Diabetic Retinopathy Detection through image analysis using Deep Convolutional Neural Networks

Jordi DE LA TORRE, Aida VALLS, Domenec PUIG
Departament d'Enginyeria Informàtica i Matemàtiques
Universitat Rovira i Virgili, Tarragona
e-mail: jordi.delatorre@gmail.com, aida.valls@urv.cat, domenec.puig@urv.cat

Abstract. Diabetic Retinopathy is one of the main causes of blindness and visual impairment for diabetic population. The detection and diagnosis of the disease is usually done with the help of retinal images taken with a mydriatric camera. Manual inspection of these images requires lots of human resources that could be significantly reduced with the use of computer-aided automatic detection tools.

In this paper we propose an automatic retina image classifier that using supervised deep learning techniques is able to classify retinal images in five standard levels of severity. In each level of severity different irregularities appear on the image, due to micro-aneurisms, hemorrages, exudates and edemas. This problem has been approached before using traditional computer vision techniques based on manual feature extraction. Differently, we explore the use of the recent machine learning approach of deep convolutional neural networks, which has given good results in other image classification problems.

From a training dataset of around 35000 human classified images, different convolutional neural networks with different input size images are tested in order to find the model that performs better over a test set of around 53000 images. Results show that is possible to get a classification accuracy above 0.70 (quadratic weighted kappa index) using only the information contained in the image of the retina of the studied eye.

Keywords. deep learning, convolutional neural networks, supervised learning, computer vision, diabetic retinopathy

1. Introduction

Diabetic Retinopathy (DR) is a leading disabling chronic disease and one of the main causes of blindness and visual impairment in developed countries for diabetic patients. Studies reported that 90% of the cases can be prevented through early detection and treatment [1]. Eye screening through retinal images is used by physicians to detect the lesions related with this disease. Due to the increasing number of diabetic people (which is expected to grow to by 2050), the amount of images to be manually analyzed is becoming unaffordable. Moreover, training new personnel for this type of image-based diagnosis is long, because it requires to acquire expertise by daily practise. In 2003 the medical community established a standardized classification based on four severity stages [2] determined by the type and number of lesions (as micro-aneurysms, hemorrhages and exudates). The five classes in the DR severity scale are:

0. - No apparent retinopathy
1. - Mild Non-Proliferative Diabetic Retinopathy (NPDR)
2. - Moderate NPDR
3. - Severe NPDR
4. - Proliferative DR

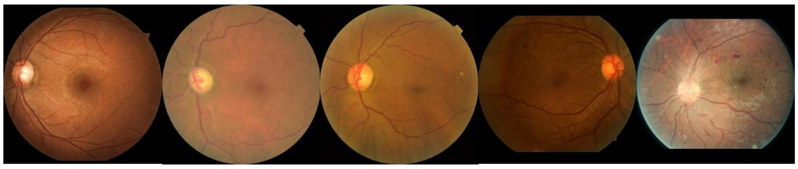


Figure 1. Image samples of the five different DR severity classes sorted from 0 (left) to 4 (right)

In the last decade, several attempts to automatize the DR diagnosis through images of the eye fundus have been tested. They are basically focused on applying well-known pattern recognition models. In this paper, we want to apply a Deep Convolutional Neural Network (DCNN) model, as it has been proven to be a very effective algorithm to solve general image classification problems. DCNN is a supervised learning model for automatic classification that does not require any pretreatment of the images, nor any feature analysis. Deep learning techniques are focused on learning multiple levels of representation and abstraction that help to make sense of the hidden information in data such as images. In this way having a complete set of correctly classified images and without having any a priori understanding of the features required to make the classification, the system is able to learn the properties of the image that minimize a defined cost function that is direct or indirectly related with the classification score index that has to be optimized. In this article we show that a DCNN is able to learn from data the most important features to make the classification of retinal images into the five DR categories, without the need of a hand-crafted feature extraction process.

The paper is organized as follows: we present the background of this work, next we explain the characteristics of the available data and why deep learning techniques can be applied over them, next we explain the methodology used for solving the problem, we show the obtained results and finally we expose the conclusions and further steps for improving the results.

2. Background

Traditional models of pattern recognition in images since the late 50s have been based on extracting hand-crafted fixed engineered features or fixed kernels from the image and, then, using a trainable classifier on top of those features to get the final classification. Using this model, the problem of the DR detection has been tackled by feature extraction using on hand models targeted to the detection of microaneurisms, haemorrhages and exudates in retinal images (e.g. [3], [1], etc). This type of approach requires a good understanding of the mechanism of the disease to be able to find the important features present in the image. This knowledge is specific of the problem to be solved, requires a lot of labor time and is task specific and thereby not reusable for other different classification problems.

Deep learning is a new powerful method for supervised learning. By adding more layers and more units within a layer of a neural network, we can represent functions of increasing complexity. Training must be done on sufficiently large annotated image datasets. In computer vision classification problems, neural networks have largely displaced the traditional approaches based on handcrafted features. They have proved to be the best available method for solving the biggest classification challenges, like for example IMAGENET[4]. In this previous type of image analysis problems, distinct objects should be recognized on different types of images. In the case of DR classification, all images are very similar (as they are retina photos) and very small and vague lesions have to be detected at different locations in order to find the correct severity category of DR. This paper, thus, wants to study the performance of DCNNs for this kind of medical image analysis.

3. Data

Before defining the methodology to use, it would be better to first understand the nature of the available data. Deep learning techniques are a very effective method for classification but it is not useful in all circumstances. Specifically, supervised deep learning techniques require big enough datasets (thousands of samples per class order of magnitude) to perform well.

The dataset consists of two independent high resolution image sets (train and test). For every patient right and left eye images are reported. All the images are classified by ophthalmologists according to the standard severity scale presented before [2]. The images are taken in variable conditions: by different cameras, in different illumination conditions and have different resolutions. The training set contains a total of 35.126 images; 25.810 of class 0, 2.443 of class 1, 5.292 of class 3, 873 of class 3 and 708 of class 4. The test set contains a total of 53.576 images; 39.533 of class 0, 3.762 of class 1, 7.861 of class 2, 1.214 of class 3 and 1.206 of class 4.

4. Methodology for retinal image classification

All the state-of-the-art architectures for supervised deep learning over images (e.g. AlexNet[5], GoogleNet[6] and VGGNet [7]) are based on convolutional neural networks (CNNs). A set of convolutional layers with dimensional down-sizing blocks between them, also known as pooling layers, extract the statistical information from the data (feature extraction) that is passed to the posterior layers to construct more elaborate abstractions (features of features) that are useful for the classification. As a final stage a fully connected set of layers perform the classification based on the information coming from the last layers of the convolutional network. This kind of structure provides an end to end learning process, where either the classes or the features are learned from data with no human intervention.

Next we explain the different phases of the DCNN construction for performing diabetic retinopathy detection based on the available data: evaluation function, data preprocessing and data augmentation, model, training, testing and probabilistic combination of the models of both eyes.

4.1. Evaluation function

The performance of the classification model is measured using the quadratic weighted kappa (QWK) index 1. QWK, also known as Cohen's Kappa, measures inter-rater agreement for categorical items in multi-class classification problems[8], penalizing the discrepancy quadratically with the distance from the correct class. It is generally thought to be a more robust measure than simple percent agreement calculation, since QWK takes into account the agreement occurring by chance. This metric typically varies from 0 (random agreement) to 1 (complete agreement). Negative values are also possible, the maximum possible negative value (-1) indicates a complete disagreement between classes. Being $O_{i,j}$ the observed values, $E_{i,j}$ the expected ones and C number of classes, QWK is defined as:

$$\kappa = 1 - \frac{\sum_{i=1}^{C} \sum_{j=1}^{C} \omega_{i,j} O_{i,j}}{\sum_{i=1}^{C} \sum_{j=1}^{C} \omega_{i,j} E_{i,j}} \quad \text{where} \quad \omega_{i,j} = \frac{(i-j)^2}{(C-1)^2} \tag{1}$$

The QWK coefficient is used to compare the performance of different prediction models and with the performance of the human experts [9]. Individual human experts report values of QWK of about 0.80 in the prediction of the correct class in the DR disease. This value is considered excellent because small discrepancies between the class prediction does not affect the treatment of the disease. The most important is the differentiation between presence or absence of the disease.

4.2. Data pre-processing and Data Augmentation

Firstly, DCNNs require large data-sets in order to avoid overfitting. A class balanced data-set is also desirable as well [10]. One of the proven approaches that has good results to get more data from small data-sets is to use data augmentation techniques[5]. The data augmentation is done in two stages: first a copy of the training examples of the small classes is done until they have the same number of images as the biggest class. This generates an equilibrated training set. After this first step, to every image the next transformations are applied in order to diversify the training examples:

- Cropping, random uniform
- Rotation (0° to 360°), random uniform
- Mirroring (X and Y axes), random uniform
- Brightness correction, random gaussian (σ = 0.1)
- Contrast correction, random gaussian (σ = 0.1)

All these transformations are applied to every image of the balanced training set and redone for every training epoch. This transformations assure that every image of the training set is different between each other for every epoch, making the final prediction invariant to rotation, brightness and contrast over the training set.

Secondly, DCNNs perform better when then input data is normalized by having mean equal to 0 and standard deviation equal to 1, on each channel (RGB). Thus, a normalization must be done on each image.

4.3. Model

The well-known LeCun et al. architecture has been taken [11]. It is composed by a series of convolutional layers followed by activation function blocks, with some max-pooling dimensional reducing blocks. At the end, a final fully connected layer performs the classification and, finally, a softmax output layer gives a probability estimation of every class (Figure 2).

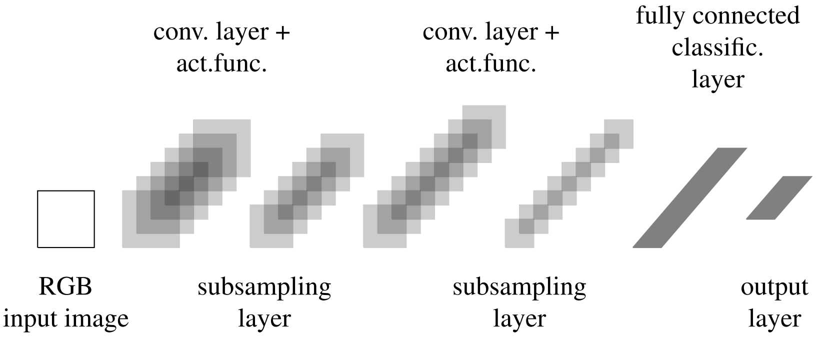


Figure 2. Architecture of a 4 layer CNN: two convolutional layers, one fully connected classification layer and the final output layer

Some of the parameters of this model are the following: input size, number of layers, number of filters per convolution layer, size of the convolution, number and size of classification layers, activation function, optimization and regularization methods to be used. Due to the high number of parameters involved, these systems have a lot of versatility and there is not a unique solution to a problem. Different configurations can perform well solving the same problem.

4.4. Training procedure

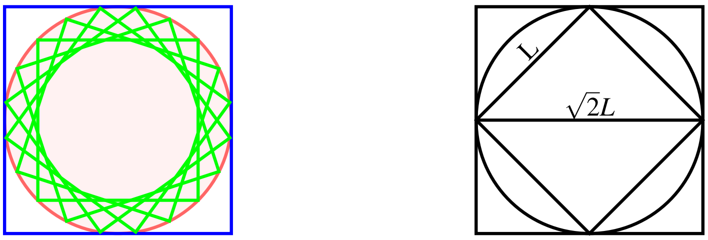
As a multi-class classification problem, a log-loss function is used to perform the optimization in the learning stage. The original training set is splitted in two random subsets: one with 90% of the data and other with 10%. The last one is used as a cross validation set for hyperparameter selection. QWK results are calculated every epoch either for the training or the test set. The model chosen is the one that maximizes the QWK over the cross validation set. In all models a ReLU or leaky ReLU[12] activation function is used. In all layers a batch normalization [13] is applied before the activation function in order to reduce the gradient vanishing problem that occurs in deep networks, reduce the internal covariance shift and improve the regularization. As an additional regularizer a dropout [14] (p=0.5) is performed before the final classification layer. L2 regularization has been tested with no significative improvements in the final classification results. A random initialization based in the Kaiming&He approach[15] is used for all the networks. When deep networks are randomly initialized using fixed standard deviations (e.g., 0.01), very deep models (e.g., >8 conv layers) have difficulties to converge, as reported by the VGG team [16]. Kaiming method takes into account the particularities of rectifier non-linearities. Initializing the weights of every layer randomly with 0 mean and $\sqrt{2/n_l}$ standard deviation (being n_l the number of connections of layer l) we get a desired zero-mean Gaussian distribution in the weight distribution. Biases are initialized to zero. All models are optimized using stochastic gradient descent with Nesterov momentum.

4.5. Testing procedure

The network is trained with a data augmented scheme that include rotations. Presumably an ensemble[17] that includes rotated versions of the original image would perform better that the single original image. Using a trained network with a significant accuracy, different ensemble schemes are tested over the cross validation set to identify the one that maximizes the classification accuracy.

Testing scheme	Predictions	κ_{CV}
Baseline: Original image (crop center)	1	0.669
Original + rot 180° (crop center)	2	0.683
Center, Left top, Left bottom, Right top, Right bottom cropped	5	0.684
Original + Hflip + Vflip + rot 180°	4	0.686
Five 72° rotated (crop center)	5	0.699
All	14	0.701

Table 1. Testing scheme performance results



(a) Ensemble of five crops

(b) Cropped vs image size

Figure 3. Testing ensemble used for evaluation

Table 1 shows the accuracy of different testing schemes for a given model. The ensemble that performs the best is the one formed by all the 14 different evaluations. The Five 72° scheme is only 0.002 points under the best tested scheme and reduces significantly the computation time required. Geometric means perform slightly better (average of 0.005 points of κ) than arithmetic means. The testing scheme chosen is the Five 72° with geometric means due to its good compromise between computation costs and performance. In figure 3 we show a representation of the scheme. Blue square represents the image size, red circle the retinal image area and the five green squares, the areas of the image that are feeded to the neural network at test time. As the diagram show, most of the useful information for the classification is considered by one of the 5 different inputs.

4.6. Probabilistic combination of the models of both eyes

Diabetic Retinopathy usually affects both eyes, specially when the illness is in high severity stages. We analyze the frequency of co-ocurrence of the five DR categories in the images of our dataset. The dataset is big enough to infer from the frequencies of co-occurrence of the classes, the conditional probabilities of having one class in one eye given another class in the other.

In table 2 we show the frequencies of occurrence of all the possible combinations of classes in both eyes. Notice that the larger frequencies are found in the diagonal, indicating that is usual that both eyes have similar levels of severity.

Eyes	C0	C1	C2	C3	C4	Sum
C0	12155	407	295	3	11	12871
C1	435	600	171	2	4	1212
C2	336	222	1998	96	50	2702
C3	3	1	87	307	27	425
C4	10	1	39	40	263	353
Sum	12939	1231	2590	448	355	17563

Table 2. Frequencies of combined occurrence of classes in both eyes (left: rows, right: columns)

Using the frequentist interpretation of probability that defines an event's probability as the limit of its relative frequency in a large number of trials, we use this frequencies as an estimation for the calculation of the conditional probability. In table 3 we show the values of all the calculated conditional probabilities of $P(Left|Right)$. The same can be done for the matrix $P(Right|Left)$.

Eyes	C0	C1	C2	C3	C4
C0	0.93940	0.33062	0.11389	0.00669	0.03098
C1	0.03361	0.48740	0.06602	0.00446	0.01126
C2	0.02596	0.18034	0.77142	0.21428	0.14084
C3	0.00023	0.00081	0.03359	0.68526	0.07605
C4	0.00077	0.00081	0.01505	0.08928	0.74084

Table 3. Probability of occurrence of Left eye class (rows) given the occurrence of the Right eye class (columns)

Using the Bayes rule, we can estimate the probability distribution of eye A, using the probability distribution of eye B given by our model. Being $P(Left)$ and $P(Right)$, the probability distributions obtained by our predictive model with the left image and the right image, respectively, we can estimate P_L and P_R using Eq. 2. To merge the information obtained from out model $P(X)$ with the estimated coming from the other eye we calculate the arithmetic mean. The class with maximum value is the one selected for each eye.

$$\begin{aligned} P_L &= P(Left|Right)P(Right) \\ P_R &= P(Right|Left)P(Left) \end{aligned} \tag{2}$$

5. Experiments

Different experiments have been conducted to analyze the quality of the classification with different parameters of the DCNN. First, we perform an study of the best image size. Due to the type of image classification that is done, it is crucial to choose the right size of the input images in order to detect the important features involved in the RD severity detection. As explained before, cropping is part of the data augmentation scheme. The original size is chosen from the NN input size as $\sqrt{2}$ times the input size. In this way, due to the circular nature of the retina, we maximize the useful information of the square cropped from the center of the image (see Fig. 3). The input sizes tested are:

- Resizing to 181x181 cropped to 128x128 NN input
- Resizing to 362x362 cropped to 256x256 NN input
- Resizing to 543x543 cropped to 384x384 NN input
- Resizing to 724x724 cropped to 512x512 NN input

Different number of layers have also been studied for each image size. In table 4 are presented the higher classification rates obtained from the best models obtained for different input sizes. The number of layers of the best models are also shown.

Layers	Input size	κ_{test}	FN	FP
12	(3,128,128)	0.488	11.6%	11.5%
14	(3,256,256)	0.636	4.4%	28.7%
16	(3,384,384)	0.668	7.9%	14.9%
16	(3,512,512)	0.725	5.0%	11.9%

Table 4. Best classification results for different input sizes. FP (false positives), FN (false negatives)

As the input size and the complexity of the network is increased the results obtained become better. Greater sizes give more definition of the underlying features in the image and the increased complexity of the network, increasing the number of layers, allows the construction of abstractions in the form of features of features that improves the classification accuracy. In table 6 we show the architecture details of the best model, a very deep network of 16 layers. This indicates that the distinction of the severity categories of DR in images in not an easy classification problem. With the 512x512 model we achieve a $\kappa_{test} = 0.725$ using only the information contained in the examined retine image.

Next, we study if the inclusion of the co-ocurrence information of both eyes is able to improve the quality of the classifier. Table 5 shows the final accuracy obtained combining both eyes for all the sizes. In the 512x512 model the accuracy is increased in about 0.03 κ points, a value that makes our model perform near human level expertise with a final $\kappa_{test} = 0.752$.

Layers	Input size	κ_{test}	FN	FP
12	(3,128,128)	0.555	11.2%	12.9%
14	(3,256,256)	0.661	4.4%	28.7%
16	(3,384,384)	0.722	11.2%	4.0%
16	(3,512,512)	0.752	6.5%	7.0%

Table 5. Best classification results adding the probabilistic information of both eyes

Layer	Type	Characteristics	Output Size
	Input	3 RGB channels	(3,512,512)
1	Convolution	8 filters 3x3 1,1 stride, 1,1 padding	(8,512,512)
2	Convolution	16 filters 3x3 1,1 stride, 1,1 padding	(16,512,512)
3	Convolution	16 filters 3x3 1,1 stride, 1,1 padding	(16,512,512)
-	Max pooling	2,2 size, 2,2 stride	(16,256,256)
4	Convolution	32 filters 3x3 1,1 stride, 1,1 padding	(32,256,256)
5	Convolution	32 filters 3x3 1,1 stride, 1,1 padding	(32,256,256)
-	Max pooling	2,2 size, 2,2 stride	(32,128,128)
6	Convolution	64 filters 3x3 1,1 stride, 1,1 padding	(64,128,128)
7	Convolution	64 filters 3x3 1,1 stride, 1,1 padding	(64,128,128)
-	Max pooling	2,2 size, 2,2 stride	(64,64,64)
8	Convolution	128 filters 3x3 1,1 stride, 1,1 padding	(128,64,64)
9	Convolution	128 filters 3x3 1,1 stride, 1,1 padding	(128,64,64)
-	Max pooling	2,2 size, 2,2 stride	(128,32,32)
10	Convolution	128 filters 3x3 1,1 stride, 1,1 padding	(128,32,32)
11	Convolution	128 filters 3x3 1,1 stride, 1,1 padding	(128,32,32)
-	Max pooling	2,2 size, 2,2 stride	(128,16,16)
12	Convolution	128 filters 3x3 1,1 stride, 1,1 padding	(128,16,16)
13	Convolution	128 filters 3x3 1,1 stride, 1,1 padding	(128,16,16)
-	Max pooling	2,2 size, 2,2 stride	(128,8,8)
14	Convolution	256 filters 3x3 1,1 stride, 1,1 padding	(256,8,8)
15	Fully connected	256 elements	(256)
16	Softmax	256 to 5 elements	5

Table 6. Best performing DCNN architecture

6. Conclusions and Future Work

In this paper is shown that deep learning techniques that have been proven to be very effective for solving general classification problems, are also a good technique for solving medical imaging problems like the diabetic retinopathy detection. Having enough data this method is able to perform near human level expertise.

Future work will be centered on testing the newer schemes such as residual networks, the use of alternative cost functions that encode the prior information of the ordering of the classes and other more elaborated methods for combining the information coming from both eyes.

Acknowledgments

This work is supported by the URV grant 2014PFR-URV-B2-60 and the Spanish research projects PI15/01150 and PI12/01535 (Instituto de Salud Carlos III). The authors would like to thank to the Kaggle platform as organizer of machine learning competitions hosting the Diabetic Retinopathy Challenge, to the California Healthcare Foundation as sponsor of the competition and EyePACS for providing the images used in this study.

References

[1] Jordina Torrents-Barrena, Jaime Melendez, Aida Valls, Pere Romero, and Domenec Puig. Screening for diabetic retinopathy through retinal colour fundus images using convolutional neural networks. In *Artificial Intelligence Research and Development - Proceedings of the 18th International Conference of the Catalan Association for Artificial Intelligence, Valencia, Catalonia, Spain, October 21-23, 2015.*, pages 259–262, 2015.

[2] Wilkinson CP, Ferris FL 3rd, Klein RE, Lee PP, Agardh CD, Davis M, Dills D, Kampik A, Pararajasegaram R, and Verdaguer JT. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*, 110(9):1677–82, 2003.

[3] L. R. Sudha and S. Thirupurasundari. Analysis and detection of haemorrhages and exhudates in retinal images. *International Journal of Scientific and Research Publications*, 4:1–5, 2014.

[4] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[6] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[8] Joseph L. Fleiss, Jacob Cohen, and B. S. Everitt. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72(5):323–327, 1969.

[9] Moss SE, Klein R, Kessler SD, and Richie KA. Comparison between ophthalmoscopy and fundus photography in determining severity of diabetic retinopathy. *Ophthalmology*, 92(1):62–67, 1985.

[10] Mazurowski MA, Habas PA, Zurada JM, Lo JY, Baker JA, and Tourassi GD. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Networks*, 21(2-3):427–436, 2008.

[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998.

[12] George E. Dahl, Tara N. Sainath, and Geoffrey E. Hinton. Improving deep neural networks for LVCSR using rectified linear units and dropout. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8609–8613. IEEE, 2013.

[13] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Deep Learning Symposium (NIPS 2015)*, 2015.

[14] P. Baldi and P. Sadowski. Understanding dropout. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 26, pages 2814–2822, 2013.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[17] Zhi-Hua Zhou, Jianxin Wu, and Wei Tang. Ensembling neural networks: Many could be better than all. *Artificial Intelligence*, 137:239–263, 2002.