Retinal Disease Classification using CNN

Abstract

Retinal diseases impact millions of people every year and have always been one of the top priorities in the field of medicine and surgery as can be seen in the ever-growing research efforts dedicated towards retinal diseases. Occular fundus impairment is the cause behind many retinal diseases. These types of retinal diseases can be identified by the appearance of the fundus (the portion of the eye opposite the pupil). In recent years, neural networks are being used in the field of disease identification, achieving higher accuracy with every research effort. Given how well CNNs work with image analysis and classification, CNNs are being used in such applications which require localized feature identification in input images. In this project, our goal is to automatically classify the diseases using the Fundus image analysis without performing any explicit segmentation or feature extraction. We took the task to classify 46 retinal diseases using the RFMID dataset. We have experimented with multi-layered CNN, transfer learning techniques and UNet to observe the metrics in each of them. We have observed that weighted binary cross-entropy loss function performed better than standard binary cross-entropy loss for our dataset. We have observed that for our dataset, we were able to achieve maximum AUC-ROC score of 0.9312 and PR-metric of 0.6231 with U-Net architecture.

1. Introduction

Vision is one of the most vital senses which helps humans perceive the outside world on a much greater scale. Retinal diseases such as Dibetic Retinopathy, Age Related Maccular Degeneration and Glaucoma are the rampant ratinal diseases that can cause irreversible damage if timely diagnoses and treatment are not received [4]. This has been one of the primary factors behind development in the techniques for retinal disease identification and treatment.

Healthcare systems do not get their due attention in rural areas; as a result basic primary healthcare facilities are not at the disposal of the people. In such cases, availability of a practiced healthcare professional seems far from reality. With limited capacity, the assigned healthcare practitioner might not be able to cater well to the geographical areas under them. In such scenarios, an automated health inspection system capable of performing preliminary diagnosis can go a long way.

Artificial Neural Networks have achieved great accuracies in the targeting problem statements that involve supervised learning. Given how well they perform with identifying patterns and classifying images related to the field of medicine, artificial neural networks are being employed in applications such as automated disease detection systems.

Fundus is the portion of eye opposite to the pupil. Othalmologists use retinal fundus images to diagnose retinal problems. Early detection can improve chances of cure and also prevent blindness. Additionally, it makes the treatment easier and it adds several years of healthy vision to the patient's life. However, early detection of the minor changes in the patterns of fundus images might be missed by the human expert. Neural networks can help in identification early onset of retinal diseases. With such a very real impact of the application, use of neural network for retinal disease detection is the need of the hour.

2. Related Work

Ocular impairment detection often requires diagnosis based on study of fundus images. CNNs excel in identification of localized attributes of images and hence are well suited when working with image data [2]. Such automated disease identification can aide medical practitioners in dedicating due attention to fundus samples identified as diseased by the automated retinal disease identification systems and reduce the cases of false negatives.

Many of the previous studies have mainly been focused to detect Diabetic Retinopathy through deep learing [3] [12] [16]. Some researchers have also used optical coherence tomography images to detect retinal diseases [5].

Developments in CNNs have made it feasible to extract predictive features of the data given large labelled dataset [10]. Since the dataset we have used is relatively new, little work has been done on it to benchmark the dataset. This paper [11] suggests multiple models which the authors tried and got AUC-ROC value of above 95%. Since the models suggested are pretty complex, we have taken a more manageable approach inspired from [7] as our baseline, which works on fundus images to classify images as against the forty-six diseased classes.

We have also explored best suitable activation functions

for output layer in the customized CNN model from the activation functions available for neural networks as discussed in [13]

3. Approach

Our model takes as input fundus images and performs preliminary pre-processing on them. These pre-processed images are then fed to the Neural Network to begin the training. After series of combination of Convolutional layers, Max Pooling, Dropout layers and Batch normalization, the model predicts labels which are then compared against the ground truth labels. The model then backprops to optimize its parameters in order to reduce the loss. A more detailed description of our approach is mentioned as follows:

3.1. Consulting Opthalmologist

We started off by consulting Dr. Neeraj Gupta so as to get a better analysis of the dataset.

The Opthalmologist explained the attributes observed by an expert when alayzing fundus images: In Figure 1 The first image shows whitening around optic disc called ischaemia and disappearance of retinal arteries is because of anterior ischaemic optic neuropathy. The second image shows the whitening of retinal vessel, called Sclerosis of vessel, because of branch of retinal vein occlusion caused by Diabetic Retinopathy or Hypertensive Retinopathy. The third image contains multiple haemorrhages (encircled) because of leakage of vessels in Diabetic Retinopathy.

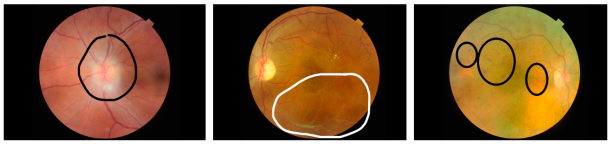

Figure 1. Features observed by an opthalmologist

3.2. Pre-processing

The original dimensions of the images were 4288x2848x3 or 2144x1424x3. As these images were too large to process in the available resources, we resized them to a more compact size of 500x300x3.

To make our model more robust, we performed data augmentation via following operations:

- Re-scaled the data to the range (0,1)
- Rotated the image by 40 degrees
- Performed a horizontal flip of the images

For the validation data and test data we only performed re-scaling of the data to range (0,1).

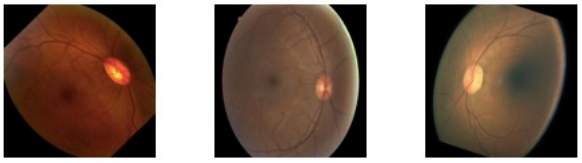

Figure 2. Augmented data samples before re-scaling

3.3. Base Model Architecture

We have made modifications to the architecture proposed in the base paper [8].The base model contained two convolutional layer, each followed by max-pooling and two fully connected layers at the end.

We have modified the base model by adding one more convolutional layer and we have included batch normalization layer [6] after each convolutional layer followed by max-pooling. At last, the model contains two Fully connected layers feeding into a final dense layer producing forty-six output, one for each label. The architecture diagram 3 depicts the same.

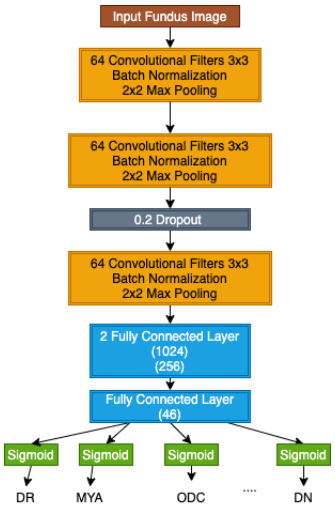

Figure 3. Modified Base Model Architecture

4. Experiments

4.1. Dataset

We used the RFMiD dataset [14] of fundus images mentioned in the base paper [8]. The dataset consists of forty-six labels of retinal diseases. After initial analysis, we found that the dataset was imbalanced with as less as single sample for some classes. Additionally, in order to not let more number of target labels weakening the performance [9], we performed initial experiments with six classes with highest number of corresponding samples in order to mitigate the issue of bias towards some labels. In this approach, the distribution of the data sample was as stated Figure 4.

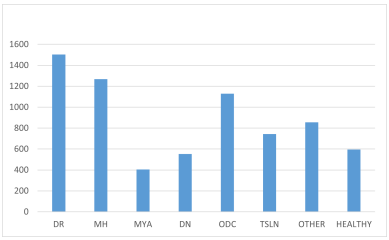

Figure 4. Taxonomy of Data after Augmentation

However, on later experiments, we saw that our model performance was almost unhampered when using data for all forty-six labels. Hence, our final metrics have been calculated based on the model performance for all the labels.

4.2. Details

4.2.1 Initial Experiments

Initially to tackle the problem of biased dataset, we had reduced the number of classes to 8, the taxonomy of which is stated in figure 4. On this dataset we first tried a two layered CNN Model with two fully connected layers and we were able to achieve 50% accuracy on the data without augmentation and 57% with augmentation. Sometimes the loss was going to NAN therefore the need for Batch Norm was necessary.

4.2.2 Later Experiments

As we learnt that using entire dataset for all labels was not hampering performance of our mode, we changed our loss function from binary cross entropy to weighted binary cross entropy. We even allowed multiple labels to be assigned to a single image.The figure 7 shows the training metrices of the base model.

4.2.3 Transfer Learning Approach

We also applied transfer learning approaches with two models independently. For this we used VGG-16 and ResNet. We were able to achieve accuracy of 99% with both VGG-16 and ResNet. These also showed impressive AUC-ROC scores. The figure 5 below shows the flowchart of the transfer learning approach employed by us.

4.2.4 UNet Architecture

U-Net is an enhancement over CNNs specifically designed to work with biomedical image data [15]. Its architecture has two main parts namely encoder and decoder. The encoder reduces the spatial dimensions in every layer and increases the channels. On the other hand, the decoder increases the spatial dimensions while reducing the channels. The encoder is all about the convolutional layers followed by pooling operation. It is used to extract the factors in the image. The decoder uses transposed convolution to permit localization. Hence, we also experimented our dataset on the UNet architecture 6.

4.3. Last Activation Function:Sigmoid

Since our problem statement is multi-categorical classification, it implies every fundus image can have multiple labels of diseases assigned to it. Since softmax predicts the probability of each label from the last layer output such that sum of probabilities of all the labels is one, this is not a valid use case for our problem statement. Hence, we have chosen sigmoid, which is an element wise activation function, as our final activation function:

$$S(x) = \frac{1}{1 + e^x}$$

where
$S(x)$ Sigmoid function.
x image score for a label

4.4. Loss Function

Due to class imbalance we explored some loss functions and found that wWeighted Binary Cross Entropy loss function to be the most suited for our model:

$$J_{wbce} = -\frac{1}{M}\sum_{m=1}^{M}[w \times y_m \times \log\left(h_\theta(x_m)\right) + (1 - y_m) \times \log\left(1 - h_\theta(x_m)\right)]$$

where
M number of training examples.
w weight
y_m target label for training example m
x_m input for training example m
h_θ model with neural network weight theta

We compared Binary Cross Entropy Loss and Weighted Binary Cross Entropy Loss and found that weighted binary performs better since binary cross entropy doesn't penalize for mistakes for smaller classes as weighted binary cross entropy does.

4.5. Metrics

Metrics explored:

- Model Loss: This is an average of value by which the model is far from the ground truth. Training loss is metric against which the model tunes its parameters to eventually achieves least possible test-time loss.

Model	ROC AUC	PR AUC	Accuracy	Recall	Precision	Hamming Loss	F1
Basemodel	0.9290	0.5771	0.8422	0.7438	0.2975	0.0803	0.4138
VGG-16	0.9297	0.5419	0.99	0.8264	0.2073	0.1331	0.3265
ResNet-50	0.9286	0.5346	0.997	0.8400	0.1928	0.1468	0.3097
U-Net	0.9312	0.6231	0.995	0.8511	0.1997	0.1421	0.3162

Table 1. Comparison of Metrics on Test Set

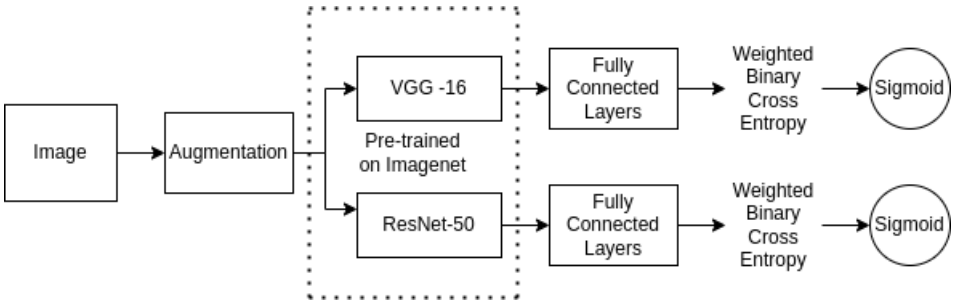

Figure 5. Transfer Learning Models Flowchart

- Precision: This is an evaluation of performance of the model with respect to the positive predictions for the positive labelled data. Precision in essence is ratio between number of true positives and total number of positive prediction.
- PR Metric: It visualizes Precision and Recall under single curve.
- AUC-ROC Metric: It is visualization of trade-off between True Positive Rate and False Positve Rate. An ideal ROC curve is pushed toward top left corner.
- Hamming Loss: It is the ratio of total number of wrong labels and total number of labels and is calculated as hamming distance between predicted labels and true labels. In a multi-label scenario like ours, hamming loss only penalizes individual labels. [1]

System with low Precision indicates high false positive rate. A retinal disease classification with low precision will end up classifying fundus images as unhealthy even when they are not. Alternatively, it might also result in more labels assigned to each sample than expected by the ground truth. Hence a model with low precision score is far less preferred in this scenario.

4.6. Quantitative Evaluation

The results of the models have been stated in 1. Additionally calculate the average hamming score. UNet gave 0.38, basemodel gave 0.41, VGG-16 gave 0.35 and ResNet gave 0.36 scores.

Base Model 7 gave best Precision, F1 score and Hamming Loss than other models.The UNet Model 10 gave best ROC AUC score, PR AUC score, Recall but had the lowest F1 score. Metrics of The transfer learning model of ResNet 9 and The transfer learning model of VGG-16 8 have also been stated.

Observing the graphs we could state that ResNet architecture had lowest variance in its metrics where as VGG and BaseModel had the highest.

5. Conclusion

We were able to see an improvement in performance as compared to the base model when using layers like batch normalization and dropout. Additionally, we saw an increase in accuracy when using different augmentation techniques as the model became more robust to different image orientations and styles.

With use of transfer learning techniques such as VGG-16 and ResNet, we saw high increase in accuracy as compared to our base model. Since these models are pre-trained on diverse dataset such as ImageNet, they carry with them a foundational learning about feature recognition, which exceeds the performance achievable by a model trained from scratch especially in a limited resources setting like ours.

We also learnt how multi-class classification loss function can be different from multi-categorical multi-class classification loss function. We applied conditional learning rate decay after every 2 epochs. With this,we observed a smoother convergence of the model.

5.1. Future Scope

The first area of improvement is volume of data. With bigger quantity of data, the model will have plenty of training for its parameters and will be able to achieve higher accuracy.

The accuracy of the model can further be improved with the use of transformers. Transformers employ attention

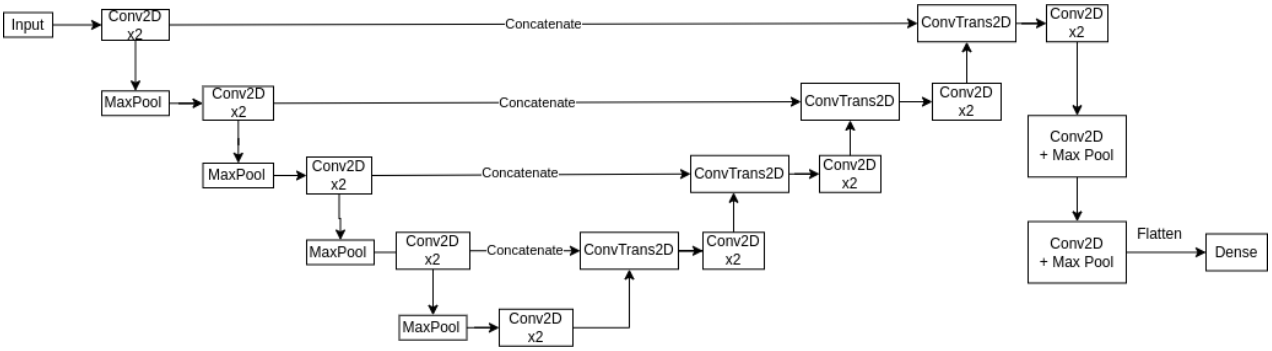

Figure 6. U-Net Architecture

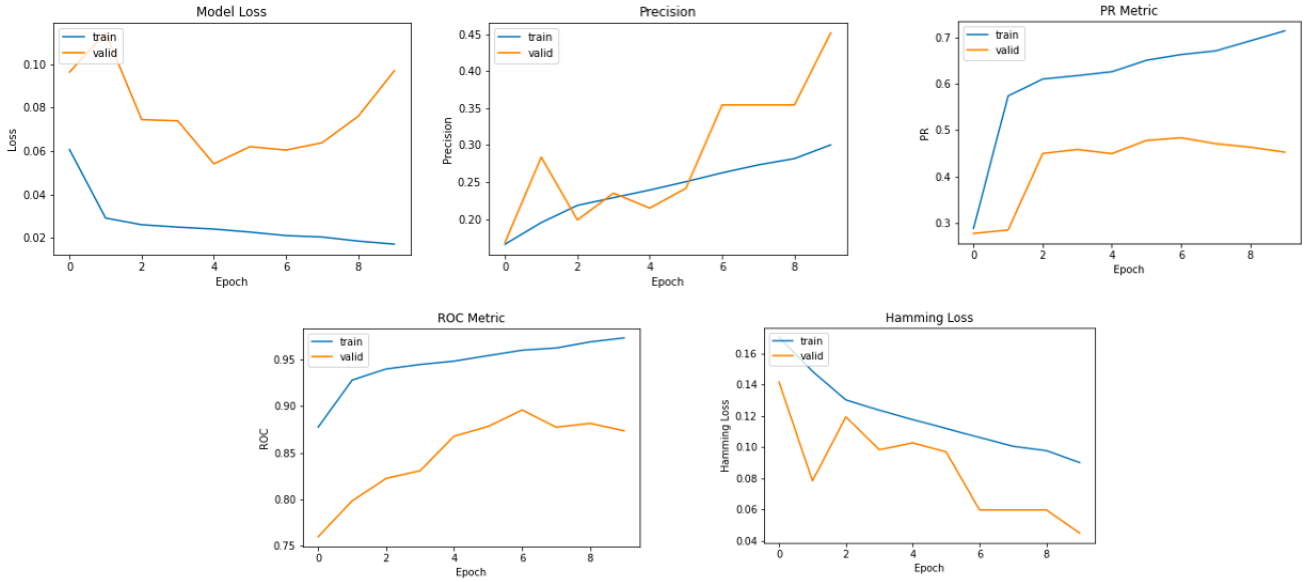

Figure 7. Base Model Training Matrices

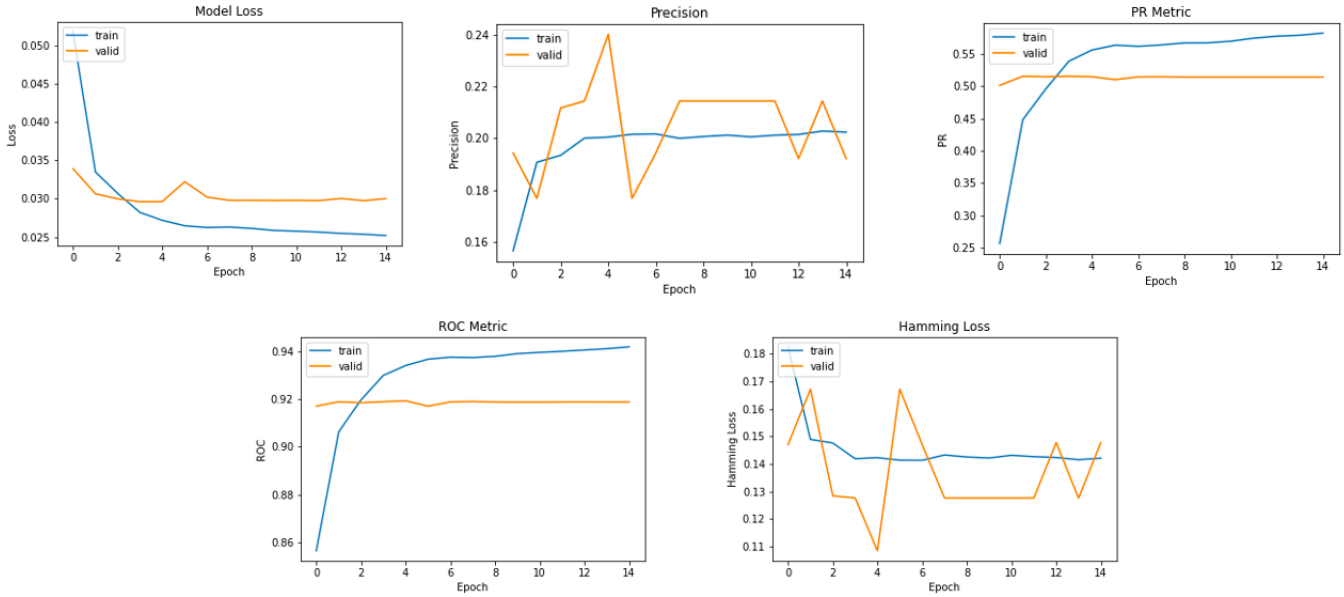

Figure 8. VGG Training Matrices

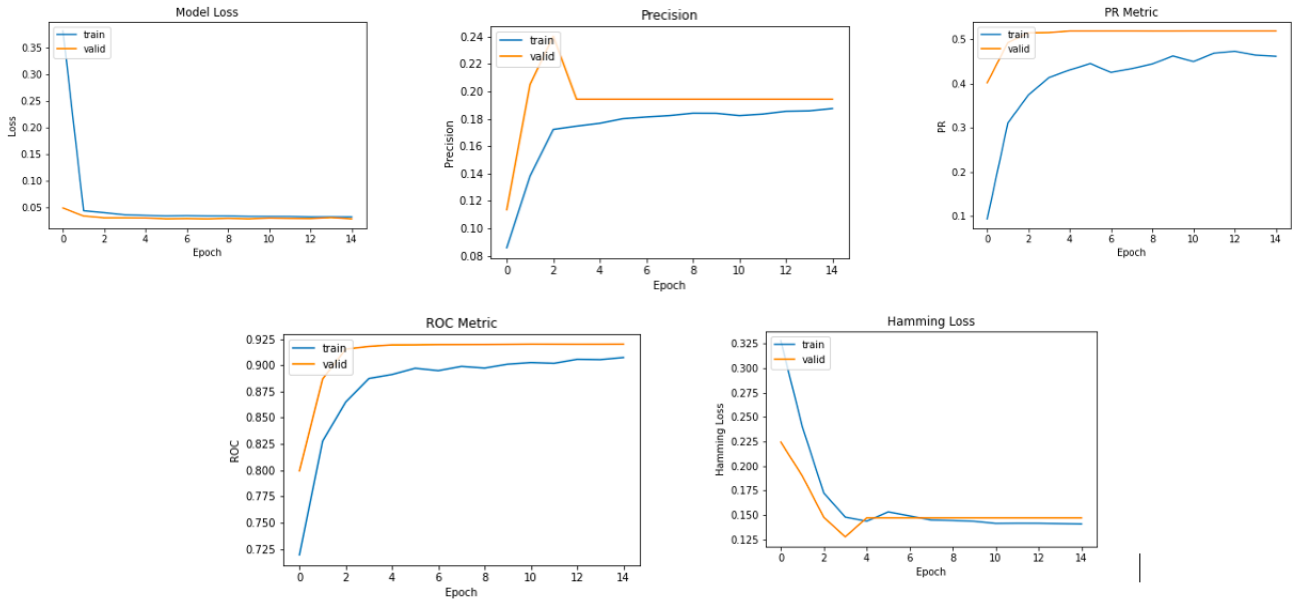

Figure 9. ResNet Training Matrices

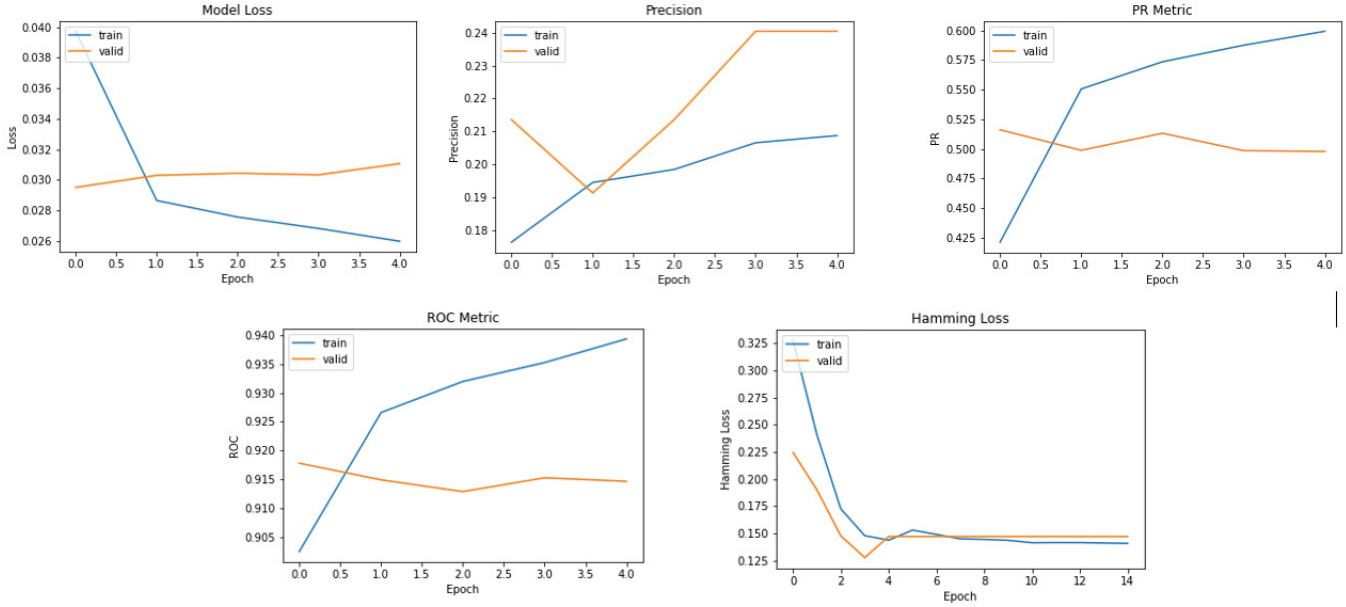

Figure 10. U-Net Training Matrices

mechanism which can prove vital when identifying localized features inside complex images such as biomedical images.

References

[1] tensorflow.org. 4

[2] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. Understanding of a convolutional neural network. In 2017 International Conference on Engineering and Technology (ICET), pages 1–6, 2017. 1

[3] Wejdan L. Alyoubi, Wafaa M. Shalash, and Maysoon F. Abulkhair. Diabetic retinopathy detection through deep learning techniques: A review. Informatics in Medicine Unlocked, 20:100377, 2020. 1

[4] Neely DC, Bray KJ, Huisingh CE, Clark ME, McGwin G, and Owsley C. Prevalence of undiagnosed age-related macular degeneration in primary eye care. JAMA Ophthalmol, 2017. 1

[5] Li F, Chen H, Liu Z, Zhang XD, Jiang MS, Wu ZZ, and Zhou KQ. Deep learning-based automated detection of retinal diseases using optical coherence tomography images.

Biomedical optics express, 10:6204–6226, 2019. 1

[6] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv.org, 2015. 2

[7] Lorick Jain, H V Srinivasa Murthy, Chirayush Patel, and Devansh Bansal. Retinal eye disease detection using deep learning. In 2018 Fourteenth International Conference on Information Processing (ICINPRO)- Retinal Eye Disease Detection Using Deep Learning, pages 1–6, 2018. 1

[8] Lorick Jain, H V Srinivasa Murthy, Chirayush Patel, and Devansh Bansal. Retinal eye disease detection using deep learning. In 2018 Fourteenth International Conference on Information Processing (ICINPRO), pages 1–6, 2018. 2

[9] Choi JY, Yoo TK, Seo JG, Kwak J, Um TT, and Rim TH. Multi-categorical deep learning neural network to classify retinal images: A pilot study employing small database. PLoS ONE 12, 2017. 2

[10] Y. LeCun, Y.and Bengio and Hinton G. Deep learning. PLoS ONE 12, 2017. 1

[11] Dominik Müller, Iñaki Soto-Rey, and Frank Kramer. Multi-disease detection in retinal imaging based on ensembling heterogeneous deep learning models, 2021. 1

[12] Tahira Nazir, Aun Irtaza, Ali Javed, Hafiz Malik, Dildar Hussain, and Rizwan Ali Naqvi. Retinal image analysis for diabetes-based eye disease detection using deep learning. Applied Sciences, 10(18), 2020. 1

[13] Chigozie Enyinna Nwankpa, Winifred Ijomah, Anthony Gachagan, and Stephen Marshall. Understanding of a convolutional neural network. pages 1–6, 2018. 2

[14] Samiksha Pachade, Prasanna Porwal, Dhanshree Thulkar, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, Luca Giancardo, Gwenolé Quellec, and Fabrice Mériaudeau. Retinal fundus multi-disease image dataset (rfmid): A dataset for multi-disease detection research. Data, 6(2), 2021. 2

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. 2015. 3

[16] Ayaka Sugeno, Yasuyuki Ishikawa, Toshio Ohshima, and Rieko Muramatsu. Simple methods for the lesion detection and severity grading of diabetic retinopathy by image processing and transfer learning. Computers in Biology and Medicine, 137:104795, 2021. 1