The Alan Turing Institute

Data Study Group Final Report: University of Birmingham

13 – 24 Sep 2021

Automating perfusion assessment of sublingual microcirculation in critical illness

https://doi.org/10.5281/zenodo.6799096

Contents

1 Executive summary

1.1 Challenge overview

Achieving and maintaining normal sub-lingual blood flow in small ($\leq 20\mu m$) vessels, termed as microcirculation, is essential for critically ill intensive care patients since this is where the delivery of blood and oxygen to tissues occurs. However, historically, most clinical trials & treatments have focused on blood flow in the larger blood vessels (macrocirculation) [20, 21, 14], largely due to the greater ease with which this can be practically measured.

Technological advances have enabled video recordings of the sub-lingual microcirculation (i.e. from under the tongue) to be obtained using dark-field microscopy (DFM). However, the difficulties in analysing these videos has hindered the uptake and utilisation of this imaging modality. Currently, these short video sequences are analysed mostly by hand, to quantify the vessel density and flow within vessels within the field of view. The manual analysis and vessel segmentation is an extremely labour intensive procedure, which can take up to one hour to score a single video [12].

This challenge aimed to establish whether a single validated measure of microcirculatory perfusion (microcirculatory flow index) can be predicted directly from a DFM video sequence, without intermediate manual analysis steps. Automatic analysis that can be carried out in (near) real-time would facilitate the incorporation of microcirulatory targets into clinical trials by enabling the impact of interventions to be quantified and enacted upon with the aim of optimising the microcirculation and improving patient outcomes.

1.2 Brief data overview

The data comprised of 800 grayscale videos of different length and quality obtained via DFM from 52 patients monitored over 4 days. Figure 1 shows a couple of example frames selected from such videos. For several patients, video sequences were recorded of the sub-lingual microcirculation, while each recording was taken at a rate of 25 frames per second with different lengths. These videos have been manually

analysed to obtain perfusion parameters per short clip, labelling each data point. For detailed information please see Chapter 2.

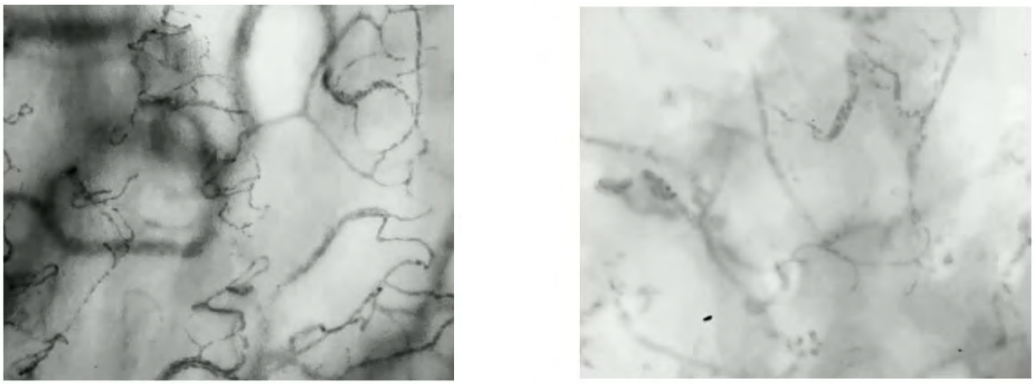


Figure 1: Example frames from sub-lingual microcirculation videos.

1.3 Main objectives

The main objective of whether a perfusion index can be predicted from a DFM video sequence was separated into two sub-objectives:

1. understand and explore the video data to identify factors that influence the automatic prediction (e.g. video stability, brightness, etc.).
2. develop techniques to predict a single measure of perfusion.

1.4 State-of-the-art literature

State-of-the-art methods in automated monitoring of microcirculation largely depend on intravital video microscopy (IVM) and data science-based solutions. Mahmoud et. al [10] have used a novel two-step image processing algorithm using a trained Convolutional Neural Network (CNN) to functionally analyse IVM microscopic images without the need for manual analysis. The first step used an adaptation of the well-established Steger Unbiased Detector of Curvilinear Structures (SUDCS) algorithm to segment the vessel structures. While in the second

step the authors used a 3D-CNN algorithm to determine whether a vessel segment carries blood flow or not.

Demir et al. [2] presented an experimental algorithm that automatically extracted microvascular network and quantitatively measures changes in the microcirculation. It involved two key parts: video processing and vessel segmentation. Microcirculatory videos were first stabilised using the Gaussian gradient algorithm, while the local contrast and clarity was improved by contrast limited adaptive histogram equalisation (CLAHE) approach. The vessel segmentation was evaluated using functional capillary density (FCD) values [2].

CapillaryNet is a fully automated state-of-the-art system to quantify skin nutritive capillary density and red blood cell velocity from handheld microscopy videos. The methodological details can be found here [6]. Similarly, Rizzuto et al. [16] performed video analysis of RBCs perfusion in a microfluidic device using a novel transfer learning approach. It involved the application of AlexNet with support vector machines (SVM) for healthy cell classification.

Current literature shows that medical image analysis often suffers from overfitting in CNNs due to typically small datasets [17]. Data augmentation is used as an approach to overcome this problem through applying various transformations to the training data images such as translation, rotation, flipping and zooming. This provides a larger training dataset and forces the model to be invariant to transformations that may be present in the real world, hence reducing overfitting in the network. Other regularisation methods, such as dropout and batch normalisation, have also been developed to try to extend CNNs for application on smaller datasets [17].

Instead of training a CNN from scratch on a new dataset, transfer learning in computer vision aims to leverage models pre-trained on large datasets, such as ImageNet with millions of available images, to transfer the knowledge learned from one task to another. In [15], it was shown that transferring knowledge from models built in ImageNet greatly improved performance in chest X-ray detection across a wide range of pre-trained models. Transfer learning can be done either as feature extraction, which simply uses the feature output from a pre-trained model and feeds it into a new classifier, or fine-tuning, which retrains the last few layers of the

pre-trained model. Popular pre-trained models applied in transfer learning medical images analysis include AlexNet, VGGNet, ResNet and DenseNet [22].

1.5 Approach

This work pursued the task of automating the perfusion assessment in three stages.

1. Modelling challenges, see Chapter 3
2. Video pre-processing, see Chapter 4
3. Machine learning based experiments, see Chapter 5

1.6 Main conclusions

The work in this report demonstrates that a machine learning based model is able to predict perfusion from a DFM video sequence. The accuracy of these models varies between 60-70%. We identify that the machine learning models require more data to be able to improve accuracy.

1.7 Limitations

Several limitations were identified and are listed below.

1. The dataset is comprised of a smaller number of data points than would normally be used to train neural networks, the accuracy of which increases with a larger dataset. As a result, several data augmentation methods have been applied.
2. To feed videos into deep neural network architectures, pre-processing steps were explored. For example, one of the techniques compressed the video into a single image by applying absolute temporal gradient for all the frames and then summed over all the gradients with subsequent histogram equalisation. This may result in information loss.
3. Manual probe placement results in varying stability of the recordings adding to the noise in the data. We identified methods to detect

the most unstable recordings that could be used to provide instant feedback in the clinic.

4. Due to time and computational constraints, most model fits were not repeated, nor were any confidence values calculated for any metrics. This in conjunction with the small test set means result comparison may not be statistically significant.

5. Vascular perfusion scores are not found as discrete groups in real-world since patients can appear with conditions on a continuous scale. However, the available labels in the data classify patients in discrete perfusion scores.

1.8 Recommendations and future work

The work in this report has shown that deep learning based methods are able to make predictions that are better than random prediction, which shows the promise of automated analysis techniques. However, more data is required for future work since deep learning methods are sensitive to data size and class imbalance (see Section 3.1). Furthermore, there was an inherent class imbalance present in the data particularly affecting lower MFI patients.
On the methods front, more sophisticated methods for modelling the time dimension may aid predictive accuracy while generative models may offer the potential to produce synthetic data which could be used to augment the small dataset.
The video sequences were found to suffer from stability and quality issues, for which image processing methods were explored in this work to counter such issues. Future work may explore the development of a dark-field microscopy pen with an integrated inertial motion unit that can provide instant feedback to the clinician as towhether the recording is of sufficient quality.

2 Data overview

2.1 Dataset description and handling

The dataset provided for this study was obtained from 52 patients which were monitored over four consecutive days, recording a sub-lingual flow video four times per day. This led to about 800 videos of different lengths and qualities. Due to the probe handling variation during recording, some videos were too unsteady to be subsequently used for model training and predictions purposes. The dataset provided also included a subset of stabilised videos where the recording was more steady. Each video name contained information about the patient anonymous ID, the day, the equipment used, etc. To simplify the later analysis, the videos were renamed with a unique 4-digit identifier and collected in a single folder. Additionally, a comma separated value (CSV) file was created that links to the unique video name. The CSV file also incorporated essential information regarding the labelling, namely the perfusion parameters as well as other parameters. The perfusion parameters were derived from a proprietary software Automated Vascular Analysis (AVA, developed by MicroVision Medical, Amsterdam, The Netherlands), and are listed below. The MFIq and MFIv values (described below) were used for image-based automated analysis algorithm training and validation.

TVD: total vessel density (mm2/mm2), this is the amount of the image taken up by vessels.
PVD: perfused vessel density (mm2/mm2), this is the density of vessels with velocity scores of 2 or 3 (where 0 is no flow, 1 is intermittent, 2 is sluggish and 3 is continuous).
PPV: partial perfused vessel parameter, this is the percentage of the ratio of PVD and TVD.

Both of the following parameters, MFIq and MFIv, can be considered as a 'perfusion index'. Either both or just a single are treated as the **output labels** for each data point.

MFIq: microcirculatory flow index by quadrant, where the image is divided

into quadrants and each quadrant is scored for the predominant velocity (0,1,2,3) and the average is taken.
MFIv: microcirculatory flow index, where all capillary segments are mapped manually. Each capillary segment is given a score (0,1,2,3).

After analysing the entire data, some mismatches were identified. Considering all the videos as well as the perfusion parameter, there were some cases in which a video had no perfusion parameters and vice versa. Reasons for missing parameters were that parameters are only generated for the best videos per patient per day so that the AVA software was not able to compute parameters due to video quality issues (see section 2.2).

2.2 Video quality issues

Due to the manual placement of the imaging probe on the patient's tongue, the exact placement of the probe on the tongue will vary each time. Furthermore, a proportion of the recordings will also present as unsteady due to human error. Ideally, the method needs to be robust enough to account for this additional noise or omit recordings that are deemed too unstable. Incorporating instant feedback on recording steadiness would allow the clinician to know when to repeat the recording and increase data quality.

2.3 Data summary

2.3.1 The output variables: MFIq and MFIv

The output variables whose values we are trying to predict in this challenge are, for each video clip, the microcirculatory flow index of vessels (MFIv) and the microcirculatory flow index by quadrant of the video (MFIq).

2.3.2 Relationship between MFIq and MFIv

Performing a linear regression on the data available tells us that the MFIq and MFIv values of each video are strongly correlated. The line of linear regression is given by MFIv $= 0.50868 + 0.77862 \cdot$ MFIq. The R-squared

value for the linear regression is 0.849 and p-value $<$ 0.001, which is usually interpreted as good evidence for a linear relationship between the two variables. Figure 2b tells us that the true value of MFIv is 0.125 or

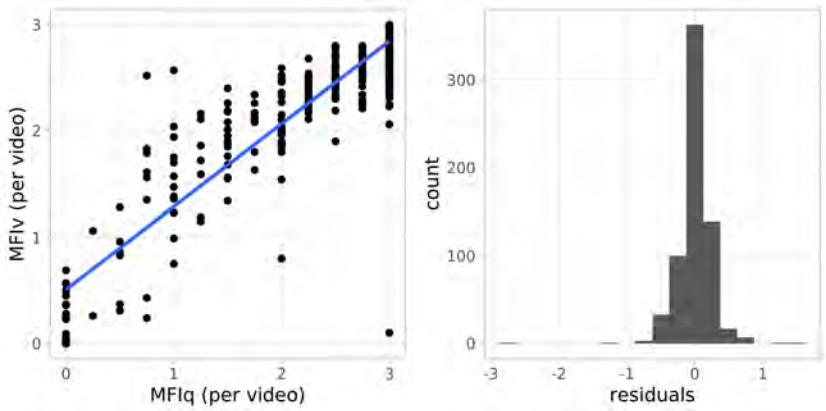


Figure 2: Linear regression between output variables (on individual videos) and residual differences.

less away from the predicted value for roughly half of the data available (bars' width is 0.25). It would be interesting to know whether the values with the data points with the highest residuals are also the hardest to predict using ML. Lastly, where multiple videos were available for a single patient on a given day, the average MFI values are considered in Figure 3. This shows that averaging multiple measures produce an even stronger linear relationship and removes "outliers" concerning the linear regression. The equation of the line of best fit is now given by MFIv $= 0.38281 + 0.82597 \cdot$ MFIq with an R-squared value of 0.932 and p-value $<$ 0.001.

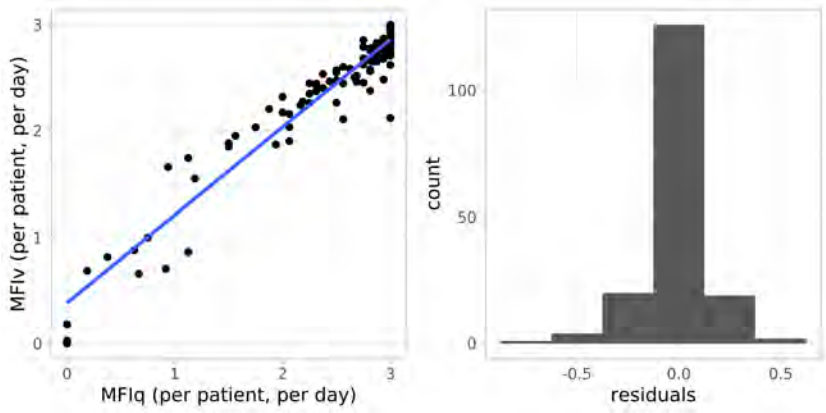


Figure 3: Linear regression between output variables (per patient) and residual differences.

3 Modelling challenges

3.1 Class imbalance

The dataset received suffers from class imbalance, with the healthy class samples outnumbering the diseased class samples. When a class imbalance exists in the training data, the prior probability of the majority class can cause models to over-classify the majority group, also resulting in observations in minority groups being misclassified more often than the majority group [9]. This problem is exacerbated as the dataset size decreases and the class skew increases. Metrics, such as accuracy, can also become misleading; for example, if the majority class is 99% of the data, then always predicting the majority class label will give an accuracy of close to 99%. Metrics should therefore be taken into consideration alongside the data distribution.

From the available data, it was found that the data distribution is skewed heavily towards MFI values of 3 (i.e. patients with healthy blood flow). There are much fewer data observations of patients between MFI 0 and 2 (considered poor blood flow). We assumed that we can set the modelling problem up as a classification task, binning the MFI values into classes 0, 1, 2 & 3 representing values (v) as $0>v<0.5$, $0.5>=v<1.5$, $1.5<=v<2.5$ and $v>2.5$. Fig. 4 shows the class proportions for MFIv and MFIq for the entire dataset where it is evident that healthy blood flow (class 3) is the majority class with a proportion of approximately 75%. With the small size of the overall dataset, of 666 usable videos, there are very few examples of low MFI scores (0 & 1) for the model to be trained and tested on. The class imbalance combined with the relatively small dataset makes the classification very difficult.

There have been various methods proposed to combat the class imbalance problem in the deep learning literature. A thorough review can be found in [9], and a brief overview of methods will be provided here. Class imbalance methods can be broadly split into three categories. The first, *data-level* methods, use data sampling methods such as random under-sampling (RUS), which discards samples from the majority class, and random over-sampling (ROS), which duplicates samples from the minority classes. When done in its simplest form, RUS reduces the amount of training data in the model, which is already scarce in this

dataset, whilst ROS can cause overfitting [9]. Variants of RUS and ROS have been proposed, such as two-phase learning, which first trains using RUS and then fine-tunes using all of the data, but the performance of variants may not consistently outperform simple RUS and ROS, with ROS outperforming RUS in most cases [13]. In *algorithm* methods, modifications are made to the learning algorithm, usually in the form of weighting or cost function adjustments, to reduce bias towards the majority group. Methods such as output thresholding, which divides the network outputs class by its prior probability, and new cost functions such as mean false error, which splits the typical mean squared error cost into false positive and negative error, have been shown to improve classification performance. Lastly, *hybrid* methods combine data-level and algorithm-level methods. There has not been a conclusive comparison across class imbalance treatment methods with many variations of datasets and class skews. Choosing the correct type of class imbalance treatment is therefore an open problem, and currently trial-and-error is required.

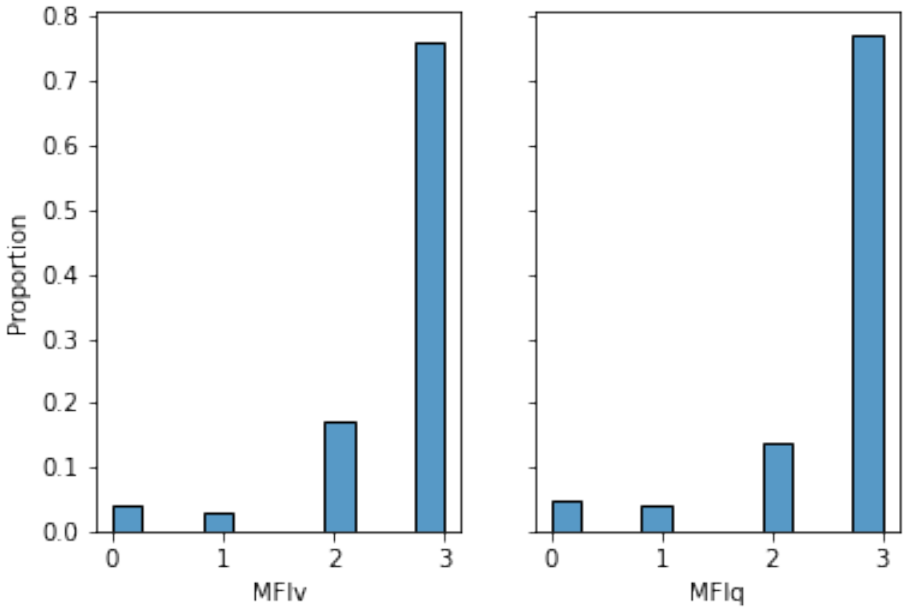


Figure 4: Class proportion for MFIv & MFIq, in the available data.

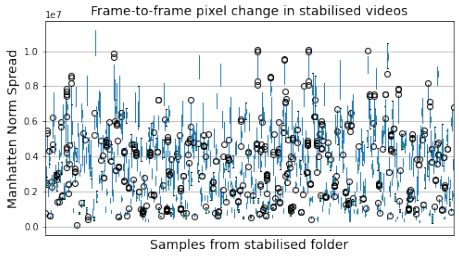
3.2 Video quality

Another challenge is that images are often unsteady due to probe movement by the clinician/user, as the probe is a handheld tool. This additional noise could hinder our analysis. We aimed to investigate whether a potential scoring mechanism of the video 'steadiness' could allow us to identify and remove the most unstable videos and increase the accuracy of our classification. The primary challenge in implementing this is that labels do not currently exist for 'unsteady' or 'steady' images. Nevertheless, our investigation of these approaches offers useful insight for future research.

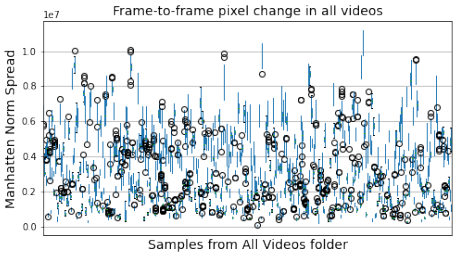
3.2.1 Approach 1: Pixel differencing

Multiple approaches were carried out. The first approach is a differencing method. A video comprises of n frames with width (w) and height (h). Each frame of a video was converted into a w x h numpy array of pixel values. The arrays were stacked into a single matrix per video, with final size (w x h x n). For each stack (video), we calculated the difference between each array (frame) and the previous array, using the the Manhattan norm (the sum of the absolute values) or zero norm (the number of elements not equal to zero) reflective of a change in pixels between successive frames. This resulted in a list of values of length n-1 for each video, representing the difference between each successive pair of frames. We hypothesised that unstable videos would have a larger mean change from frame to frame. We compared this metric in stabilised and unstabilised videos (Figure 5b and 5a).

The mean and standard deviation in each video has been calculated with distribution shown in Figure 6a. We then identified the videos showing the highest mean change in pixels (the top 3%) as shown in Figure 6b, and therefore in theory, large flux in image presented. We then looked at these images using the optical flow method described in the following section. A caveat of this approach is that it may not work as expected if the change in pixels due to shaking is on an equal level to the change in pixels due to blood flow. Indeed, using the optical flow method, it appeared that although a high proportion of these identified videos were also identified as unstable, videos showing clear changes in blood flow also appeared in



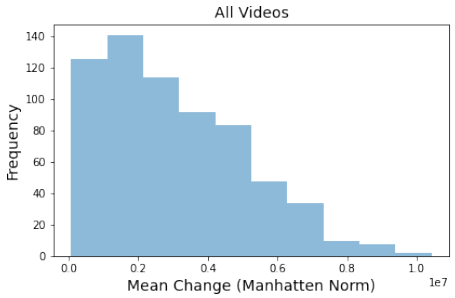
(a) Assessing all 'stabilised' videos provided.



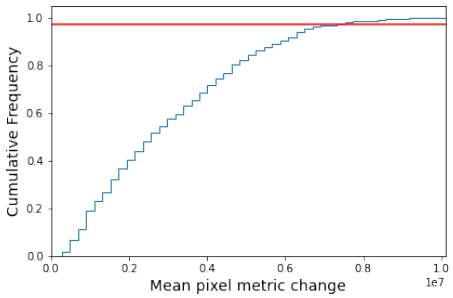
(b) Assessing all 'sorted' videos provided

Figure 5: Assessing change in image stability using pixel differencing and calculating Manhattan norm between consecutive frames.

this subset.



(a) Distribution of mean change from frame to frame for all sorted videos.



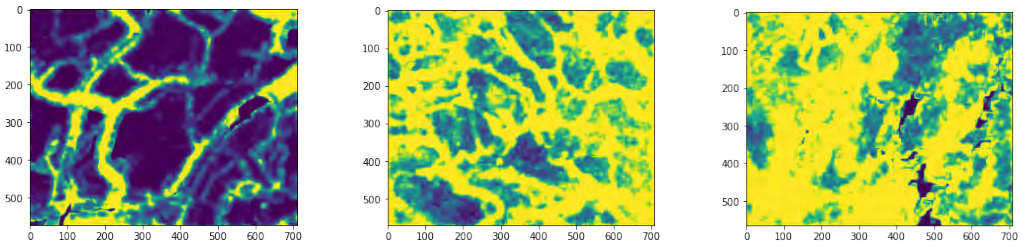
(b) Identifying top 3% of videos with a high pixel change.

Figure 6: Assessing frame to frame change in pixels for all videos.

3.2.2 Optical flow to determine video stability

An alternative method involves using optical flow to estimate the stability of each video, which may better identify unsteadiness rather than blood flow movement changes. A good quality video presents with minimal jitter or changes in brightness. When processed, a good quality video gives us a clear CLAHE image with fewer artefacts. When an image with few artefacts is used as a binary mask for an image, the optical flow tracking is localised to the vessel pathways. However, the presence of artefacts

causes the optical flow to be smeared across the whole image, as shown in Figure 7 below. These illustrations are originally grayscale images, but here they have been represented using contrasting colours to highlight their differences. All three images shown in the pictures below are for healthy patients with MFIv $= 3$. The first image clearly shows the vessel structures and the blood flow in them, while it becomes increasingly harder to discern the vessel structures in subsequent images. From preliminary experiments with the dataset, it was observed that when the video has a smear pattern that covers over 75% of the screen, the cytocam measurements are highly unstable videos. The next steps would be to produce and input dataset with the videos that become smeared removed, and assess whether this improves classification accuracy enough to warrant removal or avoidance of these types of videos in the future.



(a) Low Smear (b) Intermediate Smear (c) High Smear

Figure 7: Assessment of the video using calculation of optical flow and combining the lines indicating flow direction and magnitude into one image. A smeared image suggests an unsteady video.

3.3 Limitations and future implementation

A caveat in this approach of removing unsteady videos to increase the accuracy of subsequent analysis is that it reduces the already limited sample size. However, assessing whether the removal of unsteady videos improves analysis accuracy will, in practice, be crucial information to improve the performance of the medical device. For example, if the clinician can receive instant feedback on video quality they can be advised whether to take a new recording immediately. Therefore it will be

of great use to define at what point poor video quality results in insufficient accuracy when classifying.

In practice, a potentially quicker and simpler way to assess stabilisation would be to incorporate an inertial motion unit (IMU) into the cytocam probe. This IMU device is a sensor that is highly sensitive to changes in motion. By assessing pitch/roll etc that occurred throughout a video recording, the user could be supplied with immediate feedback on user motion. Within the research and development phase, correlation of IMU recorded movement with the stability of output images in terms of our metrics could help to determine IMU cutoff thresholds. This information will then be needed to assess which method would be most appropriate to use based on accuracy gain versus costs. This prior investigation of the effect of video quality on outcome accuracy is therefore an important step towards producing a cost and time-efficient method to improve video quality.

4 Data pre-processing

Multiple methods of pre-processing have been assesed before feeding the data into analysis frameworks.

4.1 Temporal gradient

To start with, a temporal gradient of consecutive frames and averaged over all the frames of a video has been considered. The temporal gradient of two consecutive frames seems informative however averaging over all the frame may not. Figure 8 depicts an example of such analysis. The black and white image on the left is a frame of video. The middle image is the temporal gradient of two randomly selected successive frames of the video. The image on the right is the result of averaging all the temporal gradients between successive frames of the whole video. The red region in the image is a region with high average pixel intensity at

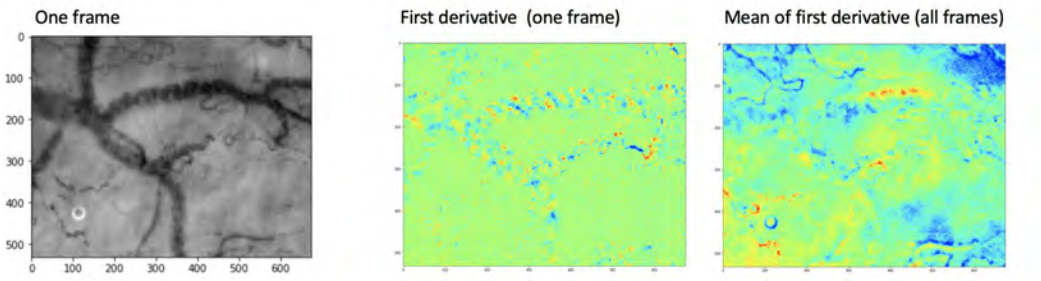
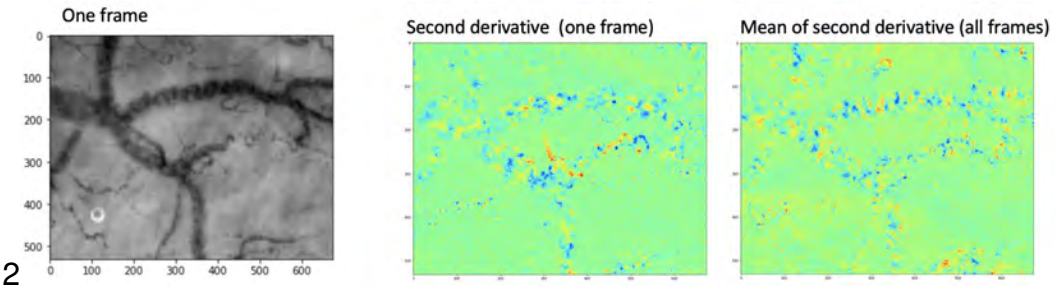


Figure 8: One frame with first temporal gradient and the average gradient over all consecutive frames.

initial frame and low average intensity in the next frame. On the other hand, a blue region in the image is a region with low average pixel intensity at initial frame and high average intensity in the next one. Taking the sum or the mean of temporal gradients does not make lots of intuitive sense because the blue and the red regions appear randomly and can cancel each other out and increase noise. However, looking at single temporal gradient frames, there might exist some interesting signals which a deep learning algorithm might be able to extract. The intuition here is that, the size and distances between the red and blue dots can

signal the strength and the velocity of the blood flow moving through the vessels. The size signals the width of the vessels and the distance between successive red/blue patches signal the flow. The hypothesis is that, on average, size and distance between different red/blue patches are different in sick vs. healthy patients. A deep learning or computer vision technique might be able to learn these patterns.

One upside of such an approach is the fact that each video with N frames results in N-1 temporal gradients (data points). This is in contrast to MAD method which transformed each video into a single frame (data point). This means that number of images for training is large. One possible non-deep learning approach is to use masking as masking only select the parts of the image which contain vessels. This can then be followed by analysing the size of red/blue dots and distances between them. One could also incorporate spatial Fourier analysis on the masked parts since Fourier transform gives the spatial frequency which can be translated into distance. Secondly, the second-order temporal gradient has been looked at where both the gradient of two consecutive frames as well as the average over all the frames of the video seem informative and able to capture the main visible vessels. Figure 9 depicts an example of such analysis. The black and white image on the left is a frame of video. The middle image is the second temporal gradient of three randomly selected successive frames of the video. The image on the right is the result of averaging all the second temporal gradients between successive frames of the video.



2

Figure 9: One frame with second temporal gradient and the average gradient over all consecutive frames.

4.2 Mean absolute difference and CLAHE

In this approach, an absolute temporal difference between successive frames was calculated. We then averaged all the absolute temporal differences between the frames to convert a video into a single image. If the videos are stable enough, we hypothesised that the mean absolute mean differences of a video should contain the fine-grained traces of the micro and macro-circulation patterns in the video. This means that for the images with low circulation, the traces of patterns should be less visible and distinguishable. Contrast limited adaptive histogram equalisation (CLAHE) was applied as a histogram equalisation method to remove unwanted differences in colour spaces between different videos. One disadvantage of this method is the fact that it turns the videos into single images which means that we have only a small number of data points after these operations. One possible way to solve this issue is to perform the aforementioned procedures on n consecutive frames instead of the whole video. This way, each video will give us several images.

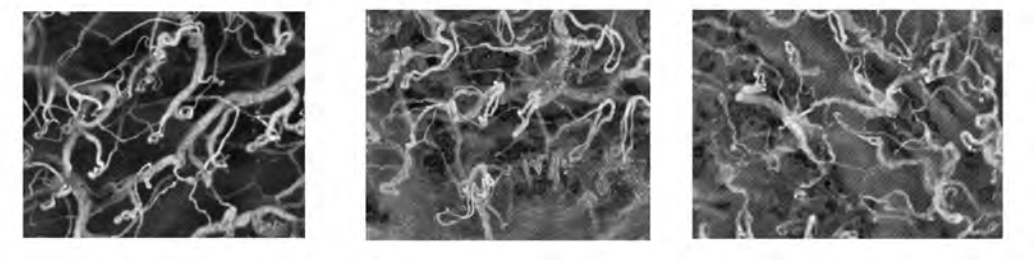


Figure 10: Results of mean absolute difference (MAD) followed by contrast limited adaptive histogram equalisation (CLAHE).

4.3 Temporal moving average

We have found evidence that where blood flows intermittently, a simple 10-frame average (corresponding to 0.5 seconds of a video) reconstructs the full paths of blood vessels while denoising the frames. Figure 11 compares an original frame (on the left) to the temporal moving average of ten consecutive frames - including the original frame (on the right). Our evidence, however, is limited to this specific example and we have not further used this approach in our study.

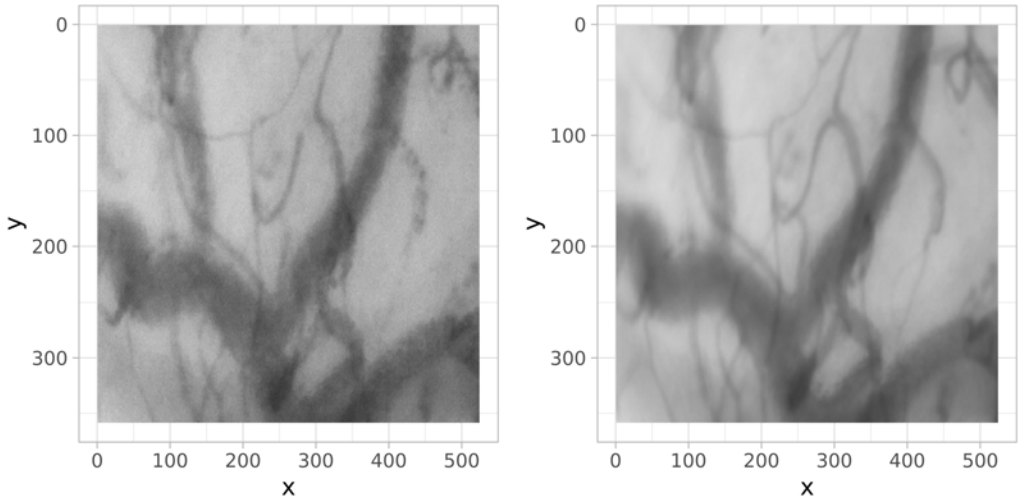


Figure 11: A single frame (left) and a 10-frame moving average (right). Where blood flows intermittently, the moving average seems to be able to reconstruct the full path of blood vessels.

4.4 Ridge detection

This method identifies the 'ridges' in the image, which are defined as curves whose pixel points are local maximum. The aim of this approach is to separate the main region of interest (ROI): micro-vessels from main vessels and their surrounding tissues as only the status of microcirculation will play a role in the classification problem. Specifically, in the first step, we find ridges on each input video frame via computing the eigenvalues of matrix of second-order derivative of an image, also known as hessian matrix. To increase the signal-to-noise (SNR) ratio of the initially segmented maps, we subsequently apply a morphological transformation, i.e. a dilation function to each segmented frame. Technically, dilation operation is where one will "expand" the edges of the image. The way these work is we work with a sliding kernel and we observe an impact of the kernel size for the final image quality in Figure 12. While one slide goes around, and if all of the pixels are black, then we get black, otherwise, we obtain a white output.

Furthermore, we threshold (T=130) the denoised frames and sum all frames up to get the segmented blood flow maps. Lastly, as shown in

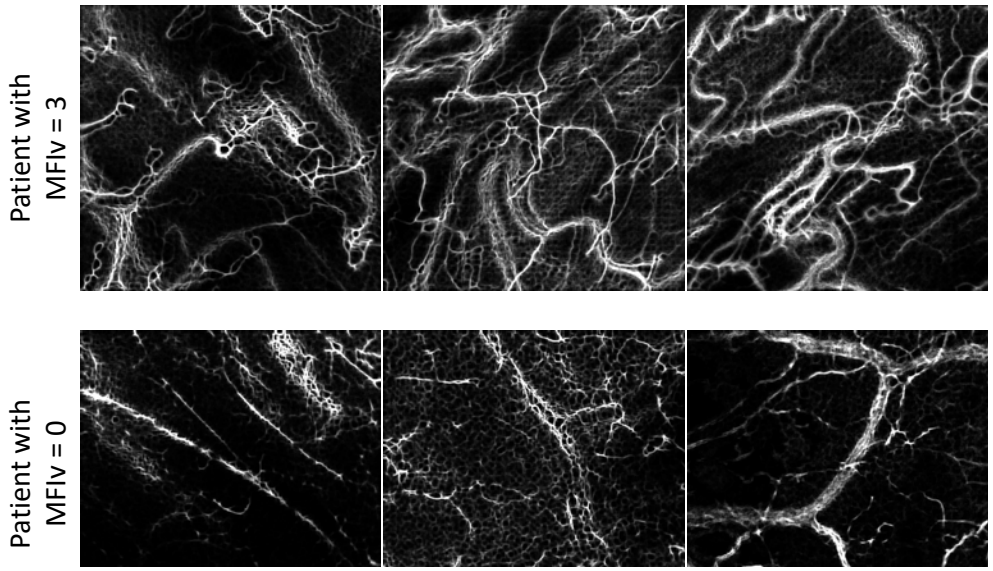


Figure 12: Extracted blood flow maps from input videos via ridge detection.

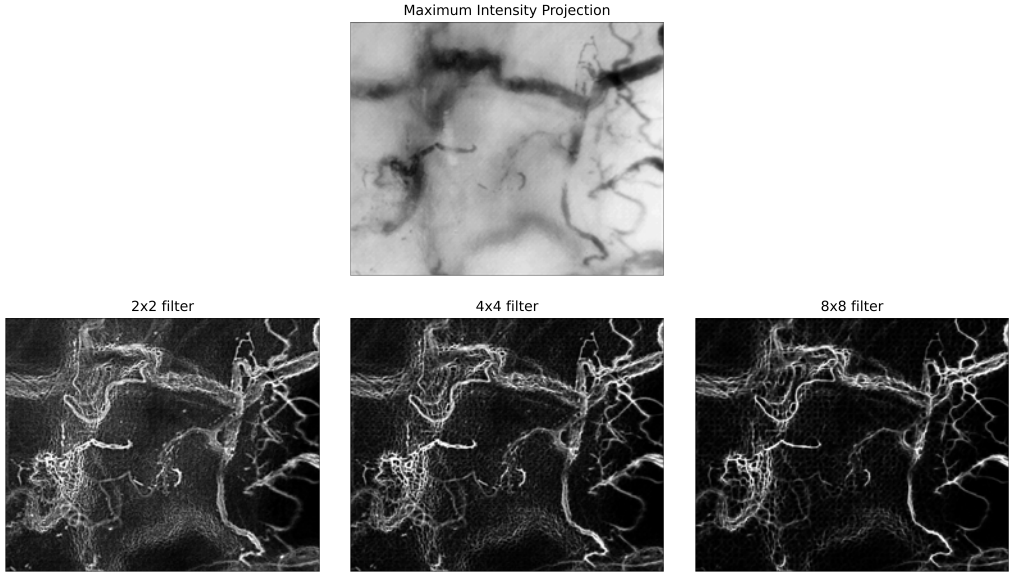


Figure 13: The denoising effect of morphological operation with change in kernel size.

Figure 12, to highlight the blood flow, we invert the intensity of segmented and therefore get a bright-field blood flow microscopy image from the input video. In Figure 12, we plot the segmented blood flow of 6 distinct patients (1st row: healthy and 2nd row: no flow). Within this study, we observe a significant impact of the input kernel size in the morphological transformation function. We visualised this effect in Figure 13, the denoising effect as well as the lost of fine details of the map become stronger when we increase the size of kernel in morphological operations.

4.5 Optical flow

The main intuition behind this approach was to understand if the blood flow between successive frames can be represented in one image, where the intensity of each pixel represents how much blood has flowed through that point in the image. A higher pixel intensity would indicate a higher blood flow through that area, and a lower intensity would represent little to no blood flow. However, this is not very straightforward as the pixel intensities in each image are affected by brightness, video stability and relative movement between the measuring instrument and the patient. To address these problems we use a three-fold approach - first the CLAHE image for each video is converted into a binary image where the vessel pathways have a value of 1 and the rest get a value of 0. This image is then used as a binary mask for each frame in the denoised video so that only the motion of pixels in the capillaries is retained. Finally, the optical flow between successive frames is calculated using the Gunnar Farneback algorithm[4]. All these frames are then blended to produce one final output image that represents the optical flow across the span of the entire video. This entire process is depicted in Figure 14.

The output is a gray scale image, with brighter points indicating a higher change in pixel intensity over time, and darker pixels indicating little to no changes in pixel intensity. A salient and contiguous cluster of pixels intensities resembling the blood flow across the vessel structures are observed in the image for MFIv $= 3$. However, this is not the case for the image with MFIv $= 0$ as there are no salient pixel clusters representing a flow of blood across the vessel. This conforms with the intuition that higher optical flow must be observed in cases where blood flow is

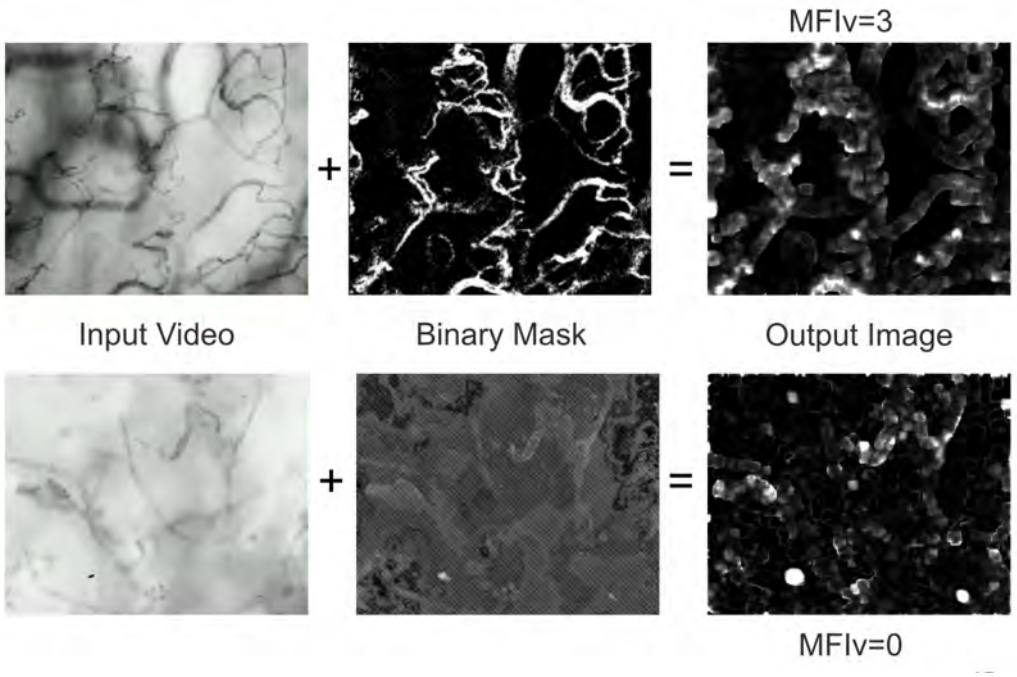


Figure 14: The optical flow calculation results from vessels by using binary mask corresponding to the blood vessels.

contiguous. In theory, the output image of MFIv $=$ 3 should be 0, however there is still a scattered distribution of pixels in the output image. This is due to the excessive fluctuations in brightness, jitter and air bubbles present that result in noisy measurement. Given more time, better noise filtering and adaptive video masking techniques can be further explored address this problem.

5 Experiments and results

This section describes our various set of experiments on data preparation, analysis by a simple image processing based approach, dimensionality reduction, and analysis by deep learning based approaches. The experiments in this work were implemented in Python utilising various relevant data analysis libraries including NumPy, Matplotlib, OpenCV, TensorFlow, Keras, and Pandas. The experiments were run remotely on University of Birmingham's provided Baskerville system (https://www.baskerville.ac.uk/).

5.1 Data preparation

5.1.1 Train, validation & test splits

After data cleaning (as described in Sections 2.1 & 3.1), the overall dataset size was 666. This data has been split into train, validation and testing with splits of 70% (458), 15% (104) and 15% (104) respectively. When splitting, patients were stratified so as to avoid data leakage. For example, Patient 1's videos were only in the training dataset, whilst Patient 2's might have only been in the validation dataset, with no videos in other splits. The results provided in the following sections have consistent train, validation and test datasets to facilitate fair comparison. Class distributions for each split are shown in Figure 15 and show that whilst class imbalance remains across all splits, each split does contain members of each class.

5.1.2 Class weights

Given the significant class imbalance previously described in the dataset, class weights were calculated and implemented in the neural network classifiers (Sections 5.4.1, 5.4.2 and 5.4.3) to minimise the impact of class imbalance.

Class weights ω_i were calculated for each class i according to the label distribution on the training set using the formula shown in Equation 1 where P_i represents the number of positive cases of class i, N_i represents the number of negative cases of class i, and n_c is the number of classes.

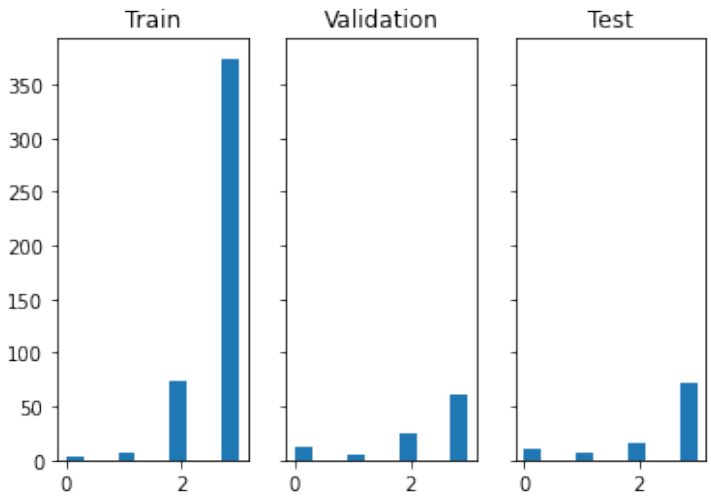


Figure 15: Class distribution per training, validation & test splits.

Calculated class weights are shown in Table 1.

$$\omega_i = \frac{N_i + P_i}{n_c \times P_i}, \tag{1}$$

Table 1: Class weights, estimated as per class distribution in training data, for use in network learning process.

Class	**Weights**
0	31.58
1	18.95
2	1.53
3	0.31

5.1.3 Data augmentation

Data augmentation is the practice of increasing the sample size of a dataset through the use of transformations such as translations, rotation, and zooming and it is commonly used in computer vision literature, including in medical imaging [7]. Within the context of microvascular imaging using DFM, augmentation performed using translation or rotation

is considered analogous to the natural changes observed in acquiring subsequent images due to factors such as variation in probe-tongue positioning. Importantly, transformations such as zooming would not be plausible given the fixed focal length of the DFM device.

The following data augmentation approaches were applied:
Filtering For machine learning techniques that analyse the flow of pixels such as optical flow, it is vital to have videos with reduced noise. To prepare videos for that challenge they are disassembled into frames, denoised and assembled back to a video. The denoising is adjusted to appropriately remove noise without removing image details. The number of surrounding images to use for the target image is evaluated to be three.
Rotation & flipping This technique included rotating the videos clock- or counter-clockwise as well as flipping it horizontally or vertically. In this way, more input data was generated.

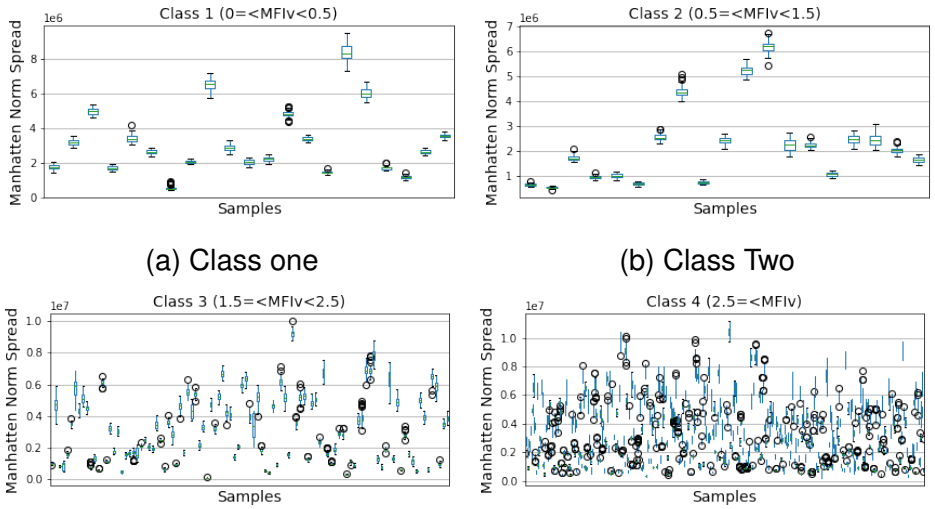
5.2 Image processing based methods

5.2.1 Pixel change

As part of this work, we assumed that the complexity of the problem requires deep learning methods to solve. However, we also assessed whether it is possible to distinguish classes based on changes in pixel intensity and in pixel movements. Using the methods described in 3.2.1, we assessed whether change in pixels from frame to frame varied more or less within the different classes. Figure 16 highlights the large difference in the number of samples within each class but ANOVA (with unequal sample number) shows that there is not a significant difference between classes ($p=0.245$) (see Figure 17). We also looked at temporal changes (Figure 18) but no clear pattern was identified.

5.2.2 Pixel intensity change

Here, we are providing a simple analysis of pixel intensity. We are investigating whether a simple approach can inform video classification. We answer two questions. First, do pixels' intensity changes from one frame to the next in the original videos correlate to the MFIv value? Second, does the pixel intensity of single images obtained through the



(a) Class one

(b) Class Two

(c) Class 3

(d) Class 4

Figure 16: Mean difference between frames divided by MFIv values (v) with classes 0,1,2 and 3 defined as $0>v<0.5$, $0.5>=v<1.5$, $1.5<=v<2.5$ and $v>2.5$ respectively.

augmentation of the original videos, such as through segmentation and optical flow analysis, correlate to the MFIv values?

Mean pixel intensity change. From the analysis of all videos provided, we observe no correlation between the average pixel intensity change (from one frame to the next) and the MFIv values. Furthermore, the average pixel intensity change is equal to 0. This implies that, on average, any pixel intensity increase will be offset by a pixel intensity decrease (consider the large amount of pixels involved: each frame has roughly 500 x 500 = 250,000 pixels and each video has at least 50 frames, totalling to at least 12,500,000 pixel intensity variations measured in each video).

Brief conclusion We find that a simple statistical analysis of frames yields no-correlation relationship between pixel changes from one frame to the next and the MFIv values in a video.

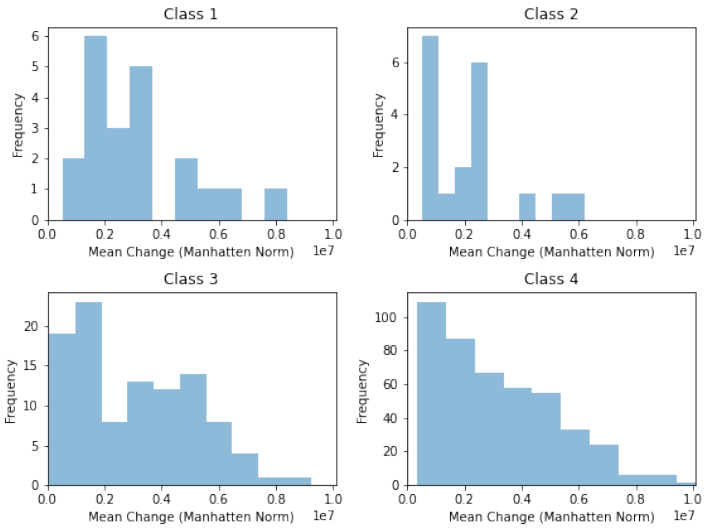


Figure 17: Mean difference between frames divided by MFIv values (v) with classes 0,1,2 and 3 defined as 0>v<0.5, 0.5>=v<1.5, 1.5<=v<2.5 and v>2.5 respectively. ANOVA between means: p= 0.245

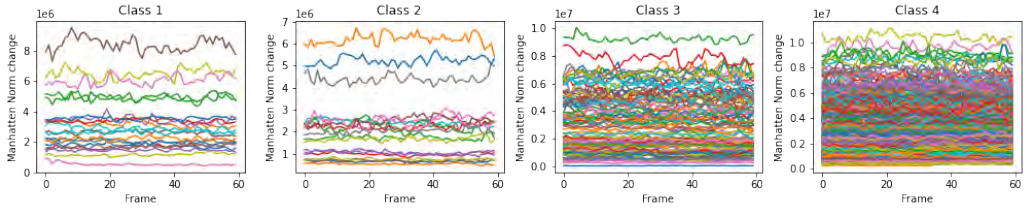


Figure 18: Manhattan Norm difference from frame to frame divided by MFIv values (v) with classes 0,1,2 and 3 defined as 0>v<0.5, 0.5>=v<1.5, 1.5<=v<2.5 and v>2.5 respectively.

5.3 Dimensionality reduction based methods

Dimensionality reduction algorithms seek to reduce the number of dimensions of data by performing transformations to a lower-dimensional space whilst retaining as much of the variance in the data as possible. As a result, they have a longstanding history of use for visualising, clustering and modelling high-dimensional data [1]. To explore the utility of dimensionality reduction on our dataset, three linear dimensionality

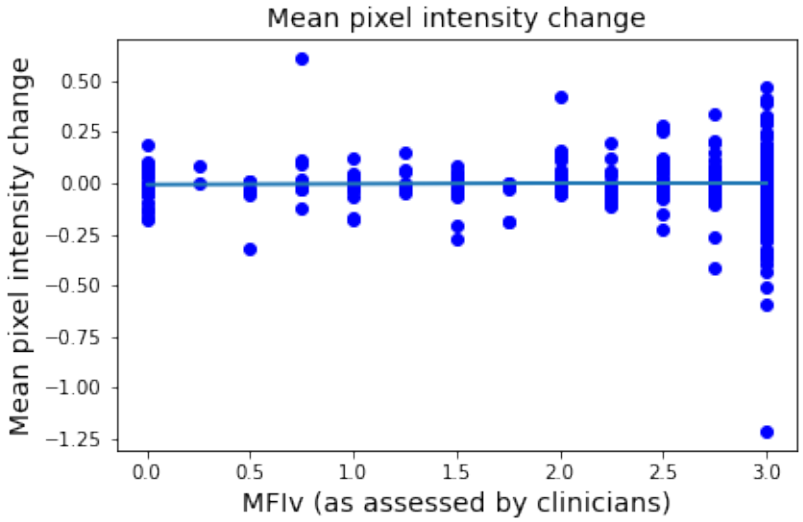


Figure 19: Mean pixel intensity change obtained by first a) generating new frames by subtracting the value of pixels of consecutive frames and then b) for each video, taking the mean value of all new frames.

reduction algorithms were explored: principal component analysis (PCA), linear discriminant analysis (LDA) and neighborhood components analysis (NCA).

The inputs consisted of a $(50, 176)$ dimensional vector for each video, derived from a $(224, 224)$ centre-crop taken from the 2D CLAHE transformation described in Section 4.2. The dataset was then scaled to standard mean and unit variance before each algorithm was fitted on the training set. The fitted model was then used to transform the training set to a lower-dimensional space. This transformed training data was plotted to visualise the degree of class separation and used to train a k-nearest neigbhours (KNN) classifier which was then evaluated on a held-out test set that underwent the same transformations learned in the fitting stage. For each dimensionality reduction algorithm, the top two and three components were visualised and evaluated with a KNN classifier (for $k = 1, 3, 5$) with mean accuracy reported on the held-out test set. As seen in Figure 20, none of the methods produced a clear visual separation between classes using either the first two or three components. For all dimension reduction methods, the greatest mean accuracy from the KNN

classifier was reported with $k = 5$. Using the top two components PCA was the best performing method with a mean test accuracy of 62%, despite providing perhaps the lowest visual separation, although this assessment is subjective. Suprisingly, the addition of the third component worsened classification performance for PCA (60% vs 62%), however, it improved predictive performance for both LDA and NCA with LDA reporting the highest mean test accuracy of 63%.

5.4 Deep learning based methods

In the following sub-sections, we describe various experiments and their results with deep learning based approaches.

5.4.1 2D-CNN with absolute temporal gradient image

The pre-processing in Section 4.2 converted the video frames into a single 2D image, aiming to encapsulate the temporal behaviour of the blood flow whilst doing so. These images can then be fed into a CNN architecture (Figure 21), which aims to extract the spatial information in the images to classify the MFIv into either classes 0, 1, 2 or 3.

Within this experiment, various combinations of pre-trained models for transfer learning were tested. The class imbalance treatment implemented was class weights, as described in Section 5.1.2. Data augmentation was also applied during training with horizontal flipping, vertical flipping and random rotations with a range of 10 degrees. These were chosen as they make sense in a clinical and microscopy setting. Zooming was not chosen as an augmentation method as the dark field microscope has a fixed focal length, and contrast methods were not chosen as they may distort the images in a non-realistic manner. Time restrictions prevented testing whether additional augmentation strategies would have benefited model performance, and also testing further class imbalance treatments.

The chosen pre-trained models were VGG16, InceptionV3, ResNet101 and EfficientNetB7 with default image input sizes [19]. The top prediction layers of the pre-trained models were discarded, and appended was a global average pooling layer followed by a softmax layer, which

normalises the outputs into a probability distribution between 0 and 1, with the output class then being assigned to the maximum probability. Initially, the pre-trained layer weights were frozen, and the model was fit using the Adam optimiser with a default learning rate of 10^{-3}, with the epochs set to 200 with early stopping to avoid overfitting. Fine-tuning was then carried out after initial convergence, by unfreezing some layers in the pre-trained model and re-training the model using a smaller learning rate of 10^{-5}. For VGG, the entire base model was unfrozen. For InceptionV3, the last 2 convolution blocks were unfrozen. For EfficientNet, block 7 was unfrozen. For ResNet101, the last convolutional block was unfrozen. The choice of layer unfreezing was based on computational considerations as well as prior knowledge of each model.

To illustrate the importance of transfer learning with little data, VGG16 was trained first from scratch and then using the pre-trained weights, without the use of class weighting. Figure 22 compares the loss of the two initialisations. The loss with random weights shows erratic learning with no convergence, whilst the VGG loss with ImageNet weights shows clear learning and convergence after approximately 20 epochs. Figure 23 shows further evidence of the benefit of using pre-trained weights, with the confusion matrix for random weights showing that the model simply predicts the majority class at each test instance, whilst with ImageNet initialisation the model shows learning of all classes.

Table 2 shows the results from the 2D CNN with the pre-processed images. Precision and recall for each class are given, with average accuracy and the average weighted F1 score across all classes. The average accuracy needs to be taken into consideration with the distribution of the majority class in the test set, as simply predicting an MFI value of 3 each time gives an accuracy of 68.3%. The method column refers to whether fine-tuning, class weights or both were applied to the model.

It is evident that not using class weights to combat class imbalance produces good precision and recall for the majority of class 3 in comparison to other classes, but is inept at identifying poor blood flows. It doesn't however beat the simple baseline of 68%. Using class weights results in much better recall and precision for the minority classes, except ResNet101, which still has poor performance across all minority classes. Interestingly, fine-tuning did not provide any significant performance gain

Table 2: Results of various popular 2D-CNN experiments using pre-processed images as their input.

Class				0		1		2		3	
Model	**Method**	**Accuracy**	**F1**	**Prec**	**Rec**	**Prec**	**Rec**	**Prec**	**Rec**	**Prec**	**Rec**
VGG16	None	0.65	0.34	1	0.18	0	0	0.25	0.27	0.73	0.87
VGG16	Fine-tune	0.60	0.31	0.33	0.09	0	0	0.26	0.47	0.76	0.76
VGG16	Class weights	0.62	0.43	0.7	0.64	0	0	0.23	0.40	0.78	0.72
VGG16	Class weights + Fine-tune	0.50	0.31	0.50	0.18	0	0	0.21	0.40	0.72	0.70
InceptionV3	None	0.65	0.25	0	0	0	0	0.21	0.20	0.73	0.92
InceptionV3	Fine-tune	0.66	0.28	0	0	0	0	0.29	0.33	0.74	0.90
InceptionV3	Class weights	0.66	0.44	0.70	0.64	0	0	0.24	0.33	0.79	0.80
InceptionV3	Class weights + Fine-tune	0.63	0.45	0.55	0.55	0.33	0.14	0.20	0.23	0.79	0.76
EfficentNetB7	None	0.65	0.32	1	0.09	0.17	0	0.31	0.33	0.73	0.87
EfficientNetB7	Fine-tune	0.66	0.30	1	0.09	0	0	0.27	0.20	0.72	0.90
EfficientNetB7	Class weights	0.62	0.48	0.47	0.64	0.40	0.29	0.33	0.21	0.89	0.69
EfficientNetB7	Class weights + Fine-tune	0.63	0.51	0.54	0.64	0.40	0.29	0.21	0.40	0.86	0.72
ResNet101	None	0.61	0.25	0	0	0	0	0.19	0.27	0.72	0.83
ResNet101	Fine-tune	0.61	0.26	0	0	0	0	0.21	0.40	0.75	0.80
ResNet101	Class weights	0.57	0.26	0	0	0	0	0.20	0.40	0.80	0.75
ResNet101	Class weights + Fine-tune	0.56	026	0	0	0	0	0.21	0.40	0.80	0.72

across all models. This is likely due to the model not being capable of learning very task-specific features with the small dataset provided. Further work could investigate this fine-tuning process, for example, the number of layers to freeze and the learning rates used. The fine-tuning process is likely to be important as the dataset size increases, with the perfusion data differing from the ImageNet data and therefore requiring different features to be learned.

The prediction for a model with the best average F1 score over all classes was EfficientNetB7 with a score of 0.51. It is assumed by the authors this is because EfficientNetB7 accepts an image size of $(600, 600)$, which is much closer to the large image sizes in the perfusion data, whereas the other models accept smaller image sizes (e.g. $(224, 224)$). It is possible that scaling down the images results in a loss of information, but a range of image size inputs across models would need to be trialled to confirm this. The confusion matrix for this model for all classes is given in Fig. 24. The recall and F1 for MFI 0 & 3 are better than classes 1 & 2, possibly due to the distribution assumption set using class weights. Another possible reasoning in the worse prediction of MFI 1 & 2 is that clinicians have less certainty in scoring these MFI scores, resulting in variance across patients and therefore additional model noise. The results obtained by this classification are much better than randomly choosing

any class, suggesting that the CNN is appropriately learning features from the pre-processed images that it can then use to correctly classify the perfusion index. Additional data, particularly of lower MFI patients, could result in a much better classification across all classes. Although the clinician metrics are unknown, it is assumed that clinicians perform much better than this model in predicting the perfusion index so further work is required.

5.4.2 2D-CNN with segmented blood flow maps

The motivation behind this 2-step classification after segmentation approach is largely becuase only small vessels are relevant to microcirculation classification. Giant vessels and their surrounding tissue play a minor or no role in the classification problem. The workflow of this approach is depicted in Figure 25a: we perform a ridge-based vessel detection followed by a sum operation along the time dimension of all frames in an input video for both micro-vessel ROI localisation and dimensionality reduction. Currently, we trained VGG-16 on top of the segmented blood flow maps. The challenges we have are the training set is rather small and the label distribution is extremely imbalanced. To address the problem of the limited training set, we fine tuned our classification network on the pre-trained weights from ImageNet classification. Additionally, we performed different kinds of data augmentation techniques such as random cropping, affine transformation, random contrast, and random zoom in/out. To address the class imbalance problem, we used class weighting as described earlier.

According to patient-wise stratification, we split the initial dataset into training (458 videos), validation (104 videos) and test sets (104 videos). The result of our 2D CNN trained with segmented blood flow maps is shown in Figure 25 b and c. Although the class weight addresses the class imbalance problem, we still observe the CNN model was biased and tend to vote for the majority class (MFIv=3) frequently (see Figure 25 c).

5.4.3 2D-CNN + RNN

To harness both 2D spatial information from each frame, and temporal data from the image sequence, a hybrid neural network architecture was constructed using a CNN and recurrent neural network (RNN) for spatial and temporal information respectively. To automatically extract features from each 2D frame within a video, we used transfer learning with the InceptionV3 CNN [18] architecture, that has been pre-trained on the ImageNet dataset [3]. This is similar to the approaches outlined in Sections 5.4.1 & 5.4.2 however differs in being performed on a frame-by-frame basis, rather than on images derived from averaging across frames in a video.

Despite ImageNet being a generic imaging dataset, rather than medical imaging specific, pre-training models on it have been shown to confer enhanced performance when applied to medical imaging classification problems, such as chest radiographs [15], giving rise to the popularity of this form of transfer learning within the medical imaging literature. Therefore we can reasonably expect such an approach to be beneficial, even with a highly specialised form of an image such as those produced by DFM.

In a typical image classification application, as shown in Sections 5.4.1 & 5.4.2, the top layer of InceptionV3 would be trained to predict the desired classes using a softmax activation function to produce an n-dimensional output, where n is the number of classes and represents a valid probability distribution. In the case of the CNN-RNN hybrid architecture, we remove the top layer and output a 2048-dimensional latent representation of the input image, with the hypothesis that this may confer information useful for classification when input to the sequence model. The RNN architecture consists of two Gated Recurrent Unit (GRU) layers and a final fully connected layer with softmax activation function to produce class probabilities. In simple terms, the GRU can be understood to represent a form of selective memory that learns which features from the sequence should be retained and used for prediction via minimising a loss function of categorical cross-entropy with the Adam optimisier. Similar to previous experiments the same training, validation & test splits and class weights are used.

Three experiments were performed, the first using a sequence of 20 frames and raw frames as the inputs, the second using a sequence of 100 frames and the third using 20 frames with images that have undergone ridge detection as described in Section 4.4. Results for these three experiments are shown in Table 3 with the confusion matrices and learning curves shown in Figure 26.

Table 3: Results of CNN + RNN, which aims to exploit spatial and temporal information.

Class:	**Avg.**		**0**		**1**		**2**		**3**	
Input	**Acc**	**F1**	**Prec**	**Rec**	**Prec**	**Rec**	**Prec**	**Rec**	**Prec**	**Rec**
20 frames, Raw	0.69	0.61	0.60	0.27	0	0	0.50	0.07	0.70	0.96
100 frames, Raw	0.65	0.59	0	0	0.50	0.14	0.15	0.13	0.73	0.92
20 frames, Ridge	0.69	0.58	0	0	0	0	0.50	0.07	0.70	1

We can see that results are broadly similar for all modelling approaches, with a fractionally higher weighted F1 score for the 20 frame sequences with raw images. Despite using class weights in an attempt to mitigate the class imbalance, we can see that the classifier is still highly biased towards prediction of the majority class. Interestingly the images processed with ridge detection displayed the worst performance, despite appearing easier to identify on visual inspection. This may be due to the fact that pre-processing results in a loss of information and deep learning algorithms typically perform well on complex raw data from which they are able to extract meaningful features.

When considering the reported results, it is important to note that owing to time & computational constraints, multiple model fits were not performed, and thus prediction uncertainty and reliability cannot be quantified. Anecdotal experience during the model construction process would suggest that the current models exhibit high variability between fits and therefore the reported metrics shown in Table 3 may not be consistent.

A final approach of using the weights learned through training in Section 5.4.1 in the CNN, in attempt to extract more meaningful features, was tested and whilst the results are not formally documented they did not show a performance gain over those shown in Table 3.

5.4.4 2D-CNN + RNN + data augmentation

The 2D CNN with RNN is a promising approach for the challenge. Using the augmentation technique described in section 5.4.4, the distribution from Figure 29 is generated. Apart from the class with a MFIq value of 1, the distribution seems more balanced, having a threshold of 1.8. To generate more data, the quadrant videos are again augmented by rotating and flipping, obtaining a set of circa 1500 quadrant videos. The results are summarised in Figure 27. Looking at the results from this study it becomes obvious that the algorithm tends to predict an MFI of 3 for every video on bigger datasets. It emphasises the challenge of clearly identifying no-flow videos.

Cropping To get the most out of the provided data another augmentation technique applied is cropping. Since a MFIq score is given for each

quadrant a much better-balanced dataset can be created by cropping each video and assigning the individual MFIq score.

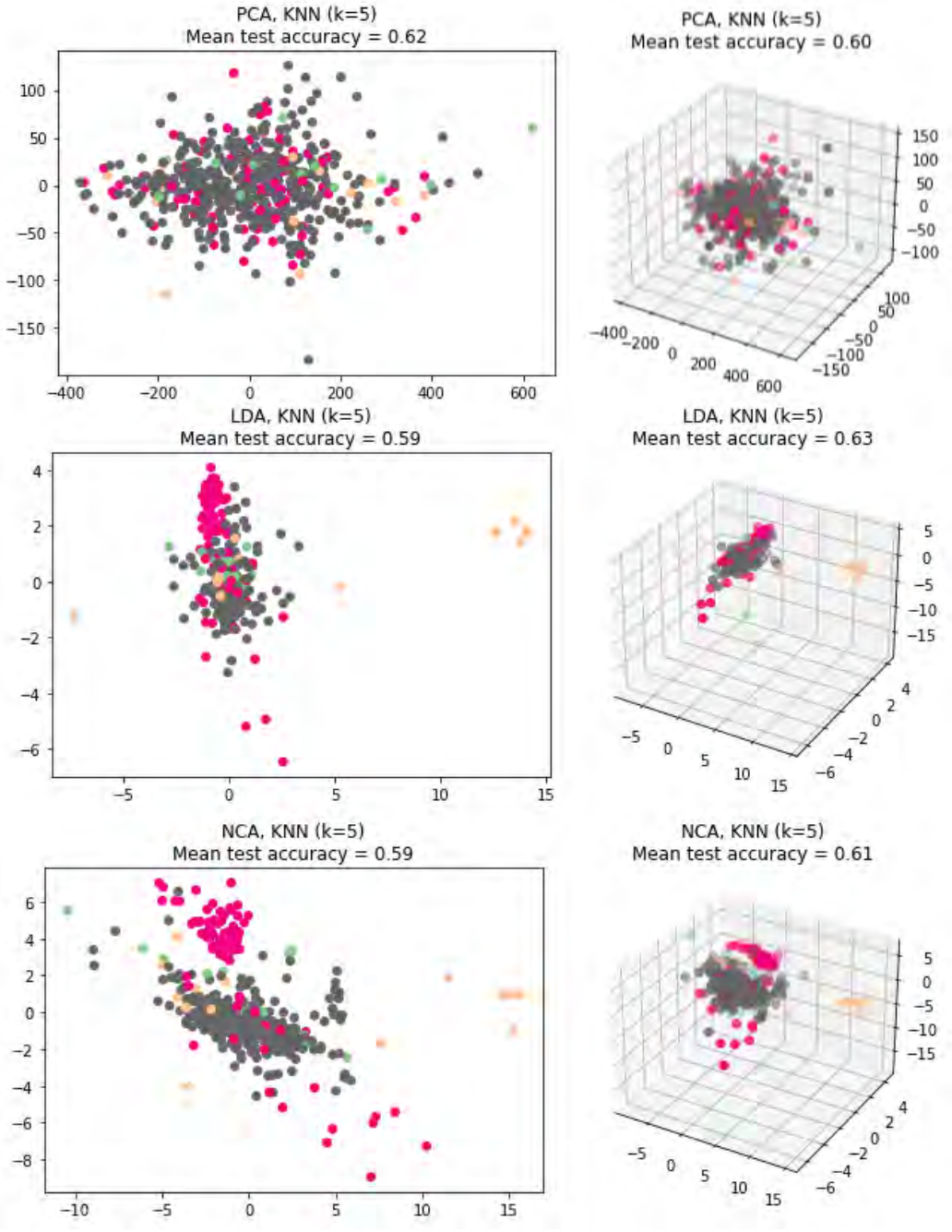


Figure 20: Dimensionality reduction approaches, exploring PCA, LDA, and NCA with 2D and 3D visualisation and the subsequent classification performance by k-NN classifier.

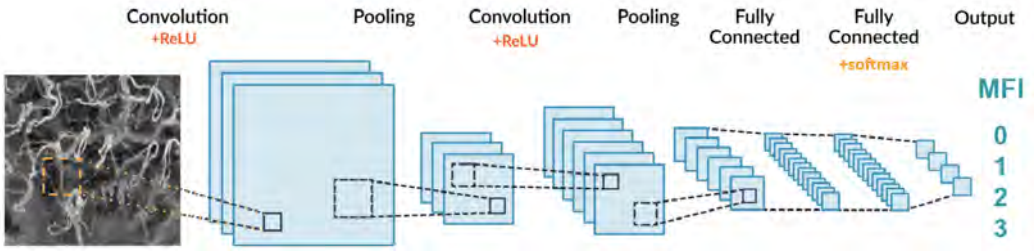
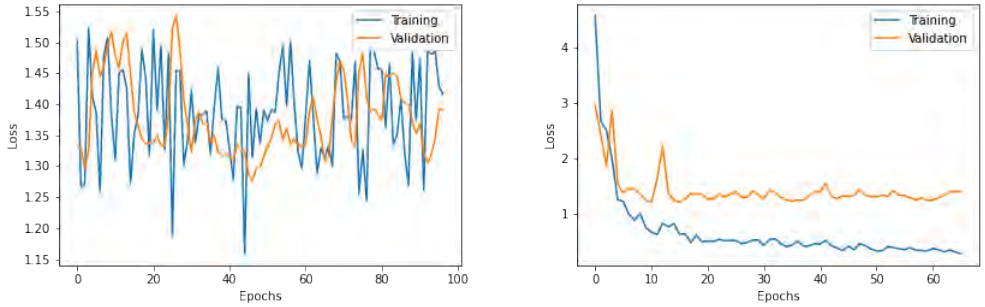
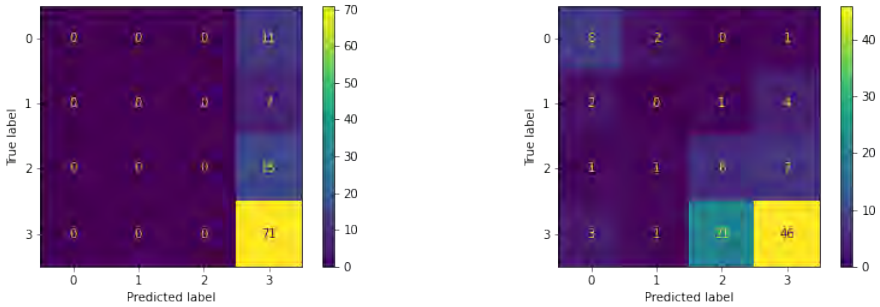


Figure 21: Illustration of CNN processing scheme. (modified from *https://medium.com/analytics-vidhya/understanding-convolution-operations-in-cnn-1914045816d4*)



(a) VGG with random weights (b) VGG with ImageNet weights

Figure 22: VGG network training loss with random and ImageNet weights initialisation.



(a) VGG with random weights (b) VGG with ImageNet weights

Figure 23: VGG confusion matrix with random and ImageNet weights initialisation.

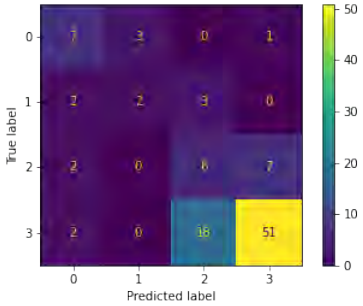
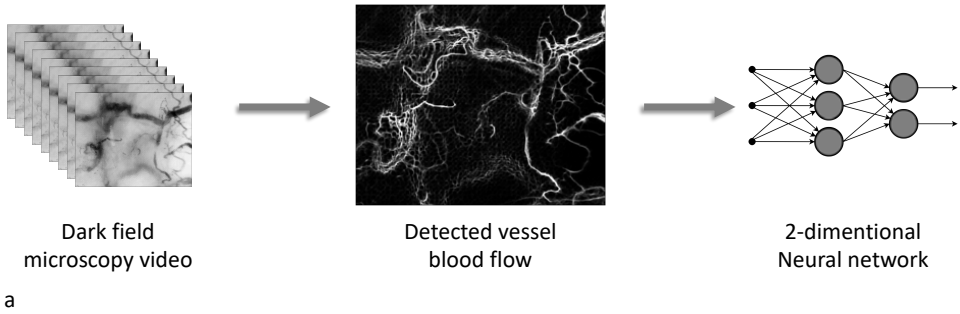
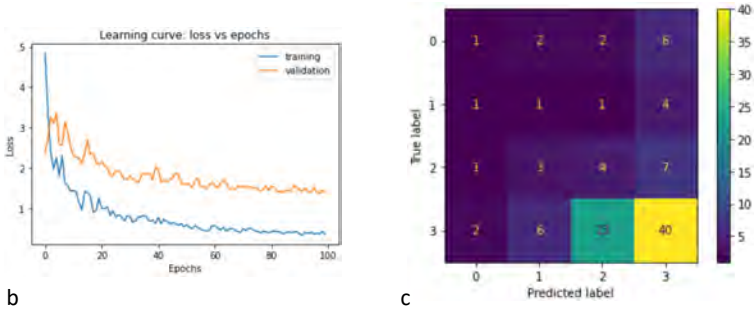


Figure 24: Confusion matrix for EfficientNetB7 with class weights and fine-tuning.



a



b

c

Figure 25: (a) the workflow of the proposed 2D CNN with segmented blood flow maps as input. (b) the loss curve of fine-tuning a VGG-16 model pre-trained on ImageNet dataset. (c) we plot the confusion matrix to visualise the multi-class classification performance.

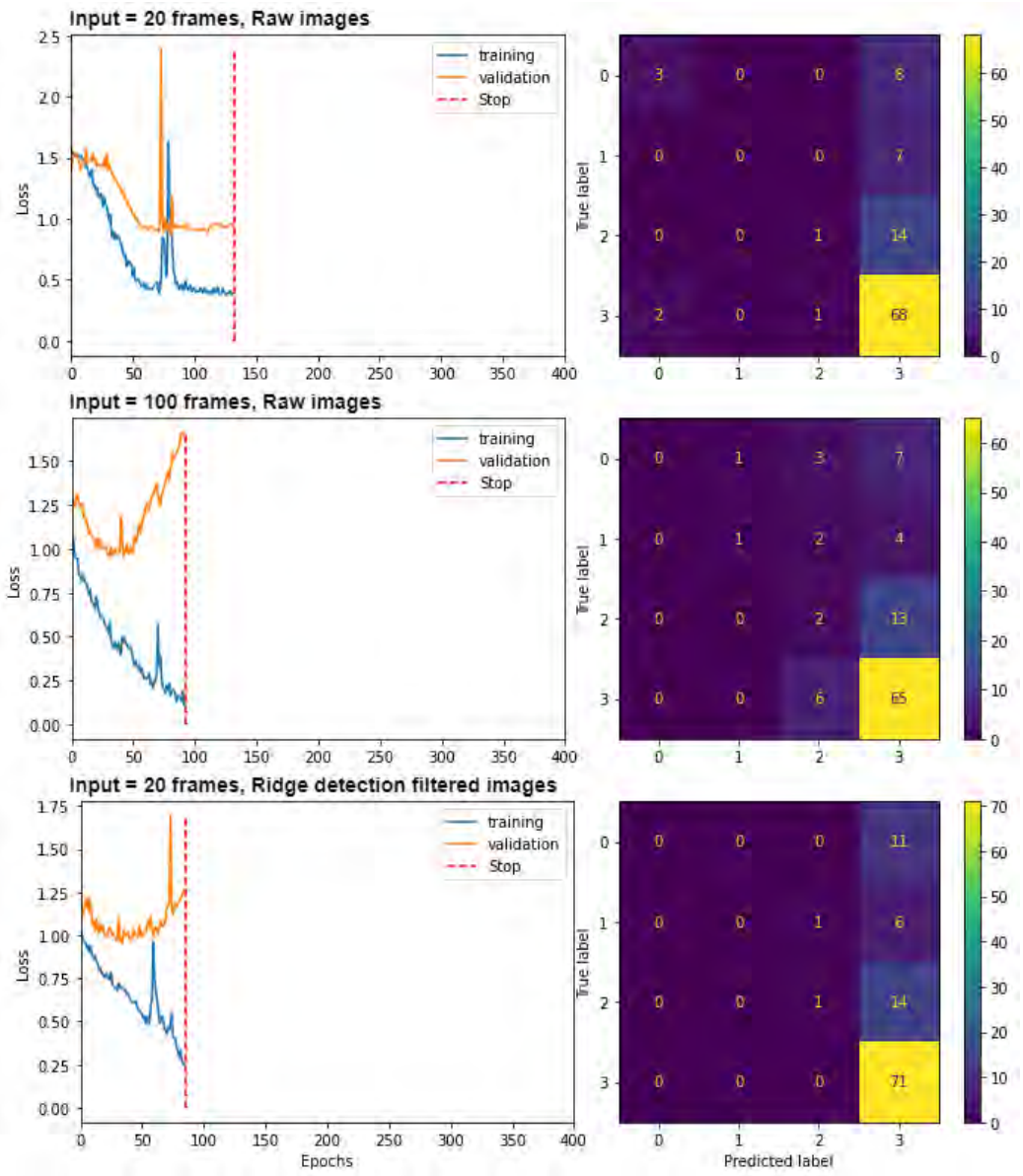


Figure 26: Results of CNN+RNN experiments, which aim to exploit both spatial and temporal information.

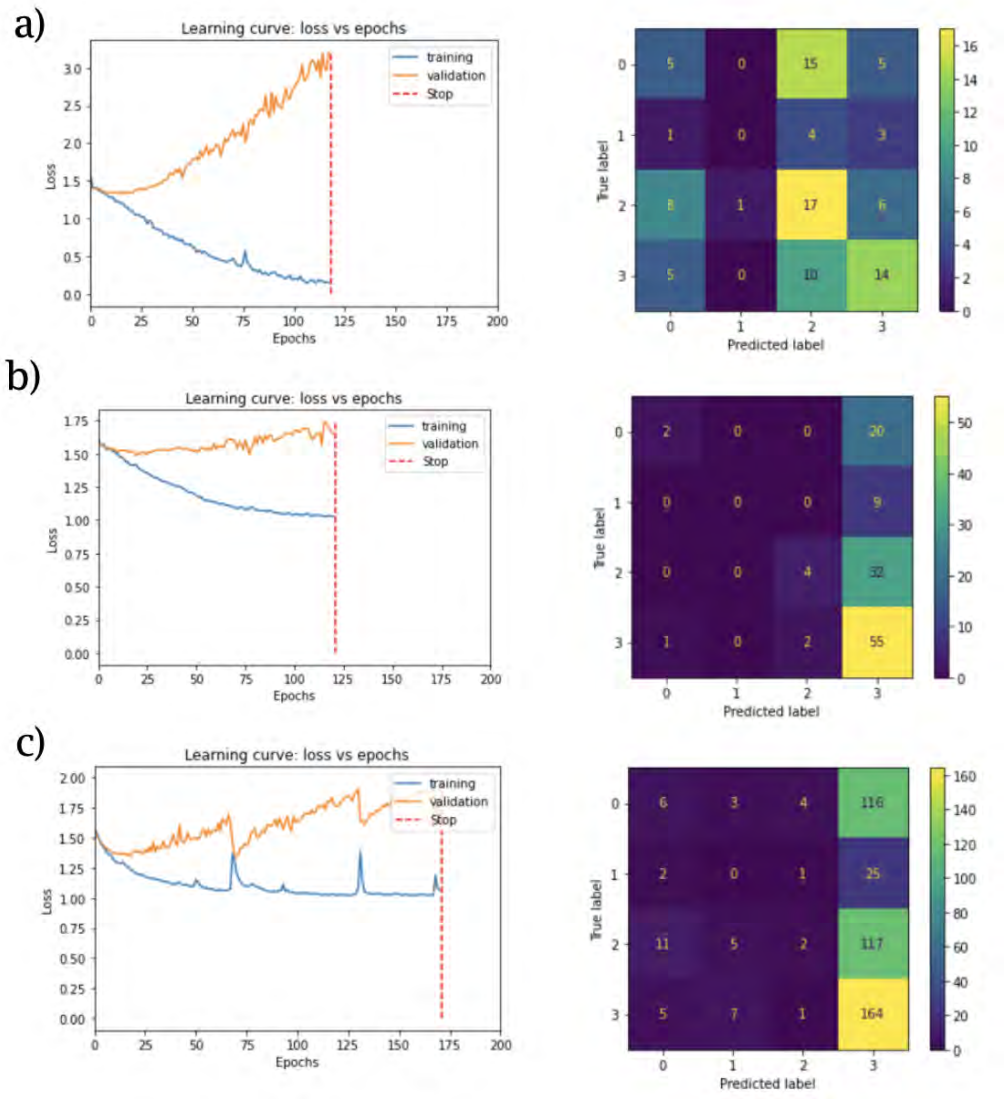


Figure 27: The result of applying the CNN + RNN on augmented data: a) using only the quadrant videos, b) using a subset of all augmented quadrant videos (including rotating and flipping) and c) using the all augmented videos (including rotating and flipping).

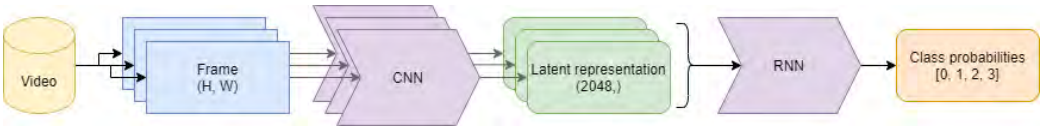


Figure 28: Illustration of CNN + RNN model design, note that this network aims to exploit both spatial and temporal information.

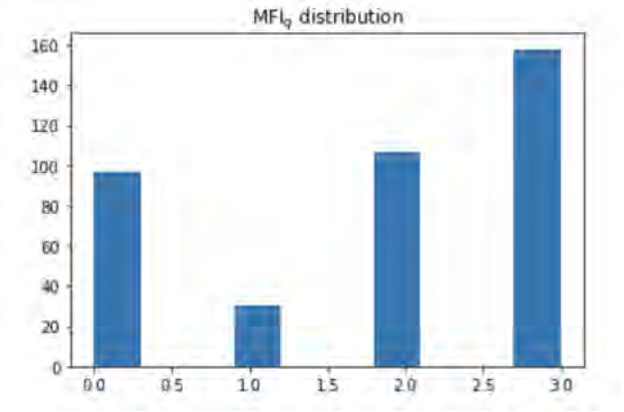


Figure 29: MFIq distribution from the quadrant cropping.

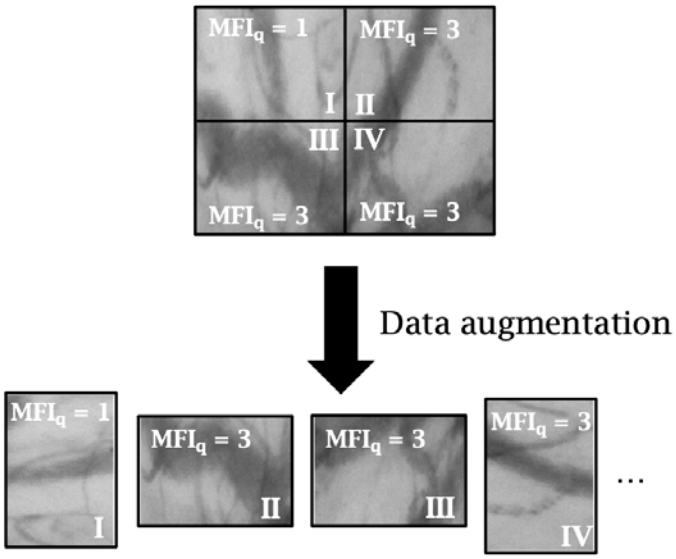


Figure 30: Quadrant based rating and a demonstration of data augmentation.

6 Future work and research avenues

6.1 Data

As seen in the performance of our classification algorithms, the imbalanced dataset lead to frequent prediction of the majority class (MFI=3). This is particularly problematic in a clinical context where identification of the minority classes (MFI=0,1) is of greater relevance. Several attempts were made to mitigate this imbalance, including augmentation and class weights, however neither was entirely satisfactory. Selective augmentation of the minority class was discussed, however this would be expected to reduce variance within the minority class, vs the unaugmented majority class, which may hamper generalisation of the trained algorithm to both test set and real-world minority classes.

Whilst something of a machine learning cliche, it is likely that the greatest improvement to classification performance would arise from additional data. Whilst public datasets are available for other medical imaging modalities, such as chest radiographs [8], we were unable to identify another source of microcirculatory videos. A potential solution to this could be collaboration with other research groups investigating this modality, for example the Xtreme Everest 2 research consortium has published results of the MFI for 133 patients measured at three time points [5]. Collaboration could augment not only the volume of data but help address the class imbalance as MFI has been shown to be reduced in the high altitude environments studied by this group [11].

6.2 Modelling

Given the intrinsic temporal quality of 'flow', it is possible that more sophisticated methods for modelling the time dimension may aid predictive accuracy. Such approaches may include a neural network with separable spatial and temporal convolutions which is able to take the sequence of frames as input, encompassing both stages of the CNN-RNN model. Limited by the number of labels available, we consider unsupervised approaches such as variational autoencoders or normalising flow models which are able to learn a latent representation of

the input may offer the possibility of density estimation. Learning such a distribution from highly imbalanced data, as seen in our experiments, could then allow for classification of the minority class to be reconceptualised as an outlier detection problem. This may suit the clinical application of an algorithm where greater importance may be placed on automatic identification of the minority/abnormal/pathological class. Additionally generative models may offer the potential to produce synthetic data which could be used to augment the small dataset for other models such as those described in Section 5.

6.3 Alternative approaches

We approached this problem as a classification problem, when in reality the patients present with scores on a continuous scale. We assume that the MFIv scores 0.0 to 3.0 can be grouped into 4 classes - but we have little evidence to say that change from a score of 0.0 to 0.51 is equal to change from 2.49 to 3.0. Our principle component analysis results 5.3 also suggest no clear separability of classes. Future work could either: a) identify appropriate boundaries between scores based on clustering of an automatically measured metric from a specific feature in the image, or b) investigate an alternative approach that can extract features and provide a continuous output metric which clinicians can then apply boundaries too. For example, if a score 1.0 to 50 was produced or 1 to 5, clinicians can still set their boundaries and clinical pathways based on the original video presentations.

6.4 Re-evaluating results

Above, we have trained CNNs to classify images according to four classes i.e. MFIv scores equal to 0, 1, 2 and 3. However, the CNN in first instance assigns to each input, either a video or an image, a probability for its belonging to each of the classes. For example, this is how probabilities were assigned to each MFIv class in one of our CNN + RNN runs:

	p_0	p_1	p_2	p_3
1	0.008762	0.007320	0.090191	0.893727
2	0.992866	0.005654	0.000817	0.000663
3	0.007621	0.020966	0.432366	0.539047
4	0.010822	0.017582	0.161700	0.809896
5	0.008873	0.007911	0.095709	0.887506
6	0.010770	0.009339	0.112526	0.867364
7	0.007669	0.013442	0.138968	0.839921
...	...	...	...	...

Below, we show a comparison between outputting the class with the highest probability and interpolating between each class using the probabilities as weight, where the final MFIv prediction is calculated as

$$\textit{MFIv (predicted)} = p_0 \cdot 0 + p_1 \cdot 1 + p_2 \cdot 2 + p_3 \cdot 3.$$

	p_0	p_1	p_2	p_3	**class**	**weighted**
1	0.008762	0.007320	0.090191	0.893727	3	2.8688
2	0.992866	0.005654	0.000817	0.000663	0	0.0092
3	0.007621	0.020966	0.432366	0.539047	3	2.5028
4	0.010822	0.017582	0.161700	0.809896	3	2.7706
5	0.008873	0.007911	0.095709	0.887506	3	2.8618
6	0.010770	0.009339	0.112526	0.867364	3	2.8364
7	0.007669	0.013442	0.138968	0.839921	3	2.8109
...	...	...	...	...	...	...

Next, we compare this set of two predictions (”class” and ”weighted”) with the true underlying MFIv assigned by clinicians to each video.

	class	weighted	true MFIv
1	3	2.8688	2.86
2	0	0.0092	0.25
3	3	2.5028	2.47
4	3	2.7706	0.03
5	3	2.8618	2.92
6	3	2.8364	2.52
...	...	...	...

This set of predictions is made on a test set of 104 videos. To conclude, we can understand whether predictions would improve using a weighted approach by computing and comparing the mean squared error for the two predictions. In the scenario considered here, we obtain MSE(class, MFIv) = 0.93, whereas MSE(weighted, MFIv) = 0.79. (Note that comparing the true MFIv rounded to the nearst integer to the predicted classes (as done in the confusion matrix) would yield an even higher MSE at 1.01 for this scenario.)

6.5 Video stability & quality

In this report, we addressed the challenge of predicting perfusion indices from a single DFM video sequence using machine learning techniques and exploratory data analysis. One of the main contributions were the exploration of various algorithms with a strong analysis of CNN for this problem. The main focus of the report and the entire team was to deliver a proof-of-concept which we could obtain through the aforementioned algorithms. We further proposed methods to improve the video quality and leading the way to optimal input videos that can be used in implementing a video stability analysis tool. Overall, as discussed in 3.3, this research could inform the development of a DFM pen with an integrated IMU that can provide instant feedback to the clinician as to whether the recording is of sufficient quality.

7 Team members

7.1 Participants

Chris Tomlinson Chris is an Anaesthetics & Intensive Care registrar undertaking a PhD at the UCL CDT in AI-enabled Healthcare Systems. He has interest in microcirculation after working on the Xtreme Everest 2 expedition which used sidestream darkfield imaging (similar to DFM) to study hypoxic acclimatisation in Sherpa's & Westerners to identify biomarkers of (mal-) adaptation that may inspire novel diagnostic/ therapeutic targets for the cellular hypoxia seen in critical illness. Chris's contributions to this DSG challenge include exploring class weights adjustment to address class imbalance, future work directions, and implementing dimensionality reduction and 2D-CNN+ RNN.

Jan Gröls Jan is a PhD student at the University of Bath. He graduated from the laboratory of fluid separations, TU Dortmund, with an MSc in Chemical Engineering. He is working together with Dr Castro-Dominguez to develop a machine learning assisted approach for new pharmaceutical drugs. Jan's contributions to this DSG challenge include exploring data augmentation approaches.

Ramit Debnath Ramit is a computational social scientist and a Gates Scholar based at the University of Cambridge. He uses data science to inform public policy on climate and energy justice issues, some of his recent work is being presented at COP26. He has a particular interest in developing novel data-driven methods that can enhance social decision-making to address energy poverty and climate change. Ramit's contributions to this DSG challenge include conducting and writing literature review.

Sarah Johnson Sarah will soon start work as a Post Doc at Stanford University, working within the Digital Athlete Program, which focuses on applying machine learning and modelling techniques to solve problems within sport. She has previously worked for Dynamic Metrics, a gait analysis company, as a researcher combining data analysis of sensor data with musculoskeletal modelling and classification techniques. Sarah's contributions to this DSG challenge include studying video quality issues, intensity difference analysis, limitations & future work directions,

and exploring data pre-processing methods.

Max Barton Max is a PhD candidate at the University of Manchester. His research interests include applying anomaly detection methodologies to identify defective operations, and building soft sensors to predict qualities of interests for use in industrial processing. Max's contributions to this DSG challenge include studying future recommendations & class imbalance issues, implementing 2D-CNN with temporal gradient image, and exploring modelling challenges.

Tianyu Han Tianyu is a PhD candidate in Physics at the RWTH Aachen University. He is a Research Assistant at Physics of Molecular Imaging Institute, where he develops mathematical and machine learning models for pathology detection, classification, and state progression analysis. In close collaboration with university hospital RWTH Aachen and Fraunhofer MEVIS, he has developed a medical data sharing platform using advanced generative adversarial networks combined with federated learning to prevent the privacy leakage of patient information during the process of medical data processing. Tianyu's contributions to this DSG challenge include implementing ridge detection and 2D-CNN with blood flow maps.

Seyedeh Nazanin Khatami is a Postdoctoral Research Fellow at Harvard Medical School and Mass General Hospital. She works on multiple research projects including but not limited to developing data-driven Machine Learning models to address opioid crisis in the US, modeling Tuberculosis disease progression and transmission model in people with HIV in low and middle income countries, and cost-effectiveness analysis of interventions. She received her PhD from University of Massachusetts Amherst where she developed mathematical and computational models for disease prediction, prevention, and control analysis with the focus on reinforcement learning algorithms to evaluate phased public health decisions for infectious diseases like HIV and COVID-19. Seyedeh's contributions to this DSG challenge include implementing temporal gradient and CLAHE.

Giacomo Baldo Giacomo completed his PhD in the School of Mathematics at the University of Leeds studying evolutionary dynamics and emergent phenomena in collective behaviour. Giacomo's contributions to this DSG challenge include exploring moving average and pixel intensity analysis, while writing about data summary and results re-evaluation.

7.2 Facilitators

Aniketh Ramesh Aniketh is a doctoral student at the University of Birmingham, working on facilitating human-interaction with variable autonomy multi-robot systems. His work is highly interdisciplinary, combining insights from Multi-Agent AI, swarm intelligence, human robot interaction and human factors. Currently, he is exploring the merits of quantifying a robot's health using a set of robot vitals, analogous to a human's vital signs. Aniketh's contributions to this DSG challenge include facilitating the challenge, along with Diego, while also exploring optical flow approach and writing future work directions.

Diego Cammarano Diego is a Computer Scientist with a BSc/MSc earned at Sapienza University of Rome. He has recently worked at the European Medicines Agency and the European Central Bank on data-driven projects related to the monitoring of the clinical trials, the medicine marketing authorisation across EU and the statistical data management from National Central Banks and other international organisations. Diego's contributions to this DSG challenge include facilitating the challenge, along with Aniketh.

7.3 Principal Investigator

Kashif Rajpoot Kashif is an Associate Professor of Computer Science and Programme Director for Computer Science and AI programmes at the University of Birmingham's Dubai campus. His research focuses on developing computational and artificial intelligence solutions for data science problems in healthcare. Kashif's contributions to this DSG include preparing and guiding the challenge before and during the DSG.

References

[1] Altuna Akalin. *Computational Genomics with R*. URL: https://compmgenomr.github.io/book/ (visited on 09/21/2021).

[2] Sumeyra U Demir et al. “An automated method for analysis of microcirculation videos for accurate assessment of tissue perfusion”. In: *BMC Medical Imaging* 12.37 (2012).

[3] J. Deng et al. “ImageNet: A Large-Scale Hierarchical Image Database”. In: *CVPR09*. 2009.

[4] Gunnar Farnebäck. “Two-frame motion estimation based on polynomial expansion”. In: *Scandinavian conference on Image analysis*. Springer, 2003, pp. 363–370.

[5] Edward Gilbert-Kawai et al. “Sublingual microcirculatory blood flow and vessel density in Sherpas at high altitude”. In: *Journal of Applied Physiology* 122.4 (Apr. 2017), pp. 1011–1018. ISSN: 8750-7587. DOI: 10.1152/japplphysiol.00970.2016. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5407196/ (visited on 09/23/2021).

[6] Maged Helmy et al. “CapillaryNet: An Automated System to Quantify Skin Capillary Density and Red Blood Cell Velocity from Handheld Vital Microscopy”. In: *arXiv:2104.11574 [cs]* (Aug. 2021). arXiv: 2104.11574. URL: http://arxiv.org/abs/2104.11574 (visited on 09/15/2021).

[7] Zeshan Hussain et al. “Differential Data Augmentation Techniques for Medical Imaging Classification Tasks”. In: *AMIA Annual Symposium Proceedings* 2017 (Apr. 2018), pp. 979–984. ISSN: 1942-597X. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977656/ (visited on 09/23/2021).

[8] Jeremy Irvin et al. “CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison”. en. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01 (July 2019). Number: 01, pp. 590–597. ISSN: 2374-3468. DOI: 10.1609/aaai.v33i01.3301590. URL: https://ojs.aaai.org/index.php/AAAI/article/view/3834 (visited on 09/23/2021).

[9] Justin Johnson and Taghi Khoshgoftaar. “Survey on deep learning with class imbalance”. In: *Journal of Big Data* 6.1 (), pp. 1–54.

[10] Ossama Mahmoud, Mahmoud El-Sakka, and Barry G. H. Janssen. “Two-step machine learning method for the rapid analysis of microvascular flow in intravital video microscopy”. en. In: *Scientific Reports* 11.1 (Dec. 2021), p. 10047. ISSN: 2045-2322. DOI: 10.1038/s41598-021-89469-w. URL: http://www.nature.com/articles/s41598-021-89469-w (visited on 09/14/2021).

[11] Daniel S. Martin et al. “Changes in sublingual microcirculatory flow index and vessel density on ascent to altitude”. eng. In: *Experimental Physiology* 95.8 (Aug. 2010), pp. 880–891. ISSN: 1469-445X. DOI: 10.1113/expphysiol.2009.051656.

[12] Michael J. Massey and Nathan I. Shapiro. “A guide to human in vivo microcirculatory flow image analysis”. en. In: *Critical Care* 20.1 (Feb. 2016), p. 35. ISSN: 1364-8535. DOI: 10.1186/s13054-016-1213-9. URL: https://doi.org/10.1186/s13054-016-1213-9 (visited on 09/13/2021).

[13] Buda Mateusz, Maki Atsuto, and Maciej Mazurowskiac. “A systematic study of the class imbalance problem in convolutional neural networks”. In: *Neural Networks* 106 (), pp. 249–259.

[14] Paul R. Mouncey et al. “Trial of Early, Goal-Directed Resuscitation for Septic Shock”. en. In: *New England Journal of Medicine* 372.14 (Apr. 2015), pp. 1301–1311. ISSN: 0028-4793, 1533-4406. DOI: 10.1056/NEJMoa1500896. URL: http://www.nejm.org/doi/10.1056/NEJMoa1500896 (visited on 09/14/2021).

[15] Pranav Rajpurkar. “Deep learning for medical image interpretation”. English. PhD thesis. Standford University, June 2021. URL: https://cs.stanford.edu/~pranavsr/files/thesis.pdf.

[16] Valeria Rizzuto et al. “Combining microfluidics with machine learning algorithms for RBC classification in rare hereditary hemolytic anemia”. en. In: *Scientific Reports* 11.1 (June 2021), p. 13553. ISSN: 2045-2322. DOI: 10.1038/s41598-021-92747-2. URL: https://www.nature.com/articles/s41598-021-92747-2 (visited on 09/15/2021).

[17] Connor Shorten and Taghi M. Khoshgoftaar. “A survey on Image Data Augmentation for Deep Learning”. en. In: *Journal of Big Data* 6.1 (Dec. 2019), p. 60. ISSN: 2196-1115. DOI: 10.1186/s40537-019-0197-0. URL: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0197-0 (visited on 09/15/2021).

[18] Christian Szegedy et al. “Rethinking the Inception Architecture for Computer Vision”. In: *arXiv:1512.00567 [cs]* (Dec. 2015). arXiv: 1512.00567. URL: http://arxiv.org/abs/1512.00567 (visited on 09/15/2021).

[19] Keras Team. *Keras Applications*. en. URL: https://keras.io/api/applications/ (visited on 09/23/2021).

[20] The ARISE Investigators and the ANZICS Clinical Trials Group. “Goal-Directed Resuscitation for Patients with Early Septic Shock”. en. In: *New England Journal of Medicine* 371.16 (Oct. 2014), pp. 1496–1506. ISSN: 0028-4793, 1533-4406. DOI: 10.1056/NEJMoa1404380. URL: http://www.nejm.org/doi/10.1056/NEJMoa1404380 (visited on 09/14/2021).

[21] The ProCESS Investigators. “A Randomized Trial of Protocol-Based Care for Early Septic Shock”. en. In: *New England Journal of Medicine* 370.18 (May 2014), pp. 1683–1693. ISSN: 0028-4793, 1533-4406. DOI: 10.1056/NEJMoa1401602. URL: http://www.nejm.org/doi/10.1056/NEJMoa1401602 (visited on 09/14/2021).

[22] Jian Wang et al. “A Review of Deep Learning on Medical Image Analysis”. en. In: *Mobile Networks and Applications* 26.1 (Feb. 2021), pp. 351–380. ISSN: 1572-8153. DOI: 10.1007/s11036-020-01672-7. URL: https://doi.org/10.1007/s11036-020-01672-7 (visited on 09/15/2021).

turing.ac.uk
@turinginst