# Uncertainty Estimation in Selective Classification on Skin Lesions

AC40001

Mid-term Report

Professor S. Mckenna &

Daniel Blackley - 160007728

# 1. Summary and Aims

Machine learning algorithms have been shown to perform exceptionally well on the classification of skin cancer, matching the accuracy of experts in the field[1]. Due to the nature of the medical domain being safety critical, it is important that any predictions made by these algorithms have a quantifiable measure of uncertainty and correctly consider the cost of misclassifying certain Skin Lesions[10].

I hope to investigate epistemic uncertainty in machine learning Algorithms by comparing a baseline Softmax response[8] against Yarin Gal's Monte Carlo Dropout[4] (MC Droppout) method, using the EfficientNet architecture[11] and Entropy across predictions as a measure for uncertainty[6].

# 2. Background Research

The two most common types of uncertainty that you can have with regards to machine learning are epistemic and aleatoric[2]. Aleatoric uncertainty is a measure of unnatural (i.e., artificially added white noise) and out of sample data, whereas epistemic uncertainty is uncertainty in regards to things that a machine could in theory learn, but has not due to a lack of training samples. There are many papers looking into both types of uncertainty, and many different methods of discerning a measure of uncertainty from machine learning algorithms[5][6].

Bayesian Networks usually learn a distribution over the weights[3], which requires significant modification to the training procedure and can be computationally expensive, There are many alternatives to traditional Bayesian Networks; One popular method is called Deep Ensembles[3], which, on a basic level, consists of training multiple networks and combining them into one more powerful Network. Another Method that was proposed by Yarin Gal was MC Dropout, it uses dropout at run time, which he proved was equivalent to approximate Varational Inference[4], this method is less computationally expensive and is also easy to implement due to the prevalence of dropout as a regularisation technique. Another, even simpler method, is using 1 – the max Softmax probability, this was shown to be a rather effective method for calculating uncertainty[8] and serves as a good baseline to compare other methods to[5][7]. In my project I plan to compare MC Dropout to a baseline Softmax response, primarily due to the comparatively computationally inexpensive nature of the proposed methods. Both have shown to have comparable results on various datasets[7], and Skin Lesion classification[5].

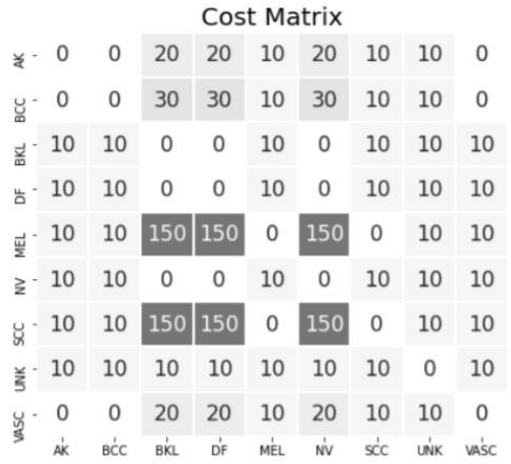


Figure 1: Cost matrix showing the cost of misclassification with Predictions on the Y axis and True labels on the X axis, taken from [9]

Not all classifications of Skin Lesions are equal, some misclassifications can be particularly more dangerous than others. A cost matrix can help counteract this[9][10], Figure 1 shows the implementation that is planned to be used.

# 3. Main Features

This research will be done using a cost sensitive and risk aware EfficientNet model with Compound scaling 0 (EfficientNetb0) and an added 512 Neuron Linear Layer using the Pytorch framework[12], from this we hope to retrieve useful evaluation metrics that can help compare the Softmax baseline and MC Dropout to provide an informative report on which model provides a better prediction of uncertainty. Dropout will be applied across the last 512 Neuron layer with a rate of 50% during testing, while no Dropout will be applied across any layers to get the Softmax baseline. This

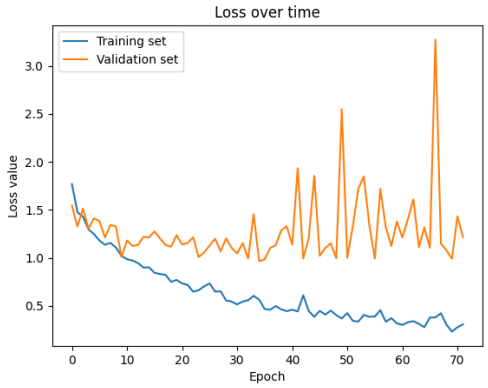


Figure 2a – Learning curve of best model

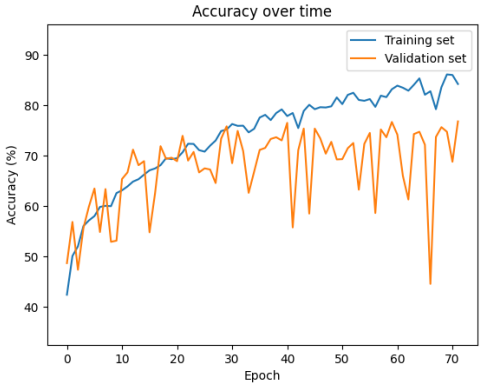


Figure 2b – Accuracy while training

research will use the ISIC 2019 dataset[13][14][15], The dataset will be split, 60% will be used to train the model, 10% will be used as to validate the model during training and to select the model that predicts best on unseen data and 20% will be used to test the best model.

# 4. Progress to date

There is currently an implementation of an EfficientNetb0 model capable of outputting both a basic Softmax prediction as well as a range of predictions by applying dropout at run time. These models get an accuracy of roughly 77% on the Testing Data without any sort of risk consideration. Figure 2a illustrates the Learning curve while Figure 2b shows the accuracy over the epochs. In terms of how this model was obtained, the model was trained on the training data until the learning loss hit a gradient of roughly 0 (80 epochs) and the model with the best validation accuracy was saved.

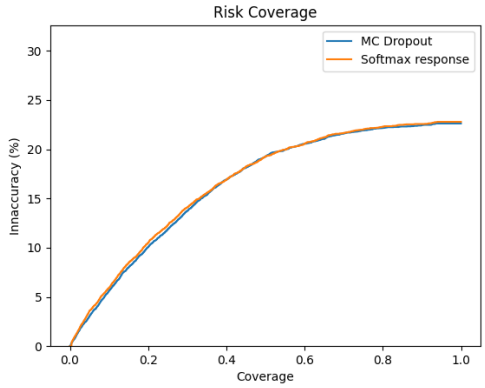


Figure 3 – Risk Coverage Curve

I have also obtained some preliminary metrics for the evaluation of uncertainty; Figure 3 shows a risk coverage curve and Figure 4a and 4b shows the entropy across incorrect and correct predictions.

In terms of discoveries and challenges, Pytorch has been particularly easy to work with, after reading so much into the theory on the subject It seemed daunting to implement ideas like gradient descent, but found out the Pytorch already supplies plenty of optimisers, loss functions and deep learning architecture for you to easily implement and adapt.

# 5. Personal Reflection

During this report, I have tried to speak in third person as much as possible to mimic academic writing, but due to the nature of this section I will be breaking that to write in first person.

I've certainly noticed the progress I've made so far, there is definitely a considerable difference in my knowledge and skillset between at the start of semester 1 and now. In terms of if I am on track, I am literally, according to Figure 5, on track for getting work done on time. In terms of the quality of work/ the grade I hope to achieve, I think it's going to be particularly hard to tell until I begin writing up my findings. I think my understanding of some things has been confused, but the current cycle of implementation, meeting up and then finding out which parts can be improved on has been the fastest way to see exactly which parts I do not understand. Even now I presume there will be some mistakes in the earlier sections of this report that will be interesting to go over.

# 6. Plans for Remainder of Project

Figure 5 is the Gantt chart that has been created to ensure this research continues according to schedule. The next major milestone will be finishing up implementation, which should occur in the coming weeks, and then begin looking into writing up my report. After that the report has been split into more minor milestones as shown in Figure 5, (i.e. literature review, data analysis). Currently the cost matrix still needs to be implemented, as the current model does not consider the cost of misclassification. There are also extra plans to Implement another method called Bayes By Backprop[16] and there has been some uncertainty about which method is the best measure for uncertainty[6] which may be interesting to look into, but these plans are to be considered extra features and won't be implemented until I have a good comparison of MC dropout and the Softmax Response.

This Gantt chart is likely to be a more of a rough guideline as the challenge of writing the report is difficult to gauge, due to this being my first time writing something of this nature. The best way to minimise this challenge is going to be looking at how other papers are interpreting their results and paying close attention to any technical details that they talk about, hopefully catching any problems before they occur.

# References


[1] Tschandl P, Codella N, Akay BN, Argenziano G, Braun RP, Cabo H, Gutman D, Halpern A, Helba B, Hofmann-Wellenhof R, Lallas A, Lapins J, Longo C, Malvehy J, Marchetti MA, Marghoob A, Menzies S, Oakley A, Paoli J, Puig S, Rinner C, Rosendahl C, Scope A, Sinz C, Soyer HP, Thomas L, Zalaudek I, Kittler H. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented /skin lesion classification: an open, web-based, international, diagnostic study. Lancet Oncol. 2019 Jul;20(7):938-947. doi: 10.1016/S1470-2045(19)30333-X. Epub 2019 Jun 12. PMID: 31201137.

[2] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? Structural safety, 31(2):105–112, 2009.

[3] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scal-able Predictive Uncertainty Estimation using Deep Ensembles. 2017. arXiv:1612 .01474 [stat.ML].

[4] Yarin Gal and Zoubin Ghahramani.Dropout as a Bayesian Approximation: Repre-senting Model Uncertainty in Deep Learning. 2016. arXiv:1506.02142 [stat.ML]

[5] Leibig C. et al. "Leveraging uncertainty information from deep neural networksfor disease detection". In: Sci Rep 7.17816 (2017). DOI:https://doi.org/10.1038/s41598-017-17876-z.

[6] Marc Combalia et al. "Uncertainty Estimation in Deep Neural Networks for Der-moscopic Image Classification". In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. June 2020.

[7] Geifman, Yonatan & El-Yaniv, Ran. (2017). Selective Classification for Deep Neural Networks.

[8] Dan Hendrycks and Kevin Gimpel. "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks". In:CoRRabs/1610.02136(2016). arXiv:1610.02136. URL:http://arxiv.org/abs/1610.02136.

[9] Vasileios Aidonis. "Cost-Sensitive Deep Learning for Skin Lesion Diagnosis". 2020. University of Dundee.

[10] Di Zhuang, Keyu Chen, and J. Morris Chang. CS-AF: A Cost-sensitive Multi-classifier Active Fusion Framework for Skin Lesion Classification. 2020. arXiv:2004.12064[cs.CV].

[11 ]Mingxing Tan and Quoc V. Le.EfficientNet: Rethinking Model Scaling for Convolu-tional Neural Networks. 2020. arXiv:1905.11946 [cs.LG].

[12] Adam Paszkeetal. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: Advances in Neural Information Processing Systems 32. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 8024–8035. URL:

http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[13] Tschandl P., Rosendahl C. & Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data 5, 180161 doi.10.1038/sdata.2018.161 (2018)

[14] Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, Allan Halpern: “Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC)”, 2017; arXiv:1710.05006.

[15] Marc Combalia, Noel C. F. Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Allan C. Halpern, Susana Puig, Josep Malvehy: “BCN20000: Dermoscopic Lesions in the Wild”, 2019; arXiv:1908.02288.

[16] Charles Blundell et al.Weight Uncertainty in Neural Networks. 2015. arXiv:1505.05424 [stat.ML].

# Appendix

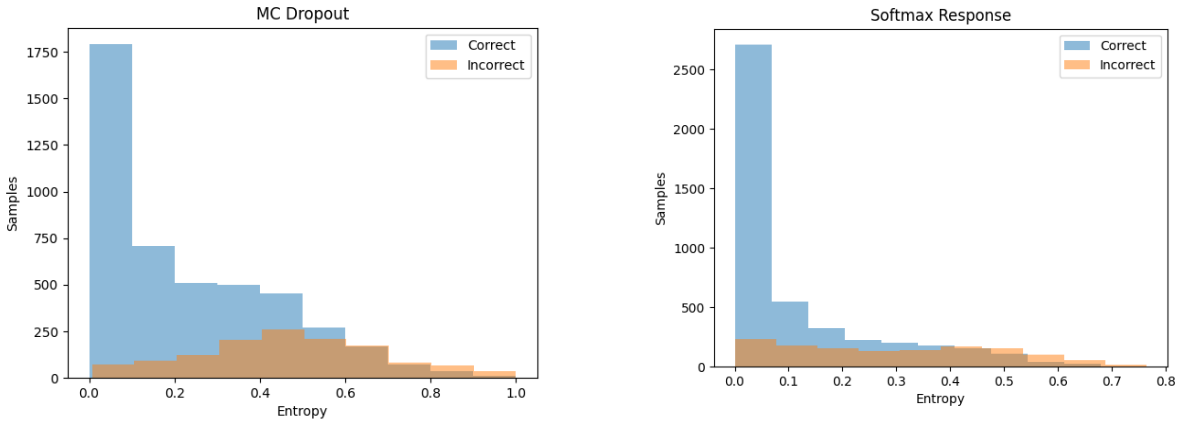


Figure 4a/b, histograms showing number of correct/incorrect samples and their entropies

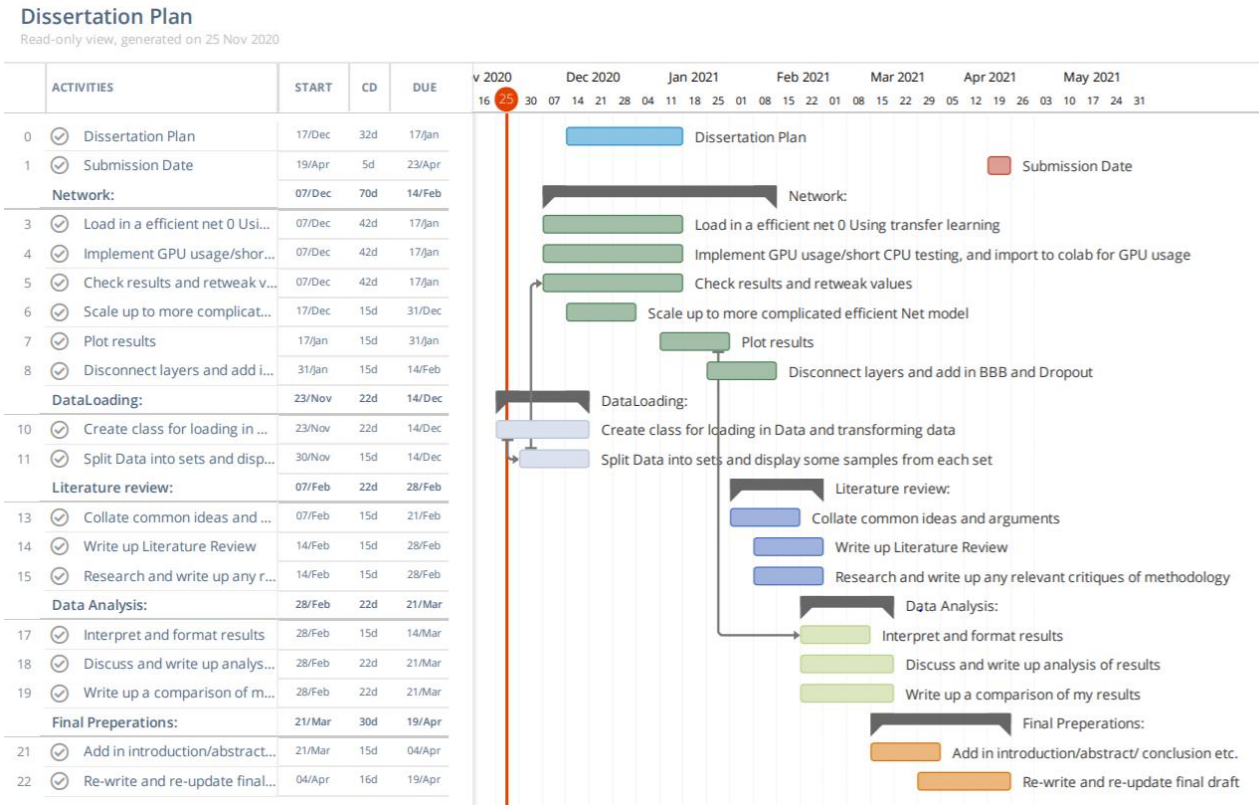


Figure 5, Gantt chart showing dates and predicted deadlines.