Evaluation of Grad-CAM for explaining Deep Learning's decisions on various medical imaging datasets

Ifigeneia Athanasoula[1], Ioannis D. Apostolopoulos[2], and Peter P. Groumpos

[1] Department of Electrical and Computer Technology Engineering, University of Patras, Patras, Greece; ece6932@upnet.gr, groumpos.ece.upatras.gr
[2] Department of Medical Physics, School of Medicine, University of Patras, Patras, Greece; ece7216@upnet.gr

Abstract. Deep Learning (DL) is a well-established pipeline for feature extraction in medical and non-medical imaging tasks, such as object detection, segmentation, and classification. However, DL faces the issue of explainability, which prohibits reliable utilisation in everyday clinical practice. The present study employs the well-established Grad-CAM algorithm to assess the decisions of a Deep Learning framework in various medical image classification tasks. Seven datasets are utilised, involving images from SPECT, CT, Microscopy, and X-Ray, which correspond to numerous diseases, including Lung Cancer, Coronary Artery Disease, and COVID-19. The main conclusion of the research is that DL with Grad-CAM might reveal important image features. However, it is observed that on many occasions, Grad-CAM shows the model's inefficiency in discovering the right locations, even in the classification accuracy is at a top level.

Keywords: Deep Learning; Grad-CAM; Explainability

1 Introduction

There is much discussion about the quality and usefulness of the image features extracted by Deep Learning (DL) methods, especially in medical imaging. DL holds first place in terms of the number of features mined and quantified throughout the processing layers of a deep network; however, few studies have performed a deep analysis of the clinical, prognostic and diagnostic value of these features. The application of DL in processing medical images, especially in developing predictive and diagnostic models, is hindered by the fact that it is neither explainable nor transparent. Moreover, the model's supervision of the decision-making process is unknown to engineers, let alone to medical staff. On the other hand, the discovery of vital image biomarkers has recently catalysed the emergence of many techniques based on Machine Learning (ML) and others that export features in a non-automatic way [1, 2]. Most related research in the field is focused on predicting clinical outcomes by analysing pre-defined potential biomarkers extracted from medical and clinical image data. The inherent ability of DL to reveal such biomarkers is not formally validated. Especially for clinical image analysis, the automatic procedures must be straightfor-

ward, transparent, explainable and reproducible [3]. Every scientific conclusion derived from an Artificial Intelligence (AI) method, which may affect and redefine the medical pipeline in many domains, must rely on established techniques, analytical experiments, and clinical evaluation. In medical imaging, the definition of quantifiable and reproducible significant biomarkers evolves into a particular field of research. It utilises the full range of tools developed, such as ML, DL, and Radiomics.

The present study intends to assess DL methods from a biomedical engineering perspective rather than a statistical and mathematical evaluation based solely on predefine accuracy metrics. The study evaluates the potential importance of medical imaging features extracted by the successful DL model called FF-VGG19 [4].

2 Methods

2.1 Datasets

The present study utilises seven medical imaging datasets of various modalities and domains. **Error! Reference source not found.** summarises the characteristics of each dataset.

Table 1. Datasets utilised for the experiments

Dataset Name	Modality and domain	Classes (number of images)
PET/CT	PET/CT scans, including Solitary Pulmonary Nodules (SPNs)	Benign (61), Malignant (111)
LIDC-IDRI *[5]*	CT scans, including SPNs	Benign (620), *Malignant (616) SPNs*
COV_X *[6]*	X-ray scans of respiratory diseases	Pulmonary Edema (293), Pleural Effusion (311), Obstructive Pulmonary Disease (315), Pulmonary Fibrosis (280), COVID-19 (455), Bacterial and Viral Pneumonia (910), Normal cases (1341)
COV_CT *[7]*	CT scans of respiratory diseases	COVID-19 (349), Normal (397)
Cells *[8]*	Microscopic Images of cancerous cells	Eosinophil (424), Typical Lymphocyte (3937), Monocyte (1789), Myeloblast (3268), Neutrophil (4419)
MPIP *[9]*	SPECT Myocardial Perfusion Imaging Polar Maps of Coronary Artery Disease (CAD)	CAD (136), Normal (80)
MNIST *[10]*	Skin images for skin cancer recognition	Actinic Keratoses (327), Basal cell carcinoma (514), Benign keratosis (1099), Melanoma (1113), Melanocytic nevi (6705)

2.2 Deep Learning and Grad-CAM

Based on the architecture of VGG19, a modification is proposed to obtain more information from the high-level feature-extracting layers. This modification is entitled Feature-Fusion VGG19 (FF-VGG19) [4]. The layers of the network are completely trainable.

The Grad-CAM [11] algorithm intends to identify the areas of the input image having a critical effect on the classification decision of the classifier placed at the top of the CNN. Hence, its functionalities are fully exploited in object detection tasks, where a specific image area contains the desired object.

3 Results

We present the quantitative results in **Table 2**. It is observed that FF-VGG19 obtains good accuracy in classifying the LIDC-IDRI, COV_X, and Cells datasets. Sub-optimal metrics are reported for the rest of the datasets.

Table 2. Results

Dataset	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC	F1 – score
PET/CT	69.77	69.76	65.76	76.64	60.93
LIDC-IDRI	82.52	76.29	88.70	90.67	80.80
COV_X	84.94	-	-	95.25	-
COV_CT	78.57	87.96	70.27	83.48	79.23
Cells	91.61	-	-	94.46	-
MPIP	66.08	85.29	33.75	90.70	39.26
MNIST	77.93	-	-	84.58	-

In **Fig. 1**, we present the results of the Grad-CAM algorithm. The Grad-cam algorithm traces the convolutional layers and neurons participating strongly in formulating the final classification outcome. In this way, FF-VGG19 reveals the important areas of the original images, wherein the decisive features are discovered.

For the SPNs depicted in **Fig. 1**, subfigure a, it is noticed that the benign SPNs are highlighted in red, and their surroundings are not highlighted at all, in contrast with the malignant SPNs, where the red colourmap covers the entire image. It is demonstrated that the model seeks information in the correct areas, although the discovered information could not be visualised, at least with the Grad-cam algorithm. It is also noticed that the COVID-19 X-ray images are classified as such based on discovered patterns in the high respiratory. In contrast, normal cases are classified based on features in other locations. That demonstrates that CNN is looking for information in the right direction, as it has been proven that the novel Coronavirus affects specific areas of the respiratory system. However, it is also observed that some irrelevant features are extracted and misjudged as decisive ones.

The detection of COVID-19 in CT images is questioned in the scientific community. For this reason, the highlighted areas of the CT images (subfigure c) are unreliable. However, the CNN also highlights irrelevant areas (e.g. c1 outer left picture, c2 outer left picture). Therefore, the model is assumed to be confused by the incomplete and irrelevant-to-covid-19 information discovered in CT cases. Since the CT_COV dataset included few images, no further assumptions are made at this stage.

For the Myocardial Perfusion polar maps, the CNN is based on green areas of the initial image, which is the correct method. However, the initial image is not informative in a significant proportion of the subjects and leads to false positives.

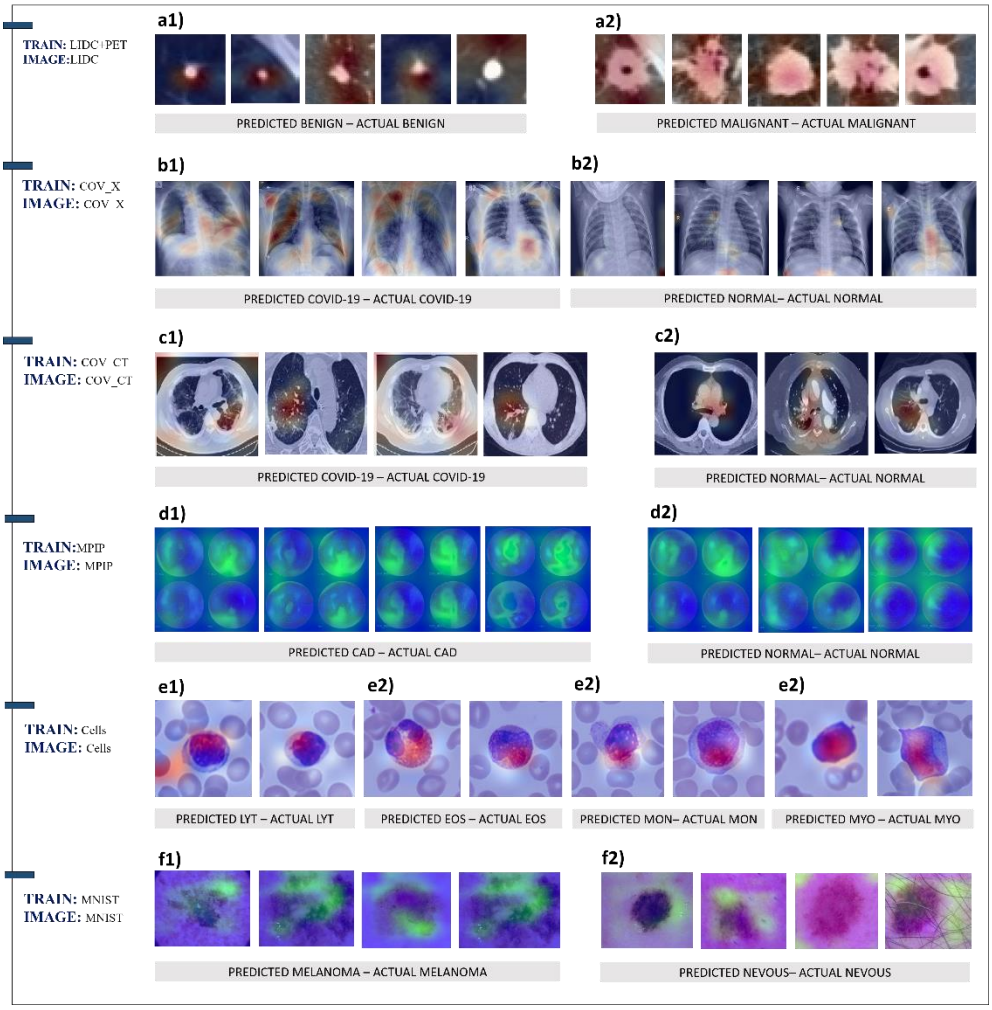

Fig. 1. Grad-CAM results

4 Conclusions

This analysis yields strong evidence that medical-specific features (i.e. image biomarkers) are extracted when training the proposed networks from scratch and observed in large-scale datasets, especially in pathological findings. The outcome derives from obvious features (e.g. colour, size, shape). Moreover, the study performs a preliminary analysis of the Grad-CAM algorithm to inspect the suggested regions.

5 References

1. Traverso, A., Wee, L., Dekker, A., Gillies, R.: Repeatability and Reproducibility of Radiomic Features: A Systematic Review. Int. J. Radiat. Oncol. 102, 1143–1158 (2018). https://doi.org/10.1016/j.ijrobp.2018.05.053.
2. Jain, D., Singh, V.: Feature selection and classification systems for chronic disease prediction: A review. Egypt. Inform. J. 19, 179–189 (2018). https://doi.org/10.1016/j.eij.2018.03.002.
3. Razzak, MI, Naz, S., Zaib, A.: Deep Learning for Medical Image Processing: Overview, Challenges and the Future. In: Dey, N., Ashour, AS, and Borra, S. (eds.) Classification in BioApps. pp. 323–350. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-65981-7_12.
4. Apostolopoulos, I.D., Papathanasiou, N.D., Panayiotakis, G.S.: Classification of lung nodule malignancy in computed tomography imaging utilising generative adversarial networks and semi-supervised transfer learning. Biocybern. Biomed. Eng. 41, 1243–1257 (2021). https://doi.org/10.1016/j.bbe.2021.08.006.
5. Armato, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., Kazerooni, E.A., MacMahon, H., van Beek, E.J.R., Yankelevitz, D., Biancardi, A.M., Bland, P.H., Brown, M.S., Engelmann, R.M., Laderach, G.E., Max, D., Pais, R.C., Qing, D.P.-Y., Roberts, R.Y., Smith, A.R., Starkey, A., Batra, P., Caligiuri, P., Farooqi, A., Gladish, G.W., Jude, C.M., Munden, R.F., Petkovska, I., Quint, L.E., Schwartz, L.H., Sundaram, B., Dodd, L.E., Fenimore, C., Gur, D., Petrick, N., Freymann, J., Kirby, J., Hughes, B., Vande Casteele, A., Gupte, S., Sallam, M., Heath, M.D., Kuhn, M.H., Dharaiya, E., Burns, R., Fryd, D.S., Salganicoff, M., Anand, V., Shreter, U., Vastagh, S., Croft, B.Y., Clarke, L.P.: The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans: The LIDC/IDRI thoracic CT database of lung nodules. Med. Phys. 38, 915–931 (2011). https://doi.org/10.1118/1.3528204.
6. Cohen, J.P., Morrison, P., Dao, L.: COVID-19 image data collection. ArXiv Prepr. ArXiv200311597. (2020).

7. Zhao, J., Zhang, Y., He, X., Xie, P.: COVID-CT-Dataset: A CT Scan Dataset about COVID-19. ArXiv200313865 Cs Eess Stat. (2020).
8. Matek, C., Schwarz, S., Marr, C., Spiekermann, K.: A Single-cell Morphological Dataset of Leukocytes from AML Patients and Non-malignant Controls, https://wiki.cancerimagingarchive.net/x/fgWkAw, (2019). https://doi.org/10.7937/TCIA.2019.36F5O9LD.
9. Apostolopoulos, I.D., Papathanasiou, N.D., Spyridonidis, T., Apostolopoulos, D.J.: Automatic characterisation of myocardial perfusion imaging polar maps employing deep learning and data augmentation. Hell. J. Nucl. Med. 23, 125–132 (2020).
10. 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 (2018). https://doi.org/10.1038/sdata.2018.161.
11. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localisation. Int. J. Comput. Vis. 128, 336–359 (2020). https://doi.org/10.1007/s11263-019-01228-7.