Ensemble Deep Convolutional Neural Network to Identify Fractured Limbs using CT Scans

Anup Khanal, Rodrigue Rizk, and KC Santosh
Applied AI Research Lab, Department of Computer Science, University of South Dakota, Vermillion, SD 57069
anup.khanal@coyotes.usd.edu, rodrigue.rizk@usd.edu, and santosh.kc@usd.edu

***Abstract*—Accurate classification between fractured and intact bones in Computed Tomography (CT) scan serves as a precursor to further treatment planning. CNN is no exception to handle this, and as an example AlexNet ranked top in the ImageNet challenge (2012). To overcome generalization errors, we propose to ensemble deep convolutional neural networks to check how well fractured limbs can be analyzed. It primarily includes voting (soft and hard), stacking, bagging, and feature soup on a backbone consisting of VGG19, ResNet152, Inception, MobileNet, and DenseNet169. On a clinically annotated dataset of size 5,567 CT scans, we achieved the highest accuracy of 0.977, precision of 0.959, recall of 0.960, F1-score of 0.960, and AUC of 0.971. To the best of our knowledge, this is the first time this dataset has been used to classify fractured and intact bones.**

***Index Terms*—fractured limbs, ensemble technique, CNN**

I. Introduction

The number of cases of Road Traffic Accidents (RTAs) continues to increase each year [1], and approx. 20% of orthopedic beds are occupied with traumatic cases resulting from RTAs at any given time [2]. In some complex cases, such as comminuted fractures, a single bone may break into multiple pieces, potentially including dislocations. In such cases, experts typically prefer to use Computed Tomography (CT) scans to better understand anatomic structure and depth information as CT scans provides them the ability to discriminate between various soft tissues and quantitatively measure individual bone tissue density using the attenuation of x-ray beams [3]. However, due to increasing RTA cases, experts are under significant pressure, and as a result, their analyses can be error-prone. To mitigate the possibility of errors and reduce the time required for diagnosis, we propose an ensemble deep convolutional neural network that automatically classifies fractured CT slices containing fractured bones, thereby assisting radiologists in diagnosis and facilitating further treatment planning.

II. Ensemble deep convolutional neural network

Inspired by the fact that ensemble techniques are able to mitigate regularization errors that may arise when high test accuracy is desired [4], we study industrial standard ensemble techniques on the backbone of pre-trained deep convolutional neural networks (DCNNs) [5]. Our proposed method involves several crucial steps: a) image conversion and labeling, b) selection of pre-trained backbone DCNNs, c) fine-tuning of backbone models, and d) Model ensemble.

We first converted CT scans [6] DICOM images to 8-bit JPEG format via linear scaling. We selected state-of-the-art DCNNs, namely VGG19, ResNet152, Inception, MobileNet, and DenseNet169. Transfer learning was used for weight initialization instead of training the models from scratch [7]. Using three dense layers with a softmax output layer, classification was performed. To optimize DCNNs for our specific dataset, we retrained/updated weights in some of the layers, excluding the batch normalization layers. The number of fine-tuned layers varied from 10 for VGG19 to 200 for DenseNet169. We then employed five ensemble methods: voting (soft and hard), bagging, stacking, and feature soup. Our findings reveal that ensemble DCNN can potentially provide an efficient and reliable diagnosis aid for medical professionals. To make it reproducible, our code is available in https://github.com/2ai-lab/Ensemble-DCNN

III. Experiments

A. Our results

On a dataset of size 5,567 CT scans and following a split-based approach (80|20 - train|validate), the highest accuracy in classifying fractured bones is 0.977 (see Table I). The training was performed on a single node of a High-Performance Computer (HPC) [1]

For better understanding, our detailed results are provided in Table I, and it uses industry standard evaluation metrics: accuracy, precision, recall, F1-score, and AUC. Soft voting provided better performance. Bagging came next to soft voting (in case of precision). As voting strategy plays a crucial role in bagging scheme, we tried three voting methods (soft voting, hard voting, and stacking), and the best one (soft voting) was selected as the final method for bagging. In summary, the results show that the soft voting and stacking methods are effective ensemble methods for classifying CT images of fractured limbs.

To the best of our knowledge, no previous research has attempted to classify between fractured and intact bones using CT images. Therefore, no comparison is available.

[1]Computations were performed on Lawrence HPC at the University of South Dakota, funded by NSF Award(NSF.1626516)

979-8-3503-3984-0/23/$31.00 ©2023 IEEE
DOI 10.1109/CAI54212.2023.00075

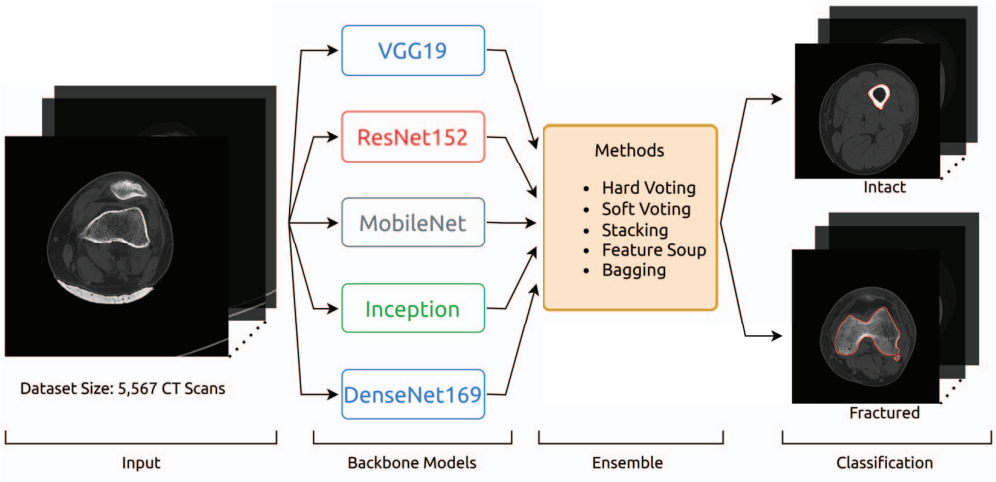

Fig. 1: Ensemble schema on five different backbone DCNNs. Each DCNN either serves as an independent classifier whose final prediction is counted as a vote or as a feature extractor in the ensemble process. Exact same set of input data is provided to each of the DCNNs.

TABLE I: Performance of different ensemble methods

Method	Accuracy	Precision	Recall	F1-Score	AUC
Hard Voting	0.974	0.937	0.955	0.946	0.967
Soft Voting	**0.977**	0.946	**0.960**	**0.960**	**0.971**
Stacking	0.976	0.950	0.950	0.950	0.967
Bagging	0.973	**0.959**	0.925	0.942	0.956
Feature soup	0.970	0.931	0.945	0.938	0.962

TABLE II: Performance of individual backbone model

Backbone Model	Accuracy	Precision	Recall	F1-Score	AUC
VGG19	0.964	0.930	0.920	0.925	0.949
ResNet152	0.956	0.923	0.891	0.906	0.933
MobileNet	0.952	0.912	0.930	0.921	0.941
Inception	0.955	0.922	0.886	0.904	0.931
DenseNet169	0.947	0.894	0.886	0.890	0.926

B. *Ablation study*

In this section, three important studies are considered: data augmentation, the performance of each model used for ensembling, and model contribution checking.

With data augmentation (geometric transformations: shear, flip, etc.) [8], no improvement has been observed. The result of the second study comparing the performance of each individual backbone model on the test set is presented in Table II. We then aimed at finding which DCNN has contributed in decision-making. Following the voting-based scheme, our findings indicate that VGG19 had a higher influence on the overall performance of the ensemble model. While removal of ResNet152, MobileNet, Inception, or DenseNet169 did not have a significant impact on the overall performance of soft voting.

IV. Conclusion and future works

In this paper, we have ensemble deep convolutional neural network to accurately identify CT scan slices that contain fractured bones in the limbs. This approach serves as a valuable aid to radiologists, enabling them to expedite the diagnosis process. On a clinically annotated dataset of size 5,567 CT scans, we have achieved the highest AUC of 0.971. Our future works include expanding our research to explore the use of state-of-the-art object detection techniques and algorithms to precisely label bone tissues as well as study the general applicability of the proposed methods in other domains.

References

[1] S Srivastava, A Bagga and R Singh Shekhawat, "Review of the Machine Learning Techniques in Road Crashes," 2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM), Dubai, United Arab Emirates, 2020, pp. 373-376, doi: 10.1109/ICCAKM46823.2020.9051506.

[2] S Ansari, F Akhdar, M Mandoorah, and K Moutaery, "Causes and effects of road traffic accidents in Saudi Arabia," *Public Health*, vol. 114, no. 1, pp. 37-39, Jan. 2000.

[3] G N Hounsfield, "Computed Medical Imaging," *Medical Physics*, vol. 7, no. 4, pp. 283–290, 1980.

[4] MA Ganaie, M Hu, AK Malik, M Tanveer, and PN Suganthan, "Ensemble deep learning: A review," *Engineering Applications of Artificial Intelligence*, vol. 115, pp. 105151, 2022, doi: 10.1016/j.engappai.2022.105151. ISSN 0952-1976.

[5] A Kumar, J Kim, D Lyndon, M Fulham and D Feng, "An Ensemble of Fine-Tuned Convolutional Neural Networks for Medical Image Classification," in *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 1, pp. 31-40, Jan. 2017, doi: 10.1109/JBHI.2016.2635663.

[6] DD Ruikar, KC Santosh, RS Hegadi, L Rupnar, and VA Choudhary, "5K+ CT images on fractured limbs: A dataset for Medical Imaging Research," *Journal of Medical Systems*, vol. 45, no. 4, 2021.

[7] H Tang and X Cen, "A Survey of Transfer Learning Applied in Medical Image Recognition," 2021 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA), Dalian, China, 2021, pp. 94-97, doi: 10.1109/AEECA52519.2021.9574368.

[8] S Dutta, P Prakash, and C G Matthews, "Impact of data augmentation techniques on a deep learning based medical imaging task," in Proc. SPIE 11318, Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications, Houston, TX, USA, 2020, p. 113180M, Mar. 2020, doi: 10.1117/12.2549806