Incorporate robustness into a vanilla convolutional neural network for skin cancer lesion classification

Ramses Moreno De La Cruz
Department of Computing and Mathematics
Manchester Metropolitan University
Manchester, UK
23624012@stu.mmu.ac.uk

Abstract— **In the context of binary image classification of skin cancer lesions, the robustness of a convolutional neural network is crucial for its success in the ever-changing environmental conditions that focus on skin cancer detection in the medical field. Therefore, this paper implemented a vanilla neural network with several combinations of dropout rates as a regularization technique in 224x224 skin cancer lesion images. It achieved an average classification F1-score of 96.06% for training and 91.78% for testing in several scenarios set by the dropout rates on the Fully Connected Layers. The aim was to build a robust model with great generalization capabilities and, at the same time, prevent overfitting. However, the Mann-Whitney U statistical test concluded that the dropout regularization applied did not have a statistically significant impact on the F1-score of the vanilla convolutional neural network. In other words, the dropout regularization technique applied on the Fully Connected Layers did not achieve its goal of building a more robust model that is less prone to overfitting.**

I. Introduction

The input of the algorithm used was skin cancer lesion images. Then a regularized vanilla convolutional neural network was used, using the dropout technique, to predict or classify if the skin cancer lesion present in the image is considered benign or malignant.

Ever-changing environmental conditions that surround the medical field studying skin cancer, such as the skewed distributions of data due to a predominant set of images of certain skin colours, the lack of instances with or without hair and tattoos, and the noise and/or high discrepancies in the quality of the images caused by variations in resolution, focus, contrast, and brightness specific to each capturing equipment. Additionally, the considerable variability in the features or properties related to the appearance of the lesion (e.g., border, colour, asymmetry, etc.) and the tendency of convolutional neural networks to overfit, motivate this project. Its aim is to develop better and deeper insights into the construction process of robust convolutional neural networks that can generalize without compromising the respective performance metric (i.e., F1-score).

II. Related Work

The following papers or research describe interesting implementations of dropout rates in several architectures that differ in the level of complexity and task compared to the vanilla convolutional neural network applied in this project.

Paper 1: Skin Lesion Classification With Deep Convolutional Neural Network: Process Development and Validation [1]

- Strength: model (pretrained models). Weakness: complex fine-tuning process to avoid overfitting. Comparison: SOTA pretrained models (Inceptionv3 and DenseNet-201), 7 types of skin lesions (HAM10000 data), and a dropout rate of 0.5.

Paper 2: Dermatologist-Level Classification of Skin Cancer Using Cascaded Ensembling of Convolutional Neural Network and Handcrafted Features Based Deep Neural Network [2]

- Strength: data (handcrafted features). Weakness: model (complex architecture). Comparison: Cascaded Ensembling (architecture), and dropout rates of 0.1, 0.3, 0.2, and 0.4.

Paper 3: A deep convolutional neural network-based pigmented skin lesion classification application and experts evaluation [3]

- Strength: it was tested by dermatologists. Weakness: 75 × 100 pixels sized input images. Comparison: 7 types of skin lesions (HAM10000 data), and a dropout rate of 0.25

Paper 4: Multiclass skin cancer classification using EfficientNets – a first step towards preventing skin cancer [4]

- Strength: model (pretrained model). Weakness: complex finetuning process to avoid overfitting. Comparison: SOTA pretrained model (EfficientNets B0-B7), 7 types of skin lesions (HAM10000 data), and a dropout rate of 0.5.

Paper 5: Melanoma Skin Lesions Classification using Deep Convolutional Neural Network with Transfer Learning [5]

- Strength: data augmentation. Weakness: small pooling layer (2x2). Comparison: 7 types of skin lesions (HAM10000 data), and a dropout rate of 0.25.

III. Dataset

About the dataset used, the following information summarize the key insights of it:

A. Source: Kaggle

- https://www.kaggle.com/datasets/bhaveshmittal/melanoma-cancer-dataset
- License: CC0 - Public Domain

B. *General information (before data preparation and/or preprocessing):*

- Images: photos of skin cancer lesions categorized as benign or malignant, set in their respective files of training data (11879 images) and test data (2000 images).
- Resolution: 224x224 pixels.
- File type: JPG image.

C. *Specific information (after data preparation and/or preprocessing):*

- Balanced dataset: ratio 1.1 (benign/ malignant images).
- Noise: there was no noise present in the dataset.
- Missing data: despite the fact that the dataset is advertised as 13900 images on Kaggle, after loading it, only 13879 images were present.
- Data although data augmentation techniques were applied, such as random crops, affine, flips, saturation, contrast, brightness, and hue, they were not taken into account during the experiment.
- Data normalization: it was not applied in this project.
- Data distribution used in this project (see Table 1.) and some data instances without augmentation (see Fig. 1)

TABLE I. DISTRIBUTION OF THE DATASET

Data Subset	**Distribution of the Dataset**		
	Number of Samples	***Benign***	***Malignant***
Training	7127	3774	3353
Validation	4752	2516	2236
Test	2000	1000	1000
Total	13879	7290	6589

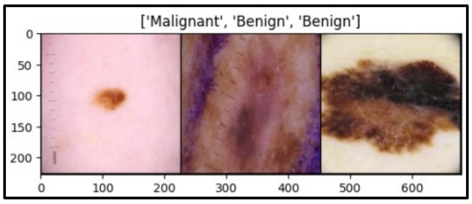


Fig. 1. Examples of skin cancer lesions with their labels

IV. METHODOLOGY

Basically, this project consists of implementing several combinations of dropout rates (i.e., a regularization technique) in the Fully Connected Layers of a vanilla convolutional neural network. The goal is to increase the robustness and generalization abilities of the model, measured through the F1-score, to classify the skin cancer lesion images as benign or malignant (see Fig 2.).

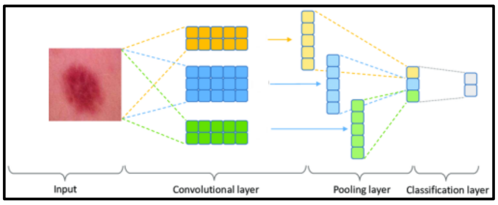


Fig. 2. Architecture of a vanilla convolutional neural network. Adapted from [6]

The learning algorithm applied was a vanilla convolutional neural network. This type of algorithm had the following key components:

- Input Layer: this layer receives the skin cancer lesion images as a tensor represented with the width, height, and the 3 color channels of the image (i.e., red, green, and blue, also called RGB).
- Convolutional Layers: these layers apply convolutional operations to the input with the aim to extract several features or patterns of the input images. "Kernels" (3x3 pixels), "stride" (1 pixel shift), and "padding" (1 extra pixel) were applied to control the size of the output of the convolution process.
- Activation Function: "ReLU" (Rectified Linear Unit) was the chosen activation function to help the network learn complex patterns by introducing non-linearity. It was preferred because it avoids vanishing gradient problems, as its gradients are either 0 or 1.
- Pooling Layer: "MaxPool2d" was the preferred choice of layers over average pooling to extract dominant features from input and reduce spatial dimensions. This aids in the classification process of skin cancer lesion images.
- Fully Connected Layers: After the convolutional phase (which uses several sets of convolutional layers together with "ReLU" functions and "MaxPool2d" layers), the output is flattened and fed as input to several fully connected layers. These fully connected layers combine the final extracted features from the convolutional phase to perform the final classification of the skin cancer lesion as benign or malignant.
- Output Layer: This final layer (located at the end of the Fully Connected Layers) produces a single value for the binary classification of the skin cancer lesion as benign or malignant.
- Loss Function: The "CrossEntropyLoss" was the function chosen to measure the difference between the predicted output and the true labels during the training phase for the skin cancer lesion images. Despite the fact that there are specific and better loss functions such as "Binary Cross-Entropy Loss" and "Log Loss" for binary classification tasks, the decision to use "CrossEntropyLoss" is justified for checking and spotting experimental performance problems and convergence issues.
- Optimizer: Adaptive Moment Estimation (also called "Adam Optimizer") was the optimization technique for gradient descent (i.e., to update the network's weights and minimize the value of the loss function "CrossEntropyLoss").
- Regularization Technique: Several dropout combinations (i.e., scenarios) were exclusively applied in the Fully Connected Layers to prevent overfitting, improve generalization, and increase the robustness of the model. The effects of this technique were measured through the F1-score. This method involved setting a fraction of neurons (i.e., the "dropout rate") to zero in a random manner during the training phase (see Fig. 3).

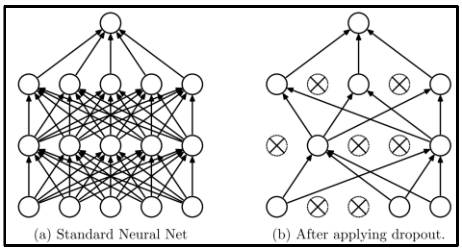


Fig. 3. Scheme of the dropout regularization. Adapted from [7]

The information corresponding to the formulation of the research hypothesis is as follows:

- Mann-Whitney U test: This statistical test is convenient for comparing the median of two groups (specifically, for sample groups with less than 10 instances in each group, as was the case in this project) [8].
- Null hypothesis (H0): The two samples being compared (the F1-score of a convolutional neural network without regularization and the F1-score of a convolutional neural network with regularization) have the same median or their medians are not significantly different. Consequently, the dropout regularization applied did not significantly impact the F1-score of the vanilla convolutional neural network, and the goal of building a robust model that is less prone to overfitting was not achieved.
- Alternative hypothesis (H1): The two samples being compared (the F1-score of a convolutional neural network without regularization and the F1-score of a convolutional neural network with regularization) have significantly different medians. Therefore, the dropout regularization applied had impacted the F1-score of the vanilla convolutional neural network significantly, and the aim to build a robust model and less prone to overfitting was achieved.

V. Experimental Results And Discussion

The main machine learning libraries and modules used to build this project were Python, Pytorch, Pandas, Numpy, Matplotlib, and Scipy.

Furthermore, the code was run in the MMU library machines that have the following features:

- Intel (R) Core (TM) i7-10700 CPU @ 2.90GHz
- CPU(s): 12 and Logical cores: 16

The vanilla convolutional neural was applied considering the hyperparameters described in the Table 2.

TABLE II. HYPERPARAMETERS

Hyperparameters			
	Value Set	***Concept***	***Insight***
learning rate	0.0003	Determines the step size at each iteration	-Learning rate too high: in all scenarios the training loss decreases but the validation loss plateaus or increases after epoch N°15 (see Fig. 4) -Learning schedulers were not used.
epoch	25	One complete pass through all the training images	-Epoch number slightly high: in all scenarios the accuracy is not lower than 90% -Early stopping techniques were not used.
batch size	32	Number of training images that pass through the layers before updating the weights of the model.	- Considering hardware constraints, 32 was the chosen value.

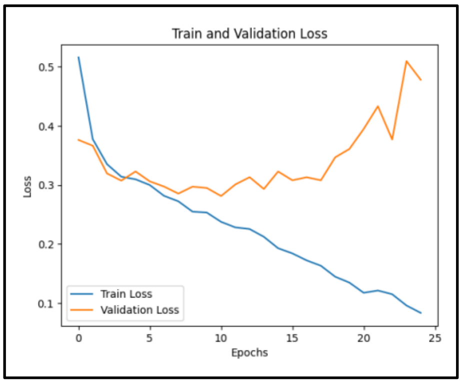


Fig. 4. Loss vs Epochs (Scenario 2)

Several metrics were computed (Confusion Matrix, Precision, Recall, AUC ROC, and Accuracy); however, the main metric used to assess the outputs of the experiment was the F1-score for the following reasons:

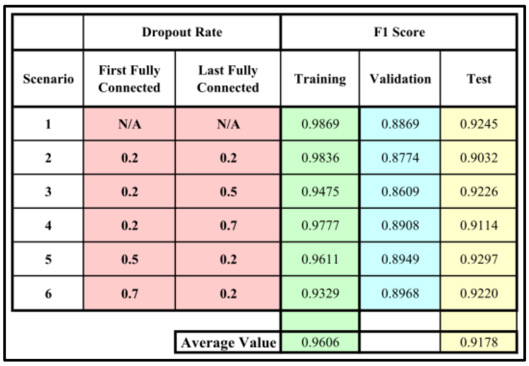
- F1-score: it is the harmonic mean of Precision (i.e., accuracy of positive predictions) and Recall (i.e., a measure of positive cases well identified). The respective formulas are as follows [9]:

$$where \rightarrow TP: True\ Positive, FP: False\ Positive, FN: False\ Negative$$

$$\boldsymbol{Precision} = \frac{TP}{TP+FP}\ \boldsymbol{Recall} = \frac{TP}{TP+FN}\boldsymbol{F1\ score} = \frac{2*Precision*Recall}{Precision+Recall}$$

- Based on the previous definition and in the context of skin cancer lesion classification, it is extremely important to use the F1-score to measure the performance of the model in identifying malignant cases (Recall) and avoiding incorrect diagnoses (Precision)

The outputs of the vanilla convolutional neural network using the F1-score as the main metric to assess the performance of the model in the training, validation, and test datasets (see Fig. 5 and Fig. 6) are based on six (6) scenarios in which several combinations of dropout rates were implemented in the first and last Fully Connected Layer, respectively. It is important to notice that the selection and combination of the dropout rate values (0.2, 0.5, and 0.7) in both layers were based on the idea of impacting the network's capacity to understand and capture patterns in the data, and consequently improving its robustness, generalization, and reducing overfitting tendencies observed when running the model without regularization (described in Scenario 1).



	Dropout Rate		F1 Score		
Scenario	First Fully Connected	Last Fully Connected	Training	Validation	Test
1	N/A	N/A	0.9869	0.8869	0.9245
2	0.2	0.2	0.9836	0.8774	0.9032
3	0.2	0.5	0.9475	0.8609	0.9226
4	0.2	0.7	0.9777	0.8908	0.9114
5	0.5	0.2	0.9611	0.8949	0.9297
6	0.7	0.2	0.9329	0.8968	0.9220
		Average Value	0.9606		0.9178

Fig. 5. F1-score in the scenarios set by the dropout rate

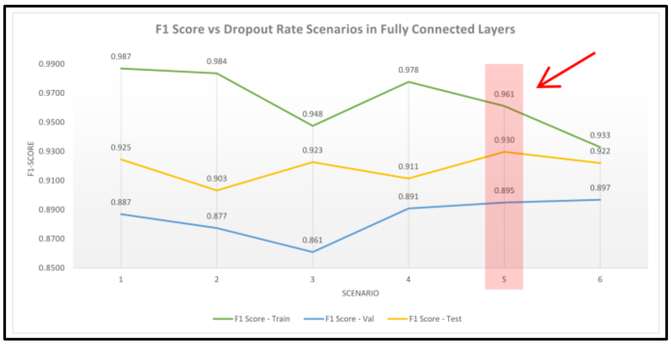
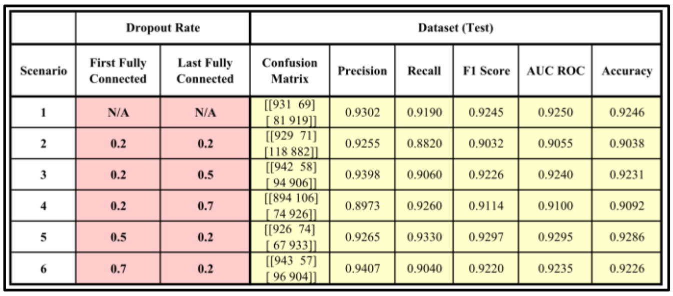


Fig. 6. Line plot of F1-score in the scenarios set by the dropout rate

It is important to remark that the best model (based exclusively on the F1-score on the test set) was found in "scenario 5" (where the dropout rate is 0.5 and 0.2 in the first and last Fully Connected Layer respectively), as shown in Fig. 6. In addition, the other metrics (Confusion Matrix, Precision, Recall, AUC ROC, and Accuracy) as a result of the experiment (for the test dataset) are shown in Fig. 7.



	Dropout Rate		Dataset (Test)					
Scenario	First Fully Connected	Last Fully Connected	Confusion Matrix	Precision	Recall	F1 Score	AUC ROC	Accuracy
1	N/A	N/A	[[931 69] [81 919]]	0.9302	0.9190	0.9245	0.9250	0.9246
2	0.2	0.2	[[929 71] [118 882]]	0.9255	0.8820	0.9032	0.9055	0.9038
3	0.2	0.5	[[942 58] [94 906]]	0.9398	0.9060	0.9226	0.9240	0.9231
4	0.2	0.7	[[894 106] [74 926]]	0.8973	0.9260	0.9114	0.9100	0.9092
5	0.5	0.2	[[926 74] [67 933]]	0.9265	0.9330	0.9297	0.9295	0.9286
6	0.7	0.2	[[943 57] [96 904]]	0.9407	0.9040	0.9220	0.9235	0.9226

Fig. 7. Other metric in the scenarios set by the dropout rate (test dataset)

Finally, the F1-score was used as the main metric to assess the performance of the model in the training, validation, and test datasets. The hypothesis research test called Mann-Whitney U test was then applied. The key outputs of the hypothesis test in the training, validation, and test datasets are shown in Table 3.

TABLE III. MANN-WHITNEY U TEST (APPLIED IN F1-SCORE)

Data Subset	**Mann-Whitney U test (applied in F1-score)**		
	p-value	***Analysis***	
Training	0.33	p-value > 0.05	Fail to reject H0
Validation	1.00	p-value > 0.05	Fail to reject H0
Test	0.66	p-value > 0.05	Fail to reject H0

In the context of considering a significance level of 5% or 0.05 for comparing the probability value (i.e., the p-value) and determining if there is strong evidence to reject the H0 or not *(see Table 3.)*:

The Mann-Whitney U test (applied in F1-score) provided strong evidence to fail to reject the Null Hypothesis (p-value > 0.05) and statistically conclude that the two samples being compared (the F1-score of a convolutional neural network without regularization and the F1-score of a convolutional neural network with regularization) have medians that are not significantly different. Consequently, the dropout regularization applied had not impacted the F1-score of the vanilla convolutional neural network in a significant manner and the aim to build a robust model and less prone to overfitting was not achieved.

VI.CONCLUSION

After conducting several experiments, the hypothesis test (Mann-Whitney U test) statistically concludes that the various values and combinations of the regularization technique applied (i.e., dropout rates) in the Fully Connected Layers of a vanilla convolutional neural network were not sufficient on their own to achieve the goal of increasing the robustness and generalization abilities of the model, and also reducing its susceptibility to overfitting when classifying the skin cancer lesion images as benign or malignant.

To summarize, the vanilla convolutional neural network that achieved better generalization and robustness (F1-score: 0.9297) without degrading performance (Accuracy: 0.9285) in the test set was the network in scenario 5 (First Fully Connected Layer with dropout rate: 0.5 and Last Fully Connected Layer with dropout rate: 0.2). Probably, in this case, a moderate dropout rate such as 0.5 in the First Fully Connected Layer helped regularize the initial features or patterns, while a small dropout rate such as 0.2 in the Last Fully Connected Layer allowed for more co-adaptation of the neurons.

In this experiment, only one regularization technique (i.e. dropout rate) was applied. However, other hyperparameters such as the learning rate, batch size, and epoch, which could potentially impact the robustness and generalization abilities of the model, as well as make it more or less prone to overfitting, remained fixed. Additionally, other regularization techniques such as L1, L2, or even data augmentation were not used. Based on the results and analysis described earlier in this report, it is evident that the complexity of the task (i.e., image classification of skin cancer lesions as benign or malignant) requires a holistic approach. This approach should not only focus on using other regularization techniques or tuning hyperparameters, but also on having additional time and available hardware (GPUs) to conduct more complex experiments.

VII. References

[1] A. Ray, A. Gupta and A. Al, "Skin Lesion Classification With Deep Convolutional Neural Network: Process Development and Validation," *JMIR Dermatology,* vol. Volume 3, no. 1, 7 May 2020.

[2] A. Kumar Sharma, S. Tiwari, G. Aggarwal, N. Goenka, A. Kumar and e. a. , "Dermatologist-Level Classification of Skin Cancer Using Cascaded Ensembling of Convolutional Neural Network and Handcrafted Features Based Deep Neural Network," *IEEE Access,* vol. Volume 10, 17 February 2022.

[3] O. Sevli, "A deep convolutional neural network-based pigmented skin lesion classification application and experts evaluation," *Neural Computing and Applications,* vol. Volume 33, p. 12039–12050, 24 March 2021.

[4] K. Ali, Z. Ahmed Shaikh, A. Ayub Khan and A. Ali Laghari, "Multiclass skin cancer classification using EfficientNets – a first step towards preventing skin cancer," *Neuroscience Informatics,* vol. Volume 2, no. Issue 4, December 2022.

[5] K. Islam and e. a. , "Melanoma Skin Lesions Classification using Deep Convolutional Neural Network with Transfer Learning," in *1st International Conference on Artificial Intelligence and Data Analytics (CAIDA)*, Riyadh, Saudi Arabia , 2021.

[6] I. Krsnik, G. Glavaš, M. Krsnik, D. Miletić and I. Štajduhar, "Automatic Annotation of Narrative Radiology Reports," *MDPI,* vol. Volume 10, no. 4, 1 April 2020.

[7] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from," *The Journal of Machine Learning Research,* vol. 15, no. 1, p. 1929–1958, January 2014.

[8] A. Lund and M. Lund, "Laerd Statistics," Lund Research Ltd, [Online]. Available: https://statistics.laerd.com/spss-tutorials/mann-whitney-u-test-using-spss-statistics.php. [Accessed 23 May 2024].

[9] S. S. Chaturvedi, J. V. Tembhurne and T. Diwan, "A multi-class skin Cancer classification using deep," *Multimedia Tools and Applications ,* vol. Volume 79, p. 28477–28498, August 2020.

Link to Google Drive[1]

[1] Link to Google Drive: Submit