Data Mining and Machine Learning In Medicine

Luke Edgecombe
Robotics with Industrial Application
Heriot-Watt University
Edinburgh, Scotland
Email: le4005@hw.ac.uk

Author
Course
Heriot-Watt University
location
Email: email

Author
Course
Heriot-Watt University
location
Email: email

Author
Course
Heriot-Watt University
location
Email: email

***Abstract*—some text**

***Keywords*—Medicine, Machine learning, Classification, Optimisation, Image analysis**

I. Introduction

Some text introducing the concepts - classifiers - image classifiers - convolutional NN - model refinement

In this report, a selection of differing machine learning (ML) techniques are explored. Each technique was chosen based on its area of application. The subject area of medicine was chosen for exploration, and a selection of datasets are explored with different features and formats: (i) a collection of brain magnetic resonance imaging (MRI) images that exhibit three cancer-based pathologies; (ii) a tabular dataset containing patient demographic and healthcare-related information; (iii) ?.

To approach the ML problems, an understanding of previous literature on the topic is required for an effective workflow. Academic papers were reviewed to gain an effective skill set and background knowledge required to begin experimentation.

II. Literature

ML as a concept is the basis of the field of artificial intelligence (AI); the primary purpose of ML is the generation of algorithms that are able to learn. In this regard, research has focused on the optimisation of these systems by generation of new techniques and identifying the most optimal combinations of pre-existing software/hardware [1].

The MRI image set relies on the classification ML technique; the specifics of this are introduced here.

A. Image Classifiers

This selection was reinforced by past research on brain cancer MRIs [2]. In particular, a convolutional neural network (CNN) was deployed. This ML technique has been shown to be highly effective in image classification [3]. A review into CNNs describes the typical architecture and construction techniques [4]. As the task differs, the design of the CNN can change, although the general format remains essentially the same. This typical structure, shown in Fig. 1, is comprised of ordered convolutional and pooling layers that come together to make a feature extractor. The inputted data is converted into a feature representation. In combination with this, fully-connected neural layers are integrated with activation functions to perform the desired ML operation.

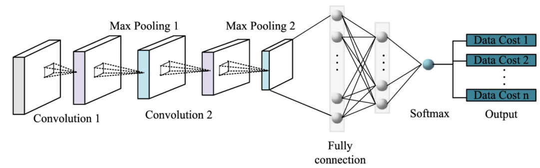


Fig. 1: Convolutional Neural Network Architecture [4]

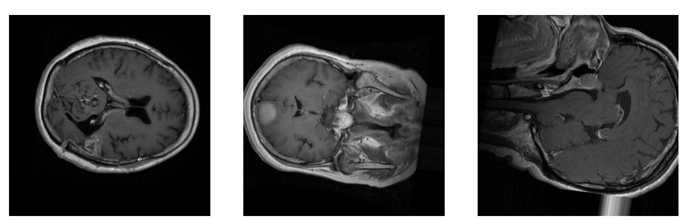
Prior to data being used in ML, in order to increase suitability and feasibility of the model training, it should receive some sort of pre-processing. In order to effectively train a model, the data should be affected in a way that benefits the robustness and accuracy of the training process [5].

1) Pre-Processing: In image classification, effective pre-processing has been shown to increase the rate of identification [6]. For very large datasets, where manual review is not possible, automatic pre-processing is a must. Studies show that image cropping and reshaping, image resizing, and background noise removal are beneficial to the training process.

III. Data Analysis and Exploration

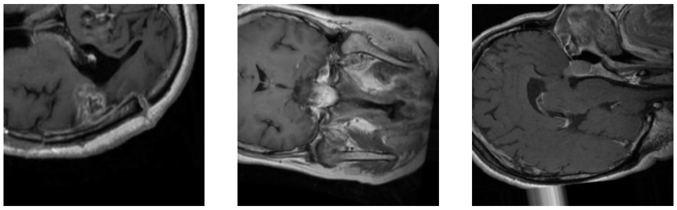
Based on the literature discussed previously, the data used in the following scenarios has undergone careful analysis in order to understand and prepare for upcoming ML.

The MRI classifying dataset was sourced from a comprehensive collection of MRI images collected from a series of hospitals in Bangladesh [7]. The data holds great value and significant effort was made to collect and label high quality images. A sample of the three classes of data present is shown in fig. 2, pictures (a to b). The proprietor has already uniformly resized the images to an equal length and height. Some further ML specific pre-processing steps are executed:



(a) Glioma (b) Meningioma (c) Tumor

Fig. 2: Sample images from MRI scans



(a) Glioma (b) Meningioma (c) Tumor

Fig. 3: Preprocessed sample images from MRI scans

- Random resized crop to 224×224 pixels
- Random horizontal flip with 50% probability
- Conversion to 3-channel greyscale
- Conversion to tensor format
- Normalisation to floating-point values in range [0,1]
- Standardisation using ImageNet mean and standard deviation

IV. Experimental setup

A. Baseline Training and Evaluation Experiments

B. Neural Networks

V. Results

References

[1] R. Abdulkadirov, P. Lyakhov, and N. Nagornov, "Survey of optimization algorithms in modern neural networks," *Mathematics*, vol. 11, no. 11, 2023. [Online]. Available: https://www.mdpi.com/2227-7390/11/11/2466

[2] J. Kang, Z. Ullah, and J. Gwak, "Mri-based brain tumor classification using ensemble of deep features and machine learning classifiers," *Sensors*, vol. 21, no. 6, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/6/2222

[3] P. Wang, E. Fan, and P. Wang, "Comparative analysis of image classification algorithms based on traditional machine learning and deep learning," *Pattern Recognition Letters*, vol. 141, pp. 61–67, jan 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865520302981

[4] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, "A review of convolutional neural networks in computer vision," *Artificial Intelligence Review*, vol. 57, no. 4, Mar 2024. [Online]. Available: https://link.springer.com/article/10.1007/s10462-024-10721-6

[5] M. Salvi, U. R. Acharya, F. Molinari, and K. M. Meiburger, "The impact of pre- and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis," *Computers in Biology and Medicine*, vol. 128, p. 104129, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482520304601

[6] P. M. Diop, J. Takamoto, Y. Nakamura, and M. Nakamura, "A machine learning approach to classification of okra," in *2020 35th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, 2020, pp. 254–257.

[7] M. M. Rahman, "Brain cancer - mri dataset," *Mendeley Data*, vol. 1, Aug 2024. [Online]. Available: https://data.mendeley.com/datasets/mk56jw9rns/1