wavelet, texture, and color using genetic algorithms (GA) and support vector machines (SVM). Several approaches classified different cataract grades from vitreous opacity using standard deviation and pixel-based structure visibility relying on decision trees. On the other hand, proponents [13] prioritized vessel vascular information for severity distinctions using template filter and SVM. These approaches select a subset of attributes in a single path, which is unsuitable for acquiring all of the intricacies of the input image, resulting in unpredictability. Experiments of [14] presented an enhanced Haar wavelet and added details through vertical, horizontal, and diagonal directional textures based on ensemble hard-voting classification exhibiting increased accuracy. Other pattern recognition is spectrum-based, utilizing two-dimensional discrete Fourier transform and linear discriminant analysis (LDA) with *AdaBoost* [15].

Deep learning (DL) gained prominence in the science of computer vision and image processing in recent years [16], and convolutional neural networks (CNN) are the most prevalent for the analysis of visual information. In contrast to standard ML models, it does not necessitate any user involvement throughout the feature extraction process. The study of [17] delivers cataract screening by getting local filters through patched-based clustering with CNN and recurrent neural networks (RNN) for retrieving higher-order features. They utilized support vector regression for grading. Author [18] presented a multi-layered CNN in evaluating cataracts, integrating feature maps from the architecture's pooling layers with time efficiency and accuracy of 93.52% and 86.69%. A complex proposal of [19] consists of a six-level classification employing a mix of CNN and random forests (RF) for assisting specialists in precisely understanding a patient's condition. The framework includes three modules yielding a 90.69% accurate extraction of fundus image characteristics. An article by [20] used the Res-Net classifier model for automatic cataract identification resulting in 95.77% accuracy, while a VGG-19 with a transfer learning approach obtained 97.47%. We noted that the numbers of existing works were based entirely on manual, traditional machine, and deep learning approaches. Current methods have significant limitations since eye specialists must confirm their vascular features, which is a tedious and time-consuming process based on their knowledge and experience subjectivity. In addition, robustness and generalization are compromised by using a small dataset.

Our scientific contribution is creating a new strategy to locate the eye's vascular features precisely by taking into account the long-distance dependencies of the fundus image and improving cataract identification with five compartmentalized grading using transfer learning and hybrid neural networks. This study's end-to-end pipeline can support ophthalmologists diagnose cataracts quickly and reliably with minimal physical involvement.

II. Methodology

The following section explains how to identify and assess cataract severity automatically - including dataset, preprocessing, data augmentation, transfer learning methods, CNN and RNN models, hyperparameter optimizations, ensemble technique, and evaluation metrics.

A. Data Acquisition and Image Preprocessing

We gathered 2500 high-resolution fundus photos from various cataract retinal archives [21 – 26]. These images are professionally annotated into five categories, ranging from 0 to 4 (see Table 1) based on severity levels, with 500 samples for each group. Using a reasonable and well-balanced dataset with huge samples is essential in improving training and validating deep learning-based models. With images compiled from different repositories, their sizes vary, and they are not adapted uniformly to the learning task. Thus, we downsized the images into equal sizes with 2048 x 2048 pixels. The luminance of the red-green-blue (RGB) image's channels was also normalized between 0 and 1 to ensure a consistent distribution before network training, hence accelerating convergence. Equation 1 depicts the *min-max* scaling:

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{1}$$

Where x and x' reflect the entire intensity range of 0 to 255 and normalized values (0 – 1) of the cataract images, while the $max(x)$ and $min(x)$ denote the original images' highest and lowest intensities. In our investigation, we discovered that the green channel (Fig. 2) makes image extraction straightforward and efficient; it gives more details due to luminosity and cuts computational time by a factor of 0.2512.

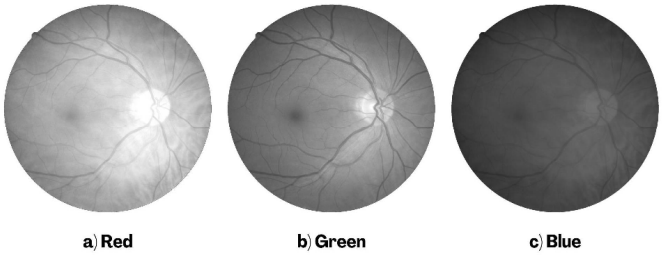

Fig. 2. Images in red (a), green (b), and blue (c) channels, where the green color space produces a lot of details for cataract feature extraction.

B. Data Augmentation

The lack of comprehensive medical image data for training is a significant obstacle in advancing pattern recognition, as it

directly impacts DL performance. Hence, to address the inadequacies, data enlargement or augmentation is necessary. We applied several geometric modifications, including resizing, rotating (0 to 180 degrees randomly), shifting (based on a reference point), and flipping (horizontal & vertical) to obtain an additional 10000 data for a total of 12500 images. Implementing this procedure is a viable solution to avoid overfitting and boost the network's predictive capability. Using 10-fold cross-validation, we split the dataset into 80/20 training and testing portions for modeling.

C. *Transfer Learning*

There is a strong correlation between the dataset's size and the quality of deep learning models. Nevertheless, in most cases, it is hard to handle huge samples of data effectively. To circumvent this problem, we employed a transfer learning (TL) mechanism pre-trained on large-scaled data such as ImageNet [27]. The premise of TL is to use previously accumulated information to enhance a model's performance to get better outcomes. Typical of a standard ML, TL-based methods are intended to handle particular concerns and other pertinent problems such as network retraining through hyperparameter optimization. It retains the primary network and leverages the pre-trained weights for its modification. Initialization weights of the network are continuously altered to acquire task-specific features. Several kinds of research [28] [29] have shown that fine-tuning methodologies apply aptly to various medical image classification applications. Fig. 3 illustrates the process flow of TL. Based on our experiment, we deployed four top-performing CNN architectures explained in the following sections.

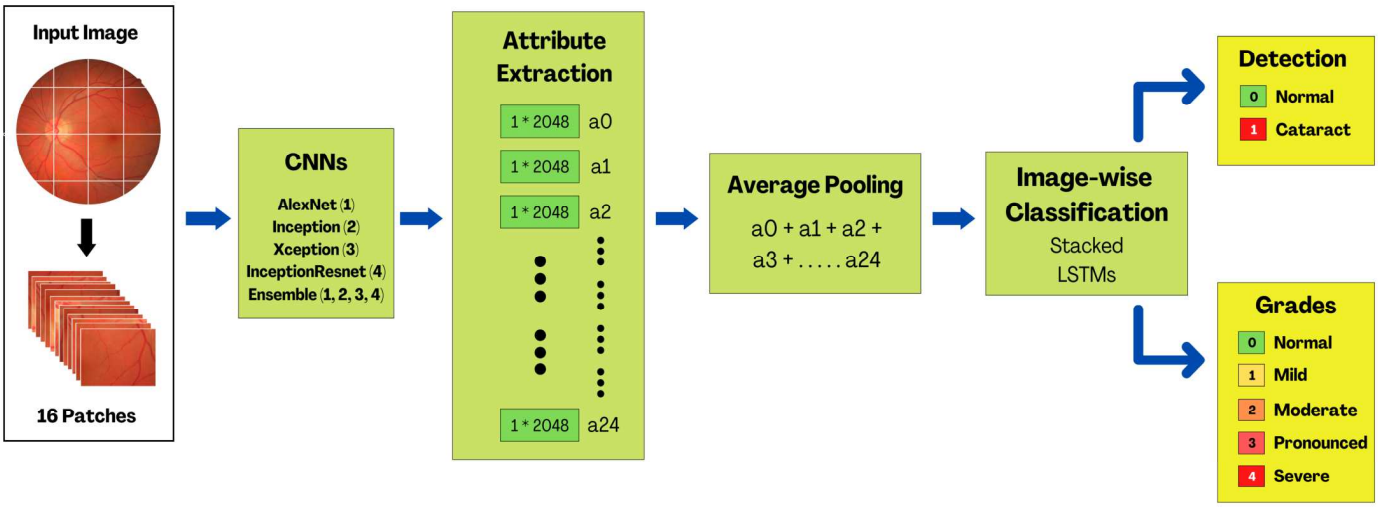

Fig. 3. Images are subdivided into 16 patches fed to various CNN models to extract features. A global average pooling combines the derived attributes and classified (image-wise) [41] by stacked LSTMs in terms of detection (normal vs. cataract) and grades (normal to severe).

D. *AlexNet*

It was the first CNN architecture to incorporate extra layers to a DL network, making it a superior design. It signifies that it has superb image-learning capability than prior models, which were confined only to one or two layers [30]. In addition, it exceeded its predecessors by substituting sigmoid functions in the hidden layers with rectified learning units (ReLU) that are computationally efficient while using smaller memory. It achieves the feats through the following. First, the five convolutional layers are the essential element of its structure since they help identify patterns through each pixel of an image which parts are meaningful and which is not. Second, the three max-pooling layers assist in identifying characteristics with fewer computations than the convolutional layers. Third, two fully connected layers with 0.5 dropout rates enable the network to forecast what object it believes to be present in an image based on all of its past estimations from the two previously stated layers. The architecture paved the groundwork for future advancements in computer vision by winning the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Fig. 4 shows the *AlexNet*'s constructs for this study.

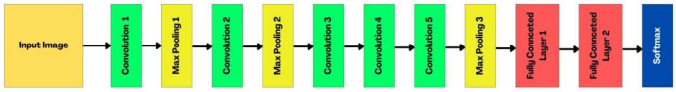

Fig. 4. AlexNet's architecture with five convolutions, three max-pools, two fully-connected layers, and a softmax activation function.

E. *Inception*

The network is a predecessor to the original inception design (*GoogleNet*), achieving superior efficiency through introducing inception modules - its most fundamental component. By reducing features using layered (1 x 1) convolution, the modules ensure faster calculations and deeper networks by utilizing auxiliary classifiers to address overfitting issues. The fundamental premise is to execute multiple filters with varying sizes concurrently rather than sequentially. Integrating an additional layer makes the architecture computationally cheap yet reliable [31]. We used a pre-trained *InceptionV3* [32] model with the configurations of 'imagenet' (weight) and input shape (224, 224, 3). The stack constitutes the global average pooling 2D, dense blocks, and batch normalization layers. We then configured two ReLU activation functions for each dense block to allow the model to learn swiftly, with better precision, and overcome the

vanishing gradients predicament. A simplified block diagram is shown in Fig. 5(a).

F. Xception

The model takes the *Inception* concepts to an extreme degree, transforming how we view neural nets and forming the backbone of all future network designs. It displaces modules with 'depthwise separable convolutions', consisting of a spatial convolution performed separately for each channel, followed by a 1 x 1 point-wise convolution across channels. This sensible design varies regarding operation sequence and non-linearity existence or absence. By integrating individual layers with subsequent output layers (shortcuts), a significant performance improvement is achieved not due to an increase in capacity but through the efficient utilization of parameters [33]. We purposively benchmarked an *Xception* model because of the limited implementation of cataract disease diagnosis on a large-scale dataset. Fig. 5(b) illustrates our research implementation of the architecture.

G. InceptionResNet

This network architecture is a hybrid of *Inception* and *ResNet (InceptionResNetV2)*, the two most renowned deep CNNs. Instead of summation, batch normalization is used for the convolutional layers. Leveraging the leftover modules enables a more significant number of *Inception* blocks, resulting in a system with outstanding precision and depth but additional computational costs [34]. As reported earlier, the training phase is the most evident issue associated with deep networks. Its conundrum is resolved using residual connections. While many filters are applied in a system, the residual is decreased to manage the training problem successfully. In an instance of more than one thousand strainers, residual fluctuations become volatile. Thus, the network cannot be adequately trained. As a direct consequence, residuals support the network in training stabilization. Fig. 5(c) represents our scaled implementation of the network.

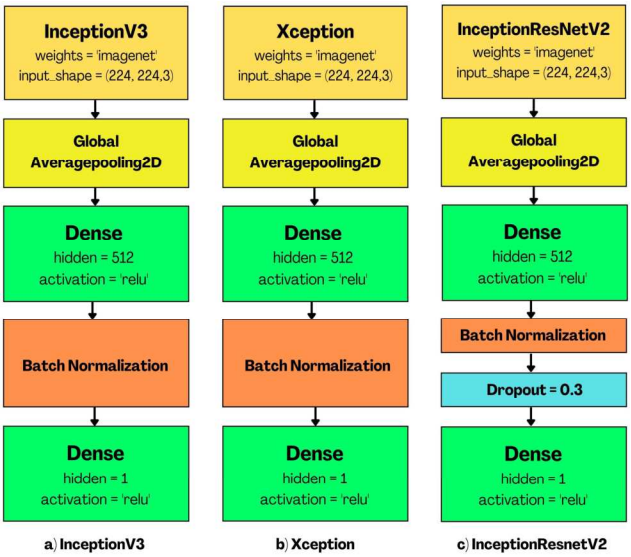

Fig. 5. Block diagrams of InceptionV3 (a), Xception (b), and InceptionResNetV2 (c).

H. Image-Wise Method

Images used in healthcare are often high-resolution photographs, which include and retain finer details. The DL pipeline used for training is supplied with these files in their entirety. Due to its enormous size, it is necessary to fragment original data into smaller pieces. The primary difficulty is combining each minor patches outcome before image classification [35]. The SVM and majority voting are the conventional solutions for this problem, which are straightforward and uncomplicated. Although this method feeds a single image into the network, it does not conserve the context of the entire picture. To preserve the contextual information, we proposed context-aware learning. It is a procedure that flattens numerous features into a single vector received from patch-wise implementation, which has a drawback for spatial characteristics. In the patch likelihood fusion method, the first level patch-wise network pulls out spatial attributes, and then the image-wise network accomplishes categorization. The approach has the foremost disadvantage of not preserving the remote contextual information.

We used a recurrent neural network (RNN) for classification in confluence with a CNN to keep track of the contextual information to fix the stated problem. Our suggested strategy is to obtain patch-wise features using CNN while the RNN collects dependencies. For cataract detection and grading, we employed a bidirectional long-short term memory (BLSTM). It improves learning performance by traversing input sequences in both forward and backward orientations, hence extending the capabilities of standard LSTMs. Fig. 6 exhibits the structure of a BLSTM.

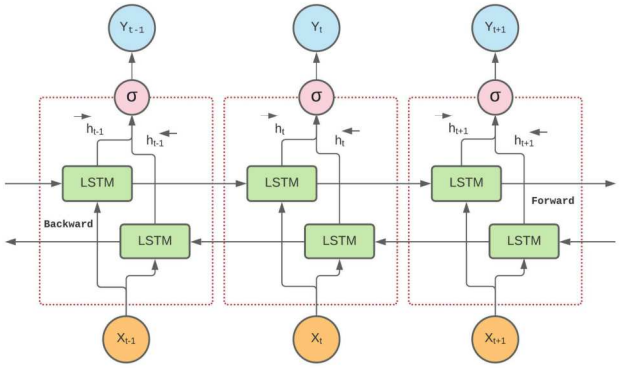

Fig. 6. A structure of a bidirectional long short-term memory network.

As headway, our study implemented a stacked LSTM (S-LSTM) for cataract detection and grading (Fig. 7). Increasing layers through stacking generate additional levels of data abstraction in decoding complicated sequences and classification tasks [36] [37]. It is achieved by integrating accumulated learned patterns throughout each layer as subsequent input to other LSTM layers. In this view, the BLSTM is a suitable option for the top layers of a DL model for the acquisition of helpful knowledge responsively (Fig. 7). In this research, we applied several CNNs to extract 25 vector features from a fundus image fed to an S-LSTM for image-wise cataract classification shown in Fig. 3.

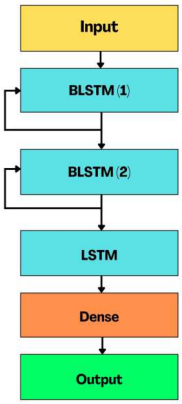

Fig. 7. Stacking configuration for cataract classification utilizing two bidirectional LSTMs and one unidirectional LSTM.

I. *Hyperparameter Optimizations*

For any machine learning to perform efficiently, hyperparameter tuning is critical. These configurations, unlike model parameters, are initialized before training. It is currently among the most daunting and neglected steps in implementing deep neural network modeling. For this study, we employed a sequential-based optimization technique [38] because of the manual configuration's complexity and time-intensive cost. Tables 2 and 3 convey the average calibrated settings for CNN and RNN models based on multiple run times.

TABLE II. NEURAL NETWORKS' OPTIMIZED HYPERPARAMETERS

Architecture	Configuration	Value
AlexNet	Learning rate	0.001
InceptionV3	Decay	0.001/epoch
Xception	Batch size	32
InceptionResNetV2	Shuffling	Per epoch
	Optimizer	ADAM
	Loss	Multiclass cross-entropy
	Epoch	120
	Environment	GPU

TABLE III. STACKED LSTM'S OPTIMIZED HYPERPARAMETERS

Training Data	Configuration	Value
Cataract Detection	Learning rate	0.001
	Batch size	16
	Dropouts	0.2
	Dense layer	1
	BLSTM (1) neurons	84
	BLSTM (2) neurons	72
	LSTM neurons	62
	Loss	Binary cross-entropy
	Optimizer	ADAM
	Epoch	200
	Activation function	RELU
Cataract Grading	Learning rate	0.001
	Batch size	16
	Dropouts	0.2
	Dense layer	1
	BLSTM (1) neurons	98
	BLSTM (2) neurons	84
	LSTM neurons	78
	Loss	Multiclass cross-entropy
	Optimizer	ADAM
	Epoch	200
	Activation function	RELU

Note: BLSTM (bidirectional), LSTM (unidirectional)

J. *Ensemble Techniques*

Ensemble modeling is an approach for determining outcomes using diverse base models. Adopting the technique diminishes prediction generalization errors. By aggregating the predictive capability of its members, it can deliver insights with more objectivity and accuracy. In addition, it minimizes bias and variance by assigning weights to attributes that contribute to reliability and robustness. We implemented a weighted ensemble algorithm, a unique form of average operation where outputs are multiplied by a weight and then linearly combined [39]. Each weight reflects the individual contribution of each model to the final output. Its distinction is that the weights are not predetermined, but their values are refined during the training phase. The only constraint is that their weight's sum must be equal to 1, but it is resolved quickly by applying a *softmax* function [40].

K. *Evaluation Criteria*

Accuracy by itself is insufficient for measuring a model's efficacy. In addition to the standard metric, we assessed the performance of different pre-trained deep learning architectures in terms of their precision, recall, specificity, Matthew's correlation coefficients (MCC), and confusion matrices. The succeeding equations summarize the evaluation criteria based on the number of TP (true positives), TN (true negatives), FP (false positives), and FN (false negatives).

$$Accuracy\ (AC) = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Precision\ (PR) = \frac{TP}{TP + FP} \tag{3}$$

$$Recall\ (RE) = \frac{TP}{TP + FN} \tag{4}$$

$$Specificity\ (SP) = \frac{TN}{TN + FP} \tag{5}$$

$$MCC = \frac{TN\ x\ TP - FN\ x\ FP}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{6}$$

III. RESULTS

We performed all tests using a high-end computer with the following specifications: Core i9-11900K processor (5 GHz & 16MB smart cache), 64GB DDR4 RAM, and an ASUS RTX3070 (1.73 GHz with 8GB DDR6) graphics card. Data preprocessing, augmentation, and neural network models were enforced with Python, TensorFlow, and Keras. The following sections detail the results.

A. *Layer Stacking Effects on Learning Stabilization*

Fig. 8 tracks the convergence loss of various S-LSTM variations versus the amount of epochs. Our experiment required more time for a stack of three LSTMs to reach equilibrium compared to lesser heaps. However, it can deliver stable results. We also observed that BLSTM on the primary layers (first and second) enhanced the model's performance.

Finally, a steadily diminishing number of neurons for each layer is beneficial (see Table 3) as the upper layers' task is to understand the overall input structure before transferring it to the successive layers for further processing.

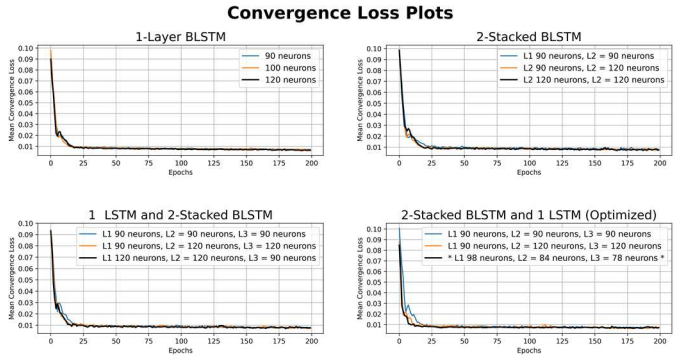

Fig. 8. Convergence loss plots of different stacking configurations.

B. Ensemble Model Weights

Table 4 presents each network's contribution to cataract detection and grading using a weighted average ensemble based on test data. The computation shows that *InceptionResnetV2* (in bold format) attains the highest throughput (0.52 & 0.45) and contributes almost half of the weight for the ensemble. This architecture implements residual connections on the network, which is essential when dealing with nonlinear patterns. *Xception* (0.22 & 0.25) and *InceptionV3* (0.14 & 0.21) shared moderately while the *AlexNet* (0.12 & 0.09) imparts the least weight.

TABLE IV. INDIVIDUAL MODEL'S CONTRIBUTORY WEIGHTS

Cataract Data (Training/Testing)	Network Contributions			
	AlexNet	InceptionV3	Xception	InceptionResNetV2
Detection [a] (2500/500)	0.12	0.14	0.22	**0.52**
Grading [b] (12500/2500)	0.09	0.21	0.25	**0.45**

a. 500 test data comprises a well-balanced dataset with 250 for each category (Normal and Cataract)
b. 2500 test data comprises a well-balanced dataset with 500 for each category (Grade 0 to 4)

C. Cataract Detection Performance

Table 5 proves that the ensemble trumps individual models with an overall accuracy of 99.20%. It also exceeded the predictive capability of other architecture with 99.60% (precision), 98.80% (recall), 99.60% (specificity), and 98.40% (MCC). The *AlexNet* ranked last, yet with satisfactory accuracy of 94.60%.

TABLE V. CATARACT DETECTION PERFORMANCE (NORMAL VS. CATARACT)

Model	Evaluation Metrics				
	Accuracy	Precision	Recall	Specificity	MCC
AlexNet [1→5]	0.946	0.970	0.917	0.978	0.894
InceptionV3 [2→5]	0.964	0.988	0.942	0.987	0.920
Xception [3→5]	0.976	0.992	0.961	0.991	0.952
InceptionResNetV2 [4→5]	0.980	0.992	0.968	0.991	0.960
Ensemble [(1,2,3,4)→5]	**0.992**	**0.996**	**0.988**	**0.996**	**0.984**

5. Stacked LSTM (see Fig. 7 and Table 3)

D. Cataract Grading Performance

The results in Table 6 demonstrate the ensemble superiority to single models with an accuracy of 97.76%. In addition, it surpassed the predictive ability of other networks by 97.78% (precision), 97.76% (recall), 99.43% (specificity), and 97.20% (MCC). Similar to cataract detection, the *AlexNet* performs the least with 94.44% grading accuracy.

TABLE VI. CATARACT GRADING PERFORMANCE (NORMAL TO SEVERE)

Model	Evaluation Metrics				
	Accuracy	Precision	Recall	Specificity	MCC
AlexNet [1→5]	0.9444	0.9449	0.9444	0.9861	0.9305
InceptionV3 [2→5]	0.9600	0.9607	0.9600	0.9890	0.9501
Xception [3→5]	0.9632	0.9634	0.9632	0.9908	0.9540
InceptionResNetV2 [4→5]	0.9712	0.9714	0.9712	0.9928	0.9640
Ensemble [(1,2,3,4)→5]	**0.9776**	**0.9778**	**0.9776**	**0.9943**	**0.9720**

5. Stacked LSTM (see Fig. 7 and Table 3)

We then quantify confusion matrices (see Table 7) to provide in-depth detail of each network's classification performances. It shows that most mean prediction errors (vice-versa) came from pronounced-severe (3.74%), followed by normal-mild (2.68%), mild-moderate (2.20%), and moderate-pronounced (1.03%). There are zero occurrences of misclassification for normal-moderate, normal-pronounced, normal-severe, mild-pronounced, mild-severe, and moderate-severe. This clearly shows each model's generalization robustness in discerning stages of cataracts, more importantly, the ensemble method.

TABLE VII. CONFUSION MATRICES FOR CATARACT GRADING (TEST DATA)

Ensemble					
	Normal	Mild	Moderate	Pronounced	Severe
Normal	489	11	0	0	0
Mild	8	490	2	0	0
Moderate	0	11	488	1	0
Pronounced	0	0	0	489	11
Severe	0	0	0	12	488
InceptionResnetV2					
	Normal	Mild	Moderate	Pronounced	Severe
Normal	485	15	0	0	0
Mild	7	487	6	0	0
Moderate	0	10	486	4	0
Pronounced	0	0	0	486	14
Severe	0	0	0	16	484
Xception					
	Normal	Mild	Moderate	Pronounced	Severe
Normal	483	16	1	0	0
Mild	9	482	9	0	0
Moderate	0	13	480	7	0
Pronounced	0	0	1	482	17
Severe	0	0	0	19	481
InceptionV3					
	Normal	Mild	Moderate	Pronounced	Severe
Normal	479	18	3	0	0
Mild	10	479	11	0	0
Moderate	0	16	476	8	0
Pronounced	0	0	0	479	21
Severe	0	0	0	23	477
AlexNet					
	Normal	Mild	Moderate	Pronounced	Severe
Normal	472	25	3	0	0
Mild	15	472	13	0	0
Moderate	0	19	471	10	0
Pronounced	0	0	0	474	26
Severe	0	0	0	28	472

E. Ensemble Training/Validation Loss and Accuracy

Fig. 9(a) indicates that the training loss commenced at an average of 0.57 logarithmic values, whereas the validation loss is at 0.75. An increasing convergence was achieved at epochs 60 to 100, and it stabilized at the 115th epoch. On the other hand, Fig. 9(b) demonstrates a gradual convergence at epochs 80 to 100, achieving the highest training accuracy of 98.77% (106th epoch) and validation accuracy of 98.68%. Both graphs empirically confirm that the ensemble model did not over or underfit in grading cataracts.

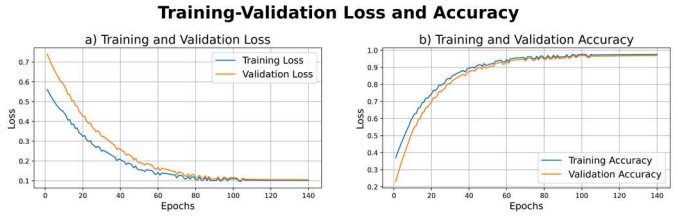

Fig. 9. Cataract classifications training & validation loss (a), and training & validation accuracy (b) for the ensemble model.

IV. Discussions

This study proved the adaptive capacity of compounded pre-trained CNNs, stacked LSTMs, and transfer learning in boosting the performance of cataract diagnosis and grading through advanced procedures. The quantitative findings indicate that the ensemble model transcends the *AlexNet*, *InceptionV3*, *Xception*, and *InceptionResnetV2* architectures, with mean collective prediction reliability for unseen data of 99.20% and 97.76%. Moreover, our pipeline cuts identification and categorization error rates by an average of 2.55% and 1.79%, respectively. In this experiment, we highlighted the integration of unique strengths and characteristics of various CNN models. The marked improvement in predictive power is the result of sequential-based optimization, layer stacking configurations, convergence loss plot, training-validation loss & accuracy analysis during the ensemble's training phase using a weighted average algorithm. Model optimization is time-consuming and expensive in terms of computational resources. Despite these disadvantages, we are convinced in our assertion that the benefits exceed the downsides. Our research's outcome advances the findings of [41 - 45], and we are confident that our deep learning-based medical image processing framework applies to other similar domains [46]. Like any research, we have experienced the challenges of noisy and insufficient fundus image quality (e.g., lousy lighting & luminance) contributing to classification divergences. This study did not investigate algorithms for offsetting these issues, such as image enhancements and reconstructions.

V. Conclusions And Future Work

Cataracts are the foremost contributor to vision impairments worldwide. If left undiagnosed and untreated, it can lead to irreversible and permanent blindness. Eye expert's planning decisions upon its treatment must be based on a timely, rapid yet dependable prognosis. Conventional procedures of cataract screening and assessment are laborious. It is a time-consuming mechanical practice vulnerable to disparities among doctor's subjective experiences for each case. In some instances, diagnosis proved to be difficult using only the naked eye due to the inadequate quality of fundus images.

We developed a minimally invasive end-to-end pipeline for identifying and classifying cataracts leveraging ensembles of pre-trained CNNs and an S-LSTM classifier with transfer learning augmentation. Our test results indicated superior accuracy and coherence between experts' cataract evaluations against the machine learning model with minor deviations. We contributed to the advancement of deep-learning medical image analysis by developing a sound and streamlined approach for the automatic recognition and grading of cataracts. This research has the potential to improve clinical practice in understanding cataract severity and their appropriate treatments. The proponents intend to incorporate image enhancement techniques and benchmark other CNNs in the future to improve the performance further.

Acknowledgement

The lead-author expresses his gratitude to Southern Luzon State University (SLSU), the Commission on Higher Education (CHED) for supporting this research, and to all contributing authors.

References

[1] X. Q. Zhang, Y. Hu, Z. J. Xiao, J. S. Fang, R. Higashita, and J. Liu, "Machine learning for cataract classification/grading on ophthalmic imaging modalities: A survey," Machine Intelligence Research, 19(3), pp. 184-208, 2022.

[2] S. R. Flaxman, R. R. Bourne, S. Resnikoff, P. Ackland, T. Braithwaite, M. V. Cicinelli, et. al, "Global causes of blindness and distance vision impairment 1990-2020: A systematic review and meta-analysis," Lancet Global Health, 5(12), pp 1221-1234, 2017.

[3] A. Li, Q. He, L. Wei, Y. Chen, S. He, Q. Zhang, and Y. Yan, "Comparison of visual acuity between phacoemulsification and extracapsular cataract extraction: A systematic review and meta-analysis," Annals of Palliative Medicine, 11(2), pp. 551-559, 2022.

[4] K. Y. Son, J. Ko, E. Kim, S. Y. Lee, M. J. Kim, J. Han, et. al, "Deep learning-based cataract detection and grading from slit-lamp and retro-illumination photographs: Model development and validation study," Ophthalmology Science, 2(2), pp. 1-9, 2022.

[5] J. G. Pearce, and T. Maddess, "The clinical interpretation of changes in intraocular pressure measurements using Goldmann applanation tonometry: A review," Journal of Glaucoma, 28(4), pp. 302-306, 2019.

[6] The Cataract Course, "Grading cataracts," An Online Resource For Learning About Cataracts, Available at: http://cataractcourse.com/cataracts-2/grading-cataracts/

[7] H. E. Gali, R. Sella, and N. A. Afshari, "Cataract grading systems: A review of past and present," Current Opinion in Ophthalmology, 30(1), pp. 13-18, 2019.

[8] I. Shaheen, and A. Tariq, "Survey analysis of automatic detection and grading of cataract using different imaging modalities," Applications of Intelligent Technologies in Healthcare, pp. 35-45, 2019.

[9] M. Rohlig, C. Schmidt, R. K. Prakasam, P. Rosenthal, H. Schumann, and O. Stachs, "Visual analysis of retinal changes with optical coherence tomography," The Visual Computer, 34(9), pp. 1209-1224, 2018.

[10] W. Fan, R. Shen, Q. Zhang, J. J. Yang, and J. Li, "Principal component analysis based cataract grading and classification," IEEE 17th

International Conference on E-Health Networking, Application and Services (HealthCom), pp. 459-462, 2015.

[11] M. Manchalwar, and K. Warhade, "Detection of cataract and conjunctivitis disease using histogram of oriented gradient," International Journal of Engineering and Technology, pp. 2400-2406, 2017.

[12] Z. Qiao, Q. Zhang, Y. Dong, and J. J. Yang, "Application of SVM based on genetic algorithm in classification of cataract fundus images," IEEE International Conference on Imaging Systems and Techniques (IST), pp. 1-5, 2017.

[13] Y. Dong, Q. Wang, and Q. Zhang, "Classification of cataract fundus image based on retinal vascular information," International Conference on Smart Health, Springer, pp. 166-173, 2016.

[14] L. Cao, H. Li, Y. Zhang, L. Zhang, and L. Xu, "Hierachical method for cataract grading based on retinal images using improved Haar wavelet, " Inf. Fusion, 53(1), pp. 196-2018, 2020.

[15] J. Zheng, L. Guo, L. Peng, J. Li, J. Yang, and Q. Liang, "Fundus image based cataract classification," IEEE International Conference on Imaging Systems and Techniques (IST), pp. 90-94, 2014.

[16] A. Imra, J. Li, Y. Pei, F. M. Mokbal, J. J. Yang, and Q. Wang, "Enhanced intelligence using collective data augmentation for CNN based cataract detection," International Conference on Frontier Computing, Springer, pp. 148-160, 2019.

[17] X. Gao, S. Lin, and T. Y. Wong, "Automatic feature learning to grade nuclear cataracts based on deep learning," IEEE Transactions of Biomedical Engineering, 62(11), pp. 2693-2701, 2015.

[18] L. Zhang, J. Li, I. Zhang, H. Han, B. Liu, J. Yang, and Q. Wang, "Automatic cataract detection and grading using deep convolutional neural network," IEEE 14th International Conference on Networking, Sensing and Control (ICNSC), pp. 60-65, 2017.

[19] J. Ran, K. Niu, Z. He, H. Zhang, and H. Song, "Cataract detection and grading based on combination of deep convolutional neural network and random forests," IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC), pp. 155-159, 2018.

[20] M. R. Hossain, S. Alfroze, N. Siddique, and M. M. Hoque, "Automatic detection of eye cataract using deep convolution neural networks (DCNNs)," IEEE Region 10 Symposium (TENSYMP), pp. 1333-1338, 2020.

[21] A. Budai, R. Bock, A. Maier, J. Hornegger, and G. Michelson, "Robust vessel segmentation in fundus images," International Journal of Biomedical Imaging, pp. 1-11, 2013.

[22] C. Hernandez-Matas, X. Zabulis, A. Triantafyllou, P. Anyfanti, S. Douma, and A. A. Argyros, "FIRE: Fundus image registration dataset," Model. Artif. Intell. Opthalmol., 1(4), pp. 16-28, 2017

[23] Z. Zhang, F. S. Yin, J. Liu, W. K. Wong, N. M. Tan, B. H. Lee, J. Cheng, and T. Y. Wong, "ORIGA-light: An online retinal fundus image database for glaucoma analysis and research," International Conference of the IEEE Engineering in Medicine and Biology, pp. 3065-3068, 2010.

[24] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, and F. Meriaudeau, "Indian diabetic retinopathy image dataset (IDRiD): A database for diabetic retinopathy screening research," Data, 3(3), p. 25, 2018.

[25] E. Decenciere, G. Cazuguel, X. Zhang, G. Thibault, J. C. Klein, F. Meyer, et. al, "TeleOphta: Machine learning and image processing methods for teleopthalmology," IRBM, 34(2), pp. 196-203, 2013.

[26] J. Staal, M. D. Abramoff, M. Niemeijer, M. A. Viergever, and B. van Ginneken, "Ridge-based vessel segmentation in color images of the retina," IEEE Transaction in Medical Imaging, 23(4), pp. 501-509, 2013.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in Neural Information Processing Systems, pp. 1097-1105, 2012.

[28] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and H. Qing, "A comprehensive survey on transfer learning," Proceedings of the IEEE, 109(1), pp. 43-76, 2021.

[29] M. T. Hagos, and S. Kant, "Transfer learning based detection of diabetic retinopathy from small dataset," arXiv:1905.07203, 2019.

[30] J. C. Tan, K. M. Lim, and C. P. Lee, "Enhanced AlexNet with super-resolution for low-resolution face recognition," IEEE 9th International Conference on Information and Communication Technology (ICoICT), pp. 302-306, 2021.

[31] M. Koklu, I. Cinar, Y. S. Taspinar, and R. Kursun, "Identification of sheep breeds by CNN-based pre-trained InceptionV3 model," IEEE 11th Mediterranean Conference on Embedded Computing (MECO), pp. 1-4, 2022.

[32] K. Liu, S. Yu, and S. Liu, "An improved InceptionV3 network for obscured ship classification in remote sensing images," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13(1), pp. 4738-4747, 2020.

[33] H. Benbrahim, and A. Behloul, "Fine-tuned Xception for image classification on tiny Imagenet," IEEE International Conference on Artificial Intelligence for Cyber Security Systems and Privacy (AI-CSP), pp. 1-4, 2021.

[34] S. Li, W. Yang, A. Zhang, H. Liu, J. Huang, C. Li, and J. Hu, "A novel method of bearing fault diagnosis in time-frequency graphs using InceptionResnet and deformable convolution networks," IEEE Access, 8(1), pp. 92743-92753, 2020.

[35] K. Nazeri, A. Aminpour, and M. Ebrahimi, "Two-stage convolutional neural network for breast cancer history image classification," International Conference on Image Analysis and Recognition, Springer, pp. 717-726, 2018.

[36] R. R. Maaliw, M. A. Ballera, Z. P. Mabunga, A. T. Mahusay, D. A. Dejelo, and M. P. Seño, "An ensemble machine learning approach for time series forecasting of COVID-19 cases," IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 633-640, 2021.

[37] R. R. Maaliw, Z. P. Mabunga, and F. T. Villa, "Time-series forecasting of COVID-19 cases using stacked long short-term memory networks," IEEE International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), pp. 435-441, 2021.

[38] R. R. Maaliw, K. A. Quing, J. A. Susa, J. F. Marqueses, A. C. Lagman, R. T. Adao, M. C. Fernando-Raguro, and R. Canlas, "Clustering and classification models for student's grit detection in e-learning," IEEE World Artificial Intelligence and Internet of Things Congress (AIIoT), pp. 39-45, 2022.

[39] T. Iqbal, and M. A. Wani, "COVID-19 and pneumonia detection using deep weighted ensemble model," IEEE 9th International Conference on Computing for Sustainable Global Development (INDIACom), pp. 337-340, 2022.

[40] X. Xu, L. Zhang, J. Li, Y. Guan, and L. Zhang, "A hybrid global-local representation CNN model for automatic cataract grading," IEEE Journal of Biomedical and Health Informatics, pp. 556-567, 2019.

[41] A. Imran, J. Li, Y. Pei, F. Akhtar, T. Mahmood, and L. Zhang, "Fundus image-based cataract classification using a hybrid convolutional and recurrent neural network," The Visual Computer, 37(8), pp. 2407-2417, 2021.

[42] M. S. Junayed, M. B. Islam, A. Sadeghzadeh, and S. Rahman, "CataractNet: An automated cataract detection system using deep learning for fundus images," IEEE Access, 9(1), pp. 128799-128808, 2021.

[43] T. Pratap, and K. Priyanka, "Computer-aided diagnosis of cataract using deep transfer learning," Biomedical Signal Processing and Control, 53(1), 101533, 2019.

[44] M. S. Khan, M. Ahmed, R. Z. Rasel, and M. M. Khan, "Cataract detection using convolutional neural network with VGG-19 model," IEEE World Artificial Intelligence and Internet of Things Congress (AIIoT), pp. 209-212, 2021.

[45] Y. Zhou, G. Li, and H. Li, "Automatic cataract classification using deep neural network with discrete state transition," IEEE Transactions on Medical Imaging, 39(2), pp. 436-446, 2019.

[46] R. R. Maaliw, J. B. Susa, A. S. Alon, A. C. Lagman, S. C. Ambat, M. B. Garcia, K. C. Piad, and M. C. Fernando-Raguro, "A deep learning approach for automatic scoliosis Cobb angle identification," IEEE World Artificial Intelligence and Internet of Things Congress (AIIoT), pp. 111-117, 2022.