International Research Journal in Global Engineering and Sciences. (IRJGES)
ISSN: 2456-172X | Vol. 4, No. 3, September - November, 2019 | Pages 87-107

An Artificial Intelligence Based Tool for Eye Disease Classification

Dr.S.Jagan,

Associate Professor, *Department of CSE, Agni College of Technology, Chennai, India*

Abstract

Human eye is affected by the different eye diseases namely Diabetic Macular Edema (DME) and Age- related Macular Degeneration (AMD). Diabetic Macular Edema (DME) is a common eye disease that causes an irreversible vision loss for diabetic patients. The AMD is further classified into Early AMD and Late AMD. DRUSEN is an eye problem caused due to aging and macular degeneration. It destroys our sharp central vision. The presence of DRUSEN is the symptom for Early AMD. Choroidal Neovascularization(CNV) is an eye problem caused due to the creation of new blood vessels in the choroid layer of the eye which leads to sudden deterioration of central vision. CNV is the symptom for the Late AMD. This work focuses on the design of an Artificial Intelligence based tool for eye disease detection and classification to detect and classify CNV, DME and DRUSEN effectively by using the Optical Coherence Tomography(OCT) images. Automatically identifying and describing the symptoms of an OCT image is really a complex and challenging task. This is done by using the pre-trained convolution neural network (CNN) models and image caption generator designed with Long Short Term Memory (LSTM). This tool will assist the Ophthalmologist in classifying the three different types of eye diseases namely DME, Early AMD and Late AMD by using the textual description generated by image captioning generator. The features extracted from the OCT images in the form of feature vectors and the partial captions generated for each images are used to fit the training model to create the next word in the sequence. This trained model is further used for eye disease detection and classification based on the text description provided by the LSTM for each images not known to the model. The trained model generates captions to classify the images under diagnosis into four different classes namely NORMAL, DME, Early AMD and Late AMD. The performance metrics of this image caption generator designed by using each of the pre-trained CNN models and LSTMs are evaluated and compared for four different classes independently to select the best image caption generator. The test results show that the performance of the image caption generator implemented by using the pre-trained model DenseNet169 and Xception are found to perform better than all other pre-trained models.

Keywords: Optical coherence tomography(OCT), Diabetic Retinopathy(DR), Deep Learning, Convolution Neural Networks(CNN), Long Short Term Memory(LSTM), Transfer Learning, Ophthalmologist, Choroidal Neovascularization(CNV), Age-related macular degeneration(AMD).

1. Introduction

In Ophthalmology, Optical Coherence Tomography (OCT) plays a vital role in the detection and classification of eye diseases for further assessment and treatment. In the present scenario, the diagnosis of eye diseases is primarily dependent and based on the clinical examination and the subjective analysis of OCT images reported by the referred expert. This paper aims for the automatic detection and classification of three different eye diseases present in the OCT images of human eye by using image caption generator designed by using pre-defined CNN models and RNN. The image caption generator model generates the textual description about the symptoms of each eye disease for detecting and classifying the OCT images. This model has to identify the relationship between different symptoms rather than identifying the symptoms present in the OCT image. The following are the three eye diseases for which text description is generated by the image caption generator to detect and classify the disease present in the OCT image.

1.Early AMD.
2.Late AMD
3. DME
4. NORMAL

Choroidal Neovascularization (CNV) involves the growth of new blood vessels that originate from the choroid through a break in the Bruch membrane into the sub–retinal pigment epithelium (sub-RPE) or sub retinal space. CNV is also closely related with the excessive amounts of vascular endothelial growth factor (VEGF). This is the symptom for Late AMD and will lead to sudden deterioration of central vision. DRUSEN, an eye problem occurs due to aging and macular degeneration. DRUSEN are tiny white or yellow accumulations of extracellular material that build up between Bruch's membrane and the retinal pigment epithelium of the eye. As the age advances or grows, it is normal to have the presence of few small hard DRUSEN. This represents the symptom for Early AMD and the presence of large number of DRUSEN in the macula destroys sharp central vision. The central vision is needed to see objects clearly and to do tasks such as driving the vehicles and reading the books. Diabetic Macular Edema (DME) is a common eye disease that causes irreversible vision loss for diabetic patients, if this is left untreated. It is mainly due to leaking of blood vessels in the retina.

The precise information about the different components or features of retina of the human such as blood vessels, the macula, the fovea and the optic disc(OD) from the OCT images aids to detect the different kinds of eye diseases or abnormalities. In the traditional method, Ophthalmologist analyzes the abnormal retinal images and derives the useful information to detect each eye disease. But, this process is really time-consuming and it leads to false prediction in sometimes due to human error. Hence, an automatic and accurate image caption generator is necessary to provide the text description about the OCT images for predicting the different kinds of eye diseases accurately.

Three different approaches namely image processing techniques; machine learning and deep learning are used to detect the different eye diseases in OCT images. Many algorithms based on image processing are available for detecting the pathologies present in the human eye from OCT images. In OCT images, the spatial local correlation among the neighboring pixels, an image processing technique, is an important source of information which is used to detect and classify different types of eye diseases. The application of image processing and computer vision to OCT image interpretation has mainly focused on the development of segmenting the retinal layer and measuring the thickness of the segmented layer for comparison with the corresponding thickness measurements made from the database of normal retinal images to identify retinal diseases. Apart from measuring the thickness of the retinal layer, research work also focused on the segmentation of fluid regions seen in the retinal OCT images such as Edema or cystic structures which are observed in advanced stages of DME and AMD or DRUSEN.

The computer-aided diagnosis (CAD) of OCT images uses ML techniques with hand-engineered features for decision-making. However, devising hand-engineered features from the OCT images is the most challenging task as it needs expertise in analyzing the variability of parameters or features in the region of interest (ROI). A machine learning (ML) technique is used to detect the different stages of Age-Related Macular Degeneration (AMD) of human eye from OCT images by using multi scale histograms of oriented gradient descriptors as feature descriptor to extract the features being given to the support vector machine based classifier for classifying the diseases into DME, Early AMD and NORMAL eye[1]. The multi scale Local binary pattern (LBP) features are used to perform multi label classification of retinal OCT images for the detection of macular pathologies by Liu et al [2]. A support vector machine (SVM) classifier using five distinct features extracted from the labeled images is proposed by Hassan et al [12] to detect and classify DME from the abnormal OCT images.

The Deep Learning (DL) model aids to overcome the challenges involved in the

meaningful feature extraction by using a cascade of different layers of non-linear processing units. In computer-aided diagnosis (CAD), both Artificial Neural Network (ANN) and Convolution Neural Network (CNN) are used to aid Ophthalmologist to diagnose precisely the particular disease or abnormality from the OCT images. But, CNN is preferred as it extracts only interested features from image and creates reduced representatives. Further, it uses learned data specific kernels instead of pre-defined kernels. In computer-aided diagnosis, deep learning is employed to automatically detect and quantify macular fluid in OCT images to identify DME, DRUSEN and RVO (retinal vein occlusion) [8]. But the computed performance metrics are not high so they are not used for diagnosis by the eye specialists. A framework based on deep learning is implemented for recognizing eye diseases on Spectral Domain Optical Coherence Tomography (SD-OCT) images through transfer learning (TL) [6]. A pre-trained CNN model, Inception-ResNet-V2 is fine tuned by using the dataset consists of DME and normal images for classification of OCT images [7]. But the drawback of this model is that it is able to classify only DME and normal images. The TL based classifier using pre-trained CNN models is employed to classify DME and AMD from the abnormal OCT images [4][5]. The different types of pre-trained CNN models are used to predict the two severity levels of DME as early DME or late DME[3]. An another CNN model, GoogleNet used transfer learning to classify the OCT images into 3 different classes namely DME, Early AMD and normal eye disease classes[9]. But they have not concentrated on the metrics computation to assess the capabilities of the model that they used for classification. The transfer learning was used in the CNN of AlexNet by for feature extraction and then SVM was used for classification into DME and normal images [10]. But the drawback of this model is that only two different types of classes are used in the classification process. The pre-trained CNN is used as an effective tool to extract features for accurate prediction of malaria parasite from thin blood smear images [18]. But this model is used only for binary classification namely infected and not infected. The pre-trained CNN models namely GoogleNet, VGGNet and ResNet are used to extract the features for the accurate prediction of tuberculosis [19]. The classifier designed using this model is also used only for binary classification

Luhui Wu et al [21] have developed an image captioning model for the diabetic retinopathy images using CNN and RNN. This model has predicted the abnormalities present in the retinal images for normal and abnormal images. But the shortcomings of this model are that they have not predicted for the five different levels of diabetic retinopathy. Min Yang et al[22] have designed an image captioning model for cross domain learning and prediction. They trained the model using images of one domain. They have used the same training model for image captioning prediction on images belonging to another domain. They used CNN-LSTM for developing their model. Yansong Feng et al [23] have used Scale Invariant Feature Transform Algorithm for representing each news images as a bag of words in the model proposed by them and they used recall and mean average precision for evaluating their model. Yang et al [24] had proposed a model that used generator for generating textual descriptions for the given visual content of the video and discriminator for controlling the accuracy of generation. Dong-Jin Kim et al [25] had proposed an image caption generator model involving CNN and Guided LSTM and they had found that this model had superior performance in terms of predicting the next word of the sequence accurately when compared with LSTM. But they have not done classification task related to OCT images. Minsi Wang et al[26] have proposed a novel parallel-fusion RNN-LSTM architecture for textual description for the image and they obtained better results and improved efficiency when compared among the dominated one. They have divided the hidden units of RNN into same equal sized parts and allowed them to run in parallel. But they have not used any specific dataset to claim the superior features of their model. Jie Wu et al [27] have proposed a cascaded RNN models for image caption generation and they have verified their model for its effectiveness and accuracy for its caption generation. Chetan Amritkar et al[28] had proposed a model using CNN for

extracting features from the image and RNN for generating clean sentence about the image and it was observed that the model frequently had given accurate descriptions about an image. Jiuxiang Gu et al[29] have introduced a language model based on CNN and identified their suitability for statistical modeling of language tasks. This language model can model the long range dependencies in the history of words, which are really important for image captioning. Xiaodong et al[30] have used deep learning based CNN and RNN for image caption generation. CNN is used for feature extraction in a sub-region of the given image and RNN is used for is used for generating image captions for the OCT images.

This paper presents the design, development of an improved image caption generator for eye disease detection and classification by providing the details of symptoms about the disease in the form of textual description for each type of disease and the name or class of the eye disease. The research findings of this work will be used in the clinical decision support system to assist the Ophthalmologist in taking the final decision in predicting the eye disease. These research findings are also used to generate the medical record for the patients automatically and also to reduce the burden of the doctors. Nine pre-trained CNN models and 7000 labeled high resolution OCT images are used for this purpose. The performance metrics of the different configurations of the image caption generator using different pre-trained CNN models along with LSTM are presented with analysis to select the best image caption model or configuration. The image caption generators designed using DenseNet201 and Xception are selected as the best models as it is more accurate than other models. These two image caption generators are validated against other proven caption generator models to build confidence on the detection and classification of Late AMD, DME and Early AMD from OCT images using the text description provided by it.

2. Data Set

***Table* I Image Class distribution in the dataset**

S.No	Total Number of Training Images	Total Number of Validation Images	Total Number of Testing Images	Eye Disease Level	Eye Disease Class
1	1750	250	8	0	Normal
2	1750	250	8	1	Early AMD
3	1750	250	8	2	Late AMD
4	1750	250	8	3	DME

The table 1 shows the distribution of different types of images in the original dataset. The training dataset used by us consists of 6000 labeled high resolution OCT images, open sourced by a free platform Kaggle for eye disease screening. This dataset is based on four different classes corresponding to the four different eye diseases as tabulated in the Table I. The dataset used for validation and testing consist of 1000 and 32 images respectively.

In our work, finally the images will be classified into any one of the 4 eye diseases namely Normal, DME, Early AMD with the presence of DRUSEN and Late AMD with the presence of CNV by using the text description about the image predicted by the combination of CNN and LSTM. . The figures from 1 to 4 shows the images related to the different eye diseases classified by our caption generator. The features in them will be different for different eye diseases. Finally, the textual description will be displayed on these images to detect and classify the different type of eye disease class.

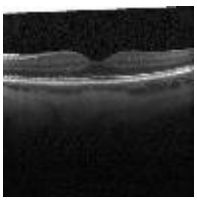

Figure 1 Normal

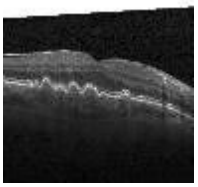

Figure 2 Early AMD

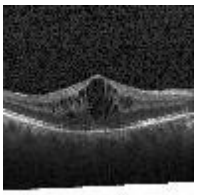

Figure 3 Late AMD

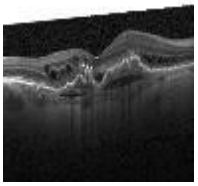

Figure 4 DME

3. Proposed Methodology

The proposed methodology shown in Figure 5 consists of four major functional blocks namely feature extractor, sequence generator, Training the Caption Generator model and new caption generator or classifier model. The feature extractor is designed by using pre-trained CNN models by removing the top layer. 7000 OCT images from the 4 eye disease classes namely Late AMD having the symptoms of CNV, DME, Early AMD having the symptoms of DRUSEN and NORMAL are given to the feature extractor for feature extraction. The extracted features corresponding to the training images and the vocabulary formed for these training images from the textual description of these images are applied as inputs to the training model to create the trained model. During the training of the model, sequence generator responsible for generating the sequence of words is used to predict the next word in the sequence along with LSTM. The image to be tested for any one of the eye disease class is given to the image caption generator as input. The caption generator will output the result in the form of textual description about the image. The description generated by the caption generator is used to find and count the number of correct and wrong predictions in the OCT images as outputted by caption generator. The number of correct and incorrect captions predicted for the four different classes are used for designing a nested dictionary in python. This nested dictionary is passed as an argument to the Confusion Matrix object in python to generate the different metrics associated with the caption generator. The results obtained from the Confusion Matrix object is observed and compared among them to identify the best image caption generator designed with the pre-trained CNN models and LSTM. The image caption generator whose performance metrics verified as the best one can be recommended to the eye specialists to diagnose the eye disease.

All the labeled OCT images are pre-processed by using resizing technique to meet the input requirements of different pre-trained deep learning models. Mostly, OCT images are resized into three different resolutions namely 299*299, 224*224 and 96*96 pixels. The purpose of Feature Extractor in the automatic caption generator is to transform the raw images into limited distinct features to reduce the

complexity in processing the images without losing the meaningful information. Pre-trained neural network models are used for this purpose. The Feature Extractor extracts the unique features in the images to form the unique vector for all the images and they are stored in features.pkl file. The feature extraction transforms the labeled OCT images into feature vectors. Feature Extraction also helps to avoid the large amount of memory requirement and computing power. The feature vectors and the vocabulary formed from the words of textual descriptions corresponding to these images are used as inputs to the training model to learn the features effectively and generate a trained model with optimum weights. The amount of time used for training the model will vary based on the pre-trained model that is used to create the feature extractor. The best trained model having the maximum validation accuracy or minimum validation loss is used for creating captions for the new images during testing. testing images.

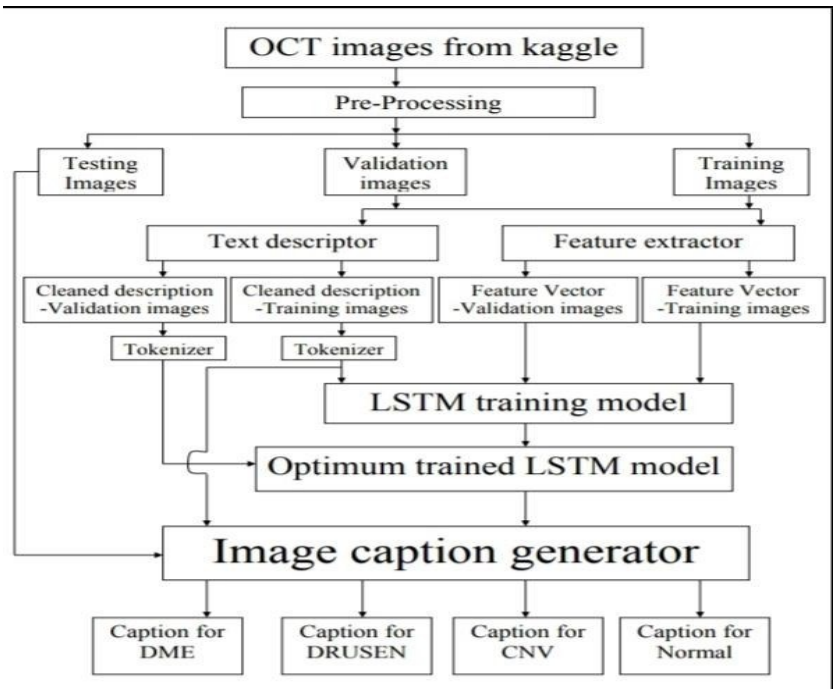

Figure 5 Proposed methodology for Image Caption Generator

Feature Extractor

The feature extractor is used to extract significant features from the pre-processed OCT images which represent the set of features of each disease class. The extracted features are used to fit the image caption generator model to learn the symptoms about the image using the different words to form the text description. In other words, extracted features along with the symptoms present in the image represented in the form of clean text are used to fit the training model. There are different methods available for extracting the features from the OCT images. They are

1). Local Invariant Features
2). Key point Localizati on
3). Scale Invariant feature Transform Detection 4). Local Feature Descriptor

These feature descriptors aforementioned will extract only some specific features in the OCT images. These features may not be sufficient to detect a specific eye disease precisely. The accuracy of these feature extractors are not sufficient to use them for feature extraction. Therefore, there is a need for a model based on deep learning that can extract all the features corresponding to a specific eye disease class. The pre- trained CNN models are selected for feature extraction with the aim of extracting all the relevant features in the OCT images by using the different convolution layers in the CNN model. The lower layers will extract the low level features and the middle layers will extract the mid level features. Finally, the end layers will use all these features to create the high level features corresponding to the particular disease class.

The feature extractor is designed by using the pre-trained weights of ImageNet and by removing the top layer in the actual pre-trained CNN model. The pre-trained CNN model is chosen over the conventional neural network and artificial neural network because it is capable of capturing and learning the features automatically from the images at different levels of hierarchy similar to the human brain by the different layers of CNN. During convolution operation, each output value is not required to be connected to every neuron in the previous layer but connected to only those receptive fields where the convolution kernel is currently applied. This specific characteristic of convolution layer which will reduce the amount of interconnection drastically is called local connectivity. Again in the pre-trained CNN , the same weights are applied over the convolution until the next update of the parameters referred to as parameter sharing. Thus, there is a drastic reduction in the number of parameters when compared with ANN where there is a connection between every pair of single input/output neuron. Clearly, CNN is more efficient than conventional neural network and ANN in terms of complexity and memory. It was proved and identified that pre-trained CNN models were a good feature extractor for a completely new task/problem. The designed feature extractor will extract useful attributes from an already pre-trained CNN with its trained weights by feeding our image data having different eye diseases on each level and to tune the CNN a bit for the specific task namely eye disease classification. The pre-trained CNN models are very efficient in these tasks when compared to neural networks. The advantage of pre-training is that there is no need to train the CNN and thus memory and time are saved. The convolution process in the CNN is capable of extracting the relevant information at low computational costs. Thus, the pre-trained CNN is selected for feature extractor after considering their merits over conventional NNs and ANN. The pre-trained CNN models will detect key points on the image and the number of key points will vary from image to image. Then feature vector is built for each image based on the number of key points used for representing the image. These features represent the internal representations of the image just before classifications were made. Then these features are represented in the form of feature vectors for all the images in the form of Numpy array. The dimension of the feature vector will vary based on the pre-trained model used for feature extraction. In this way, the feature vector is computed for OCT images in the dataset and these features are stored in the features.pkl file.

Cleaned Text Description Generator

The token.txt file is created manually with 5 different captions for each image present in the dataset used for training and validation. The image id and text description for each image is separated and stored in token.txt file. The first column of the file contains the image id and from the second column onwards it contains the description about the image in the form of symptoms in text description. The imageId is numbered from 0 to 4 after the image file. A dictionary in python is used for storing the image id of different files as keys and the description about the file as their corresponding values of the dictionary. The text description for each image in the entire dataset is cleaned and then the cleaned description is formed for each image present in the dataset. Then the cleaned description in the text file

is converted into a vocabulary of words. The text is cleaned to reduce the size of the vocabulary of words. The vocabulary should be expressive and as small as possible. If the size of the vocabulary is small, then the size of the model is small and hence this network will train faster. Finally, the dictionary of image identifiers and cleaned description are then saved in the descriptions.txt file, with one image identifier and the corresponding symptoms description in each line. Now the cleaned descriptions are ready for modeling and the order of the descriptions may vary.

Training and Evaluation Model

In this work, both training the model by using partial captions to predict the next word in the sequence in each epoch is followed by validating the model. The training model is designed by using three components namely image feature extractor, sequence processor and decoder. The feature extractor is designed by using different pre-trained models after removing the top layer in the original models. The features extracted by this pre-defined model will act as one input to the training model. The feature will be vector of 4096 elements if the pre-trained model vgg16 is used. The size of this feature vector is different for different pre-trained CNN models. A dense layer is used to process these features to generate a 256 element representation for each image. The sequence processor is implemented by using a word embedding layer followed by a Long Short-Term Memory (LSTM) recurrent neural network layer. An input sequence with a pre-defined length of 34 words is given as input to the embedding layer that uses a mask to ignore the padded values. LSTM layer with 256 memory units will follow the embedding layer in the sequence processor. Both the input models namely photo feature extractor model and LSTM will produce a 256 element vector. Both the models will use regularization in the form of 50% dropout. This is done to decrease the over fitting in the training dataset as this model configuration learns fast. The decoder model then merges the two vectors from the two input models namely feature extractor model and sequence processor model using an addition operation. The two vectors after merging is fed to Dense 256 neuron layer followed by a final output dense layer. The output dense layer makes a softmax prediction over the entire output vocabulary for predicting the next word in the sequence. The skill or capability of the model is monitored and checked by using the validation dataset. The whole model is saved to a file if the skill or capability of the model is improved at the end of each epoch on the validation dataset.

The model with the best skill on the training set is saved as the final model at the end of each run. This is done by defining a Model Checkpoint in keras and informing it to monitor the minimum loss on the validation dataset. Then the model with minimum loss is saved to the file with .h5 extension named with the epoch name, training loss and validation loss. The checkpoint can be specified in the call to the fit function to fit the training model with validation dataset as argument to the fit function. The training model will be provided one word split from the text description and the photo features during training. This model is expected to learn the next word in the sequence. When the trained model is used later to generate descriptions during testing, the generated words will be concatenated and recursively provided as input to generate a caption for a new image. Thus, the weights corresponding to each epoch where there was an improvement in validation accuracy is stored in h5 file. Finally, the weights corresponding to a particular epoch during training which gives the maximum validation accuracy or minimum loss is used for the final trained model which is used for generating image captioning for the unseen photo or image.

Evaluation Model

Once the model is fit by using the features and text description of OCT images i.e., inputs-output pairs, it is ready for evaluation for predicting its skill on the holdout test

data or validation data. The model is evaluated by generating descriptions for all the photos or images in the validation dataset. These predictions are evaluated by using a standard cost function. This involves passing in the start description token 'startseq', generating one word and then calling the trained model recursively with generated words from the previous predictions as input until the end of sequence token 'endseq' is reached or the maximum description length is reached. Thus, the trained model can be evaluated against a given validation dataset of photo descriptions and photo features. The actual and predicted descriptions are collected and evaluated collectively using the corpus BLEU score that summarizes how close the generated text is to the expected text.

Image Caption Generation Model

The trained model obtained after training in the .h5 file contains almost everything needed to generate the captions for the new OCT images. First, the Tokenizer created during the encoding of text that has the tokens describing all the training images and the maximum length of the sequence defined to generate the text description about the OCT images are needed for generating captions for the new OCT images. Next, the photo for which captions to be generated and the features to be extracted is applied to the trained model. This is done by re-defining the model by using LSTM and then adding any one of the pre-trained models to it. The pre-trained models are used to extract the features and the extracted features are used as inputs to the trained model. Captions for the OCT images are generated after the model is successfully validated. After removing the start and end sequence tokens in the sequence generated, the description having symptoms about the image along with its class label are generated for the new OCT image. This description along with the name of the class label to which the OCT image is classified will assist the Ophthalmologist in classifying the type of eye disease.

4. Experiments Carried out for Metrics Analysis to Design a Clinical Decision Support System

The trained model in the form of weights along with the image to be captioned is applied to the image caption generator for predicting captions for the image. 32 OCT images from the 4 classes are used for testing the accuracy of the caption generator. The description generated by the caption generator along with its class label is used to predict the input OCT image as either correct or incorrect. A nested dictionary in python is designed with 4 actual classes as 4 internal dictionaries as the key. The values for the nested dictionary are formed by counting the wrong and correct predictions for the 4 classes. The figure 6 shows the block diagram of PyCM with different inputs and outputs in the form of reports generated by the Confusion Matrix Object.

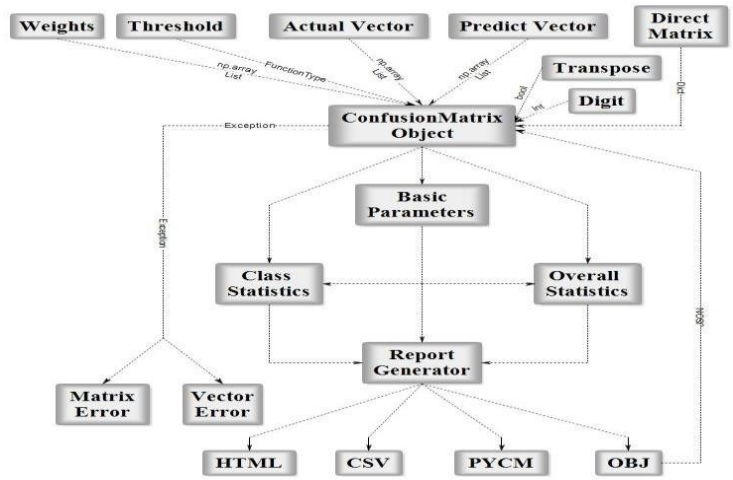

Figure 6 Block Diagram of PyCM

The designed nested matrix is applied as input to the Confusion Matrix object supported by the Python Confusion Matrix library. The Confusion Matrix object will compute all the metrics corresponding to the per class and overall class and store them in a HTML file. These metrics are computed for all the 9 image caption generator models formed by using 9 different pre-trained CNN models and the LSTM. The comparison across different caption generation models is analyzed for selecting the best caption generator model. The image caption generator model whose performance is superior is selected to aid the Ophthalmologist in predicting and classifying the disease type in OCT images.

4. Results and Discussion

Performance Analysis of the Pre-trained Models using Training Loss and Validation Loss

The performance of the caption generators using nine different pre-trained CNN models are analyzed by using the training loss and the validation loss. Training loss refers to the error or loss when fitting the model using input-output pairs. Validation loss refers to the error in the prediction by the trained model using validation images. It was found that training loss and validation loss were found to be minimum for the caption generator designed with the pre-trained model. This means that this model had trained well on the images present in the training set. When the trained model is used for validation by using the validation images, it was found that this model is having the minimum validation loss equal to 0.091. The Figure 7 shows the minimum training and validation loss for the 9 different pre-trained CNN models used in the design of image caption generator.

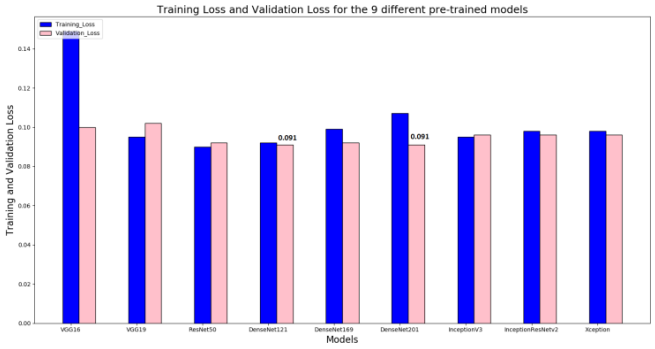

Figure 7 Minimum training and validation loss

Overall Performance Analysis of the Pre-trained CNN Models for all the Classes

Many performance metrics were considered and computed for analysis. They were cross entropy, Kappa, Kappa Std Error, Overall Accuracy, Random Accuracy, Positive Prediction Value (PPV_Macro and PPV_Micro), True Positive Rate(TPR_Macro and TPR_Micro) and standard error for performance analysis and comparison. They are all used to evaluate and assess the prediction performance of the classifier. This performance analysis will provide concrete evidence for the best classifier that can be recommended to the Ophthalmologist for assessing the eye disease.

Performance Analysis using Cross Entropy:

The cross-entropy is one of the loss function used to compare the model's prediction with the label or class which is the true probability distribution. The cross entropy is computed for each class by using the one hot encoding and the true probability distribution. The cross-entropy approaches zero as the prediction gets more and more accurate. The models DenseNet201 and Xception have good cross entropy metric when compared with the other models. Cross Entropy compares the capability of the image captioning model in

predicting the label. It is nothing but the true probability distribution for the different words present in the image captions to describe a particular label or disease class. Cross entropy is zero if the prediction capability of the model is perfect. The figure 8 shows the cross entropy for the 9 different caption models.

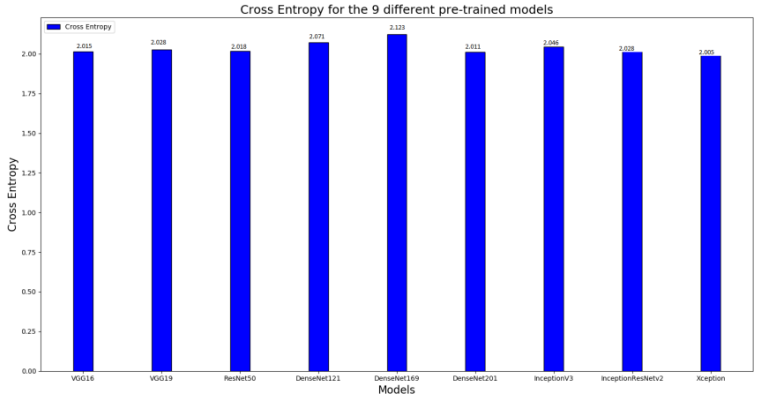

Figure 8 Cross Entropy for the caption generator designed with 9 different Pre-trained Models

Performance Analysis using Kappa:

The kappa statistics is also used to evaluate the image captioning performance of the classifier. The kappa can be calculated by using the observed accuracy and expected accuracy. In essence, the kappa statistic is a measure of how closely the instances or classes or labels classified by the caption generator designed with pre-trained CNN model matched the data labeled *as* ground truth by the experts and thus controlling the accuracy of a caption generator as measured by the expected accuracy. The strength of agreement can be recommended based on the values of kappa as very good, good, moderate, fair and poor as tabulated in the table 2.

***Table* 2 kappa value and the strength of Agreement**

Kappa Value	Strength of Agreement
<0.20	Poor
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Good
0.81-1.0	Very Good

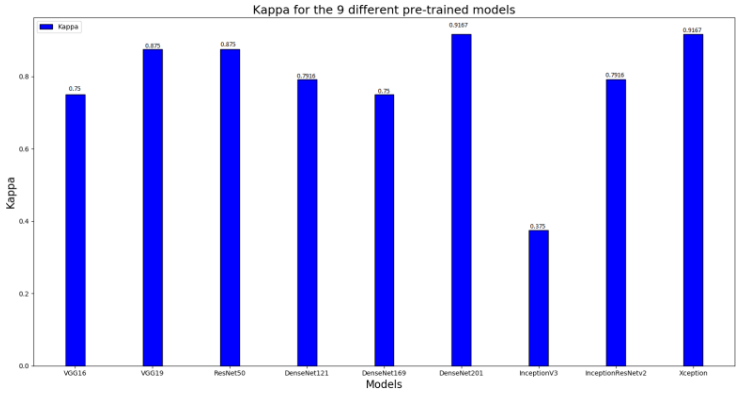

Figure 9 Kappa value for the caption generator with 9 different Pre-trained CNN Models

The Figure 9 shows that image caption generators which have used DenseNet201 and Xception as feature extractors have kappa value equal to 0.9167 and they have the very good strength of agreement.

Performance Analysis using Kappa Standard Error:

The kappa standard error is computed by subtracting the kappa value from one as kappa standard error = 1-kappa value. If the error is small and close to zero, then the model is predicted as a very good model. The kappa standard error is equal to 0.0570 for the caption generator designed with DenseNet201 and Xception as feature extractors. This gives evidence for these two models to have very good image caption generation capability to assist the Ophthalmologist in identifying the eye disease class. The figure 10 shows the kappa standard error for the caption generator designed with 9 different pre-trained CNN models.

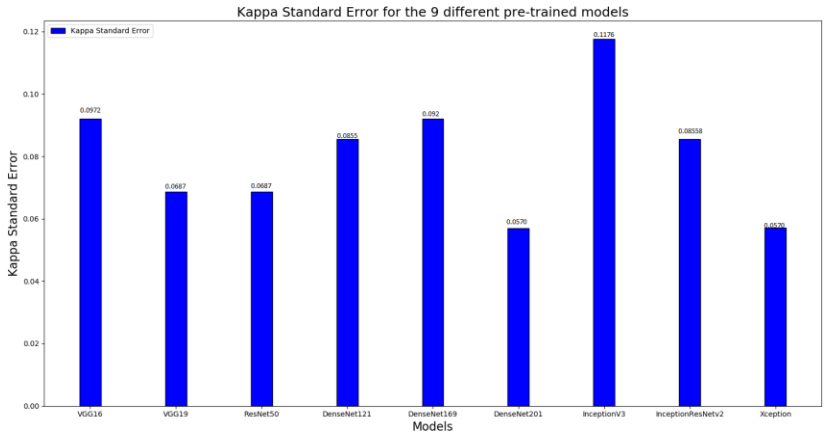

Figure 10 Kappa standard error value for the caption generator designed with 9 different Pre- trained CNN Models

Performance Analysis using Overall Accuracy:

The Overall Accuracy of the model is computed by dividing the sum of the correct predictions from all the classes of eye diseases divided by the total number of images belonging to all the classes of eye diseases used in the caption generation process. The overall accuracy is equal to 0.9375 for the caption generators that were designed with DenseNet201 and Xception as feature extractors. This is also an another evidence for suggesting these image caption generators to the ophthalmologist for

supporting their clinical decision in classifying the different types of eye diseases. The figure 11 shows the overall accuracy for the caption generator designed with 9 pre-trained CNN models.

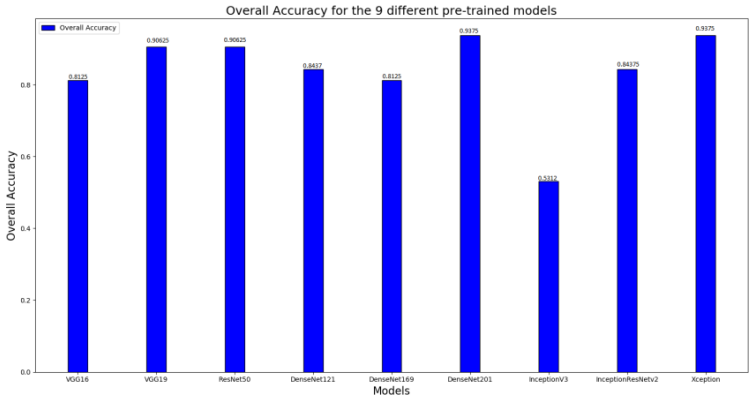

Figure 11 Overall accuracy of caption generator designed with 9 different Pre-trained CNN Models

Performance Analysis using PPV_Macro:

A macro-average will perform computation of performance metric independently for each eye disease class and then take the average among all the classes and hence all the classes are treated equally. A micro-average will aggregate the contributions of all classes to compute the average metric. In multi-class classification, micro-average is preferable if there is class imbalance. The figure 11 shows the PPV_Macro values for the caption generator designed with 9 different pre-trained CNN models.

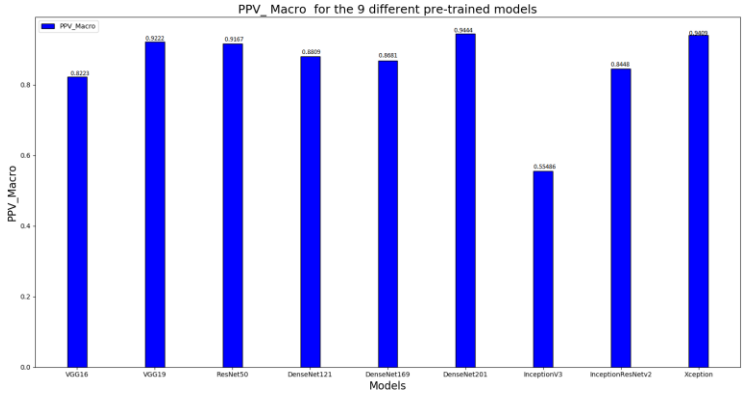

Figure 11 PPV_Macro for the caption generator designed with 9 different Pre-trained CNN Models

It is observed that caption generators designed with DenseNet201 and Xception for feature extraction have the maximum value of PPV_Macro equal to 0.94. So this metric will help us to recommend these caption generators to the Ophthalmologist to assess the different eye disease classes.

Performance Analysis using TPR_Macro and TPR_Micro:

True Positive Rate is defined as the total number of classes correctly identified as positive by considering the images in all the classes divided by the total number of True positive and False Negative. True Positive Rate is equal to 0.9375 for the image captioning system designed with DenseNet201 and Xception as feature extractors. This is also an another acknowledgement for recommending these caption

generators for taking clinical decision correctly. The figure 12 shows the TPR_Micro and TPR_Macro values for the caption generators designed with 9 different pre-trained CNN models as feature extractor.

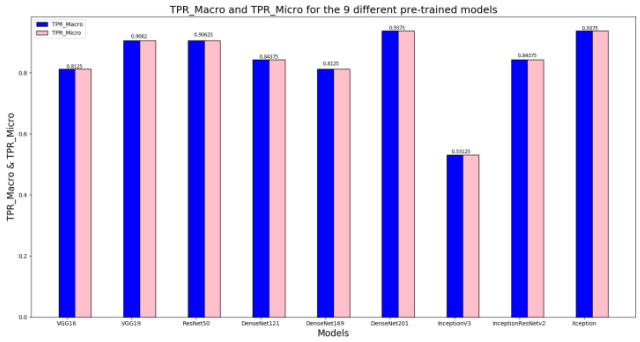

Figure 12 TPR_Micro and TPR_Macro for caption generators with 9 different Pre-trained CNN Models

Performance Analysis of the Different Pre-trained CNN Models among Different Classes

Image caption generators can be compared with each other to select the best caption generator based on the benchmark results. For each of the 4 classes, this work has considered and computed 10 different metrics. They are Accuracy, Error Rate, F1Score, F2Score, Geometric mean of precision and sensitivity, Mathew's Correlation Coefficient, Precision or Positive Predictive value, Random Accuracy, Specificity and Sensitivity. Our experiments have demonstrated that two caption generators designed using 2 different pre-trained CNN models namely DenseNet201 and Xception have shown good performance in predicting the symptoms for all the 4 classes.

Performance Analysis using Accuracy of each Eye Disease Class:

All the caption generator models except the one designed with InceptionV3 as feature extractor are perfect in predicting the Normal eye disease accurately and their accuracy of prediction is equal to or greater than 0.9375. The 4 caption generation systems designed with ResNet50, DenseNet121, DenseNet169 and Xception as feature extractors have accuracy of caption generation equal to 1.0. It is experimentally identified that the LSTM caption generation system designed with Xception and DenseNet201 for extracting the features have generated captions for all the classes almost correctly. The figure 13 shows the accuracy of 9 different caption generator models for each eye disease class.

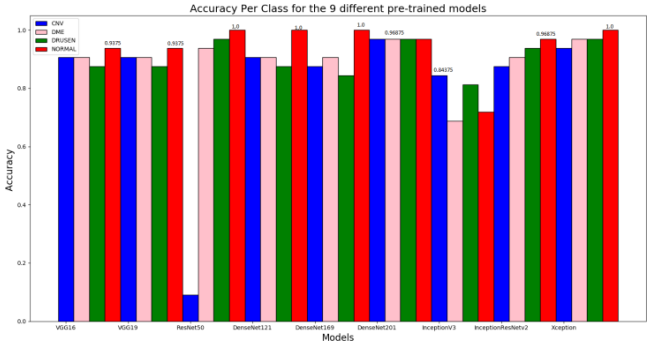

Figure 13 Accuracy of 9 different caption generators models for 4 different classes

Performance Analysis using Error Rate of each Eye Disease Class:

All the caption generator models are perfect in predicting the textual description for the Normal eye disease accurately and their error rate during prediction is very close to 0%. The 2 caption generators who have generated captions for all the classes accurately are designed with the pre-trained CNN models namely DenseNet201 Xception and Xception as feature extractors. The error rate for these models is very close to 0%. This experimental observation has favored these image caption generation systems to be recommended to the ophthalmologist for assisting their clinical decision. The figure 14 shows the error rate of 9 different caption generator models for each eye disease class.

Figure 14 Error rate of 9 different caption generator models for 4 different classes

Performance Analysis using F1Score of each eye disease class:

F1Score is defined as the harmonic mean of precision and recall. Therefore, F1Score will consider both false positives and false negatives. If we have uneven class distribution in the number of images used in the training process, then F1Score is more useful than accuracy. Four models are having F1Score equal to 1 when predicting the captions for the Normal eye disease classes accurately. The 2 models that have used DenseNet201 and Xception for extracting the features in the images have F1Score almost very close to one when predicting captions for all the 4 classes. F1Score is a significant metrics used for comparing the performance of the model when the number of images in each class is not the same when used for training the model. The figure 15 shows the F1Score of different caption generator models for each eye disease class.

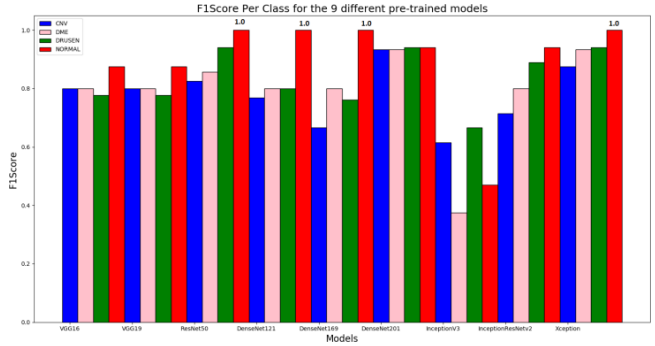

Figure 15 F1Score of different caption generator models for 4 different classes

Performance Analysis using F2Score of each Eye Disease Class:

F2Score is defined as the weighted average of precision and recall. Therefore, F2Score will consider both false positives and false negatives. If we have uneven class distribution in the number images

used in the training process, then F2Score is more useful than accuracy. F2Score is really effective in classification when the cost of false negative is much higher than the cost of false positive. 4 image caption generation models having F2Score equal to 1 when generating captions for the Normal eye disease classes accurately are designed with the pre-trained CNN models namely ResNet50, DenseNet121, DenseNet169 and Xception for extracting the features to be used during the training of LSTM. The image caption generation systems which are based on DenseNet201 and Xception are found to have F2Score close to one. The figure 16 shows the F2Score of 9 different caption generators for each eye disease class.

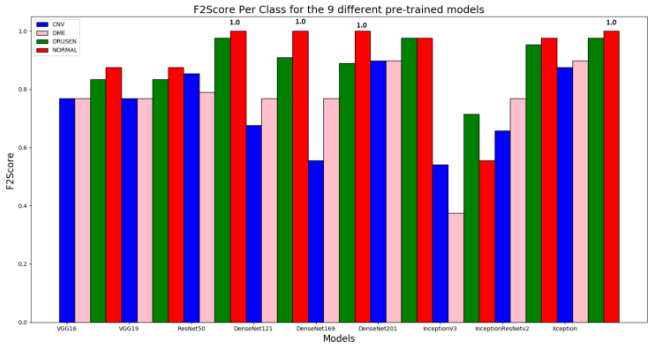

Figure 16 F2Score of 9 different caption generator models for 4 different classes

Performance Analysis using G-Mean of each Eye Disease Class:

G-Mean1 is defined as the geometric mean of sensitivity and precision. G-Mean2 is defined as the geometric mean of sensitivity and specificity. 4 different caption generator models are having G-Mean equal to 1 when predicting the textual description for the Normal eye disease classes accurately. The 2 caption generator models that have used DenseNet201and Xception in feature extraction have G-Mean almost close to one when predicting the text description for all the CNV. The figure 17 shows the G-Mean of 9 different caption generator models for each eye disease class.

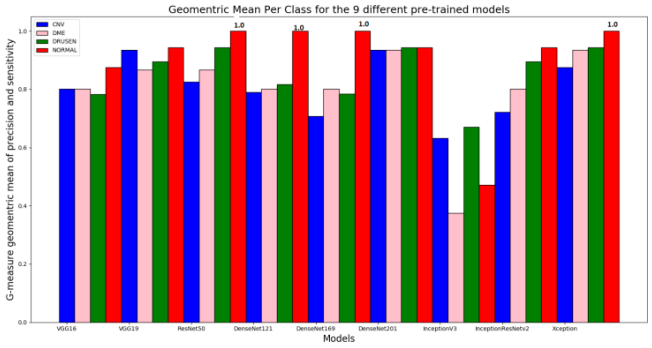

Figure 17 G-Mean of different caption generator models for 4 different classes

Performance Analysis using Mathew's Correlation Coefficient(MCC) of each Eye Disease Class:

Mathew's correlation coefficient can be used as a reference performance measure in the case of an unbalanced dataset if it is used for training the model. 4 caption generation systems are having MCC equal to 1 when predicting the textual description for the Normal eye disease classes accurately. 2 models of caption generators based on the two pre-trained CNN models namely DenseNet201 and Xception , InceptionV3 and Inception-ResNetV2 have MCC very close to one when predicting the captions for all the

classes. The figure 18 shows the Mathew's correlation coefficient of 9 different caption generator model for each eye disease class.

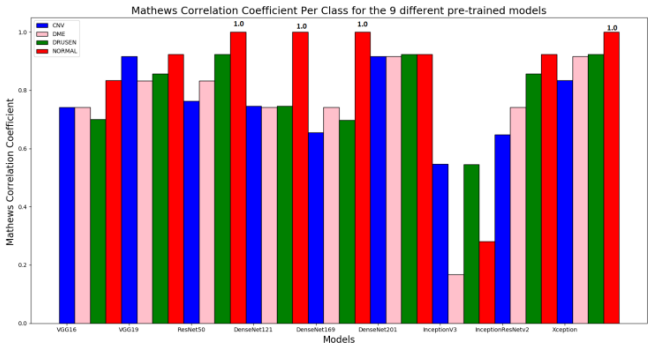

Figure 18 Mathew's Correlation Coefficient of different caption generator models for 4 different classes

Performance Analysis using Precision or Positive Predictive Value(PPV) of each Eye Disease Class:

Precision (PREC) is calculated by dividing the number of correct positive predictions by the total number of positive predictions. It is also called positive predictive value (PPV). The best precision value is 1.0, whereas the worst value is 0.0. Four caption generator models are having precision equal to 1 when predicting the image caption description for the Normal eye disease classes accurately. Two neural caption generator models that are designed with DenseNet201 and Xception as feature extractors have precision almost very close to one when predicting the captions for all the 4 classes. The figure 19 shows the precision of different caption generator model for each eye disease class.

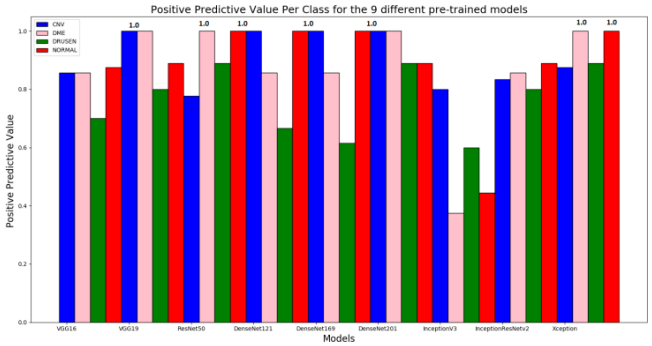

Figure 19 Precision of different caption generator models for 4 different classes

Performance Analysis using Specificity of each Eye Disease Class:

Specificity (SP) is calculated by dividing the number of correct negative predictions by the total number of negatives. It is also called true negative rate (TNR). The best value for specificity is 1.0, whereas the worst value is 0.0.All the caption models are having specificity equal to or greater than 0.8 when predicting the image description for all the classes. 4 caption generator models based on ResNet50, DenseNet121, DenseNet169 and Xception have specificity equal to one when predicting the image captions for the normal classes. 2 caption generation systems that are implemented with DenseNet201 and Xception

for feature extraction have specificity value either equal to 1 or close to 1. The figure 20 shows the specificity of different caption generation models for each eye disease class.

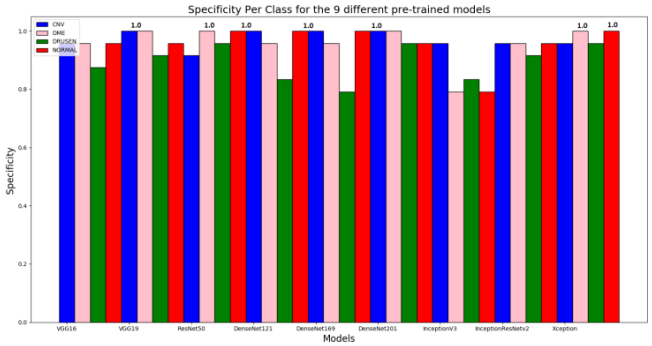

Figure 20 Specificity of the 9 different caption generation models for 4 different classes

Performance Analysis using Sensitivity of each Eye Disease Class.

Sensitivity (SN) is calculated by dividing the number of correct positive predictions by the total number of positives. It is also called recall (REC) or true positive rate (TPR). The best value for sensitivity is 1.0, whereas the worst value is 0.0.Seven caption generation models are having sensitivity equal to 1 when predicting captions for the Normal and DRUSEN eye disease classes accurately. The 2 caption generator models which have used DenseNet201and Xception as feature extractors have sensitivity either equal to one or very close to one when predicting the textual description for all the 4 classes. These experimental truths about these models have biased towards these models to recommend them for the eye specialists to predict the different types of eye disease classes. The figure 21shows the specificity of 9 different caption generators each eye disease class.

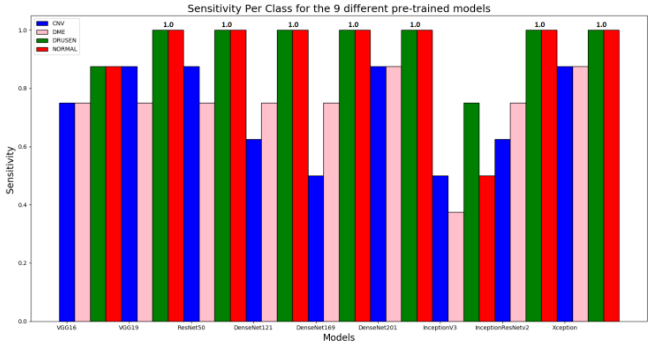

Figure 21 Sensitivity of the 9 different caption generator models for 4 different classes

5 Conclusion and Future works

This research work has focused on the design, analysis and development of an improved clinical decision support system for the Ophthalmologist by designing image feature extractors using different pre- trained CNN models to extract the features, training the LSTM model using the features and to create the trained model in the form of optimum weights. Then image caption generator is designed using pre-trained CNN models and LSTM. Then the image captions results obtained from the caption generator designed with different pre-trained models and LSTM are compared to select the best model that can be recommended to the ophthalmologist to diagnose the eye diseases for supporting their clinical decision. Based on the performance metrics, it was observed that the image caption generators that have used

DenseNet201 and Xception have provided the best performance in predicting the captions for the eye diseases correctly. Therefore, they can be used in the design of clinical decision support system to assist the ophthalmologist. This work can also be further extended by fusing the features from two different pre- trained CNN models and using their features to design a new model for caption generation. The performance of the newly created model is further analyzed to identify its suitability for designing a clinical decision support system by comparing with the existing model.

6. References

[1] Pratul P. Srinivasan, Leo A. Kim, Priyatham S. Mettu, Scott W. Cousins, Grant M. Comer, Joseph A.
[2] Izatt & Sina Farsiu (2014): Fully Automated Detection Of Diabetic Macular Edema And Dry Age-Related Macular Degeneration From Optical Coherence Tomography Images, Biomedical Optics Express, Vol 5, No 10, October 2014.
[3] S. P. K. Karri, Debjani Chakraborty, & Jyotirmoy Chaterjee(2017): Transfer Learning Based Classification Of Optical Coherence Tomography Images With Diabetic Macular Edema And Dry Age-Related Macular Degeneration, Biomedical Optics Express, Vol. 8, No. 2, Feb 2017.
[4] Thomas Schlegl, Sebastian M. Waldstein, Hrvoje Bongunovic, Franz Endstraber, Amir Sadeghipour, Ana-Maria Philip, Dominika Podkowinski, Bianca S. Gerendas, Georg Langs & Ursula Schmidt-
[5] Erfurth(2018): Fully Automated Detection And Quantification Of Macular Fluid In Oct Using Deep Learning, American Academy Of Ophthalmology, Vol 25, No 4, April 2018.
[6] Rui Zhaq, Acner Camino, Jie Wang, Ahmed M. Hagg, Yansha Lu, Steven T. Bailey, Christina J. Flaxel,
[7] Thomas S. Hwang, David Huang, Dengwang Li & Yali Jia (2017): Automated Drusen Detection In Dry Age-Related Macular Degeneration By Multiple-Depth, En Face Optical Coherence Tomography, Biomedical Optics Express, Vol. 8, No. 11 | 1 Nov 2017. Sivaramakrishnan Rajaraman, Sameer K. Antani , Mahdieh Poostchi , Kamolrat Silamut, Md. A.
[8] Hossain & Richard J. Maudeaude, Stefan Jaeger & George R. Thomas(2017): Pre-Trained Convolutional Neural Networks As Feature Extractors Toward Improved Malaria Parasite Detection In Thin Blood Smear Images, Peerj 6:E4568; Doi 10.7717/Peerj.4568.
[9] U.K. Lopes & J.F. Valiati(2017): Pre-Trained Convolutional Neural Networks As Feature Extractors For Tuberculosis Detection, Computers In Biology And Medicine, 89(2017), 135-143.
[10] Philippe Burlina, Katia D. Pacheco, Neil Joshi, David E. Freund & Neil M. Bressler(2017): Comparing Humans
[11] And Deep Learning Performance For Grading Amd: A Study In Using Universal Deep Features And Transfer Learning For Automated Amd Analysis, Computers In Biology And Medicine, 82, 80-86, 2017.
[12] Fedix Grassmann, Phd, Judith Mengelkamp, Phd, Caroline Brandl, Phd, Sebastian Harsch, Martina E. Zimmermann, Phd, Birgit Linkohr, Phd, Anntte Peters, Phd, Iris M. Heid, Phd, Christoph Palm, Phd, &
[13] Bernhard H.F. Weber(2018): Phd, A Deep Learning Algorithm For Prediction Of Age-Related Eye Disease Study Severity Scale For Age-Related Macular Degeneration From Color Fundus Photography, American Academy Of Ophthalmology, Vol.125, No.9, September 2018.
[14] Min Yang, Wei Zhao, Wei Xu, Yabing Feng, Zhou Zhao, Xiaojun Chen & Kaile, Multitask Learning For Cross-
[15] Domain Image Captioning, Ieee Transactions On Multimedia, Vol. 21, N0. 4, April 2019.
[16] Yansong Feng & Mirella Lapata, Automatic Caption Generation For News Images, Ieee Transactions On Pattern Analysis And Machine Intelligence, Vol. 35, N0. 4, Pp 797-811, April 2013.
[17] Yang , Jie Zhou, Jiangbo Ai, Yi Bin, Alan Hanjalic, Heng Tao Shen & Yanli Ji, Video Captioning By Adversarial
[18] Lstm, Ieee Transactions On Image Processing, Vol. 27, N0. 11, Pp 5600-5612, November

2018.

[19] Jie Wu & Haifeng Hu, Cascade Recurrent Neural Network For Image Caption Generation, Iet Journals & Magazines, Vol.53, No.25, Pp 1642-1643, 2017.

[20] Xiaodong He & Li Deng, Deep Learning For Image-To-Text Generation, A Technical Review, Ieee Signal Processing Magazine, Vol.34, No. 6, Pp 109-116, November 2017.

a. Conference Proceedings

[21] Jingjing Deng, Xianghua Xie, Louise Terry, Ashley Wood, Nick White, Tom H.Margrain & Rachel V.

[22] North, Age-Related Macular Degeneration Detection And Stage Classification Using Choroidal Oct Images, International Conference On Image Analysis And Recognition, Pp 707-715, Iciar 2016.

[23] Luhui Wu, Cheng Wan, Yiquan Wu & Jiang Liu, Generative Captions For Diabetic Retinopathy Images, 2017 International Conference On Security, Pattern Analysis, And Cybernetics (Spac), Pp 515-519,

[24] Genevieve C. Y. Chan, Awais Muhammad, Syed A. A. Shah, Tong B. Tang, Cheng-Kai Lu & Fabrice

[25] Meriaudeau, Transfer Learning For Diabetic Macular Edema (Dme) Detection On Optical Coherence Tomography (Oct) Images, Proc. Of The 2017 Ieee International Conference On Signal And Image Processing Applications (Ieee Icsipa 2017), Malaysia, September

[26] 12-14, 2017.
[27] Ravi M. Kamble , Genevieve C. Y. Chan , Oscar Perdomo , Fabio A. Gonzalez , Manesh Kokare, Henning
[28] Muller & Fabrice Meriaudeau, Automated Diabetic Macular Edema (Dme) Analysis Using Fine Tuning With Inception-Resnet- V2 On Oct Images, Annual International Conference Of The Ieee Engineering In Medicine And Biology Society, Jul 2018.
[29] Genevieve C. Y. Chan, Awais Muhammad, Syed A. A. Shah, Tong B. Tang, Cheng-Kai Lu & Fabrice
[30] Meriaudeau, Transfer Learning For Diabetic Macular Edema (Dme) Detection On Optical Coherence Tomography (Oct) Images, Proc. Of The 2017 Ieee International Conference On Signal And Image Processing Applications (Ieee Icsipa 2017), Malaysia, September 12-14, 2017.
[31] Aditya Kunwar, Shrey Magotra & M Partha Sarathi, Detection Of High-Risk Macular Edema Using Texture Features And Classification Using Svm Classifier, 2015 International Conference On Advances In Computing, Communications And Informatics (Icacci), 2015.
[32] Oscar Perdomo, Sebastian Otalora, Fabio A. Gonzalez, Fabrice Meriaudeau & Henning Muller, Oct-
[33] Net: A Convolutional Network For Automatic Classification Of Normal And Diabetic Macular Edema Using Sd-Oct Volumes, 2018 Ieee 15th International Symposium On Biomedical Imaging (Isbi 2018), 2018.
[34] Jingjing Deng, Xianghua Xie, Louise Terry, Ashley Wood, Nick White, Tom H.Margrain & Rachel V.
[35] North, Age-Related Macular Degeneration Detection And Stage Classification Using Choroidal Oct Images, International Conference On Image Analysis And Recognition, Pp 707-715, Iciar 2016.
[36] Luhui Wu, Cheng Wan, Yiquan Wu & Jiang Liu, Generative Captions For Diabetic Retinopathy Images, 2017 International Conference On Security, Pattern Analysis, And Cybernetics (Spac), Pp 515-519,
[37] Dong-Jin Kim, Donggeun Yoo, Bonggeun Sim & , In So Kweon, Sentence Learning On Deep Convolutional Networks For Image Caption Generation, 2016 13th International Conference On Ubiquitous Robots And Ambient Intelligence (Urai) , Pp 246-247, August 19-22, 2016 At Sofitel Xian On Renmin Square, Xian, China.
[38] Minsi Wang, Li Song, Xiaokang Yang & Chuanfei Luo, A Parallel-Fusion Rnn-Lstm Architecture For Image Caption Generation, 2016 Ieee International Conference On Image Processing (Icip), Pp 4448-4452, 25-28, September 2016.
[39] Chetan Amritkar & Vaishali Jabade, Image Caption Generation Using Deep Learning Technique, 2018 Fourth International Conference On Computing Communication Control And Automation (Iccubea), Pp 1-4 , 16-18 Aug. 2018.
[40] Jiuxiang Gu, Gang Wang, Jianfei Cai & Tsuhan Chen, An Empirical Study Of Language Cnn For Image Captioning, 2017 Ieee International Conference On Computer Vision, Pp 1231-1240, 22-29 October 2017.