Interpretable AI For Protein Expression Images and Associated Clinical Metadata

Satish Pilla, Atif Khan & Stephen McGough

School of Computing, Newcastle University, Newcastle, UK

Introduction

With a conservative frequency estimate of about 1:5,000, mitochondrial disorders are among the most prevalent inheritable diseases [1]. Diagnosis and understanding the different mitochondrial diseases are extremely difficult because they have a wide range of symptoms in each patient and affect different organs and tissues of the body [2]. However, recent studies show deep learning algorithms with interpretability and explainability, especially **Convolutional Neural Networks (CNN),** can help us automatically diagnose and evaluate different diseases by detecting the different patterns in the images.

In the study, we developed deep learning models using transfer learning to predict mitochondrial diseases and used existing machine learning interpretability and explainability AI approaches for computer vision, like **Grad-CAM [3], and Neural Disentanglement (concept whitening) [4]**, to understand the features that result in the prediction of mitochondrial disease from protein expression images.

Methods

We divided our experiment into two phases: In the first phase, we adapted existing deep learning models to classify different mitochondrial diseases for single-channel protein images and multi-channel protein images. In the second phase to evaluate the results, and to understand the underlying pathology of mitochondrial diseases, we applied Grad-CAM [3] and concept whitening [4].

To train and evaluate the model a total of 280 images were collected from Wellcome Centre Mitochondrial Research [WCMR]. Each image was collected in two different formats, 140 JPEG and 140 TIFF images.

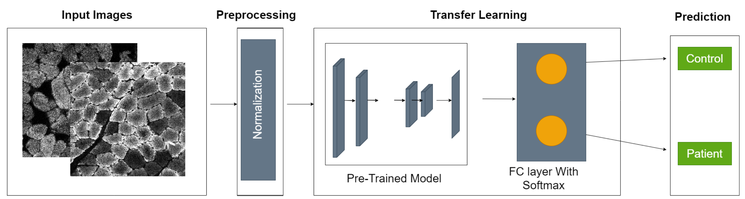


Figure 1: Overview of the methodology transfer learning for mitochondrial disease diagnosis

Evaluation Methodology

Since our main goal is to understand the underlying pathology we adapted two existing methodologies to evaluate the models.

Grad-CAM: This method draws attention to the areas of the input image that the model focused on during the classification process, indicating that the feature maps created in the final convolution layer hold the spatial information needed to effectively capture the visual pattern. These visual patterns help in classifying the classes. The layers and abstracted features from the trained model are used to apply the Grad-CAM technique. The architecture explaining the Grad-CAM technique is shown in Figure 2.

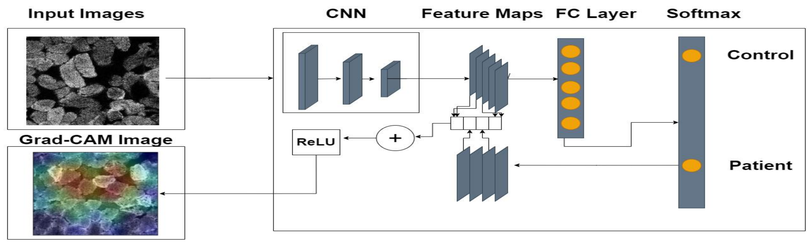


Figure 2: Architecture describing the Grad-CAM technique [3]

Concept Whitening: To make the model interpretable we applied the concept whitening. By using this approach, the CNN network's output and the features of an input image may be related more easily, making the deep learning model interpretable.

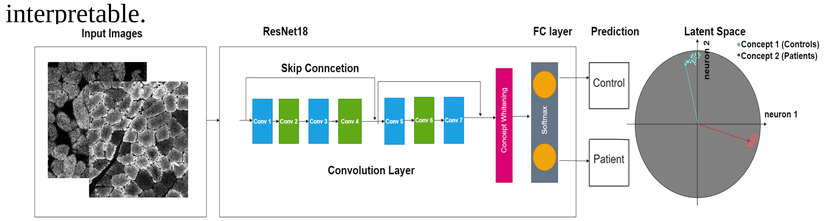


Figure 3: Architecture to describe the Concept Whitening technique [4]

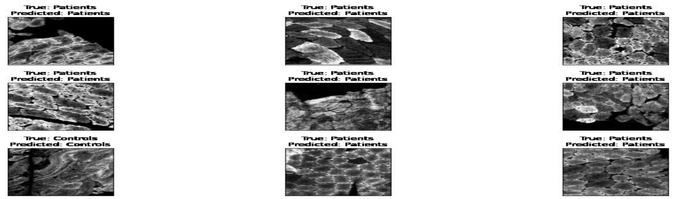
Results

we applied transfer learning with fine-tuning on the two pre-trained models (VGG and ResNet) for single-channel protein images (SDHA) and multi-channel protein images in order to predict mitochondrial diseases using protein expression images. The models' performance was verified using a number of metrics, including recall, precision, F1-score, accuracy, accuracy and loss graphs, and confusion matrix.

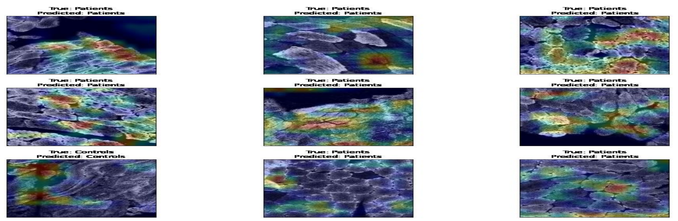
Model	Accuracy	Precision	Recall	F1 Score
VGG16 (single-channel (SDHA))	95%	96%	94%	95%
ResNet-50 (single-channel (SDHA))	90%	91%	90%	90%
ResNet-18 + CW (single-channel(SDHA))	67%	65%	65%	65%
VGG16 (multi-channel images)	95%	95%	94%	95%

Results

Predictions: The predictions of VGG16 accurately determined whether the protein images under test belonged to the control or patient for both the experiments, but ResNet50 and ResNet50+CW had the lowest prediction accuracy and misclassified the test images.



Grad-CAM technology was used on the VGG16 model of experiment 1 (single-channel protein image), which emphasized the areas that the model paid the most attention to during the feature extraction process.



Discussion

VGG16 model for single-channel and multi-channel attained the highest accuracy with 95% among other models. Grad-CAM highlighted the areas where the model was paying more attention.

Some limitations were identified with concept whitening, even though we were able to attain good accuracy the main issue was the unavailability of clinically approved metadata of the protein expression images.

References

1. Gorman, A. M. Schaefer, Y. Ng, N. Gomez, and Blakely, "Prevalence of nuclear and mitochondrial DNA mutations related to adult mitochondrial disease," Annals of Neurology, vol. 77. Wiley Online Library, pp. 753–759, 2015.
2. P. Forny, E. Footitt, J. E. Davison, A. Lam, and Woodward, "Diagnosing mitochondrial disorders remains challenging in the omics era," Neurology. Genetics, vol. 7. p. e597, 2021. doi:10.1212/NXG.0000000000000597.
3. LR. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visualexplanations from deep networks via gradient-based localization," International journal of computer vision, vol. 128. pp. 336–359, 2020. doi: 10.1007/s11263-019-01228-7.
4. Z. Chen, Y. Bei, and C. Rudin, "Concept whitening for interpretable image recognition," NatureMachine Intelligence, vol. 2, no. 12, pp. 772–782, 2020, doi: 10.1038/s42256-02000265-z.