Benign and malignant mammographic image classification based on Convolutional Neural Networks

Bin Li
Robotics Institute
Shanghai Jiao Tong University
Shanghai, China
lbin_sjtu@sjtu.edu.cn

Yunhao Ge
Robotics Institute
Shanghai Jiao Tong University
Shanghai, China
gyhandy@sjtu.edu.cn

Yanzheng Zhao, Enguang Guan, Weixin Yan*
State Key Lab of Mechanical System and Vibration
Shanghai Jiao Tong University
Shanghai, China
xiaogu4524@sjtu.edu.cn

ABSTRACT
Computerized breast cancer diagnosis system has played an import role in early cancer diagnosis. For this purpose, we apply deep learning by using convolutional neural networks (CNN) to classify abnormalities, benign or malignant, in mammographic images based on the mini Mammographic Image Analysis Society (mini-MIAS) database. Accuracy, sensitivity, and specific values are observed to evaluate the performance of the CNN. To improve the performance, we utilize image-preprocessing methods containing cropping, global contrast normalization, augmentation, local histogram equalization, and balancing preprocessing. We built four CNN models to study the impact of depth and hidden layer structure on model performance. The CNN-4d model performs best among four proposed CNN models consisting of four convolution layers with a dropout of 0.7. The CNN-4d model achieved a balance of high sensitivity (90.63%) and high specificity (87.67%), and an accuracy of 89.05%. The result of this study indicates that CNNs have promising potential in the field of intelligent medical image diagnosis.

CCS Concepts
• **Computing methodologies → Object recognition.**

Keywords
Convolutional Neural Networks; Mammographic image; Breast abnormalities; Classification; Benign or malignant.

1. INTRODUCTION
Breast cancer is the second leading cause of death among women. According to a World Health Organization (WHO) report, breast cancer accounts for 22.9% of diagnosed cancers and 13.7% of cancer-related deaths worldwide [1]. Detection of breast cancer in its early stages dramatically increases the chances of a successful treatment plan [2]. The development of computer-aided diagnosis systems (CADs) that can assist medical personnel with the early detection of tumors serves as a crucial alternative. In such systems, a high reliability in the accuracy of the classifier is a top priority [3].

ICMLC 2018, February 26–28, 2018, Macau, China

ACM ISBN 978-1-4503-6353-2/18/02…$15.00
https://doi.org/10.1145/3195106.3195163

Thanks to recent developments, convolutional neural networks (CNN) have become one of the most popular methods for image classification and a driving force behind deep learning. Many researchers have studied mammogram classification using CNN models, and obtained significant results. Gallego-Posada, et al. [4] demonstrated an application of CNN for the detection and diagnosis of breast tumors. The Mammographic Image Analysis Society (MIAS) database was used, and 64.52% test accuracy was achieved. Jadoon, et al. [5] proposed a model that uses two methods, namely CNN-DW and CNN-CT, to classify results as normal, malignant, and benign. Using the IRMA dataset, the model achieved accuracy rates ranging from 81.83%-83.74%.

In this study, we proposed a CADs, including a preprocessing method and supervised classification method. In the classification method, a novel CNN model is proposed to classify abnormalities and benign or malignant tumors. The high accuracy and other outstanding evaluating indicator shows the CNN outperforms high performance as the key step in our mammography CADs. To improve performance of the classifier, we focused on image preprocessing, which is specifically suitable for the mammographic image and structure of CNN models. Our data source is the mini Mammographic Image Analysis Society (mini-MIAS) database [6]. The mini-MIAS database contains valuable information such as the location of the center of abnormality and radius of the circle that surrounds the abnormality. This information was used to crop the original image and prepare the data by cropping the region of interest (ROI). In addition, we built four CNN models to study the influence of the CNN structure on model performance.

The remainder of this paper is organized as follows. In section 2, we provide details about our methodology for preprocessing data and the structure of CNN. In section 3, we describe the experiment performed and results obtained. Finally, section 4 concludes this work.

2. MATERIALS and METHODS

2.1 Dataset
In this study, we test the performance of CNNs to classify the exact breast abnormalities that are obtained from mammograms. Our data source is the mini-MIAS database, which consists of 323 mammogram images, each of size 1024x1024 pixels. In the MIAS database, mammogram images are divided into three classes: glandular dense, fatty, and fatty glandular. Each class is subdivided into images of normal, benign, and malignant tissue. Each abnormal image, either benign or malignant, has a type such

as calcification, mass, and asymmetry. A total of 207 normal images and 116 abnormal images (64 benign and 52 malignant) were obtained. In this study, we only use the abnormal images in the dataset to classify the benign and malignant classes.

2.2 Proposed method

Fig.1 presents an overall view of the proposed method, comprising two main stages: preprocessing and supervised training. The preprocessing stage prepares the data in the ROI through a set of transformations, so that the next stage takes advantage of relevant characteristics in the ROI. Supervised learning in the second stage involves two processes: feature learning and classification training. Both are performed by training a CNN. The convolution layers and pooling layers extract the features, while the back propagation (BP) algorithm updates the parameters in the hidden layers to achieve feature learning. In addition, the fully connected layers and the last softmax layer create the final classification based on the extracted features mentioned above. We note that the two processes are in the supervised stage, since the CNN training is guided by the labeled samples.

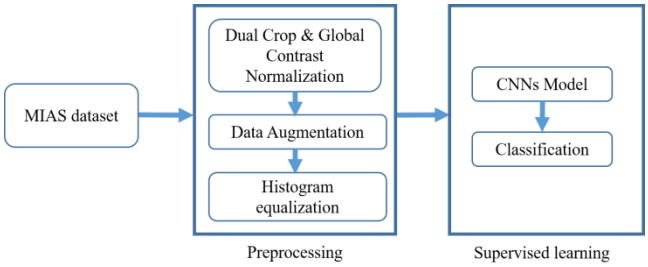

Figure 1. Workflow diagram of the proposed method.

2.2.1 Preprocessing

Preprocessing is a common stage in CADs that enhances the characteristics of the image by applying a set of transformations to improve performance. We apply cropping, global contrast normalization, augmentation, local histogram equalization, and balancing preprocessing to the datasets and show their effects on the accuracy of the final classification.

2.2.1.1 Double cropping and global contrast normalization

The first crop of each image eliminates black spaces and useless noise in the mammogram image, such as patients information and icons. After the first crop, the global contrast normalization (GCN) is conducted. Due to the digitalization process, the lighting conditions between different film images will be different, which affects all pixel values of the image. A GCN eliminates this effect by subtracting the mean of the intensities in the image from each pixel. Because the mean is of the image instead of each pixel, it can be subtracted without determining whether the current image belongs to the training, validation, or test set [7].

$X \in R^{r*r}$ is the image, and the element-wise transformation is

$$X'_{i,j} = X_{i,j} - x \tag{1}$$

where $x \in R$; $x = \frac{1}{r^2}\sum_{i,j} X_{i,j}$ is the mean of the X image intensities, and $X_{i,j} \in R$ is the intensity in the i, j pixel.

The second crop classifies a previously identified ROI in the entire film image after GCN. Using the smaller sizes of images, the ROI reduces training time. In the mini-MIAS dataset, each image provides valuable information about the type of abnormality, the coordinates of abnormality, and the approximate radius of the circle surrounding the abnormality. The abnormal ROI of each abnormal image is extracted using the x and y-coordinates of the center of the abnormal images and radius r.

We fixed the input size to ROIs of (2r, 2r) pixels. With this, ROIs can be easily extracted using the bounding box of the segmented region. Specifically, ROIs were cropped to the square bounding box of the lesions and reshaped to (2r, 2r) pixels. The ROIs area is greater than that of circle with radius r, so the lesion is centered with scaling and the surrounding region is preserved. And the diagram of preprocessing is shown in Fig. 2.

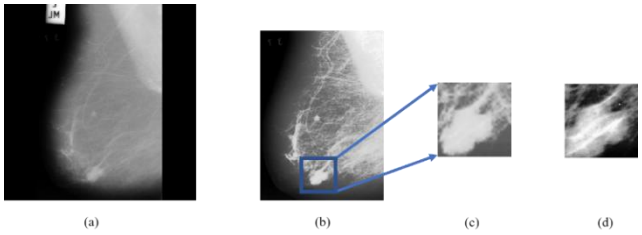

Figure 2. Diagram of preprocessing (a) origin image (b) image after first crop and global contrast (c) ROI after second crop (d) ROI after local histogram equalization

2.2.1.2 Augmentation

Data augmentation is often used in the context of deep learning and refers to the process of generating new samples from existing data, which is used to ameliorate data scarcity and prevent overfitting [8]. Transformations include rotations, translations, horizontal and vertical reflections, crops, zooms, and jittering. For tasks such as optical character recognition, Simard et al. [9] showed that elastic deformations can greatly improve performance. The main sources of variation in mammography at the lesion level are rotation, scale, translation, and amount of occluding tissue.

We propose a argument algorithm that is more suitable for each resultant mammogram, with the following characters: single-channel, low-contrast, small area-of-interest, and slow change in texture gradient. We augmented all positive examples with scale and translation transformations. Full scale or translation invariance is not desired nor required, since the candidate detector is expected to find a patch centered on the actual focal point of the lesion. The key is to perform the proper amount of translation and scaling, to generate realistic lesion candidates. The algorithm provides three options for data argumentation. First, flipping and rotating angle (including horizontal and vertical flipping) is widely used. Rotations transformations of 0, 90, 180, 270 can be used, which avoid the existence of points outside the boundaries. Second, by centering the ROI range, Δx and Δy are the difference of the ROI center and lesion center provided in the MIAS label. Variations Δx and Δy identify the translation of lesions in the ROI. Finally, the scaling ratio, enlargement, and reduction of the lesion area in the ROI can adjust the area of the surrounding region, which may preserve more texture and contrast information. By combining different argumentation method, every lesion can be shown in any specific orientation, up to 32 times, and argumentation-realistic lesion candidates can be obtained to eliminate the overfitting problem. Examples of images after translation augmentation are given in Fig. 3.

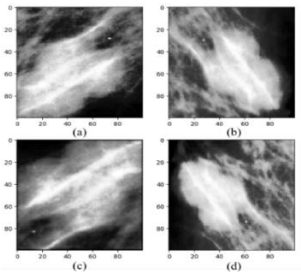

Figure 3. Examples of images after translation augmentation

2.2.1.3 Local histogram equalization

A histogram equalization process is conducted to prepare data for learning algorithms. It is widely known that feature learning and deep learning methods usually perform better when the input data has some properties such as decorrelation and normalization, mainly because such properties help gradient-based optimization techniques to converge [10]. The probability density function (PDF) both before and after the histogram equalization is $p_r(r)$ and $p_s(s)$, and *s* is the transform function.

$$p_s(s) = p_r(r)\left|\frac{dr}{ds}\right| \tag{2}$$

$$s = T(r) = (L-1)\int_0^r p_r(w)dw \tag{3}$$

2.2.1.4 Balancing Data

There were only 51 malignant mammograms among 116 abnormal images in mini-MIAS. Therefore, the number of benign images was greater than malignant ones in our datasets. Those datasets are referred to as imbalanced datasets, which may negatively affect classification of the minority class (i.e., abnormal images). To address this, we use different time argumentations on benign and malignant mammograms. We guarantee that the number of malignant mammograms is the same as the number of benign mammograms.

2.2.2 Training and classification details based on CNN

CNNs have been successfully applied in image classification and object location after Krizhevsky et al. achieved state-of-the-art performance in the ImageNet Large-Scale Visual Recognition Challenge [11]. A CNN model contains three main components: a convolution layer, pooling layer, and fully connected layer. Each layer has a different task. The convolutional layer is composed of several small matrices or "kernels" that are convolved throughout the entire input image as filters, achieving the feature extraction. Pooling layers minimize the dimensions of the feature map generated by the convolution layers, and a fully connected layer classifies which category the input image belongs to using the feature maps.

Since the mini-MIAS is not a large dataset, we didn't build a very new deep CNN model to avoid overfitting. The models created have the maximum of 10 layers. In contrast, some very deep CNN models such as ALEXNET and VGG16 that achieve state-of-the-art in many classification areas were also trained. To measure how the network structure such as depth and dropout method affect the performance of the model, we first evaluate the architecture with two convolutional layers and two max-pooling layers with a fully connected layer. This architecture is referred to as CNN-2 in the experiments. Meanwhile, we add the dropout strategy (Srivastava et al., 2014) [12] in the fully connected layer with p = 0.7 in CNN-2, referred to as CNN-2d. Using the work of Ciresan et al. [13], we adopt the smaller convolution size and deeper network strategy, adding two additional convolutional layers to CNN-2 with a reduced filter size from 25x25 to 3x3. This is referred to as CNN-4. Meanwhile, we adopt the dropout strategy in the fully connected layer with p = 0.7, referred to as CNN-4d, and the structure is shown in Fig. 4. These four CNN model parameters are described in Table 1. And we report the number of parameters for each configuration in Table 2. In spite of a large depth, the number of weights in the net is not greater than the number of weights in a shallow net with larger convolutional layer width.

We implemented the network with the Tensorflow framework, which takes advantage of GPU technology to obtain up to 140 times speedup with respect to GPU implementations. This property makes it feasible in the training of architecture with millions of parameters. All experiments were implemented on a workstation with two NVIDIA Titan X GPU with 24 GB memory each. Based on our implementation and our configuration, the CNN-4d model took 0.4ms to process every image during testing.

We employed stochastic gradient descent (SGD) with RMSProp [14], an adaption of R-Prop for SGD with Nesterov momentum [15]. We used the uniform weight filler, with a learning rate of 0.001. To address the strong class imbalance, samples were randomly chosen in the augmented dataset, where a benign instance is the same size as a malignant instance in each mini batch.

To evaluate the performance and discriminative power of the CNN model, measurements for the overall classification accuracy, sensitivity, and specificity were calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{5}$$

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

where *TP*, *FN*, *TN* and *FP* represent the true positives, false negatives, true negatives, and false positives, respectively.

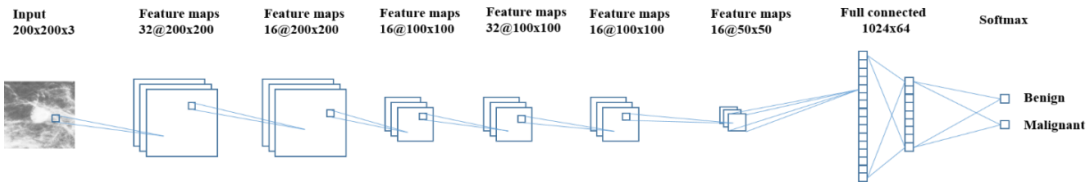

Figure 4. An illustration of the network, the proposed CNN-4d network has 3x3 local kernels and RELU as activation function in the first and second convolution layer followed by 2x2 pooling layer without overlapping. Then a fully connected layer with 1024 units with RELU activation and a fully connected layer with 64 units is stacked to finally add a softmax classifier

Table 1. CNN configurations (shown in columns). The depth of increases from the left (CNN-2) to the right (CNN-4d). The convolutional layer parameters are donated as "conv<receptive field size>-<number of channels>". The RELU activation function is not shown for brevity.

CNN Configurations			
CNN-2	**CNN-2d**	**CNN-4**	**CNN-4d**
5 weight layers	5 weight layers	7 weight layers	7 weight layers
Input (200x200x3 RGB image)			
conv25-32	conv25-64	conv3-32 conv3-16	conv3-32 conv3-16
max-pool			
conv25-16	conv3-32	conv3-32 conv3-16	conv3-32 conv3-16
max-pool			
FC-1024	FC-1024 (Dropout rate-0.7)	FC-1024	FC-1024 (Dropout rate-0.7)
FC-64	FC-64 (Dropout -0.7)	FC-64	FC-64 (Dropout -0.7)
FC-2			
soft-max			

3. RESULTS and DISCUSSION

3.1 Learned features

Recall that the CNN weights in the convolution layer and maxpool layer worked as filters over the image. Thus, we exact the output feature map and visualize them. Fig. 5 shows the output of the convolution layer in CNN-4d. This image exposes a set of edges in different orientations, as well as some texture patterns. We find that the saliency regions match the arthritic lesions that were assessed by the physicians to make the diagnostic decision, therefore visually proving the diagnostic validity of the CNN-4d model. It also reflects that the CNN-4d extracts the relevant information from the images to our problem, using the same information that physicians use to make the diagnostic decision.

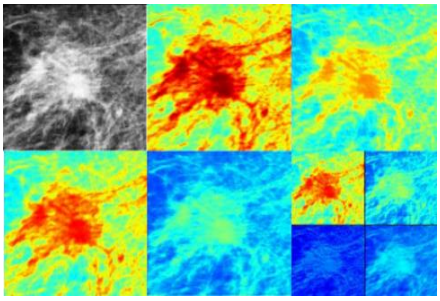

Figure 5. Examples of output feature maps of convolution layers. The top left image the origin input image, the others are the typical feature maps output of convolution layers.

Table 2. Number of parameters (in millions).

Network	CNN-2	CNN-2d	CNN-4	CNN-4d
Number of parameters/million	41.41	41.41	41.01	41.01

3.2 Classification results

The training group consisted of 547 images, including 275 benign and 272 malignant mammographic images. The test group consisted of 137 images, including 73 benign and 64 malignant mammographic images. In Table 3, the confusion matrix obtained using CNN-4d as the feature extractor is shown. Next, the sensitivity and specificity of cases can be calculated as follows.

Table 3. Confusion matrix for augmented MIAS test set predictions and feature extraction using CNN-4d

		Predict		
		Benign	**Malign**	**Total**
Actual	**Benign**	64	9	73
	Malign	6	58	64
	Total	70	67	137

The CNN-4d achieved a sensitivity of 90.63% and specificity of 87.67%. To study the effect of the dropout strategy, we compared the results of CNN-4 with CNN-4d, and found that the training accuracy of CNN-4 was 98.90%, however its testing accuracy was only 86.13%. While the training accuracy of CNN-4d was 93.43%, and its testing accuracy was 89.05%. This corresponds to 128 and 122 well-classified examples, taking into account the total number of 137. Thus, without dropout, the training accuracy is much higher than the test accuracy, indicating that the classifier may be overfitting the training data. CNN-4d has stronger generalization abilities than CNN-4. Results show that this approach of using a dropout strategy in fully connected layers decreases the level of overfitting and obtains a better performance in classifying mammograms.

Table 4 shows the diagnostic performance corresponding to the different CNN models including ALEXNET, VGG16, CNN-2, CNN-2d, CNN-4, and CNN-4d.In order to study the influence of the depth of the CNN structure, the accuracy, sensitivity, and specificity of the CNN-2d and CNN-4d were compared. As shown in Table 4, CNN-4d obtained higher accuracy (89.05%) than CNN-2d (75.91%), higher sensitivity (90.63%) than CNN-2d (90.48%), and higher specificity (87.67%) than sensitivity CNN-2d (65.51%). Results show that smaller filter sizes but deeper CNN have better performance in classifying mammograms, for the reason that deeper CNN with smaller filter can express the more complex nonlinear relationship with the same size of parameter with bigger filter with shallower CNN structure.

We also tested the classical very deep CNN models, which achieved state-of-the-art in image classification and object detection, such as ALEXNET, VGG16. These two model results were shown in Table 4. Results shows the very deep CNN models obtained a poor performance, for example, the VGG16 just obtained the accuracy of 0.6977, the sensitivity of 0.7437, we believe that the MIAS dataset is too small to trained a very deep CNN model such as VGG16 and ALEXNET, thus the parameters cannot be trained completely in very deep models. A series of

experiments were also conducted to validate that the CNN-4d model is the best one, as the models which are deeper than CNN4-d owned worse performance than CNN4-d, for the reason that the MIAS dataset is not big enough and it is improper to train CNN models where depth is higher than 4.

Table 4. Diagnostic performance of different classification models. The proposed CNN-4d achieved superior performance in terms of the four measurements. The best measurements were highlighted in bold.

	Test Dataset			Training Dataset
	Accuracy	**Sensitivity**	**Specificity**	**Mean ± std (Accuracy)**
ALEXNET	0.6558	0.6775	0.6979	0.7501±0.02
VGG16	0.6977	0.7437	0.6963	0.8031±0.02
CNN-2	0.5328	0	**1**	0.5027±0.05
CNN-2d	0.7591	0.9048	0.6551	0.8647±0.05
CNN-4	0.8613	0.7826	0.9259	**0.9890±0.03**
CNN-4d	**0.8905**	**0.9063**	0.8767	0.9343±0.03

4. CONCLUSION

In this study, a novel deep learning model in the form of a CNN trained on the mini-MIAS is proposed to classify abnormalities, both benign and malignant. To enhance the characteristics of the image and improve the performance of classification, a preprocessing algorithm is proposed that uses a series of preprocessing methods, such as cropping, GCN, local histogram equalization, and balancing preprocessing. The CNN model takes the ROI of the raw image as input, achieving the feature learning and classification of abnormalities. To satisfy the mammographic image, a data augmentation method is proposed to ameliorate data scarcity and prevent overfitting. Specific experiments are conducted to explore the influence of the CNN layer structure and kernel, or activate the function on the classification performance. CNN-4d has the best performance with a training accuracy 93.43% and testing accuracy of 89.05%. The experiment also shows that dropout strategy in fully connected layers decreases the level of overfitting and obtains better performance at classification. The same model complexity with a smaller filter size but deeper CNN has a better performance. The results inspire a specific design strategy of CNN structures that satisfies the classification of mammographic images. The model proposed in this study improves the accuracy and stability in mini-MIAS breast mammographic image classification.

5. ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China under Grant No. 51475305 and 61473192.

6. REFERENCES

[1] Jiao, Z., Gao, X., Wang, Y., & Li, J. 2016. A deep feature based framework for breast masses classification. *Neurocomputing*, 197: 221-231.

[2] K.Doi, 2007. Computer aided diagnosis in medical imaging: historical review, current status and future potential, *Comput.Med.ImagingGraph*.31(4), 198–211.

[3] Hepsag, P. U., Ozel, S. A., & Yazici, A. 2017. Using deep learning for mammography classification.

[4] Gallego-Posada, J. D., Montoya-Zapata, D. A., & Quintero-Montoya, O. L. 2016. Detection and diagnosis of breast tumors using deep Convolutional Neural Networks

[5] M. M. Jadoon, Q. Zhang, I. U. Haq, S. Butt, and A. Jadoon, 2017. Three-Class mammogram classification based on descriptive CNN features, *Hindawi Biomed Research International*, DOI=https://doi.org/10.1155/2017/3640901, 2017.

[6] Suckling, J. 1996. The mammographic image analysis society digital mammogram database.

[7] Arevalo, J., Gonzalez, F. A., Ramos-Pollan, R., Oliveira, J. L., & Guevara Lopez, M. A. 2016. Representation learning for mammography mass lesion classification with convolutional neural networks. *Comput Methods Programs Biomed*, 127: 248-257.

[8] Kooi, T., Litjens, G., van Ginneken, B., Gubern-Merida, A., Sanchez, C. I., Mann, R., den Heeten, A., & Karssemeijer, N. 2017. Large scale deep learning for computer aided detection of mammographic lesions. *Med Image Anal*, 35: 303-312.

[9] Simard, P.Y., Steinkraus, D., Platt, J.C., 2003. Best practices for convolutional neural networks applied to visual document analysis. *Document Analysis and Recognition*, pp. 958-963.

[10] K. Jarrett, K. Kavukcuoglu, M. Ranzato, Y. LeCun, 2009, What is the best multi-stage architecture for object recognition. *IEEE 12th International Conference in Computer Vision*, pp. 2146-2153, DOI=http://dx.doi.org/10.1109/ICCV.2009.5459469.

[11] Krizhevsky, A., Sutskever, I., and Hinton, G. E., 2012. ImageNet classification with deep convolutional neural networks. *NIPS*, pp. 1106-1114.

[12] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929-1958.

[13] Ciresan, D. C., Meier, U., Masci, J., Gambardella, L. M., and Schmidhuber, J. Flexible, 2011. High performance convolutionnal neural networks for image classification. In *IJCAI*, pp. 1237-1242.

[14] Dauphin, Y. N., de Vries, H., Chung, J., Bengio, Y., 2015. Rmsprop and equilibrated adaptive learning rates for non-convex optimization. arXiv:150204390.

[15] Sutskever, I., Martens, J., Dahl, G., Hinton, G., 2013. On the importance of initialization and momentum in deep learning. *International Conference on Machine Learning*, pp. 1139-1147.