A New Lightweight Architecture and a Class Imbalance Aware Loss Function for Multi-label Classification of Intracranial Hemorrhages

Prabhat Lankireddy[0000−0003−3161−3806], Chitimireddy Sindhura[0000−0003−4478−3302], and Subrahmanyam Gorthi[0000−0003−1957−6985]

Department of Electrical Engineering, Indian Institute of Technology Tirupati, Andhra Pradesh, India
prabhat.lankireddy@gmail.com,
sindhurareddy234@gmail.com,s.gorthi@iittp.ac.in

Abstract. Deep learning algorithms have proven effective in solving many medical imaging tasks in recent years. The design of lightweight neural networks is gaining importance in the medical imaging community as not many hospitals and clinics are equipped with high computational resources to deploy large deep learning algorithms. Also, medical imaging data often comes with high class imbalance and thus there is a high necessity to develop deep learning models that can address this issue. With this motivation, a resource-efficient deep learning model called Lightweight-Fully Convolutional Network (LightFCN) is developed which can be deployed in clinical settings with limited computational resources. Label Distribution Aware Margin loss (LDAM) is used in the context of medical imaging for the first time for multi-label classification with class imbalance. The proposed model has a smaller memory footprint, a smaller number of parameters, lesser inference time and fewer Floating Point Operations (FLOPS) when compared to state-of-the-art models, without compromising on performance and can be deployed in clinical settings with limited computational resources. The model and the performance of the loss function are evaluated on the task of Intracranial Hemorrhage (ICH) classification on CT scans, and the model was deployed on a Raspberry Pi 4B (8GB), on which inference times were compared. It is found that the proposed model significantly reduced the number of model parameters by a factor of 26, and reduced the inference time by a factor of 3, when compared to the popular lightweight network MobileNetV2.

Keywords: Deep Learning · Class Imbalance · Lightweight · Multi-label · Image Classification · Intracranial hemorrhage.

1 Introduction

Deep learning models perform very well on medical imaging tasks like classification and segmentation, but they are often over-parameterized, and lightweight neural network design is all about looking for a network that can do the same

task while consuming significantly less time and resources. Lightweight neural network architectures are now being given plenty of importance due to their multiple advantages, such as reduction of training time, inference time, lesser memory footprint and lesser number of computations (Floating Point Operations, or FLOPS). This is gaining importance in the medical imaging community as many hospitals and clinics are not equipped with high computational resources to deploy large deep learning algorithms, nor can they be replaced by new machinery owing to financial constraints.

Lightweight Neural Network design has been given a lot of importance in Computer Vision literature. A few popular architectures include MobileNet [7], which replaces regular convolutions with depthwise separable convolutions to reduce the number of FLOPS and inference time, and Shufflenet [18], that uses pointwise group convolutions and channel shuffling to reduce the number of FLOPS. The research focus in the medical imaging community is on resource efficient architectures, for various classification and segmentation tasks [12, 16].

Class imbalance is often prevalent in medical imaging data because classification or segmentation tasks often involve identifying a lesion or disease, and the collection of huge data with rarer lesions or diseases is impractical. This causes the data to be skewed towards the "negative" samples i.e. samples without a lesion, and when deep learning models are trained on such data, they tend to learn to classify a test sample as negative more often. One easy way to address skewed data problems is to simply remove the excess data belonging to majority classes in an effort to have approximately equal number of data samples for all classes, but that is impractical in medical imaging tasks owing to the shortage of data in many cases, as less data makes it harder to train deep learning models that are extremely data hungry. This brings the need to develop methods to address the problem of class imbalance, in order to improve the performance of deep learning models for medical imaging tasks.

Lots of work has been done in addressing class imbalance when training machine learning algorithms, but the same cannot be said for deep learning tasks. A review paper on deep learning with class imbalance [8] broadly categorised the methods that address class imbalance into three different types, which are resampling methods, re-weighting methods and loss function based methods. Resampling involves over-sampling (training with the positive-labeled data more often than negative-labeled data) or under-sampling (skipping training on excess negative-labeled data), while re-weighting involves giving a higher weight to the losses for rare class data than losses for data of the frequently occurring classes, as done in the class balanced loss [4]. There are also loss-function based methods in which a loss function is designed to address class imbalance. Few examples for this are the focal loss [10], which gives a higher weight to the samples that are classified improperly by the model, and the Label-Distribution-Aware Margin Loss (LDAM) [3] which gives a higher classification margin to the rarer class samples than the dominant class samples.
While class imbalance has been discussed in the context of multi-class classification and segmentation tasks, little to no attention has been paid to multi-label

classification tasks. In multi-label classification, there are multiple classes and more than one of them can be positive for sample data. One example of multi-label classification tasks is the identification of different types of Intracranial Hemorrhages (ICH) in brain CT scans, where more than one type of injuries can be identified at once. This approach is more time and resource efficient than using many binary-classification models to identify each kind of injury.

In this paper, we propose a lightweight architecture called Light-Fully Convoluted Neural Network (Light-FCN) for the multi-label image classification task of ICH identification, along with ways to address class imbalance in the training data to improve performance of the model.

The rest of the paper is organized as follows. Section 2 discusses the architecture of the proposed model and the loss function. The section 3 describes the dataset and the implementation details, followed by the ablation study and results. The conclusions are presented in Section 4.

2 Architecture and Loss Function

2.1 Lightweight Architecture Design

We propose a simple and lightweight architecture, *Light-FCN*, for the detection of ICH in CT scans. This architecture is inspired from the Simple Fully Convolutional Network (SFCN), proposed by Peng et al. [14], which has been found to work well in the task of Brain Age Prediction using CT scans. SFCN is developed based on the VGGNet [11] and the fully convolutional network (FCN) [17]. The success of SFCN in the task of Brain Age prediction suggests that the network is capable of extracting useful features from CT scans, which was the reason we decided to use SFCN as the baseline architecture for our task. The architecture of the proposed model is shown in the Fig. 1.

To reduce the complexity while maintaining the performance, two modifications are made to the baseline model. Firstly, the convolution layers have been replaced with a combination of depth-wise separable convolutional layers and pointwise 1×1 convolutional layers, an implementation inspired from the MobileNet architecture [7]. This has significantly reduced the total number of parameters. Secondly, inverted residual connections along with linear bottlenecks are introduced, which were adopted from MobileNetV2 [15]. The residual connections enabled easy training and improved performance without the need for deeper networks.

The Light-FCN network has 5 expansion blocks, in each of which the number of channels are increased using pointwise convolutions (called *expansion*), followed by depthwise convolutions. After that, another pointwise convolution operation is used to shrink the number of channels (called *projection*), and a residual connection is added to this output. Using expansions and projections as mentioned above significantly reduces the number of floating point operations, with a negligible reduction in performance, as the information in a feature map with large number of features can be represented using a lower dimensional subspace [15]. We use a constant bottleneck width of 32 channels in LightFCN.

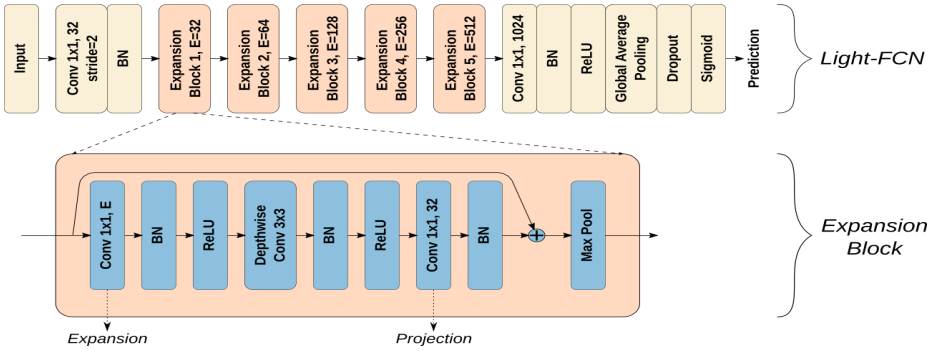
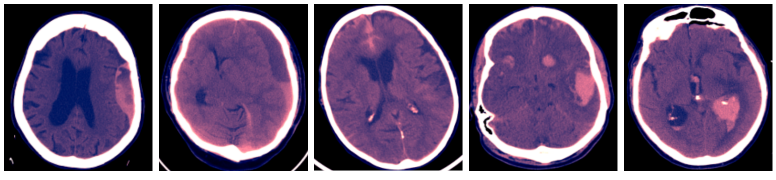


Fig. 1. Illustration of the proposed architecture Light-FCN and the Expansion block. "Conv 1x1, E" denotes a pointwise convolution layer with E output channels.

2.2 Loss Function

For multi-label classification, one can assume that all the outputs are independent of each other and devise a loss function accordingly. The sigmoid activation function at the output layer is suitable here, as it turns outputs into probabilities independent of other outputs. Another important factor to address when choosing a loss function, is the class imbalance in the training data. Certain loss functions are designed to minimize the bias that is introduced in the model due to the class imbalance present in the training data.

One such loss function is the LDAM loss function [3]. The LDAM loss introduces the concept of margin $\Delta_i = \frac{C}{n_i^{1/4}}$ (where C is a constant and n_i is the class frequency i.e. number of datapoints in the training data that belong to the i-th class.) and adds the margin to the model output if the samples corresponds to a negative ground truth (or subtracts the margin from the model if the samples correspond to a positive ground truth). To address class imbalance, a higher value of margin is used for classes that aren't well represented, leading to higher loss output. These modified loss outputs can be used along with any activation function. In this work, the sigmoid activation along with the binary crossentropy loss is chosen as we require the probabilities to be independent of each other. Then the loss is given as follows:



(a) EDH **(b) SDH** **(c) SAH** **(d) IPH** **(e) IVH**

Fig. 2. Examples of preprocessed images for each type of injury

$$L(p,q) = -\frac{1}{k}\sum_{i}\left[p_i \log\left(\frac{1}{1+e^{-(z_i-\Delta_i)}}\right) + (1-p_i)\log\left(\frac{1}{1+e^{-(z_i+\Delta_i)}}\right)\right] \tag{1}$$

where p_i is the ground truth and z_i is the output logit for the i-th class.

3 Experiments and Results

3.1 Dataset and Preprocessing

The dataset used for this work is RSNA Intracranial Hemorrhage dataset ([5]). Each CT slice was annotated with 5 binary labels indicating the presence of the following types of Intercranial Hemorrhage: 'Epidural Hemorrhage (EDH)', 'Intraparenchymal Hemorrhage (IPH)', 'Intraventricular Hemorrhage (IVH)', 'Subdural Hemorrhage (SDH)', and 'Subarachnoid Hemorrhage (SAH)'. The dataset has 21,744 CT scans in total, out of which only 7,652 had hemorrhages. As the injuries are observed in a small number of slices of a scan, only the scans that had injuries were included, in an effort to minimize class imbalance. The dataset has been split into 6122, 765 and 765 scans for training, validation and testing datasets respectively. The patient-wise and slice-wise distribution of data into train, validation and test datasets is shown in Table 1. All the 2D slices are down sampled to 256 × 256 and processed with intensity windowing method to construct a three channel RGB-like image. The three kinds of Hounsfield Units (HU) windows were considered for pre-processing as recommended in [2], each focusing on a different type of tissue: Brain window [Window Center (WC): 40, Window Width (WW): 80], Subdural window [WC: 80, WW: 200] and Soft Tissue window [WC: 40, WW: 380]. Few examples of each type of injury and the image obtained after pre-processing is shown in the Fig. 2.

Table 1. Patient-wise distribution of hemorrhages

	Training	**Validation**	**Test**
Patients	6,122	765	765
Slices	2,67,279	25,783	24,636
Without hemorrhages	1,78,052	16,211	15,502
With hemorrhages	89,227	9,572	9,134
Epidural	2,709	266	170
Intraparenchymal	29,860	3,301	2,957
Intraventricular	21,586	2,408	2,211
Subarachnoid	29,299	3,407	2,969
Subdural	38,864	4,166	4,136

3.2 Implementation Details

The proposed network has been implemented using Tensorflow [1] and has been trained on the RSNA dataset with a mini-batch size of 64. The network has been trained on a machine powered by 4 GeForce RTX 2080 Ti GPUs with 11GB memory each. No data augmentation was used during training. The Adam optimizer with an initial learning rate of 0.01 has been used to train the network. To prevent overfitting, a dropout layer with a probability of 0.1 is introduced to the final fully connected layer of the network. The values for mini-batch size, initial learning rate and dropout are obtained by carrying out hyperparameter optimization using the hyperband algorithm introduced by [9]. The KerasTuner library [13] has been used to automate the process of Hyperparameter optimization. The models are compressed using Tensorflow Lite [6], which is a framework that automates model compression for Tensorflow models and generated compressed models that can be deployed in edge devices like mobile phones or microcontrollers. The metrics that were used to measure task-related performance are Accuracy, Sensitivity, Specificity and Area under the ROC curve (AUC). To quantitatively measure computational efficiency, we obtain and compare the number of model parameters, memory footprint, inference time and the number of floating point operations per second (FLOPS). For all the latter, a lower number implies superior efficiency.

3.3 Ablation Study: From SFCN to LightFCN

Using the SFCN as the baseline model, a number of architectural modifications have been introduced to finally obtain the LightFCN model. The list of intermediate models along with the architectural modification is shown below:

1. *SFCN*: The Simple Fully Convoluted Network with no modifications.
2. *SFCN_depthwise*: All Convolution operations are replaced by Depthwise Separable Convolutions, as used in [7]. In addition to that, the first convolution layer is modified to have a stride of 2.
3. *LightFCN*: In addition to Depthwise Separable Convolutions, Linear Bottlenecks and Inverted residuals are added to the model as in [15].

All the models have been trained for 5 epochs, and the results are shown in Table 2. We can see that the architectural modifications hasn't degraded the performance by a lot, and yet managed to bring down the size by a significant amount (From 12.3 MB to 1.7 MB).

3.4 Comparing performance with other models

The proposed method is compared with the state-of-the-art lightweight architectures: MobileNetV2 [15] and SFCN [14]. The models were all trained with same training data, and tested on same test data.

The quantitative comparison of the models using the aforementioned metrics is presented in Table 3. Our model achieved performance similar to the SFCN

Table 2. Comparison of the proposed Light-FCN model with the state-of-the-art methods.

Model	Accuracy	AUC	Specificity	Sensitivity	# Params	Size
SFCN	89.59	74.72	93.88	**55.55**	1 M	12.3 MB
SFCN_depthwise	89.85	69.68	94.38	44.97	136.1K	2.65 MB
Light-FCN	**92.09**	**74.76**	**97.99**	51.52	**86.8 K**	**1.7 MB**

Table 3. Performance comparison of the proposed model Light-FCN with the state-of-the-art architectures

MobileNetV2				
ICH	**Accuracy**	**AUC**	**Sensitivity**	**Specificity**
Epidural	96.66	62.10	27.06	97.14
Intraparenchymal	91.39	65.27	30.92	99.62
Intraventricular	92.36	89.22	85.40	93.05
Subarachnoid	88.56	69.92	45.36	94.47
Subdural	87.70	73.93	53.22	94.64
SFCN				
Epidural	98.37	56.24	13.53	98.96
Intraparenchymal	94.30	80.87	63.22	98.53
Intraventricular	95.39	77.75	56.26	99.24
Subarachnoid	91.34	69.49	40.71	98.26
Subdural	88.96	70.10	41.72	98.47
LightFCN				
Epidural	99.07	59.22	18.82	99.62
Intraparenchymal	90.63	76.28	57.42	95.14
Intraventricular	94.30	80.58	63.87	97.29
Subarachnoid	88.44	71.80	49.88	93.72
Subdural	84.23	67.55	42.47	92.64

model while reducing the number of parameters by a factor of 10. The sensitivity percentage for epidural hemorrhage is low for all models as a result of minimum number of samples in the training set. The highest sensitivity for SFCN and Light-FCN is obtained for intraventricular hemorrhage.

Furthermore, the models were created using the Tensorflow framework, and were compressed using the Tensorflow Lite framework [6] into the tflite format. Tensorflow Lite compresses a trained model through the means of quantization, clustering and pruning of model parameters. The resulting compressed models were deployed on a Raspberry Pi 4B, an accessible, low-cost, feature-rich embedded system, and their inference time was recorded. The sizes of the considered compressed models and their inference times are given in Table 4. The FLOPS for each model is also listed in Table 4.

Table 4. Comparison of the compressed model size for the proposed model Light-FCN and different architectures along with their inference time (per slice) on Raspberry Pi

Model	Compressed model size	Inference time	#Params	FLOPS
MobileNetV2	8.46 MB	190.51 ms	2.3M	615M
SFCN	3.8 MB	174.79 ms	1M	2.07G
Light-FCN	**312 KB**	**61.15 ms**	**86.8K**	**189M**

Table 5. Evaluating performance of loss functions on the test dataset after training Light-FCN on the training dataset for 6 epochs.

Loss Function	Accuracy	AUC	Sensitivity	Specificity
BCE	**93.49**	90.89	51.22	98.54
Focal	91.38	86.95	29.84	**98.87**
LDAM	92.44	**91.22**	**65.34**	94.62

3.5 Comparing Loss Functions

Using the LightFCN as the baseline model, the performance of multi-label versions of the binary crossentropy loss, focal loss and LDAM loss are compared. The multi-label binary crossentropy loss doesn't address class imbalance, but the multi-label focal loss and multi-label LDAM loss try to address class imbalance. In this study, three different LightFCN models are trained using each of the losses mentioned above and the performance metrics are observed.

It was observed that the LDAM loss had the highest sensitivity and AUC on the test dataset. Moreover, the LDAM loss seems to provide a good trade-off between sensitivity and specificity, thereby increasing the true positive rate in the presence of class imbalance, as expected. Interestingly, the binary crossentropy Loss gave better experimental results than the Focal Loss in accuracy, sensitivity and AUC. However, the LDAM loss seems to perform the best overall.

4 Conclusions

Our studies reveal that the Light-FCN architecture is capable of performing well in the task of ICH identification while being significantly more computationally efficient than other state-of-the-art architectures. The multi-label LDAM loss is observed to strike a better trade-off between sensitivity and specificity and increase the true positive rate, thereby solving the class imbalance problem to an extent. Achieving performance better than the state-of-the-art while maintaining computational efficiency, developing lightweight models that can process 3-dimensional scans and coming up with ways to improve performance by taking inter-class dependency into account remains the subject of future work.

Acknowledgements This work is supported by the start-up research grant given by the Science and Engineering Research Board (SERB), India.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015), https://www.tensorflow.org/
2. Burduja, M., Ionescu, R.T., Verga, N.: Accurate and efficient intracranial hemorrhage detection and subtype classification in 3D CT scans with convolutional and long short-term memory neural networks. Sensors **20**(19), 5611 (2020)
3. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. Advances in neural information processing systems **32** (2019)
4. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9268–9277 (2019)
5. Flanders, A.E., Prevedello, L.M., Shih, G., Halabi, S.S., Kalpathy-Cramer, J., Ball, R., Mongan, J.T., Stein, A., Kitamura, F.C., Lungren, M.P., et al.: Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge. Radiology: Artificial Intelligence **2**(3), e190211 (2020)
6. Google: Tensorflow lite. https://www.tensorflow.org/lite/ (2021)
7. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
8. Johnson, J.M., Khoshgoftaar, T.M.: Survey on deep learning with class imbalance. Journal of Big Data **6**(1), 1–54 (2019)
9. Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., Talwalkar, A.: Hyperband: A novel bandit-based approach to hyperparameter optimization. The Journal of Machine Learning Research **18**(1), 6765–6816 (2017)
10. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2999–3007 (2017)
11. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
12. Masud, M.: A light-weight convolutional neural network architecture for classification of covid-19 chest x-ray images. Multimedia systems pp. 1–10 (2022)
13. O'Malley, T., Bursztein, E., Long, J., Chollet, F., Jin, H., Invernizzi, L., et al.: Kerastuner. https://github.com/keras-team/keras-tuner (2019)
14. Peng, H., Gong, W., Beckmann, C.F., Vedaldi, A., Smith, S.M.: Accurate brain age prediction with lightweight deep neural networks. Medical image analysis **68**, 101871 (2021)
15. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)
16. Shuvo, M.B., Ahommed, R., Reza, S., Hashem, M.: Cnl-unet: A novel lightweight deep learning architecture for multimodal biomedical image segmentation with

false output suppression. Biomedical Signal Processing and Control **70**, 102959 (2021)
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
18. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6848–6856 (2018)