Interpretable and Explainable AI for Protein Expression Images

Satish Pilla (B210472095)

School of Computing, MSc Data Science with Artificial Intelligence

Supervisor: Dr Stephen McGough

Motivation:

- 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.

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.

Aim:

To determine if Deep Learning (DL) can be used as a reliable method to classify mitochondrial diseases using interpretability and explainability approaches.

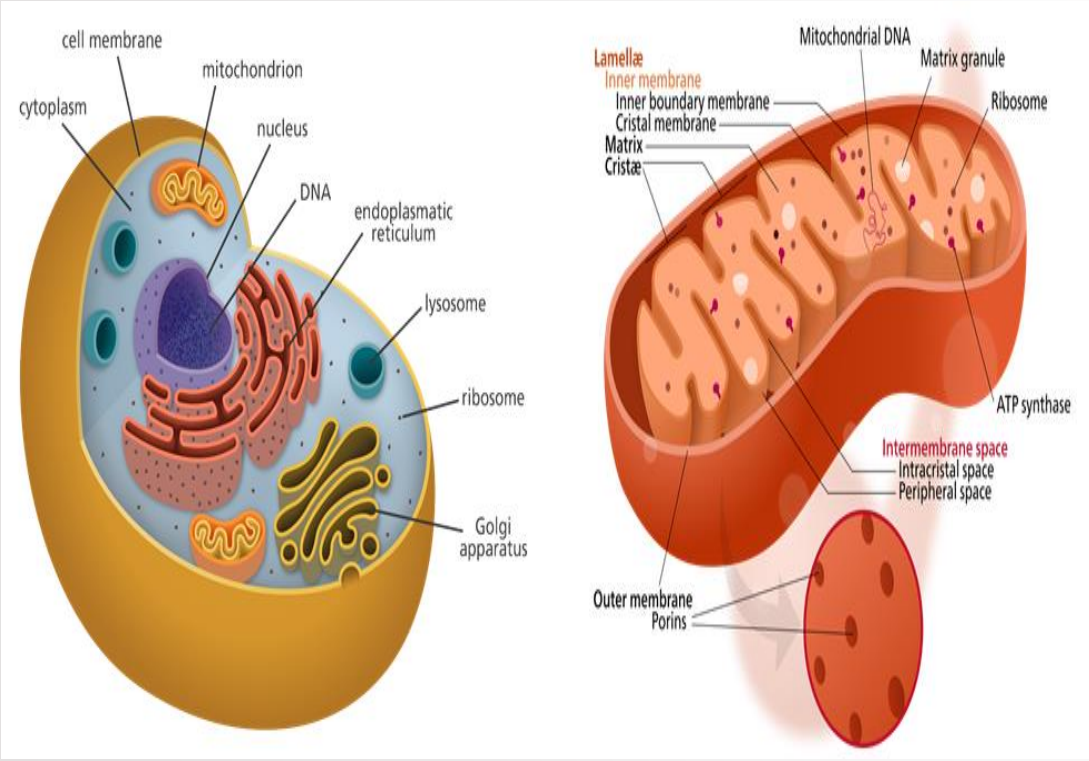
- Interpretability Method: Neural Disentanglement
- Explainability Method: Saliency Map

Objective:

This investigation has primarily two objectives:

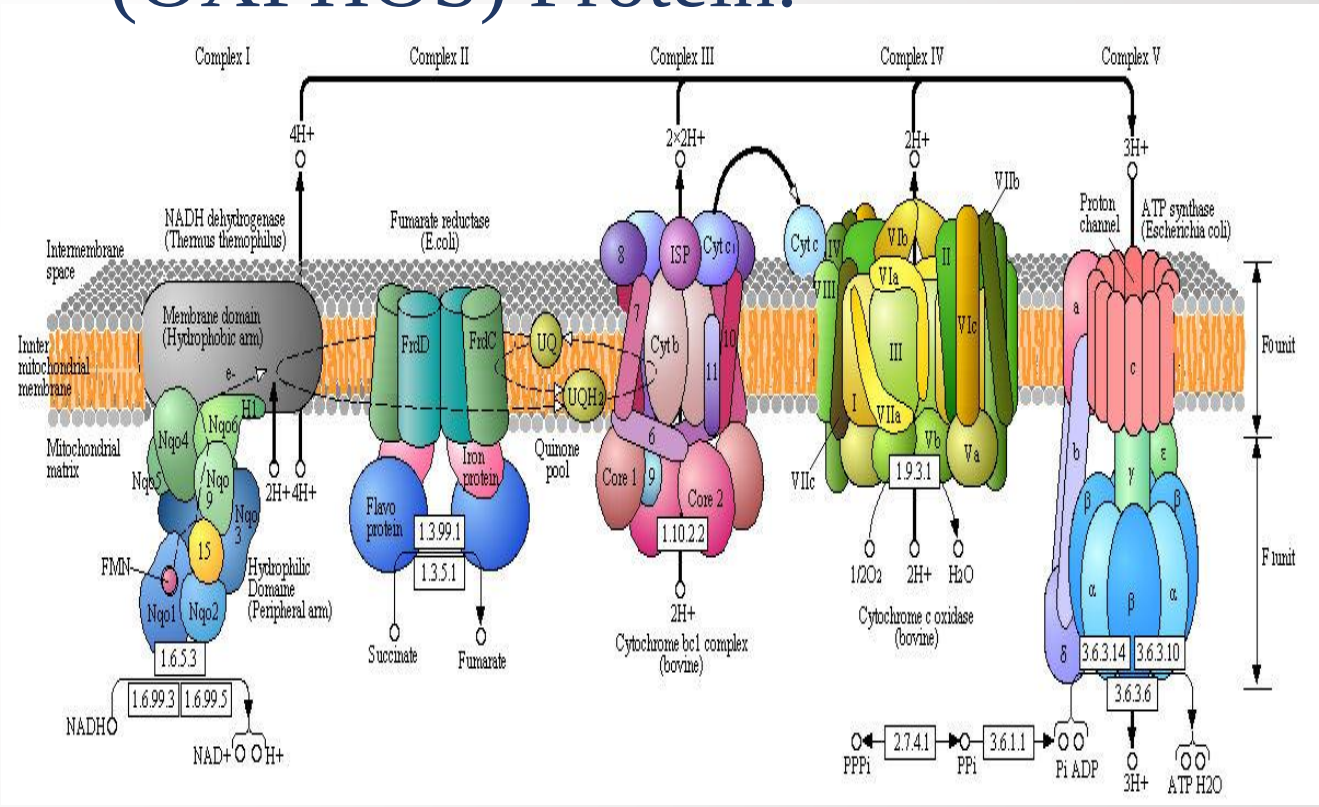
- Adapt the pre-trained models like VGG16 and ResNet-50 to classify different mitochondrial diseases.
- Investigate the existing interpretability and explainability to understand the underlying pathology of mitochondrial diseases.

Mitochondria:



Credit: The internal structure of mitochondria, Cell (2000).

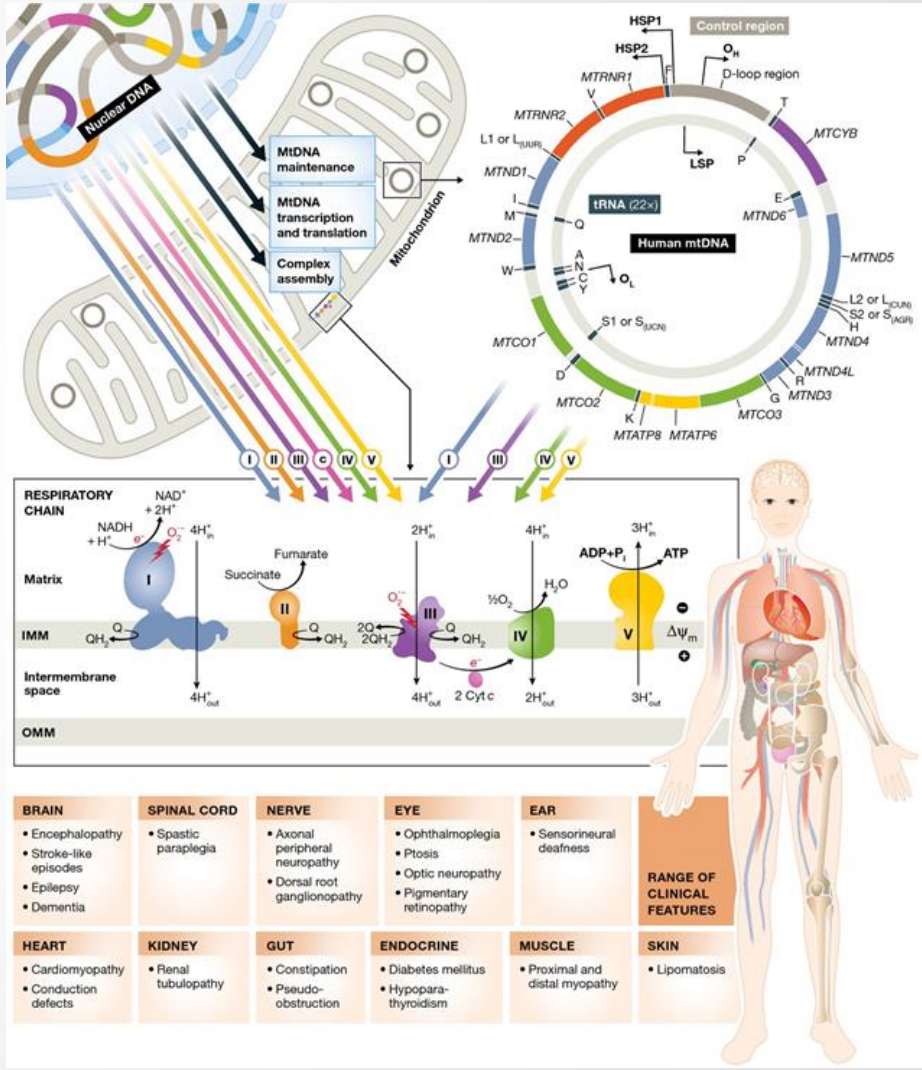
Oxidative Phosphorylation (OXPHOS) Protein:



Credit: KEGG, www.genome.jp/kegg-bin

Mitochondrial Diseases:

- Untreatable and affect 1 in 5,000
- Manifest as a result of mutations in genes that encode mitochondria
- Highly heterogeneous and adversely affect high energy demanding cells e.g. skeletal muscle cells and neurons
- Studied by finding a relationship between genome, OXPHOS proteins & disease symptoms



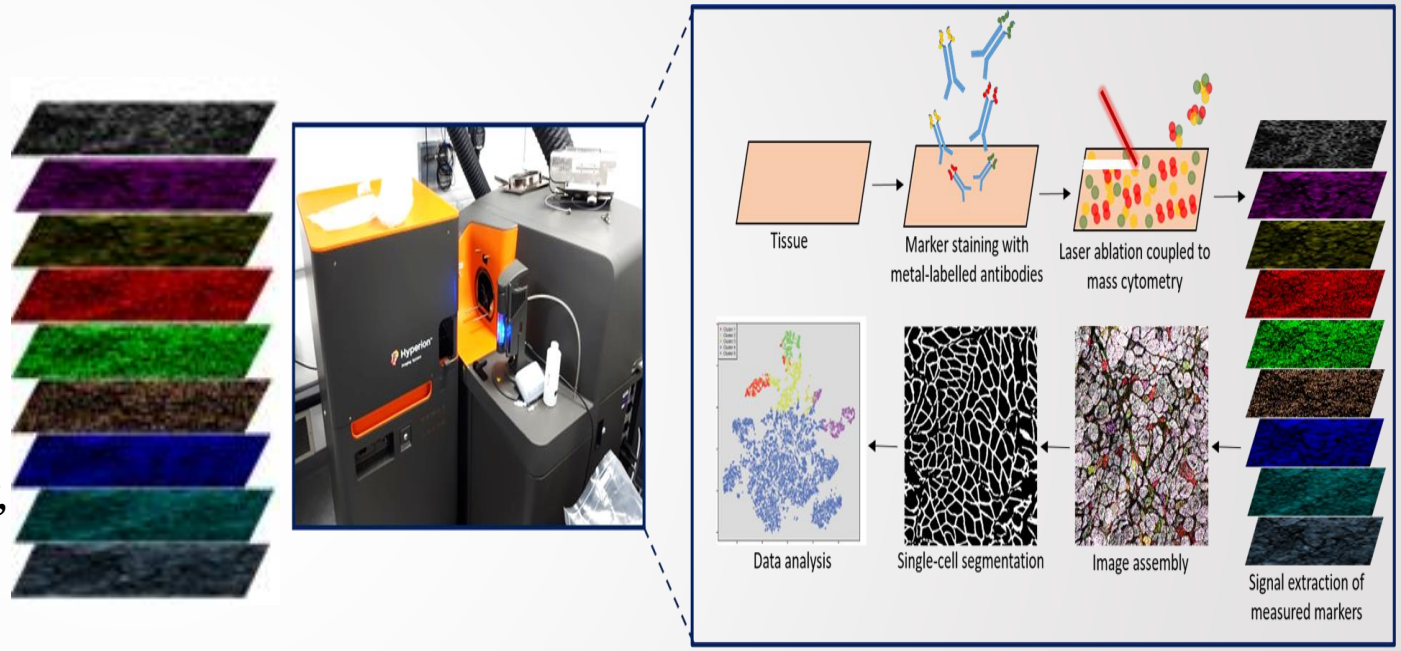
Credit: EMBO Molecular Medicine

Dataset:

Images from Image Mass Cytometry (IMC)

- Patient: 10 Samples
- Control: 4 Samples

Target Proteins: SDHA, TOM22, NDUFB8, OSCP, GRIM19, VDAC1, COX4, MTCO1, UQCRC2, and Dystrophin



IMC Process A) Hyperion imaging mass cytometer B) Diagrammatic representation of IMC experiment and associated analysis workflow

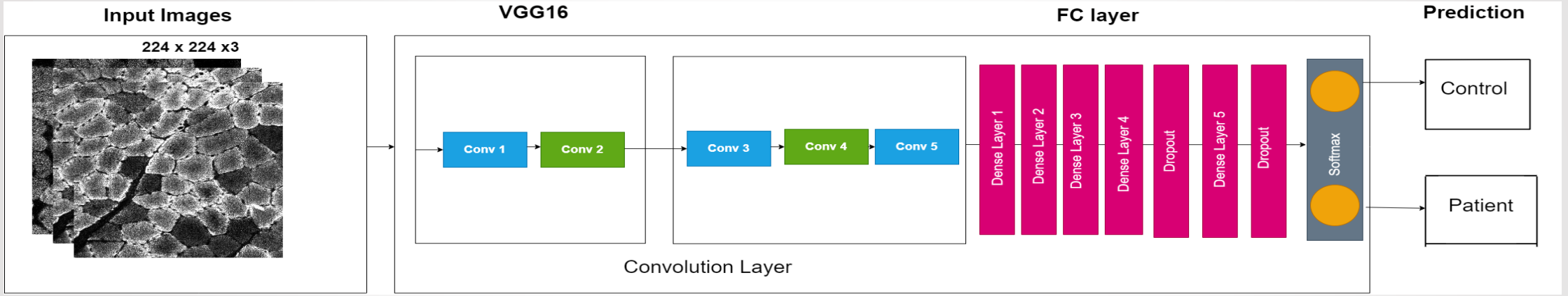
Credit: Warren et al.,(2020) Nature Communications. doi: 10.1038/s41598-020-70885-3

Data Preparation:

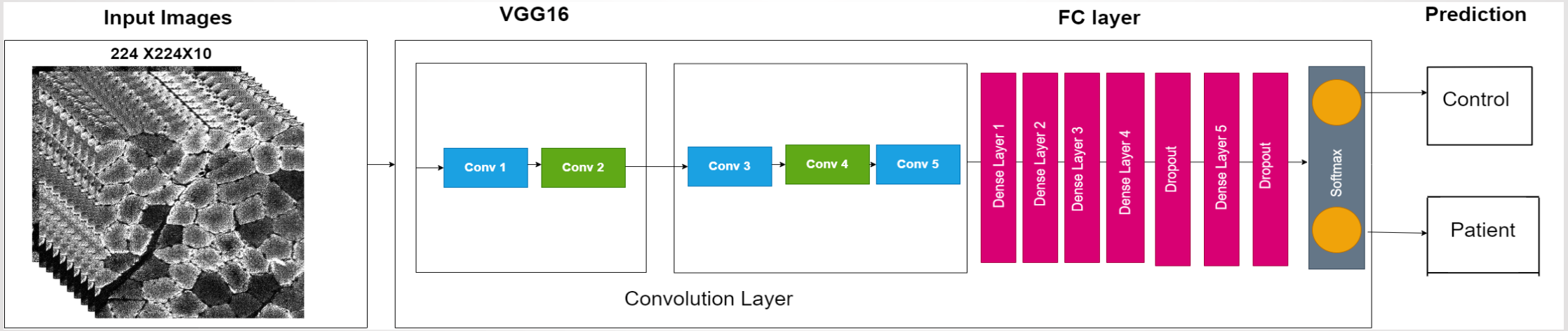
- Data-Augmentation
- Splitting the Images
- Stacking the Images
- Split into Train, Test, and Validation sets 80:10:10 ratio

Model Building:

- CNN with Transfer Learning
- VGG16 – ImageNet Weights



Summary of the VGG16 model architecture adapted in this research for single channel protein images



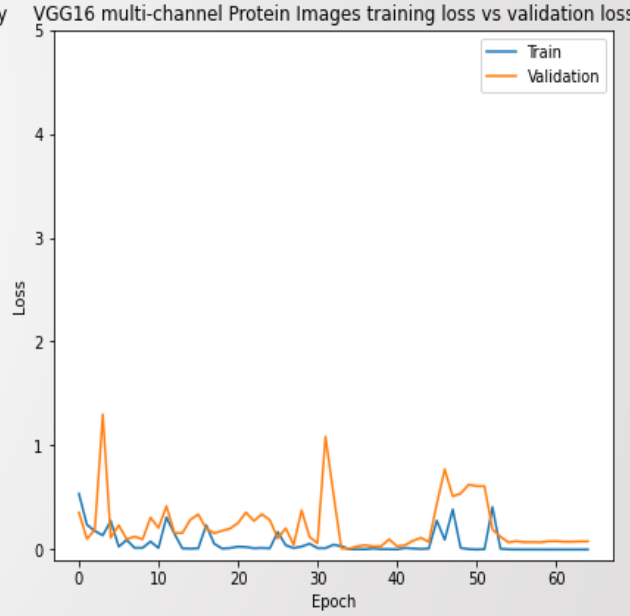
Summary of the VGG16 model architecture adapted in this research for multi-channel protein images

Model Performance:

Training Accuracy: 0.9729 Training Loss: 0.1149
Test Accuracy : 0.9550 Test Loss : 0.1649

Training accuracies vs. validation accuracies and training loss vs. validation loss for VGG16 single-channel protein images (SDHA)

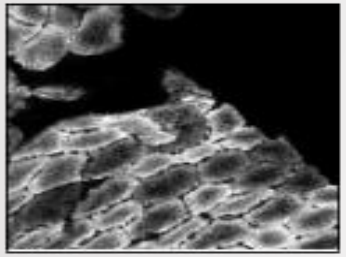
Training Accuracy: 0.9829 Training Loss: 0.1245
Test Accuracy : 0.9550 Test Loss : 0.1567



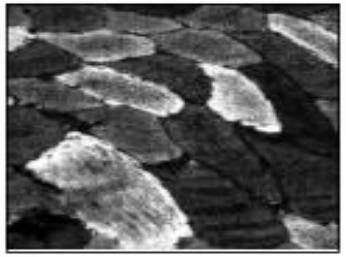
Training accuracies vs. validation accuracies and training loss vs. validation loss for VGG16 multi-channel stacked protein images

Model Prediction for Single-channel Protein Images:

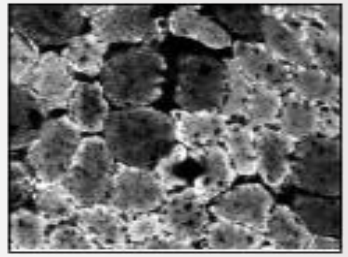
True: Patients
Predicted: Patients



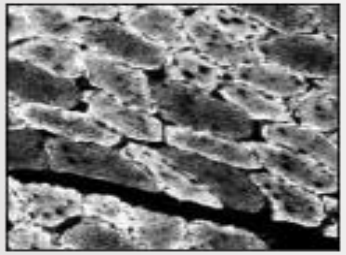
True: Patients
Predicted: Patients



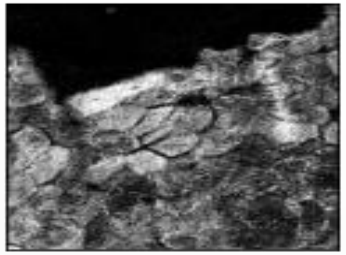
True: Patients
Predicted: Patients



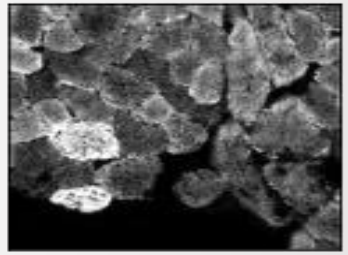
True: Patients
Predicted: Patients



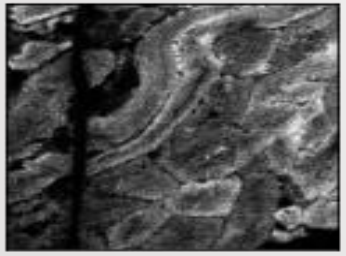
True: Patients
Predicted: Patients



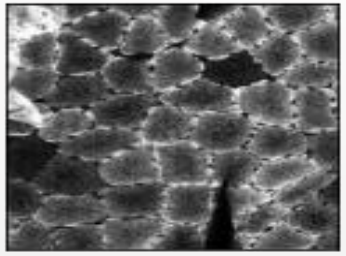
True: Patients
Predicted: Patients



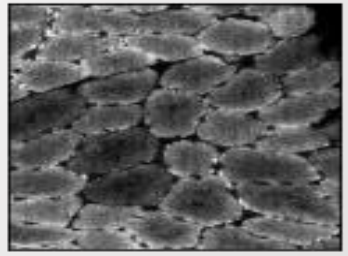
True: Controls
Predicted: Controls



True: Patients
Predicted: Patients

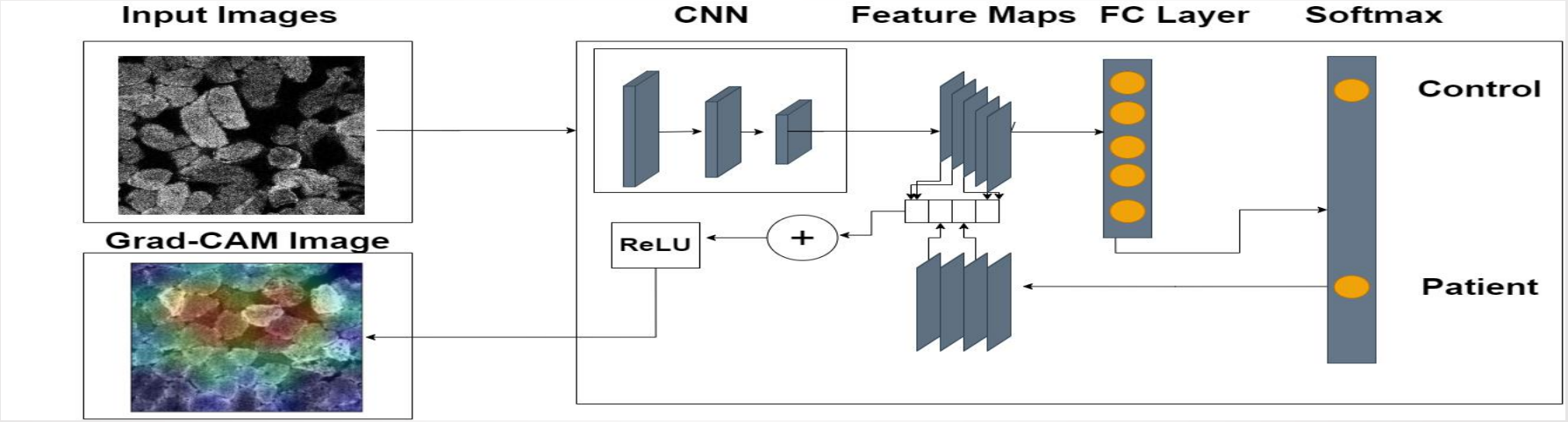


True: Patients
Predicted: Patients



Explainability Method: Grad-CAM

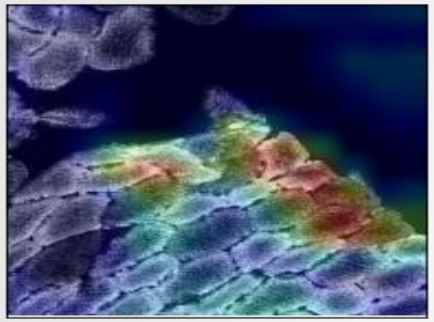
- Grad-CAM technique is used to evaluate and find which features of protein expression images are leading to the predictions of mitochondrial disease.
- This method draws attention to the areas of the input image that the model focused on during the classification process, indicating that the feature 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.



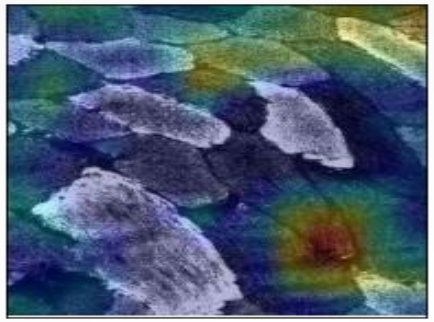
Ref: doi: 10.1007/s11263-019-01228-7

Grad-CAM Results on Test Dataset

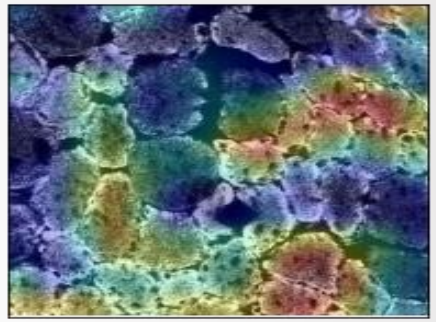
True: Patients
Predicted: Patients



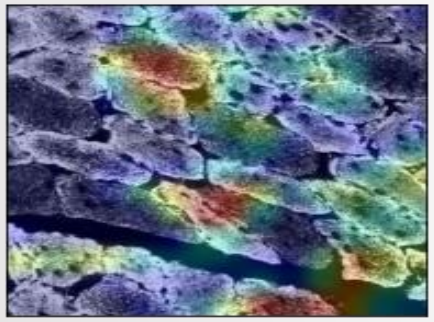
True: Patients
Predicted: Patients



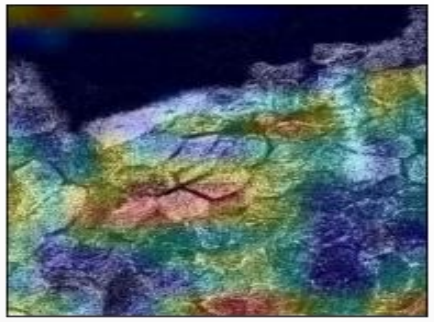
True: Patients
Predicted: Patients



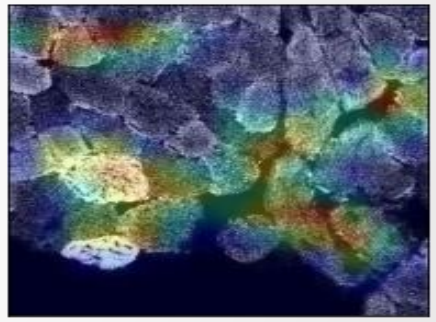
True: Patients
Predicted: Patients



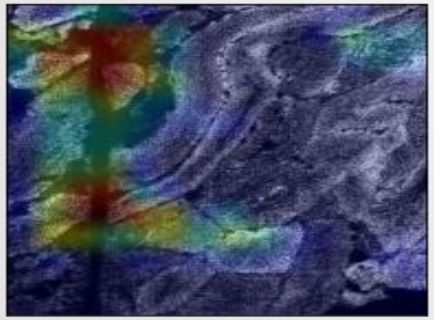
True: Patients
Predicted: Patients



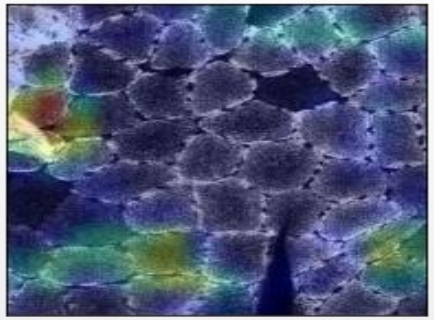
True: Patients
Predicted: Patients



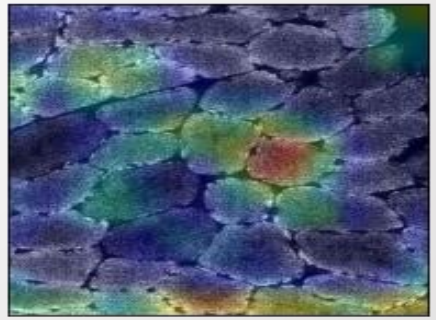
True: Controls
Predicted: Controls



True: Patients
Predicted: Patients



True: Patients
Predicted: Patients



Interpretability Method: Neural Disentanglement (Concept Whitening)

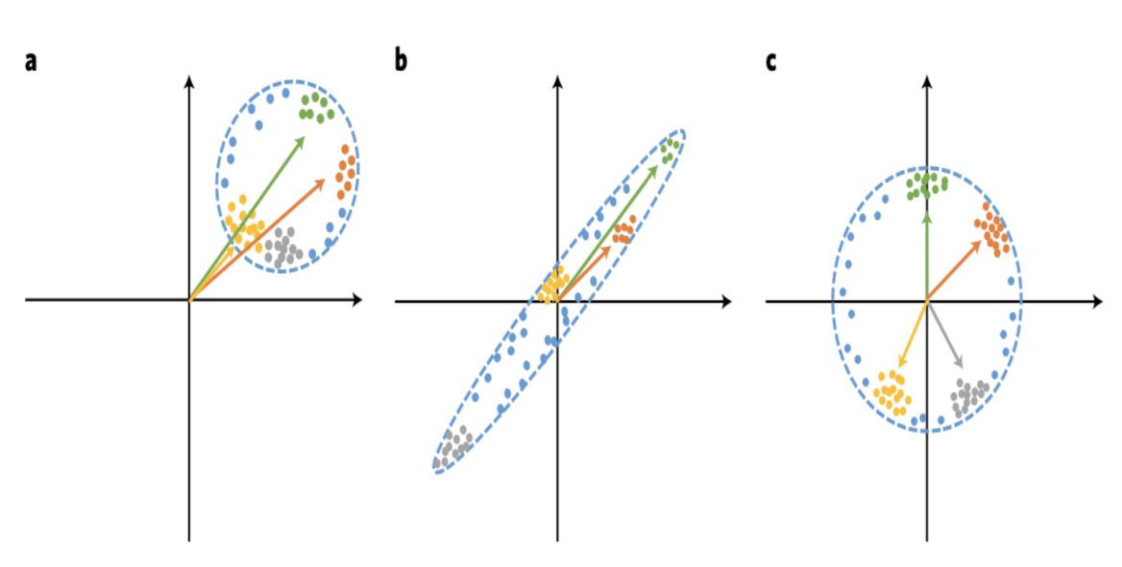
- Interpretable mechanism: aligns latent space of a layer with human-interpretable concepts.
- A method to alter existing high-performance models to become more transparent and self-explanatory.

What is a concept?

For an input image, the Concept may be anything, like in skin cancer classification where we need to predict benign or malignant, we can have concepts like age and lesion size as two concepts.

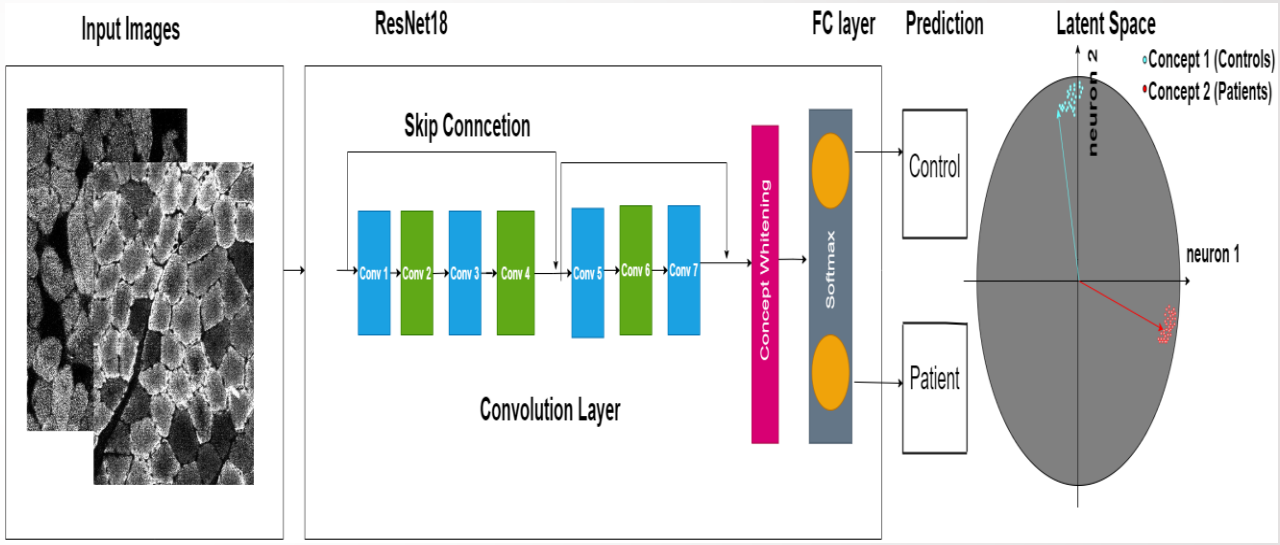
Ref: DOI: 10.1038/s42256-020-00265-z

Data Distribution in Latent Space



Data distributions in the latent space. (a) Not mean-centred; (b) standardized; (c) standardized and decorrelated (CW)

Concept Whitening



Architecture to describe the Concept Whitening technique

When a Concept Whitening module is added to a CNN,

- the latent space is whitened (decorrelated and normalized)
- the axes of the latent space are aligned with concepts of interest

Ref: DOI: 10.1038/s42256-020-00265-z

Concept Whitening Results

- Training Accuracy with ResNet CNN: 0.87
- Training Accuracy with ResNet CNN+CW: 0.67
- We saved the top 50 images, which got the greatest activation with respect to the concepts for scientists at WCMR to study and tell us whether those images belong to the right classes or not.

The sole Purpose:

From start, we didn’t have any actual metadata concepts for our experiment, we manually created our own concepts. These concepts are not clinically approved.

The sole purpose of implementing the concept whitening was to demonstrate its working to the WCMR scientists, so that in the future if the scientists are able to gather the relevant concepts related to mitochondrial diseases, the concept whitening technique can help them in a better way to understand the underlying pathology of the disease.

Conclusion

- For experiment 1 single- channel protein images, the accuracy of VGG16, ResNet-50, and ResNet-18+CW was 95.25%, 90%, and 78%, respectively.
- For Experiment 2 multi-channel protein images, VGG16 achieved 95% accuracy.
- Grad-CAM highlighted the areas where the model was paying more attention. Since the Grad-CAM technique focuses on highlighting the important pixels in the images, these highlighted regions might help scientists to focus only on those highlighted regions for a prediction and better understanding of the underlying pathology of mitochondrial diseases.
- Whereas with concept whitening, we were able to attain good accuracy but the main issue was the unavailability of clinically approved metadata of the protein expression images.

Newcastle University

Thank you!