# ALFA Pipeline – REACT-2 Study

Imperial School of Public Health, Bioengineering/BICI and i-sense EPSRC IRC in Agile Early Warning Sensing Systems for Infectious Diseases at University College London (www.i-sense.org.uk)

Nathan C. K. Wong, Sepeher Meshkinfamfard, Valerian Turbe, Matthew Whitaker, Maya Moshe, Alessia Bardanzellu, Edurado Pignatelli, Wendy Barclay, Ara Darzi, Paul Elliot, Helen Ward, Graham Cooke, Reiko Tanaka, Rachel A. McKendry, Christina Atchison and Anil A. Bharath

2021-06-29

# Contents

# 1 Executive Summary

REACT-2 Study-5: Population surveys of the prevalence of SARS-CoV-2 antibodies in the community in England. At each Round, we contacted a random sample of the population by sending a letter to named individuals aged 18 or over from the NHS GP registrations list. We then sent respondents a LFIA kit for SARS-CoV-2 antibody self-testing and asked them to perform the test at home and complete a questionnaire, including reporting of their test result. Participants were also asked to submit a photograph of their test result via an online portal using instructions and a template provided.

1. We have created a computational pipeline (ALFA - Automated Lateral Flow Analysis) that supports the review and audit of Covid-19 Home Test kits for assessing IgG seroprevalence. It uses machine learning, computer vision techniques and signal processing algorithms to analyse images of the Fortress LFIA, and subsequently classify LFIA test results as invalid, IgG negative and IgG positive.
2. In brief, our approach first involved building a large image library of participant-submitted test images from REACT-2 Study-5 as a training data set, optimising algorithms for high sensitivity and specificity, and then deploying the pipeline to assess its performance compared to traditional visual interpretation with human expert readers, and finally the REACT-2 Study-5 participants. The use of 2D CNNs for LFIAs and new image capture protocol builds on previous research by the McKendry group and i-sense EPSRC IRC at UCL (Turbé et al. 2021).
3. Across REACT-2 Study-5 rounds 1 to 5, there were 740,356 participants of which 81.7% (605,013) submitted images available for this study. We were able to use, and thus analyse, 98.4% (595,339) of these images. Images could not be used if the LFIA device or test result window was not present in the image or due to image corruption/error.
4. Weak positives were found to be a potential source of human and machine error. In tests using two substantially different proportions of weak positives, Cohen's kappa for machine read outs of IgG status against human experts ranged from 0.905 (99% CI: 0.890 – 0.919) for the set with more weak positives, to 0.966 (99% CI: 0.956-0.976) with fewer weak positives. Specificity, against human expert readers, ranged from 0.987 (99% CI: 0.962 – 0.999) to 0.994 (99% CI: 0.972 – 0.999), whilst sensitivity ranged from 0.901 (99% CI: 0.817 – 0.965) to 0.971 (99% CI: 0.929 – 0.998)
5. ALFA consistently did better than study participants. This suggests that the ALFA pipeline can support the identification of reporting mistakes by study participants, and as such reduce the number of false positives and false negatives.
6. We compared self-reported participant readings with ALFA readouts over all interpretable images of the REACT-2 study to date (over 500,000 image submissions) using Cohen's kappa to quantify the participants' ability to interpret LFIA read outs. This assessment was possible as the ALFA pipeline performed with "substantial agreement" with human experts. We found a Cohen's kappa of 0.797 (99.9% CI: 0.7966 to 0.7968) between participants' and ALFA readout of test validity

and IgG status. The results support earlier studies (Atchison et al. 2020) which concluded that participants' ability to correctly read IgG status from the LFIA device was very high. Disagreements were mostly due to weak positives being misread by participants as negative, potentially leading to slight underestimation of seroprevalence in REACT-2 Study-5.

7. The pipeline's design incorporates strategies for detecting of out-of-distribution submissions and focuses the reading algorithms on the results window of the LFIA. It provides an indication of the presence of the control line, and its status (valid/invalid test result), providing a sensitivity of 0.917 (99% CI: 0.700 - 0.990), and a specificity of 1.000 (99% CI: 1.000 - 1.000) for detecting invalid tests. Sources of error are due to partially converted control lines. The current method looks at normalized red and blue pixel intensity at the detected control region, improvement on detecting these samples could be achieved using a similar bootstrapping method to develop a CNN.
8. The ALFA pipeline is modular, and elements can be adapted and adjusted in sequence for new devices. The analysis takes between 10-14s per image, depending on the complexity of the algorithm chain. The entire chain is yet to be optimised for computational speed.
9. The use case for the pipeline is in reviewing the accuracy, at the population-level rather than the individual level, of participant-reported LFIA test results for estimating SARS-COV-2 seroprevalence. Its application could be to identify and examine trends that may highlight sources of systematic error or device failure. These can then inform improvements to study processes or user instructions.
10. We would suggest an open-source approach to maintaining the pipeline: to apply to other devices, we would need to repeat the bootstrapping approach that was used to incrementally label data for training increasingly complex systems for read out. This can take time, but we know that the strategy works.
11. We also developed tools to analyse ensembles of participant LFIA responses generated from their LFIA images, allowing identification of trends in responses accumulated over thousands of participants in addition to the individual readings provided by participants. By using this approach, it is possible to identify a consistent source of error in participants' interpretation of a specific assay component of the LFIA device. Indeed, we identified a tendency to misread the results when blood exhibited a particular form of leakage pattern on the LFIA. Because this effect might be dependent on manufacturer or batch, ensemble projection responses allow the quality control of devices and their usage to be quickly assessed at scale, with respect to the responses submitted by participants.
12. Using the above approach, we also identified significant shifts in the distribution of amplitudes of signals at the IgG response location in the LFIA read-out for participants who reported a positive test result in REACT-2 Study-5 R1, and subsequently again reported a positive test result

approximately one month later in REACT-2 Study-5 R2B. These shifts in the signal could be consistent with waning immune response, lending support to the validity of an approximate quantitative relationship between the colour intensity of the IgG band and antibody titres. If this were indeed the case, it would imply the ability of this technique to determine population-level shifts in ensemble-level responses that could provide additional insight into changes in immune response in the population. However, this work is very experimental and caution should be exercised in interpreting such results. Firstly, we cannot rule out the possibility that our findings are not due to systematic differences in the LFIA batches used in Round 1 and Round 2B. In addition, LFIAs are designed to provide a qualitative result (presence or absence of IgG). Any correlation between IgG band intensity and quantitative measurements of IgG titre would require further laboratory investigation.

13. This project has exceeded expectation, in the sense that the manpower applied to develop the pipeline has been a fraction of that of a commercial endeavour. The next steps would be to ensure sustainability and extension.

To conclude, we demonstrated the potential of a computational pipeline to accurately classify LFIA test result images, with an overall performance comparable to that of a human expert and slightly higher than that of a study participant, particularly for weak positive IgG results. Given the potential for LFIA devices to be used at scale as part of the COVID-19 response (for both antibody and antigen testing), even a small improvement in quality assurance could reduce the risk of false positive and false negative result read-outs by members of the public.

Our findings lend support for the use of machine learning-enabled automated reading of at-home self-test LFIA results for estimating SARS-CoV-2 seroprevalence. It has the potential, if used for reading antibody and antigen LFIA device results, to be a tool for quality assurance, population-level community surveillance, and validation of possible home-administered tests for travel clearance, and “infection-free” status checks to protect populations against SARS-CoV-2.

## 2 Derived Papers

Below, we provide three draft abstracts that are indicative of papers that we plan to write about this work. The potential targets for these abstracts range from “special issue” journals that cover AI applications in healthcare, to engineering journals that seek magazine-type articles of a wide readership. The abstracts are intended to be indicative, and content may be combined, depending on audience, or new interpretations that may emerge from ongoing analysis of incoming data.

## 2.1 Monitoring Self-Readings for Covid-19 Antibodies: Results on 1/2 Million Images

**Abstract** We describe how to construct an analytical pipeline that permits auditing of lateral flow tests performed in the home through images supplied by users. We developed the pipeline to support the REACT-2, Study-5 survey of SARS-CoV-2 in England, and applied it to analyse results from over 500,000 submitted images. The ALFA (Automated Lateral Flow Analysis) pipeline analyses over 98% of submitted images of multiplexed immunoassay devices. This provided an automated reading that shows excellent agreement with human experts on reading Immunoglobulin G (IgG) status from the submitted images. This agreement, assessed using Cohen's kappa, varied with the proportion of weak positive results, ranging from 0.905 (99% CI: 0.890 to 0.919) to 0.966, (99% CI: 0.956 to 0.976) in reading IgG status.

We compared self-reported participant readings with ALFA readouts over all interpretable images of the REACT-2 study (to date, over 500,000 image submissions). The results strongly support earlier pilot studies which concluded that participants' ability to read IgG status from the multiplexed device correctly was very high. We found a Cohen's kappa of 0.797 (99.9% CI: 0.7966 to 0.7968) between participants' and ALFA readout of test validity and IgG status. Disagreements in readings appear due to weak positives being missed by participants, potentially leading to slight underestimations of seroprevalence in self-reported test results.

Finally, we suggest a means of summarising readings from multiplexed LFIA devices over large surveys. By extracting projection signatures from co-registered read-out windows in an appropriate colour space, it is possible to identify trends in responses accumulated over thousands of users. By using this approach, it was possible to identify a consistent source of error in participants' interpretation of a specific assay component on the multiplexed device, due to sample leakage. Projection signatures also show evidence of waning response in participants tested twice, one month apart. Significantly, this effect is found even amongst participants self-reporting as positive, ($p < 1.e{-}5$, two-sided Mann-Whitney-Wilcoxon test against the null hypotheses of no difference between the first and second tests).

## 2.2 Ensemble Signatures from Lateral-Flow Test Kits

**Abstract** We propose a technique to support ensemble analysis of infection levels using the visual readouts from lateral flow test kits. This technique supplements the use of data-driven AI, permitting a visually interpretable overview of population-level responses. Based on easy-to-compute projection signatures, it supports response averaging and potentially yields additional finessing of participants' readings when applied across a study population. Used in the context of SARS-CoV-2 antibody testing, we found effects in normalised colour channels that showed a variation in the heights of ensemble-averaged response peaks. This finding is in line with decreasing seroprevalence assessed by proportions of positive returns, but lends additional finessing through a distribution shift in the contrast of line

responses.

The projection-based technique also allowed us to identify a source of participant-reported false-positive results, line-like presence of blood in the result window. We built this technique into an AI pipeline (ALFA - Automated Analysis of Lateral Flow Analysis) that has been designed and trained to interpret participant-submitted photographs of the Fortress SARS-CoV-2 (Flower et al. 2020), but this is adaptable to other infection monitoring scenarios.

## 2.3 The ALFA Pipeline for Visual Auditing of Lateral Flow Test Results

**Abstract** Lateral Flow Tests can be manufactured cheaply, at scale, and by multiple manufacturers to meet the needs of surge or mass testing. However, the correct usage and self-reading by members of the public benefits from monitoring. Monitoring can be done through participant-submitted photographs and self-readings by a human expert through an appropriate sampling strategy, or by algorithms at scale. In this paper, we describe the construction of a computational pipeline (ALFA - Automated Lateral Flow Analysis) based on a combination of computer vision, machine learning and signal processing algorithms that provide test interpretation matching the level of cross-validated human expert interpretation. We describe the process of bootstrapping the training of progressively larger machine learning models through heuristic algorithms. This process involves leveraging the decisions of less-sophisticated automated algorithms to rapidly improve performance, with parsimonious human annotation of discrepancies between machine and participant-submitted results. We also suggest tools to monitor changes in participant behaviour, potential device flaws in use. The techniques we propose imply the ability to determine population-level shifts in ensemble-level responses that could provide additional insight into changes in immune response.

# 3 Introduction

During the COVID-19 pandemic, there is an ongoing need to monitor population levels of immunity through individuals' antibody responses. Antibodies are a long duration biomarker of exposure, and repeated testing for them could provide early indicators of antibody waning, and insights into historical infections.

Lateral flow immunoassays (LFIAs) (Adams et al. 2020) provide a relatively convenient, inexpensive, easily distributed test that can be self-administered at home. Whilst the accuracy of commercially available LFIA antibody tests has been called into question for individual diagnoses (Adams et al. 2020), some of this variability is device-dependent. In an early comparative study, one type of device was found to have a specificity of 98.6% and sensitivity of 86%. Even though this represents moderate performance in terms of sensitivity, there is still a strong use case for public-health monitoring and estimating background rates of seroprevalence at population scales (Montesinos et al. 2020).

The REal-time Assessment of Community Transmission 2 (REACT-2) programme (Riley et al. 2020) provides robust estimates of the seroprevalence (antibody presence) in the general population in England, UK. It consists of several sub-studies to evaluate the performance of different LFIAs and their usability for the general public. These sub-studies informed the planning of REACT-2 Study-5, the series of national seroprevalence surveys. Participants receive a home test kit consisting of the Fortress LFIA and instructions, report their test results and have the option to submit an image of their LFIA. At the time of writing, the programme has collected over 500,000 responses, with future rounds of data collection planned.

The sub-studies showed that concordance between the participants (general public) and clinicians on determining LFIA results was high. For identifying negative antibody results, participants were 97% accurate; for identifying seropositive results, participants' accuracy was 93.9% (Atchison et al. 2020).

When full roll-out of regular surveys is conducted at a scale of 100,000s, there is a need for monitoring and reviewing the use of the devices, and use cases for verifying the readings. For example, changes to the instructions, or differences in participant demographics may contribute to slightly different sources of error in either using or interpreting the device readings. Such errors will be difficult and time-consuming to identify within large corpora of image samples. In this work, we propose an automated pipeline that utilises the participant-submitted images to fast-track the reviewing process.

The pipeline uses deep learning and computer vision techniques to analyse several factors in reading images of LFIA devices. The considerations in this particular work are specific to the Fortress device for SARS-CoV-2 antibody testing. However, we discuss strategies to adapt the pipeline to other devices and in as time-efficient a manner as possible.

Amongst the considerations we have had to address are:

- detecting participant-submitted images that, for various reasons, are out-of-distribution

- having a strategy to 'bootstrap' to new device shapes or characteristics, without large quantities of manual labelling
- shifts in the distribution of background seroprevalence and its effects on workload as part of an expert auditing system
- providing a compact representation of the visual test results, facilitating semi-quantitative review
- weak positive responses that are often not picked up by participants; these required specific targeting in order to find, label, and use for training

# 4 Background

## 4.1 Lateral Flow Device – Specifics

The LFIA used in the REACT-2 programme is the Fortress Diagnostics COVID-19 Total Antibody test, which contains two separate immunoassays, one for Immunoglobulin G (IgG), and another for Immunoglobulin M (IgM). The tests requires a small sample of blood to be placed into the blood well. A buffer solution is then added to the buffer region, and after 10 minutes, the reading is taken. The colour of the control line determines the test validity, whilst the test result depends on the presence of bands indicating the presence of either IgG or IgM antibodies, or both. Figure 1 shows an example of a completed test, the layout of the device, and how three categories of test results are displayed in the test result (read-out) window.

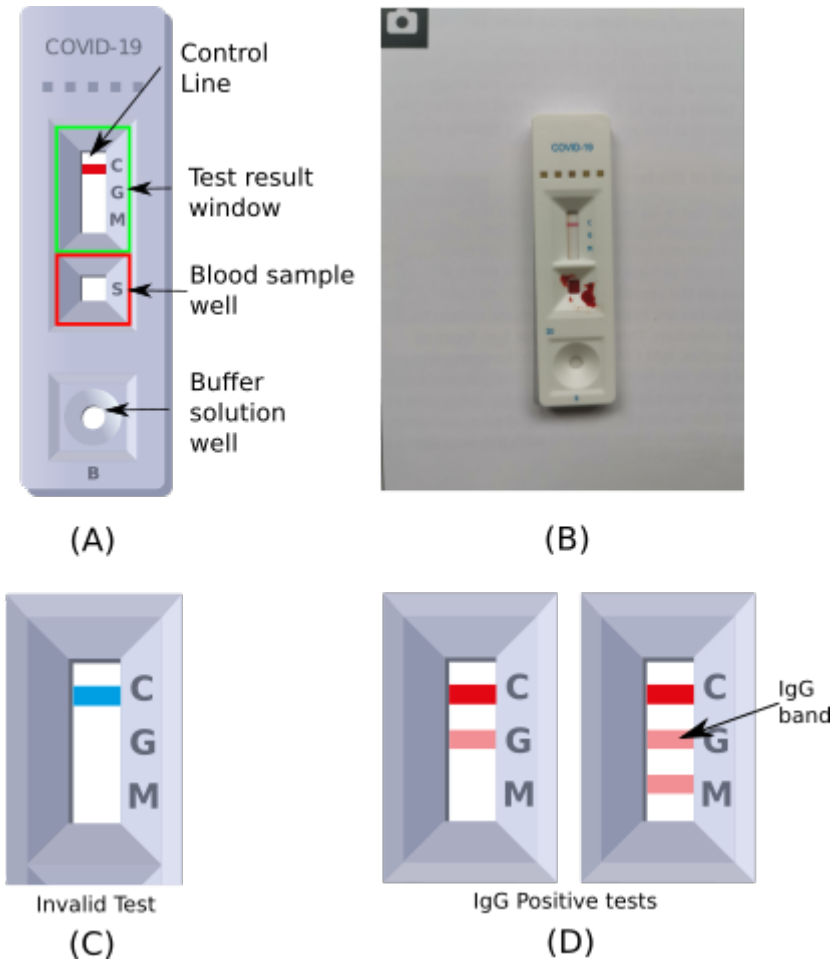


**Figure 1:** A: Key visual features of the LFIA devices are the test result window and blood sample well. This diagram also shows a negative IgG test. B: This is an example of a participant-submitted image of a negative IgG test. C: The result window has an initially blue control line, which will remain if the test is unsuccessful (invalid). D: In a successful test, the control line turns red, and if IgG antibodies are present in the blood sample, a secondary line will appear below the control. The tertiary line indicates the presence of IgM antibodies, which are not primary targets for REACT-2 detection.

## 5 Image Capture Considerations

Participant self-readings and images were collected in partnership with IPSOS MORI. Images were also, in this process, anonymised by removing specific fields from EXIF data contained within the uploaded images. Image and participant-submitted data were transferred to a dedicated REACT-2 server and then batch copied through a separate anonymisation process onto a high-throughput data store coupled to GPU-equipped servers.

Early rounds of testing (e.g. Round-1) saw substantial variation in backgrounds amongst the images submitted by participants. In subsequent rounds, the instruction booklet included a page containing a camera icon and placement guide for the device in the centre of the page, which decreased the wide variation in image appearance. A later iteration of the instruction booklet (Round-5) modified the background to ensure better framing, and the ALFA pipeline deals with both these backgrounds. This change, alongside new instructions for Round-5, aims to pave the way to increased efficiency and involved adding high-contrast markers, colour calibration squares, and a geometric layout that supports estimation of camera pose relative to the supplied background.

The participant-submitted images of the LFIA test devices varied dramatically in size from an estimated 10% of the pixel area of the captured image up to around 40% of the pixel area. In addition, since the device measures only 20mm $\times$ 73mm, and the read-out window itself is only 3mm wide $\times$ 12mm long, the proportion of area represented by the read-out window can be as small as $1/30th$ the entire pixel area of the captured image. Some participants used digital/optical zoom whilst taking a photograph of the device. We advised against having shadows falling across the read-out, which led to the design of the positioning guidance; though this contributes to the slight visual angle subtended by the read-out window, enforcing this moderate camera distancing has the additional benefit of reducing the chance of autofocus failing.

The devices used by participants to take the photographs varies. Rather than restricting participants' devices by the device manufacturer, operating system, or type, participants were allowed to capture the image from any device capable of taking a photograph, uploading the shot via a web portal. Whilst this supports a broader range of participants, it also increases the variety in image quality, resolution and encoding. The latter sometimes yielded highly compressed images. Again, the design decision was to have no constraints, simplifying the process for many participants. Due to the variety of devices and the compression applied, file sizes also varied from roughly 50 KB up to 5MB.

## 6 Datasets and Curation

6 rounds of the REACT-2 image data were harvested (denoted by R$x$ where $x$ is the round). The image datasets were collected during approximately two week periods between July 2020 and February 2021. R2B contains a sample of participants that were also tested in R1. Furthermore, the February 2021 round includes a large number of participants that had received their first-round vaccinations.

### 6.1 Derived Datasets

In order to design and build specific elements of the ALFA pipeline, image data were labelled appropriate to the particular task of the pipeline. We used Oxford Imaging Group's VIA (Group, n.d.) and custom

| Round | Period | No. of participants | No. of images | Images analysed* |
|---|---|---|---|---|
| Round-1 (R1) | 20/06/2020-10/07/2020 | 109,075 | 94,700 | 93,252 |
| Round-2 (R2) | 30/07/2020-12/08/2020 | 111,057 | 96,817 | 95,508 |
| Round-2B (R2B) | 19/08/2020-31/08/2020 | 11,517 | 9,702 | 9,500 |
| Round-3 (R3) | 15/09/2020-27/09/2020 | 166,681 | 125,499 | 123,614 |
| Round-4 (R4) | 27/10/2020-10/11/2020 | 169,927 | 135,594 | 133,225 |
| Round-5 (R5) | 25/01/2021-08/02/2021 | 172,099 | 142,701 | 140,240 |
| Total | - | 740,356 | 605,013 | 595,339 |

**Table 1:** *The difference between images available and images analysed is due to filtering based on failed segmentation (pipeline could not identify the cassette or regions of interest due to various reasons) or image corruption/error.

software to create segmentation labels and expert interpretations of read-outs. Two human experts labelled a significant quantity of data in the early rounds; in subsequent rounds, their guidance and examples were used to train other team members to label images. Segmentation labels for the LFIA devices primarily came from R1, whereas the read-out interpretations were from R1 and subsequent rounds. We summarise the curated data sets in Table 2.

**Table 2: An overview of the data sets used.**

| Dataset | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| REACT-2, Study-3, Sample A | ✓ | - | ✓ | ✓ |
| REACT-2, Study-3, Sample B | ✓ | ✓ | ✓ | ✓ |
| REACT-2, Study-5, Round-1 sample | - | - | ✓ | ✓ |
| REACT-2, Study-5, Round-2 sample | - | - | ✓ | ✓ |
| REACT-2, Study-5, Round-5 sample | - | - | - | ✓ |

| Experiment | Segmentation | | Validity | Classification 1 | | Classification 2 | |
|---|---|---|---|---|---|---|---|
| Purpose | Develop | Test | Test | Develop | Test | Develop | Test |
| No. of samples | 415 | 83 | 187 | 865 | 294 | 1700 | 237 |
| No. of IgG positive | - | - | - | 351 | 143 | 641 | 79 |
| No. of Invalid | - | - | 12 | - | - | - | - |

## 6.2 Labelling for Image Segmentation

Creating a train-test-validation dataset for segmentation was relatively straightforward; regions on the device are rectangular or square and easily bounded by rectangles or polygons. We used the VGG Image Annotator (VIA) from Oxford's Visual Geometry Group (VGG). In common with many authors, we found that a small number of images was satisfactory to produce good quality segmentation results. Segmentation failures are few and are detectable using priors on the relationship between the region shapes, areas and spatial relationship. We describe these in Section 7.1.2.

### 6.3 Labelling for Read-out

Labelling the dataset for read-out is not as easy as it may sound; even though we have participant submitted readings, creating labels for the purpose of training a deep network to provide reliable diagnostics has three complicating factors. First, the ground truth may not be easy to obtain, as in some cases, the detection of faint responses might be missed by participants. Secondly, depending on background levels of infection, examples of seropositive individuals might be relatively rare. At the time of beginning training, infection rates were still relatively low in the UK, around the 5-6% mark. This meant that to get just 100 examples of seropositive tests, one has to use around 2,000 images. Further, without strategies to address class imbalance Leevy et al. (2018), performance of a trained algorithm could be well below what is necessary to make a usable system, particularly in the face of changing population immunity levels (essentially, inducing continuous changes in the class prior distributions).

A complicating factor in labelling turned out to be the detection of test results that were weakly positive and subsequently found to be incorrectly labelled by participants. A weak positive may be characterised as a visually very faint line in the read-out window for IgG status; they are of low contrast and sometimes discontinuous. An expert human reader can detect the presence of such responses by zooming in to a moderate level, and scrolling the window across his/her field of view; in doing so, a distinctive presence of intensity variation can be discerned. Once trained to do this, the presence of the line is unmistakable to a human observer with corrected or uncorrected visual acuity in the 'normal' range. Weak positives can be difficult to detect within a large dataset, as each window needs to be carefully examined. Notably, the *proportion* of weak positive IgG results will also depend on infection rates at the time of testing, rates of waning, and the diverse nature of immune responses. Because of this, self-readings from untrained participants' in such settings are likely to be biased *against* detecting weakly positive cases; this factor became clear during subsequent rounds of data collection, impacting on machine-based accuracy as background levels of infection increased. Cross-checking of the labels – even amongst expert readers – was deemed critical to improving performance; errors in human expert reading were likely due to fatigue in labelling.

In the early stages of algorithm development, we found dramatic changes in the performance of early versions of the algorithm in subsequent rounds of data collection, most notably between R1 and R2. Leveraging discrepancies between participant readings and algorithm readings, we selected a subset of discrepancies for expert review and used such samples in training to improve read-out performance.

A final boost in performance was obtained from the samples of vaccinated participants in R5. Here, results that were reported negative were re-examined by the team; in many cases, these results turned out to be weakly positive. For these cases, a cross-checking strategy was employed to cross-validate the labelling of LFIA results for this subset of vaccinated users. This latter dataset contributed significantly to improving the sensitivity of the final system.

More specifically, we reviewed the images of 811 participants from R5, who reported negative results 21 days or more after being vaccinated. The images were reviewed by at least two independent, trained reviewers. In cases of disagreement, a third reviewer was called in to confirm the reading. Through this methodology, human experts could use majority voting to label weak positives, which were characterised by quite faint lines in the read-out window, with no other potentially confounding characteristics, such as blood leakage, image blur or lack of a control line. Using this methodology to label these positive cases enabled better training of the read-out network. Examples of these weak positives are shown in Figure 2.

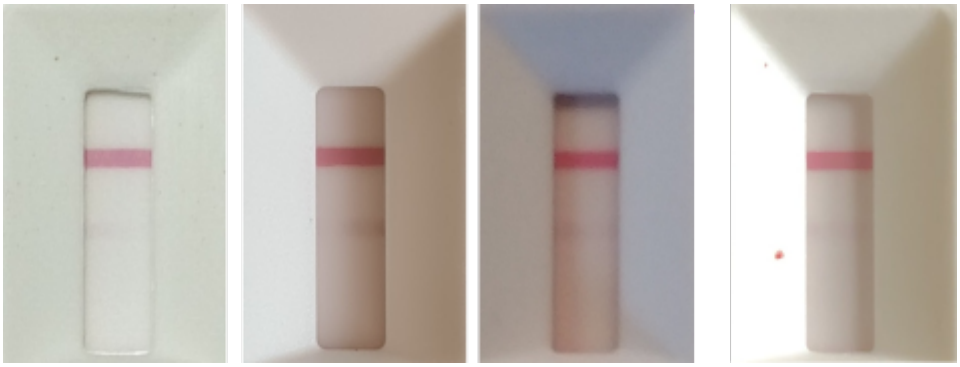

**Figure 2: Examples of weak IgG Positive samples.** 'Weak' IgG responses were highlighted as being a challenging scenario for the algorithms to classify correctly. As seen in the examples, the line is faint, and with additional issues of variable lighting, the solution was to introduce more of these cases into the training data. Many of these training examples would come from vaccinated participants of REACT-2 Study-5 Round-5.

## 7 ALFA Pipeline design

Specific considerations that we address are:

- detecting participant-submitted images that, for various reasons, are out-of-distribution
- having a strategy to 'bootstrap' to new device shapes or characteristics, without large quantities of manual labelling
- providing a compact representation of the visual test results, facilitating semi-quantitative review
- the use of transfer learning to train a system with increased read-out sensitivity

The pipeline is designed to:

- facilitate interpretability (Olah et al. 2018), following a strategy similar to that of (De Fauw et al.

2018), but with additional checks based on the uncontrolled nature of participant-supplied ('in the wild' e.g. (Murmann et al. 2019)) images.

- support incremental improvements in performance without breaking other aspects of functionality
- support the design goal of comparing IgG responses in a semi-quantitative way

A rough illustration of its content is shown in Figure 3

The choice of a pipeline in well-defined blocks facilitates its management, and the use of heuristic algorithms to bootstrap the labelling process in the absence of a large amount of ground-truth. In essence, by leveraging participant readings and hand-engineered algorithms and prioritising discrepancies between these for expert review by multiple human readers drawn from the REACT-2 team and from amongst the authors, it was possible to produce a large number of labelled examples for training. Our strategy included using increasingly complex networks to get to the point of having sufficient labelled data to train a network that performs on par with human expert readers.

Additionally, having a modular approach allows changes to specific elements of the pipeline analysis whilst avoiding both end-to-end training and re-analysis of all 1/2 million images, which would be necessary with a single network. It also permits stepwise adaptation to new device layouts in which the changes could be to the device layout itself or the positioning of test lines in the read-out window.

Finally, the interpretability of decisions made by AI systems in a diagnostic context is a emerging as a desirable design requirement. The pipeline issues metrics on different aspects of the reading process that can be used to identify both out-of-distribution images and provide some degree of confidence in the steps taken to determine the final test result.

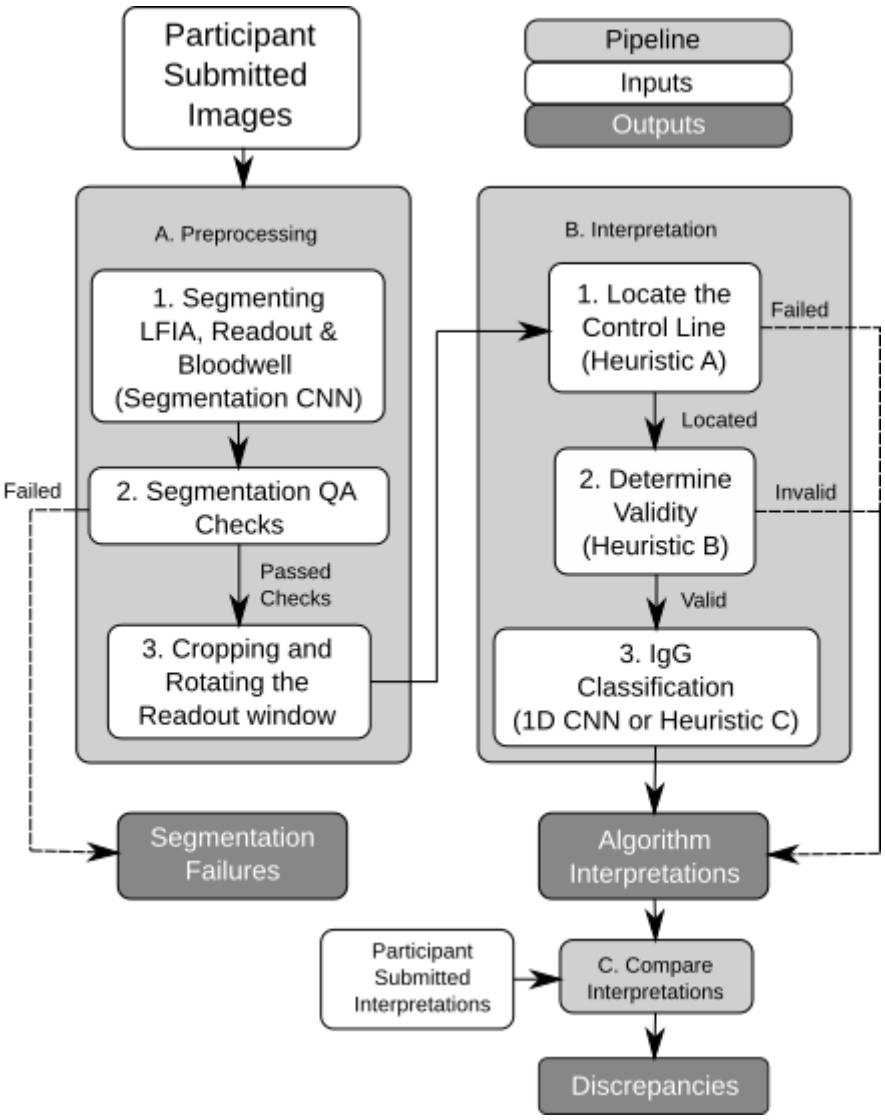


**Figure 3: Flow Diagram illustrating the ALFA Pipeline.** The pipeline takes participants' images of LFIAs, and the corresponding participants' result interpretations as input. There are three output reports. The first reports segmentation failures, including erroneous and poor-quality images. The second, an algorithm's interpretation of the LFIA result, and finally, discrepancies between participants and algorithm.

## 7.1 Pre-processing the LFIA images and removal of out-of-distribution data

Pre-processing consists of three steps (see Figure 3). First, we identify regions of interest (ROIs) through segmentation. Next, we use geometric priors to filter out images that are out-of-distribution. Finally, a simple registration process is applied to create a standard-sized representation of the read-out window, geometrically aligned with a convenient, normalised Cartesian coordinate axis.

### 7.1.1 Segmentation

We used a deep Convolutional Neural Network (CNN), based on a ResNet-50 architecture, to identify candidate regions of interest (ROI) in the participant-submitted images. These regions are i) the LFIA itself, ii) the test result (read-out) window and iii) the blood sample well. We started with a pre-trained network, implemented in Tensorflow (Abadi et al. 2016), with additional post-processing – connected components analysis and small region removal – to provide output simplification. Training of this network, known as dhSegment (Oliveira, Seguin, and Kaplan 2018), which was initially developed for document processing, involves fine-tuning the expansive part of the network. We supplied 498 LFIA images (Data set 1 shown in Table 2), on which the three regions of interest (ROIs) were manually labelled. The dataset was split into training, validation and test sets, with 415 being used for training and validation. We used the softmax cross-entropy loss function, batch sizes of 1, a learning rate of $5 \times 10^{-5}$, and 30 epochs. We assessed the performance of segmentation by calculated the Dice coefficient (see Appendix A1 in Section 13) for each of the three ROIs.

The test results showed high Dice coefficients for all ROIs: 0.985, 0.950 and 0.932 for the LFIA, read-out window and blood well, respectively, indicating that the dhSegment CNN can segment the ROIs reliably. A potential cause for the lower performance on extracting the read-out window and blood well could be the variation in blood splatter and image capture angles. Improvement is possible by gathering more labelled images, and although this is time-consuming, it does not require expertise, unlike labelling for the read-out classification.

### 7.1.2 Geometric/Semantic Checks

The second step in the pre-processing is to detect and remove out-of-distribution results from further processing. These out-of-distribution cases stem from many possible sources, and account for around 1% of submitted images. The checks are performed on geometric properties or relationships between the candidate ROIs to ensure that only successfully segmented LFIA images proceed through the pipeline. The criteria include:

- The presence of exactly one of each type of ROI
- The putative ROIs for blood well and result window lie within the bounds of the ROI for the cassette
- Geometric-priors: these are formed from subsets of the combinatorial possibilities of pairs of pixel-space measurements (area, perimeter, width and height), taking the form of non-dimensional ratios. These ratios should lie within certain bounds, which are detailed in Appendix A2 in Section 14.

The images that fail these criteria are removed from the pipeline process and reported as a 'segmentation failure' event. However, this is not necessarily the case that the segmentation process itself is flawed: the images failing these simple geometric priors include those that are erroneous – possibly due to participants uploading the incorrect image from their gallery – or poor-quality, where lighting, zoom factors, blurring, or poor choices of background contribute to results that are either meaningless for the purpose of diagnoses, or cannot be interpreted, often even by a human expert. Figure 4 shows examples of a successful segmentation and those detected as being out-of-distribution.

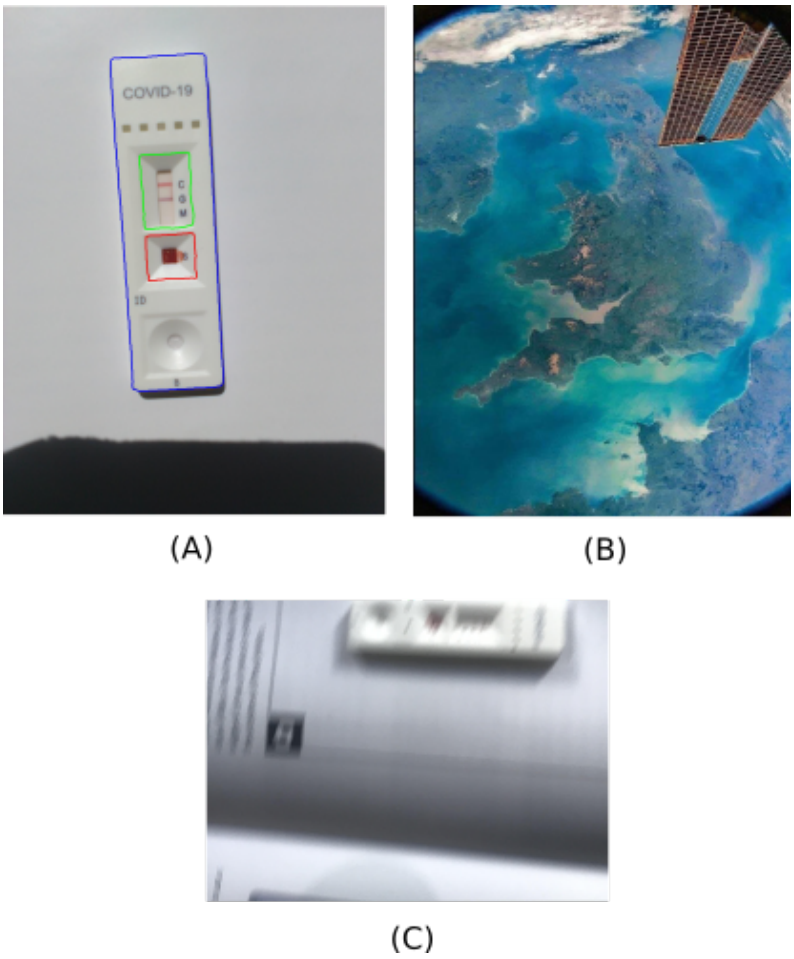


**Figure 4: Examples of good, meaningless and poor-quality participant-submitted LFIA images.** A: The LFIA images which are successfully segmented are taken at a moderate distance, have the whole device in focus, good lighting, and no shadows falling across the result window. This example also shows the output of the segmentation network. (Blue: LFIA, Green: Result window, Red: Blood well). B and C: Study participants can make mistakes when taking and submitting their LFIA images. These range from selecting an unrelated image for upload (B) to taking poor-quality LFIA images, such as being out-of-focus or unintentionally cropped (poor quality, C).

### 7.1.3 Coordinate Alignment

The coordinate alignment allows a normalised view of the read-out window to be presented both for human interpretation and automated data analysis. This step was found convenient for rapid reviewing of data for labelling and further analysis through colour-space transformations that ultimately were found applicable for a series of heuristic algorithms to interpret the read-out window.

The alignment process consists of two operations: rotation and resizing. The rotation parameter is used to apply an in-plane rotation correction to the image, which aligns the long-edge of the read-out window with the vertical image axis, resulting in the control line being positioned toward the top end. The rotation angle is estimated by constructing the vector pointing from the centre (geometric centroid of the ROI) of the read-out window to the centre of the blood well. Once rotated, we use bilinear interpolation to rescale the image region corresponding to the read-out window to be 100 pixels in length (long-length) and a variable height (short-length) to maintain the aspect ratio. The rescaling provides a convenient size for display and for constructing projection signatures. Figure 5 shows the intermittent steps.

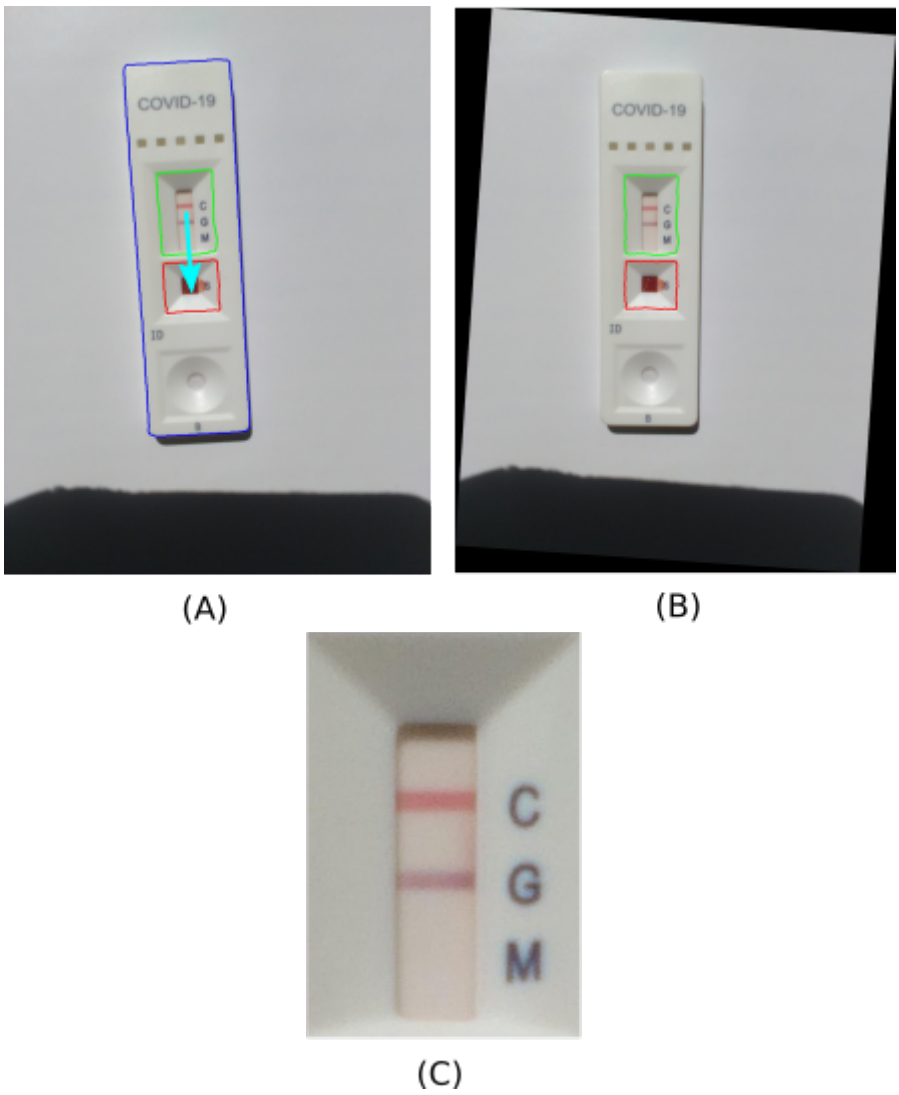


**Figure 5: Rotating and cropping the LFIA test result window.** A: The first step to align the images of the lateral flow devices is to calculate the centroids of the result window and blood well using the respective ROI coordinates. The vector from the result window centroid to the blood well centroid provides an indication of the device orientation with respect to the vertical axis. A rotation matrix to align the image of the device to a single reference axis can be formed. B: After rotating the image, the result window will be aligned to the chosen reference orientation; the same transformation can be applied to the relevant ROIs. C: The final step is to crop the test result window.

## 7.2 Read-out Algorithms (Interpreting Readings)

Having substantially reduced the size and complexity of the image data through the pre-processing steps, we are in a position to focus on the read-out window. Due to mapping the pixel data of the

read-out window into a reference axis $(x_\theta, y_\theta)$, we note that the image has a strongly linear structure. The alignment of this linear structure with the representation axis (i.e. row-column organisation of the image) means that we can apply low-complexity algorithms to perform several of the steps needed for interpreting read-out, without recourse to data-driven training.

The read-out window of the Fortress device provides a control line that switches colour from blue to red when the test has been correctly performed. Thus, a key stage that gates subsequent interpretation is the presence of the control line and the reading of its colour.

### 7.2.1 Colour Spaces

The raw RGB values of the pixel intensities corresponding to the read-out window can shift with changes in lighting, the colour calibration of the device and the presence of shadows. In practice, colour-space transformations to the raw pixel values can be a convenient way of obtaining at least partial invariance to nuisance sources of colour shift. We found it convenient to use three different colour transformations, depending on the need to be colour-selective or approximately colour-invariant.

We use ten colour channels across four colour spaces. These are (R,G,B) (original pixel values), (nR, nG, nB) (normalised colour spaces), where $nR = R/(R + G + B)$ etc, (H, S, V) and a single-channel opponent colour space, $O = nR - 2nG + nB$. Like the (H, S, V) colour space, opponent colour spaces correlate well with perceptual notions of colour. The $O$ channel, in particular, is recognisable as proportional to one channel of an Ohta (Tkalcic and Tasic 2003) colour space, which is an approximation to one channel of the Karhunen-Loeve (KL) decomposition of colours for natural imagery. The Ohta space is known to be a good choice for colour image segmentation (Kartikeyan, Sarkar, and Majumder 1998) and visual descriptors (Payne and Singh 2005).

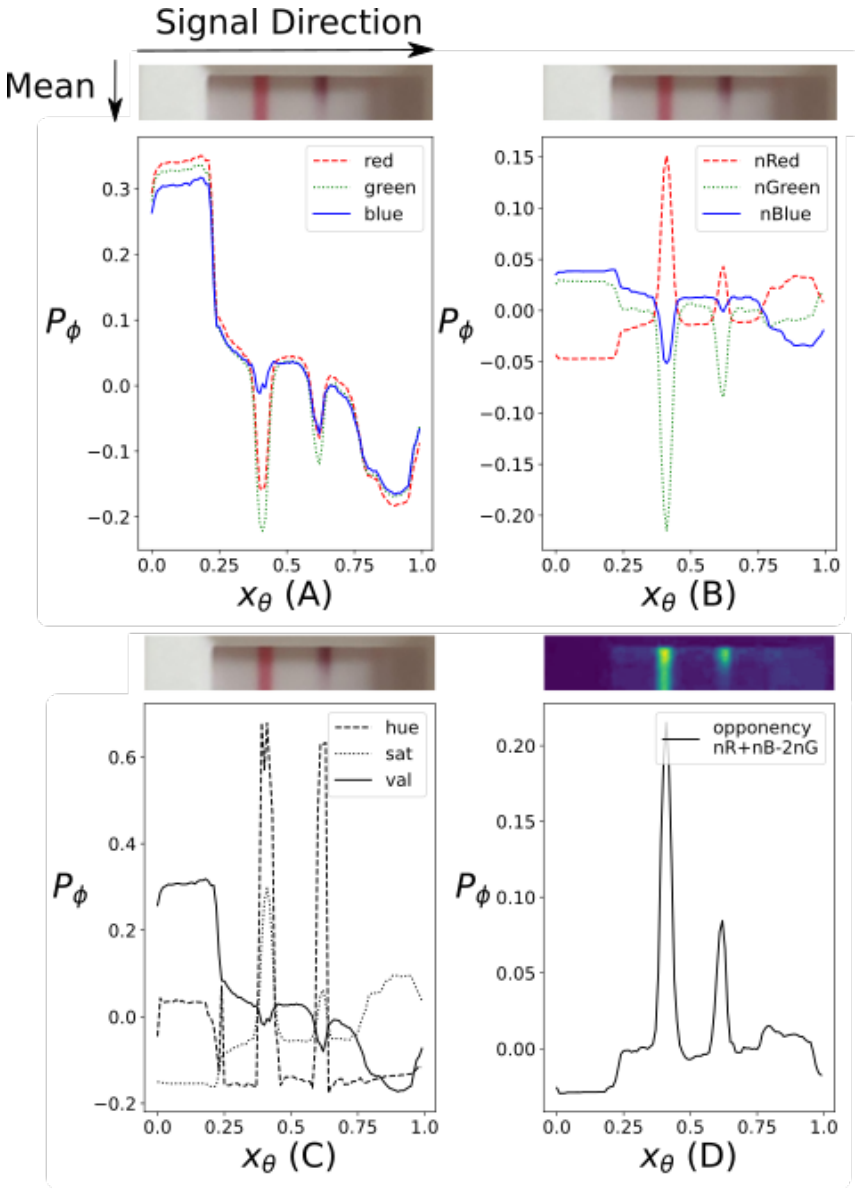


**Figure 6: Colour spaces and projection signatures.** The projection signatures are generated by averaging (taking the mean value of) the images pixel values across the short axis, which produces a signal in the long axis. A: Red, green, and blue colour channels, and their projection signatures, $P_{(R|G|B)}(x_\theta)$. B: Normalised red, green, and blue colour channels, and the signatures $P_{(nR|nG|nB)}(x_\theta)$. C: Hue, saturation, and value (intensity) channels, and signatures $P_{(H|S|V)}(x_\theta)$. D: The $O$ opponency channel, and its projection signature, $P_O(x_\theta)$.

### 7.2.2 Projection Signatures

Rather than using principal components analysis, we simply project (sum, or take an average) of the intensity data along the $y_\theta$ direction, yielding a function of $x_\theta$. These projection signatures, $P_\Phi(x_\theta)$ are produced for each channel of each of three selected colour spaces. Specifically, for any colour-spatial field in the aligned coordinate system $f_\Phi(x_\theta, y_\theta)$, for $\Phi \in \{R, G, B, nR, nG, nB, H, S, V, O\}$ we approximate the 2D to 1D projection operator:

$$P_\Phi(x_\theta) = K \int f_\Phi(x_\theta, y_\theta) dy_\theta \tag{1}$$

where $K$ is a normalising constant that compensates for different zoom factors. In practice, the integral is simply approximated by taking an average over the rows (or columns) of the two-dimensional image array representing the normalised image window corresponding to the read-out.

Illustrations of the projection signatures are provided in Figure 6. Due to the white background of the test read-out window, the presence of red or blue lines is somewhat counterintuitive in the native (R, G, B) space, manifesting as dips in intensity on the primary colour channels. The normalised channels provide a slightly more obvious depiction, and we can observe the differences in appearance between the control lines for these cases (see also Figure 7).

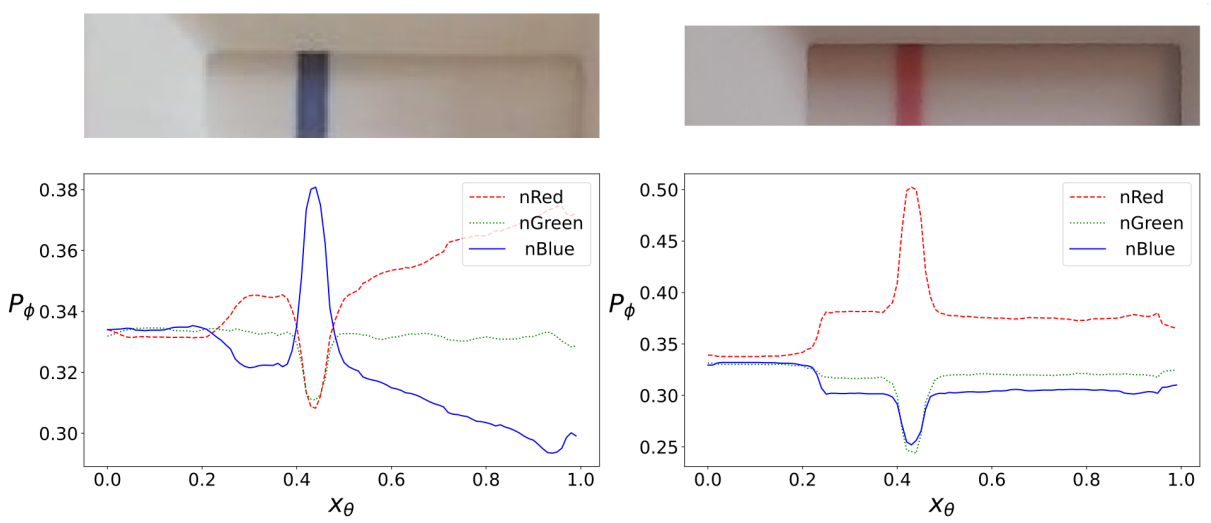


**Figure 7: Normalised RGB projection signatures for an invalid and valid LFIA.** The figure shows examples of invalid (left) and valid (right) LFIA test result windows and their respective nRGB projection signatures. Heuristic-B takes these signatures as input to determine the validity of the LFIA. The signatures are also used in the semi-quantitative analyses presented later in this report.

### 7.2.3 Detection of the Control Line

The first step to interpreting the LFIA result is to locate the control line. The pipeline uses the opponency signature (6, D) and runs a peak detection heuristic (see Figure 3, Heuristic-A) to find a peak corresponding to the control line (control-peak). Because of the rescaling operation to a reference coordinate system, Heuristic-A can focus peak detection on a defined region as all images are in the same orientation; and the geometry of the test result window is also normalised.

The control-line detection makes use of an off-the-shelf peak detection algorithm from *scikit-learn* (`find_peaks()`). This algorithm takes 5 parameters of height, width, relative height, distance, and prominence. The optimal settings for these five parameters was set by a grid search across three hyper-parameters $\alpha_h \in [0.9, 1.1]$, $\alpha_w \in [0.1, 0.5]$ and $\alpha_p \in [0.005, 0.016]$, then setting the *height* parameter of `find_peaks()` to $\mu \times \alpha_h$, *prominence* to $\mu \times \alpha_p$ the *width* parameter to $100 \times \alpha_w$; $\mu$ is the average amplitude of the projection signal. The parameter *relative_height* was fixed at 0.5, and *distance* was fixed at $100 \times 0.1$.

If Heuristic-A fails to locate the control line, then the LFIA result is deemed 'unreadable' and removed from the pipeline.

### 7.2.4 Reading the control line status

A 2nd heuristic algorithm (Heuristic-B) determines the validity by reviewing the normalised red (nR) and blue (nB) values at the control-peak location. For invalid tests, the $nB$ value is higher than $nR$. The relationship is consistently reversed for valid tests, and Figure 7 illustrates this. Invalid LFIAs are removed from the pipeline and reported as invalid test results, whilst the images deemed to be valid; move to the later stages of analysis. We used Dataset 2, Table 2, for testing Heuristic-B.

In testing, Heuristic-B classified all but one example correctly; sensitivity of 0.917 (99% CI: 0.700 - 0.990), and a specificity of 1.000 (99% CI: 1.000 - 1.000) for detecting invalids. The confidence intervals were generated using bootstrapping where we calculate the metrics on randomly oversampled samples (100,000 permutations) of the results, with replacement. A source of potential error are partially converted control lines, shown in Figure 8. Though Dataset 2 is small, and the results may not reflect the performance in the field, we can assume that the errors of Heuristic-B will be few if the proportion of partially converted control lines is small. For future work, we can collect the data that participants have labelled invalid and create ground truth, a more trivial task than labelling the IgG status of the LFIA. Heurisitic-B will then be replaced by a 1D or 2D CNN developed to classify validity.

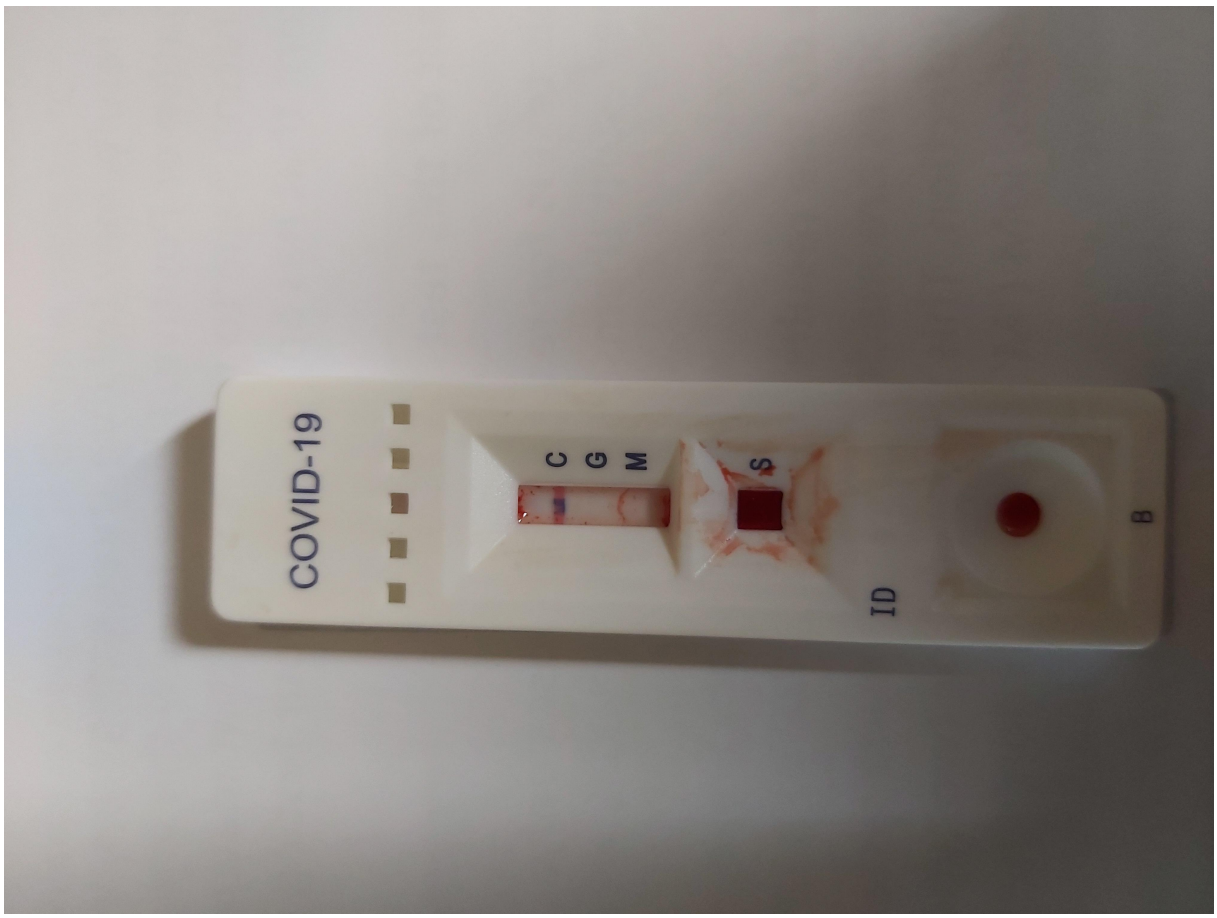


**Figure 8: Partially converted control lines.** Images containing samples such as that reproduced above are considered invalid, as the control line has not completely converted from blue to red. Invalid samples are flagged and removed from the pipeline before attempting to read the seroprevalence result.

### 7.2.5 Seroprevalence Read-out: IgG Status

Initially, with the limited amount of labelled data, we developed an algorithm (Heuristic-C) based on established peak-detection techniques to classify the IgG Status of samples. As the amount of data we labelled increased, more data-intensive methods replaced this approach, leading to the creation of first 1D and then 2D CNNs. The initial development and testing, Classification Experiment 1 (CE1), used Data set 3 in Table 2. The development set consists of samples from REACT-2 Study-3 and Study-5's R1. The test set came from Study-5's R2, removing any risk of data leakage; additionally, Study-5's R2 contained participant response data for comparison against both human experts and algorithm readings. For CE1, we did five-fold cross-validation on just the development set, ensuring that it was suitable for training; we then retrained on the whole development set and report the performance

of the final model on the test set. Reported metrics are specificity, sensitivity, overall accuracy and Cohen's kappa. These metrics are detailed in Appendix A1 in Section 13.

### 7.2.6 Heuristic approach with opponency and edge intensity signatures

Taking advantage of the linear structure of the read-out window, we designed Heuristic-C using the opponency signature and an edge intensity signature, which is only used for IgG status classification. The principles for detecting the control line are also applied; the heuristic looks for distinct peaks at identified locations. For a seropositive image, we require: one peak in opponency space, and to differentiate from possible blood leakage, two peaks in the edge-intensity space. These two peaks correspond to the two edges of the IgG band, while in the case of blood leakage, only one band typically appears. The opponency and edge signals are shown in Figure 9.

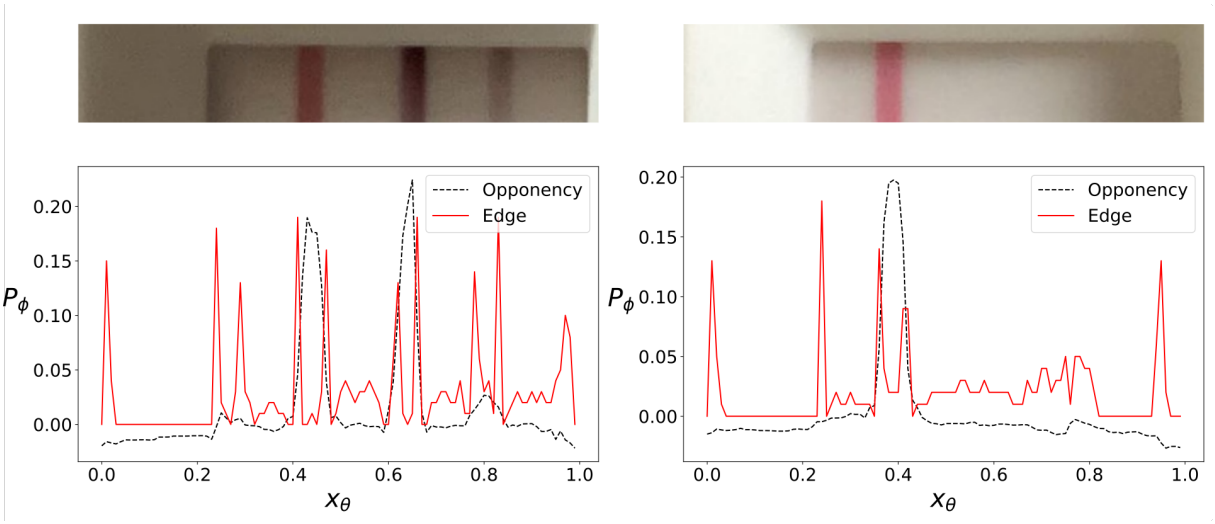


**Figure 9: Opponency and edge-intensity projection signatures for seropositive and seronegative LFIAs.** The figure shows examples of seropositive (left) and seronegative (right) LFIA test result windows and their respective opponency and edge-intensity projection signatures. Heuristic-C takes these signatures as input to determine the IgG result of the LFIA.

### 7.2.7 1D CNN with projection signatures

A popular deep-learning method is the 2D CNN; however, as seen, the LFIA read-out has a strong linear structure that can be collapsed into 1D signatures. We use ten signatures from different colour spaces, detailed in Section 7.2.2 with examples in Figure 6, as inputs to our 1D CNNs. The theoretical benefit is with fewer parameters, due to smaller networks, the models should be able to generalise better. Five 1D network architectures were developed, detailed in Table 8 in Appendix A4 (Section 16), which vary

in the number of convolutional filters, the filter lengths and the number of layers. These variations enabled us to explore how architecture parameters may affect the performance. The 1D CNNs were implemented with PyTorch (Paszke et al. 2019), and training was completed over 100 epochs, using binary cross-entropy loss, and a learning rate of 0.0013. The development set was split into a 90:10 ratio for training: validation and batch sizes set to 2. Several random generator seeds were used, providing a mean and standard deviation (across seeds).

Alongside the networks, we developed custom data augmentation routines for the projection signatures. These included shifts in $y_\theta$, scale changes and one-dimensional blur. These routines were selected as being appropriate augmentation routines for the one-dimensional signatures, reflecting the nature of the variations encountered in one-dimensional projection signatures.

#### 7.2.8 2D CNN with normalised read-out images

As more data was accumulated, we also trained a 2D CNN, implemented in Tensorflow, for read-out interpretation. This CNN utilises a MobileNetV2 (Sandler et al. 2018) architecture with pretrained weights learnt on ImageNet (Deng et al. 2009). The CNN was fine-tuned using the development set, over 100 epochs using a sparse categorical cross-entropy loss function, with a learning rate set to 0.001. Reported performance for this network is obtained from the test-set.

#### 7.2.9 Classification Experiment 1 Results

The results of CE1, shown in Table 3, are promising. With regards to specificity, all methods perform better than the study participants. However, this is not the case for sensitivity where the participants are near perfect, followed by the 2D CNN, 1D CNN models and finally, Heuristic-C. Looking at overall accuracy, the 2D CNN and participants perform similarly, both being better than the 1D CNNs and Heuristic-C. For Cohen's Kappa with the expert, we can see that the 2D CNN has performed the best, with the value representing 'almost perfect agreement.' These results are promising as they initially imply that a 2D CNN, in general, could perform at a level on par with experts and hence auto-validation for LFIA could be possible. Although the 1D CNNs are not at the standard of the 2D CNN, the method still shows promise as it has high specificity while being nearly 27-50 times smaller than the 2D CNN, see Table 4. In practice, a variety of networks, is likely to be be the most robust approach; in such settings, only specific types of disagreements between independently trained networks, would be flagged for human review.

With these results, we implemented the methods into ALFA and deployed it to analyse the 'wild' data consisting of 500,000 images. A pattern emerged, many discrepancies were flagged where participants reported seropositive, but ALFA reported seronegative. Experts reviewed a sample of the discrepancies,

**Table 3:** The results of Classification Experiment 1.

| Model/Heuristic/Participants | Specificity | Sensitivity | Accuracy | Cohen's Kappa |
|---|---|---|---|---|
| 2D CNN | 0.994 | 0.971 | 0.983 | 0.966 |
| 1D CNN Model 1 | 0.995 | 0.831 | 0.917 | 0.832 |
| 1D CNN Model 2 | 0.968 | 0.853 | 0.913 | 0.824 |
| 1D CNN Model 3 | 0.999 | 0.879 | 0.941 | 0.882 |
| 1D CNN Model 4 | 0.988 | 0.833 | 0.914 | 0.827 |
| 1D CNN Model 5 | 0.990 | 0.900 | 0.946 | 0.892 |
| Heuristic-C | 0.994 | 0.757 | 0.881 | 0.759 |
| Study Participants | 0.961 | 1 | 0.980 | 0.959 |

**Table 4:** The number of parameters/trainable parameters in the 1D CNNs and 2D CNN.

| Model | Total params | Trainable params |
|---|---|---|
| 2D CNN | 2,586,434 | 2,552,322 |
| 1D CNN (Model 1) | 90,131 | 90,131 |
| 1D CNN (Model 2) | 91,661 | 91,661 |
| 1D CNN (Model 3) | 52,051 | 52,051 |
| 1D CNN (Model 4) | 57,001 | 57,001 |
| 1D CNN (Model 5) | 51,251 | 51,251 |

discovering cases of 'weak IgG positives.' Detection of these cases is the objective of Classification Experiment 2 (CE2), which utilises samples from R5 of Study-5.

### 7.3 The Case of the Weak Positives – The Vaccination Solution

As REACT-2 Study-5 was ongoing throughout the pandemic, R5 was the first round where a small sample of participants had received their 1st vaccination dose. It was expected that the participants who received their first dose at least 21 days prior to completing the LFIA should be seropositive; however some reported seronegative tests. An expert review of these cases was completed, confirming the presence of weak IgG positive examples in addition to creating more training data. Data set 4 in Table 2 combines these new examples along with Data set 3. This new data set was used for CE2, where the networks were retrained and tested using the new data sets. Most training parameters remained the same from CE1 to CE2; however, the number of epochs increased from 100 to 150, and the training and validation batch sizes increase to 10 and 5, respectively.

Table 5 shows the results of CE2 and underlines the necessity of providing difficult cases in the training data. As the 2D CNN model trained in CE1 was the best performing method with a near-perfect agreement with an expert, we applied it to the test set of CE2. The results show that CE1's 2D CNN usually missed the weak positives; however, after retraining the model (CE2, retrained), we see improvement, that the cycle of iterative improvement is useful. This cycle involves applying the best

performing method to the 'wild' data, reviewing the discrepancies, identifying weak cases and retraining the network. Still, ground truth labelling is a labour-intensive task and requires a certain level of expertise.

Another vital message is that although the weak positives have caused the performance of the methods to decrease compared to expert reviews, we see that all methods, including Heuristic-C, perform better that the study participants.

**Table 5: The results of Classification Experiment 2 (CE2) as well as the performance of the 2D CNN, trained in Experiment 1 (CE1), on the new test set.**

| Model/Heuristic/Participants | Specificity | Sensitivity | Accuracy | Cohen's Kappa |
|---|---|---|---|---|
| 2D CNN (CE1) | 1.000 | 0.852 | 0.949 | 0.883 |
| 2D CNN (CE2, retrained) | 0.987 | 0.901 | 0.958 | 0.905 |
| 1D CNN Model 1 | 0.969 | 0.825 | 0.920 | 0.701 |
| 1D CNN Model 2 | 0.965 | 0.810 | 0.912 | 0.667 |
| 1D CNN Model 3 | 0.968 | 0.844 | 0.926 | 0.733 |
| 1D CNN Model 4 | 0.941 | 0.854 | 0.911 | 0.678 |
| 1D CNN Model 5 | 0.960 | 0.852 | 0.923 | 0.726 |
| Heuristic-C | 0.976 | 0.487 | 0.815 | 0.525 |
| Study Participants | 0.988 | 0.415 | 0.800 | 0.484 |

# 8 Deployment to ALFA and analysis of Study-5's Survey data

After settling on the best combination of algorithms and networks, we deployed them to the ALFA pipeline and analysed the 500,000+ images in the Study-5 dataset. Prior work (Ward et al. 2020) had found a declining crude prevalence of antibody positivity to SARS-CoV-2 in the first 3 rounds of Study-5; this is illustrated in Figure 10, ("REACT crude"). However, the ALFA pipeline, and more specifically CE2's 2D CNN, has found an increase in the crude prevalence in R3 from R2. A hypothesis for this discrepancy could be: as R3 coincides with the increase in daily new cases, as shown in Figure 11(data provided by GOV.UK(GOV.UK 2021)), the antibodies have not fully developed due to more recent infection (2-3 weeks to develop post infection) and hence there is a significant number of weak positives. A portion of these weak positives is being reported as negative; however as the ALFA pipeline is sufficiently sensitive to detect a proportion of the weak samples, the trend it suggests differs from user-reported curves of crude prevalence. For the same reason, weak positives could also explain why ALFA's crude prevalence is consistently higher than if relying on just the user data.

To confirm this hypothesis, it would be worth reviewing a sample of discrepancies where ALFA has deemed seropositive but the participants have reported a negative result.

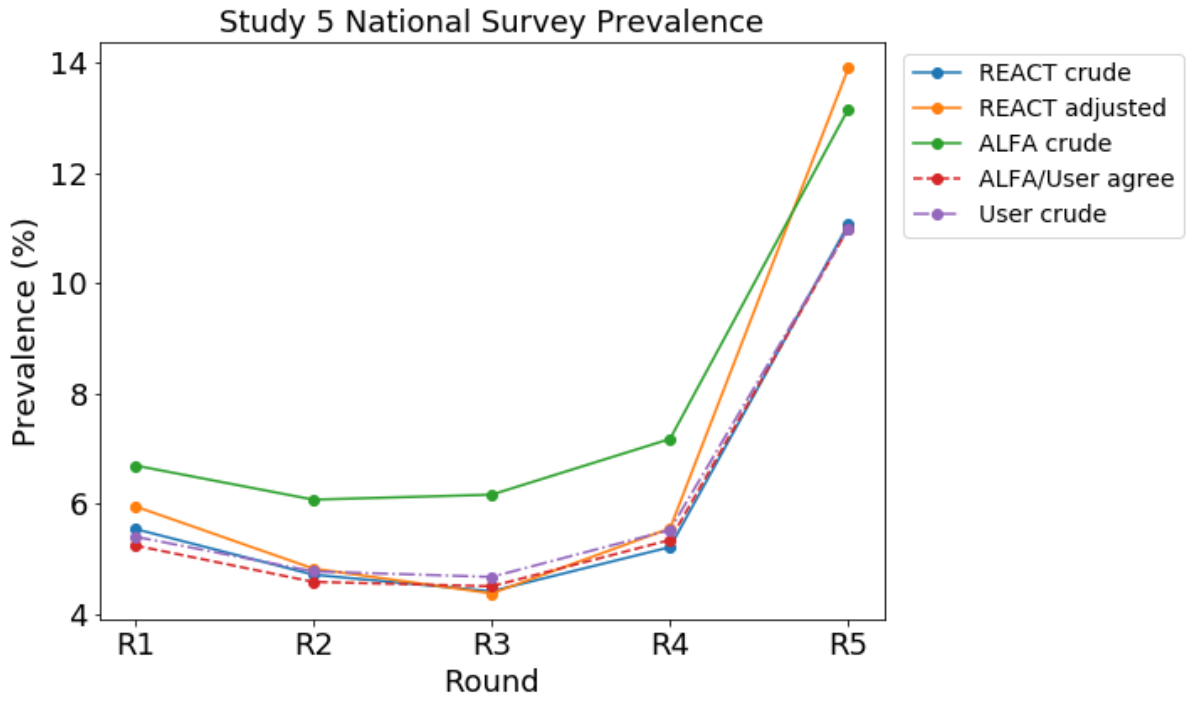


**Figure 10:** **Crude prevalences for Study-5.** "REACT crude" is the prevalence found by the REACT teams using **all** participant supplied data; while "User crude" is the prevalence based on the participants who had submitted images. "User crude" is used as comparison against "ALFA crude" which is calculated from ALFA's interpretations of the LFIA images. "ALFA/User agree" is the prevalence in samples where ALFA and the participant agrees and finally, "REACT adjusted" is the prevalence that they report after making adjustments for various measurement and sampling-related factors.

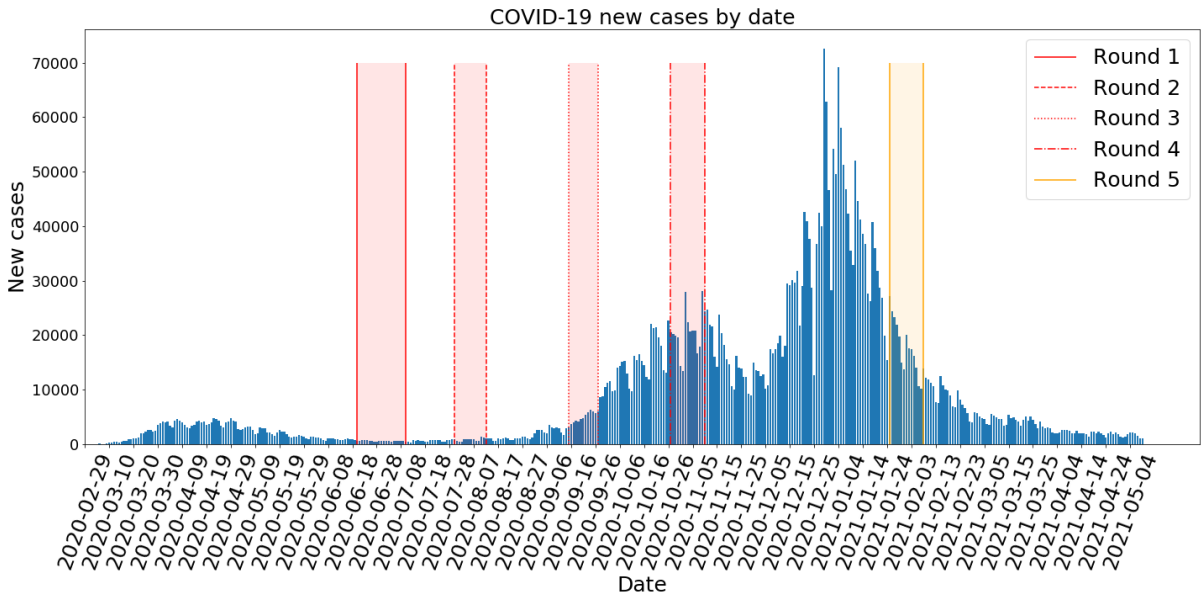


**Figure 11: The timeline of REACT-2 Study-5 in relation to the number of new cases per day during the pandemic.** The data for this figure was taken from GOV.UK's coronavirus dashboard, accessed on 10/05/2021.

## 8.1 Assessing the usability of the LFIAs

We calculated Cohen's kappa between the study participants and ALFA, which quantifies the participants' ability to interpret the LFIA outputs; this assessment is possible, as the 2D CNNs deployed in ALFA perform with 'almost perfect agreement' with the human-experts. The analysis shows that the study participants can interpret the LFIA read-out well, with kappa values indicating substantial agreement. Between rounds the kappa value does change, with a significant drop from R1 to R2 (most likely due to the origins of the training data); however if we collate all the samples, Cohen's Kappa is 0.797 (99.9% CI: 0.7966 to 0.7968). Figure 12 shows a plot of the kappa values for each round for 3 of the algorithms developed.

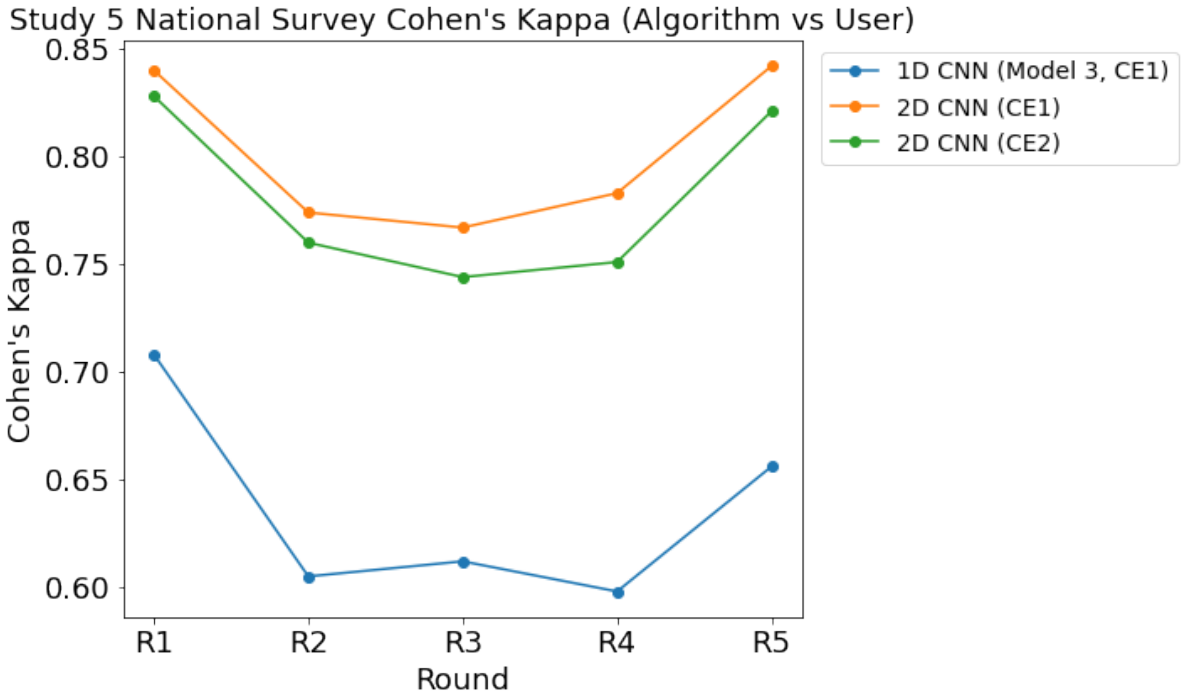


**Figure 12: Cohen's Kappa between ALFA's algorithms and study participants for Rounds 1 to 5.** 2D CNN CE2 is the most sensitive algorithm and more likely to pick up on weak positives, hence the lower Cohen's kappa compared to 2D CNN CE1. This is in contrast to the currently best 1D CNN, which is less sensitive than the 2D CNNs, producing significantly more false-negatives, and hence a lower Cohen's kappa.

## 9 Semi-Quantitative Analysis

Using the component-wise design of the pipeline to creating an analytical pipleine allows us to intercept and use intermediate results to extract information that is beyond a standard classification setting of the main pipeline. These intermediate results can provide useful analytical tools, and we have used this during development to identify consistent misreadings and occasional issues with early labelling, leading to low tolerance around the extracted read-out windows; these were subsequently rectified through re-labelling of the regions corresponding to the read-out window.

It is theoretically possible that sub-threshold responses may be detected, particularly in a cohort, which might provide cohort-level detection of weak levels of infection. In Figure 13, average projection signatures from participant-supplied diagnostics have been plotted, showing clear examples of divergence from negative responses (blue trace) associated with IgG only and IgM only positive responses.

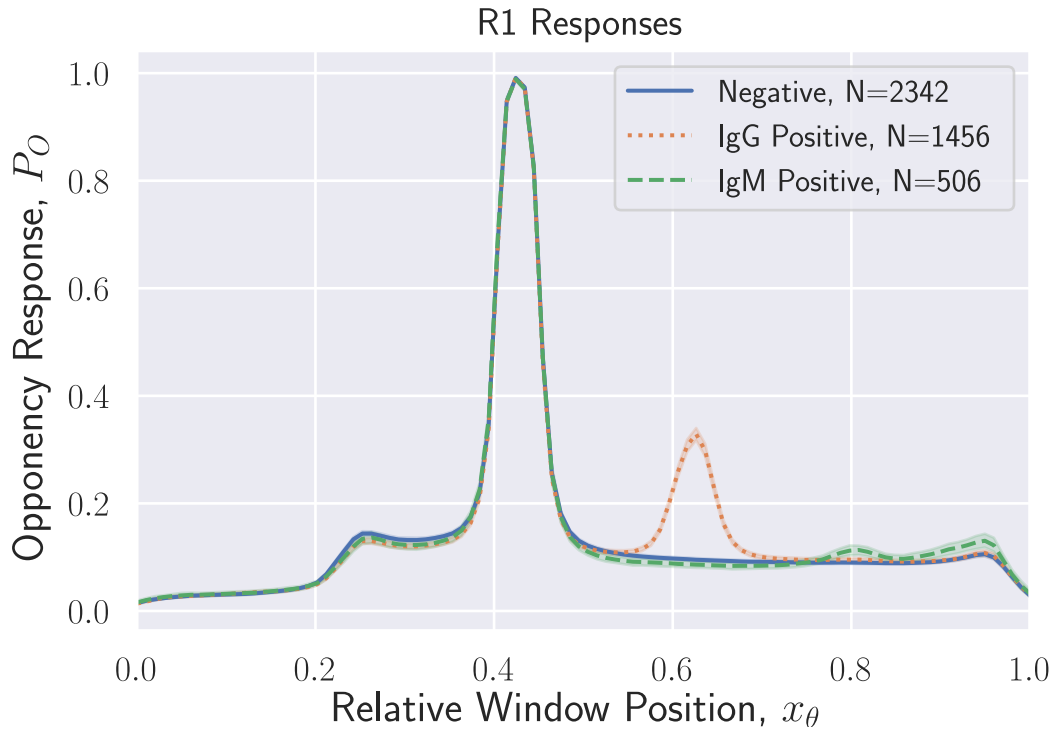


**Figure 13: Illustration of opponency projection signatures.** The opponency signal provides a condensed way of assessing the immunoassay responses acquired from the participant-submitted windows. Here, we select a subset of unverified responses from user submissions in R1, which are labelled as negative, IgG positive or IgM positive. Note that the average signatures are substantially different, and are correlated with position, $x_\theta$ along the length of the read-out window. IgG and IgM 'responses' are read out from the indicated regions; alignment along the horizontal axis tends to be dominated by the location of the control line, the peak visible in all conditional responses near to $x \approx 0.44$, with its centre indicated by the red line. Note that in this figure, the IgM response height has been amplified for the purpose of illustration; other responses - control line and IgG - are unaltered from the raw data.

## 9.1 Ensemble Analysis – Introduction

The projection signatures, illustrated in Section 7.2.2, provide a simple approach to analysing the result of the read-out window, provided that it has been detected and aligned to a reference coordinate system. In Section 7.2.2, we showed single examples from individual read-out windows, to illustrate how the projection signatures can be used for rapid processing.

One can also analyse *ensembles* of projection signatures, once the test has been deemed to be valid; in Figure 14, we show a sample of 10 signatures retrieved from valid tests (those where the control

lines are detectable and are red). Note that although the control peak can be easily identified, there is some jitter to its position. Much of this jitter is due to translational shifts in segmenting the detector window.

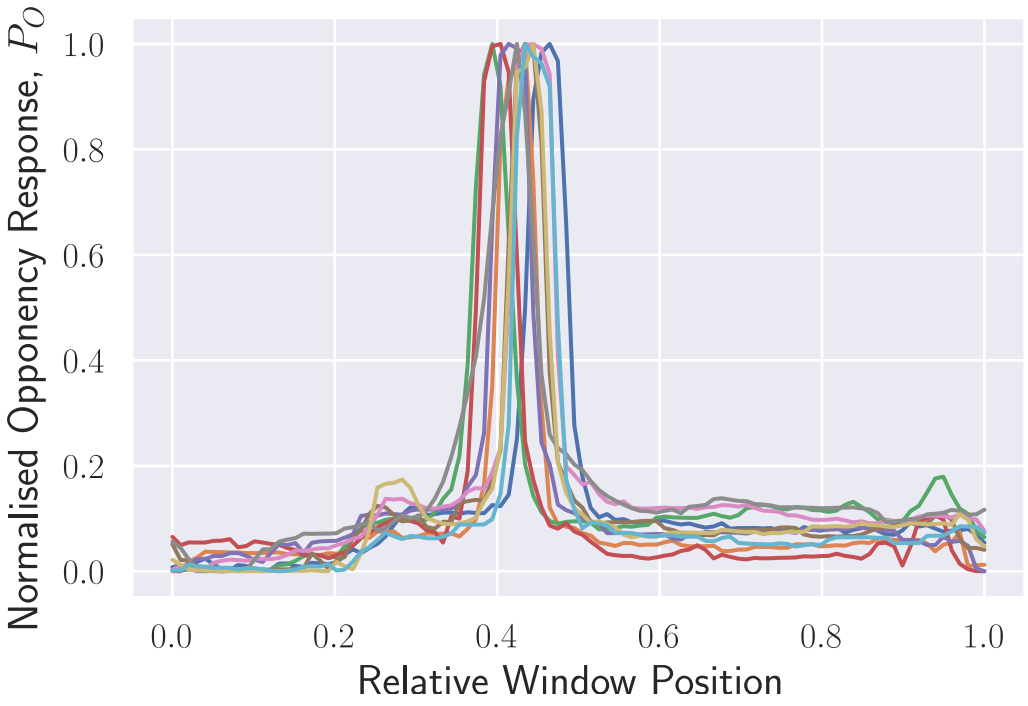


**Figure 14: Non-aligned projection signatures.** This sample of 10 projection signatures illustrates what may be described as an ensemble of projections. The jitter is caused by small errors in localising the read-out window during segmentation.

It is useful to align signals to the control peak; this allows the positions of the IgG and IgM read-outs to be better aligned, simplifying subsequent algorithms that may, for example, be used to seek population-level effects in the data, without using either a user-response or an explicit detection (e.g. the CNN-based classifiers of Section 7.2). We now illustrate the steps required to perform this analysis.

## 9.2 Aligning Signals

By computing an ensemble average over all projections from a sample of just over 12,000 images from valid tests, obtained from REACT-2, Study-5 Rounds 1 and 2B, we are able to estimate an average template for the appearance of projection signatures; this is shown in Figure 15.

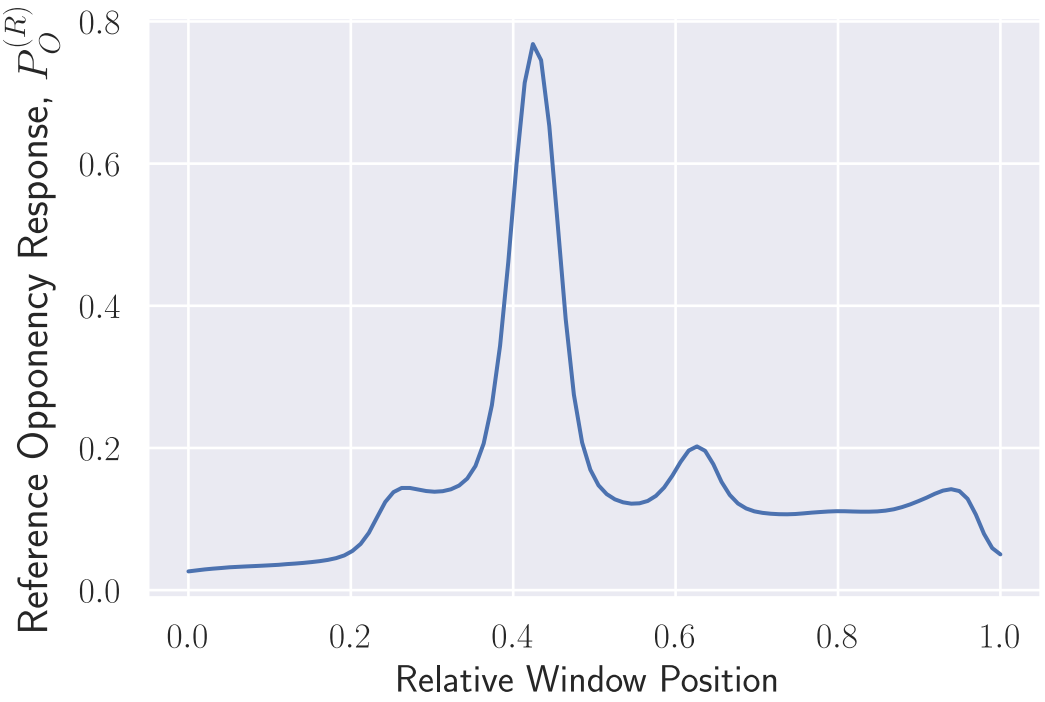


**Figure 15: Signal template, used for alignment.** This signal is produced by computing an average over an ensemble of over 12,000 un-aligned opponency projection signatures. The clear pattern produced by this averaging process yields a signal that can be used for the alignment of individual projection signals.

Cross-correlation can then be used to align individual projections to the reference window. The result of this is illustrated in Figure 16; clearly, the dominant effect that this process achieves is to register the signal peaks for the control line to a well-defined location in the position window. This alignment facilitates the reading of signal levels around the IgG and IgM locations, and averaging of projection signatures to detect changes that might be below detection level at the level of individual readings. We use the aligned signals for subsequent analysis, and in particular in looking at values of the height of these projection signatures that could be associated with changes in the immune response.

In performing cross-correlation, we also found that suggested shifts that were large were often associated with images containing anomalies, particularly associated with blood leakage. We explore this further below.

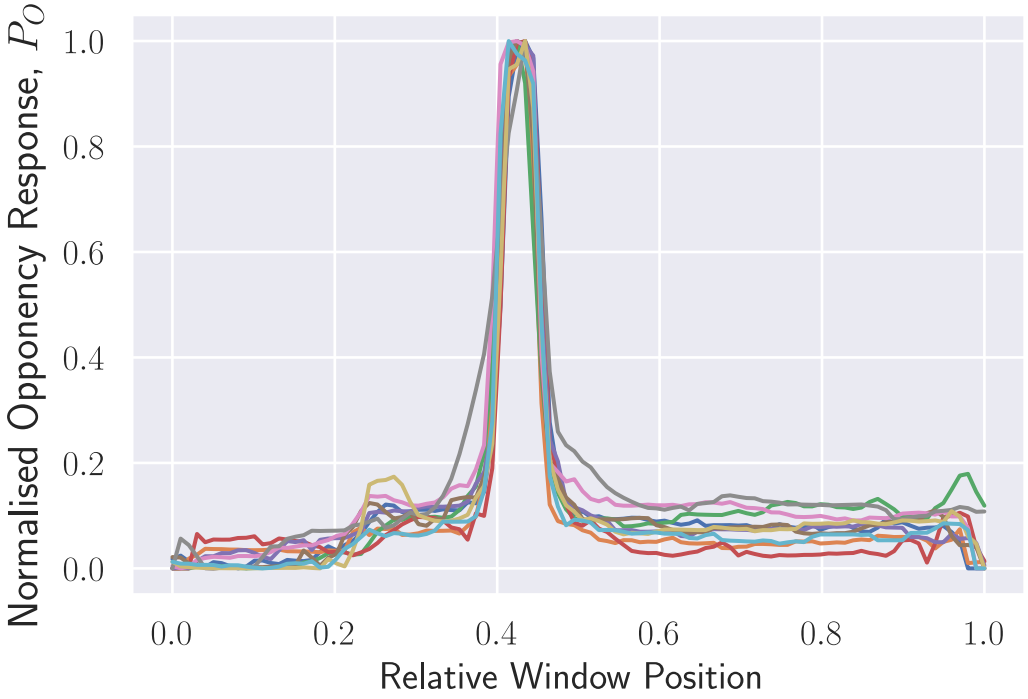


**Figure 16: Aligned Projections.** Using cross-correlation, the 10 signals shown in Figure 14 are aligned to the reference signal shown in 15. The result of this process is that the peaks of the control signals are now registered very tightly along the window position axis. This facilitates both reading of heights of peaks around the IgG and IgM locations, and averaging of projection signatures to detect changes that might be below detection level at the level of individual readings.

The process of ensemble averaging is illustrated - after the signals have been aligned - in Figure 17. The averaging can also be repeated for signals that have certain user, or expert interpretations, or even algorithm interpretations. This can be a useful diagnostic, or a tool for interpretation; indeed, we used this process to identify a key cause of error in participant-reported results.

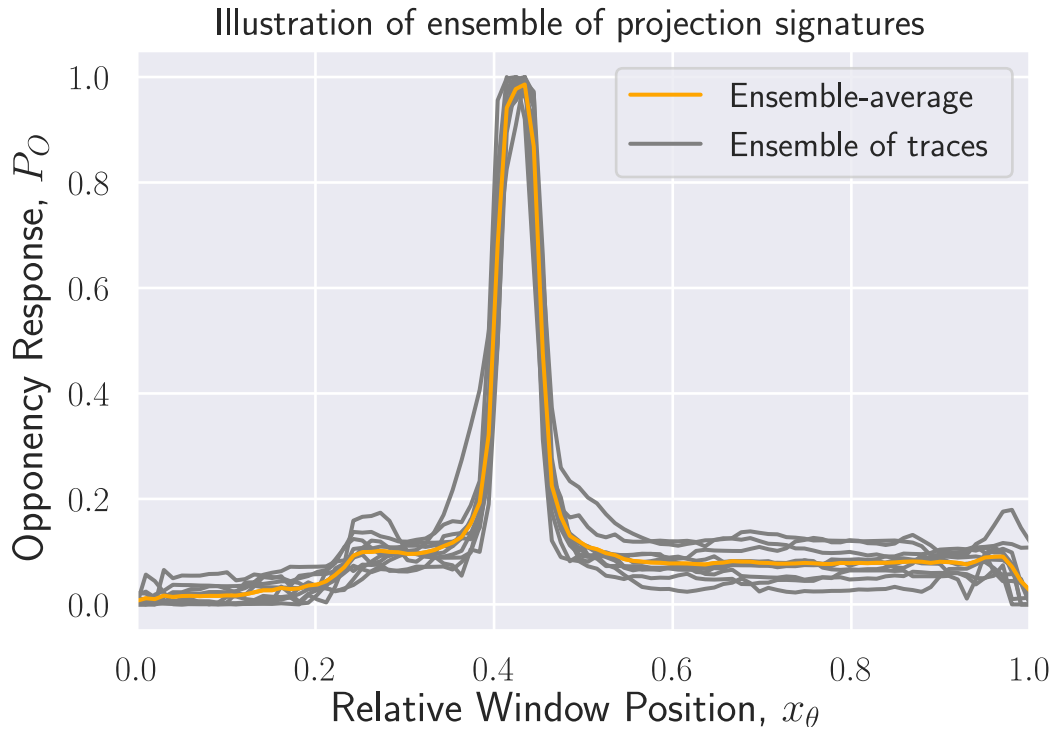


**Figure 17: Ensemble averaging of aligned signals.** This illustrates the result of calculating an ensemble average over 10 aligned traces; the smoothness of the orange trace is immediately obvious, and this tends to significantly enhance the ability to find discrepancies in results.

Two final remarks are in order about the process of alignment:

- Once individual traces have been aligned using the template derived from non-aligned projections, the template can be re-estimated. In practice, this sharpens the peak of the control line in the reference signal; we did not find using this sharper template to have a significant effect on the alignment depicted in Figure 16, so did not pursue the iterative strategy that might suggest itself.
- We found that large shifts during the alignment process, suggesting the control line peak was outside of a region spanning 1/3 the length of the read-out window and centered on the 0.44 position (see Figure 15), were often indicative of irregularities in the projection signatures, such as instances of blood leakage. We found that the presence of such large alignment shifts could be used to flag a significant proportion of anomalies in data, many of which were found to be associated with significant blood leakage into the read-out window. Checks for these large shifts were therefore used to partially filter out anomalies; further outlier detection was based on dot product similarity measures against the average signatures.

## 9.3 Detection of Waning

The response waning studies that we describe below are distinct to prior work that has investigated waning immune response through the use of home-testing. Specifically, whilst prior studies have considered the *proportion* of self-reported positive test case, here, we consider different ways of harvesting potential signs of waning, even in subjects who self-report positive test results.

In the three studies below, we consider those who reported positive test results in REACT-2, Study-5 R1; we analyse responses from a proportion of subjects at a minimum of one month later, restricting ourselves to those who *also* report positive test results in the second test.

### 9.3.1 Ensemble-averaged conditional responses

Although clear evidence has been seen for a decrease in the proportion of positive results by participants who were tested one month apart, it is possible to perform conditional ensemble-averaging, grouping signatures from specific cohorts, partitioned according to self-reported results, and generating ensemble-averaged signals.

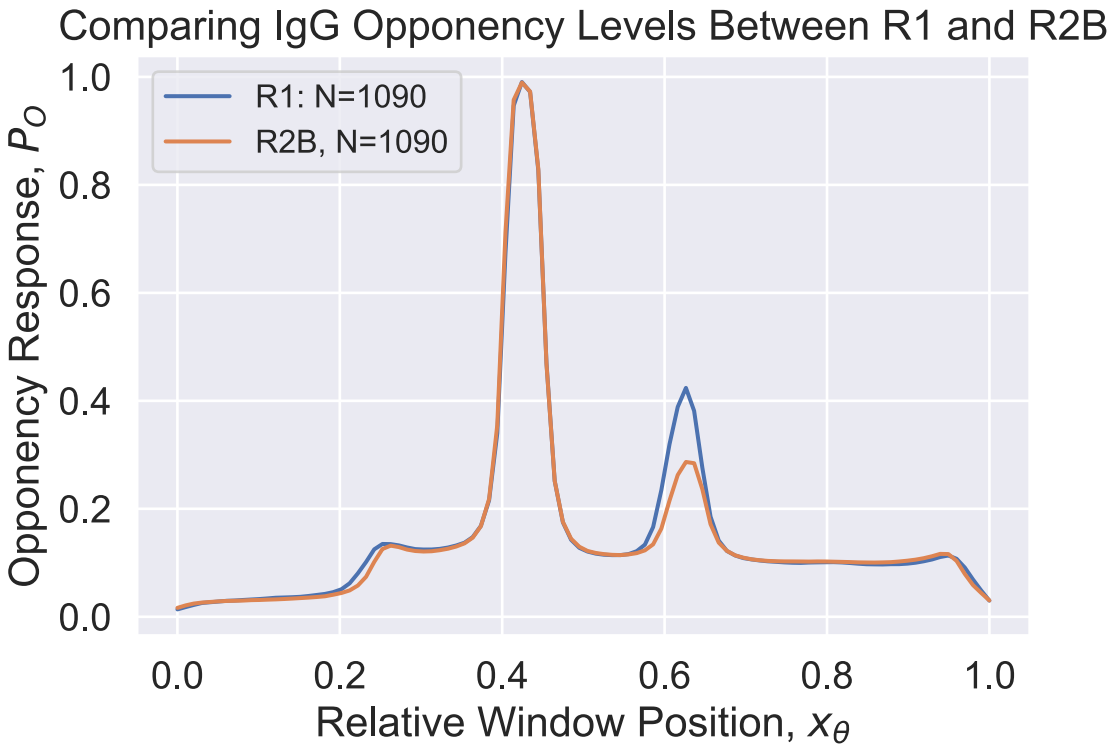


**Figure 18: Evidence of waning in ensemble averaged projections.** Here, we select a subset of unverified responses from user submissions in REACT2, Study-5, R1 and subsequent home testing in a subcohort of the same individuals. There is some evidence of a reduction in the the amplitude of IgG peaks, despite the fact that all participants self-reported positive tests. Global changes between rounds are likely to be captured in the control peaks, where we can note a small reduction in the heights of the control lines, in a direction opposite to that of the IgG responses; indeed, correcting for the control line amplitude is likely to increase the difference in IgG responses. We consider the effect using paired comparisons in the next section.

Given these differences, we captured amplitudes of the opponency signals in a region spanning approximately one tenth of the total read-out window length centred around the peaks of the registered signals corresponding to the IgG response line. In extracting these measurements, we rejected projection signatures for which alignment was considered indicative of anomalies, and also rejected outliers where the best alignment yielded signatures that had a normalised dot-product similarity score with the mean projection signature of less than 0.85. We performed kernel density estimation on the amplitudes of the remaining signals, using a Gaussian kernel, with bandwidth parameter 0.05. The distributions of these measurements are shown in Figure 19.

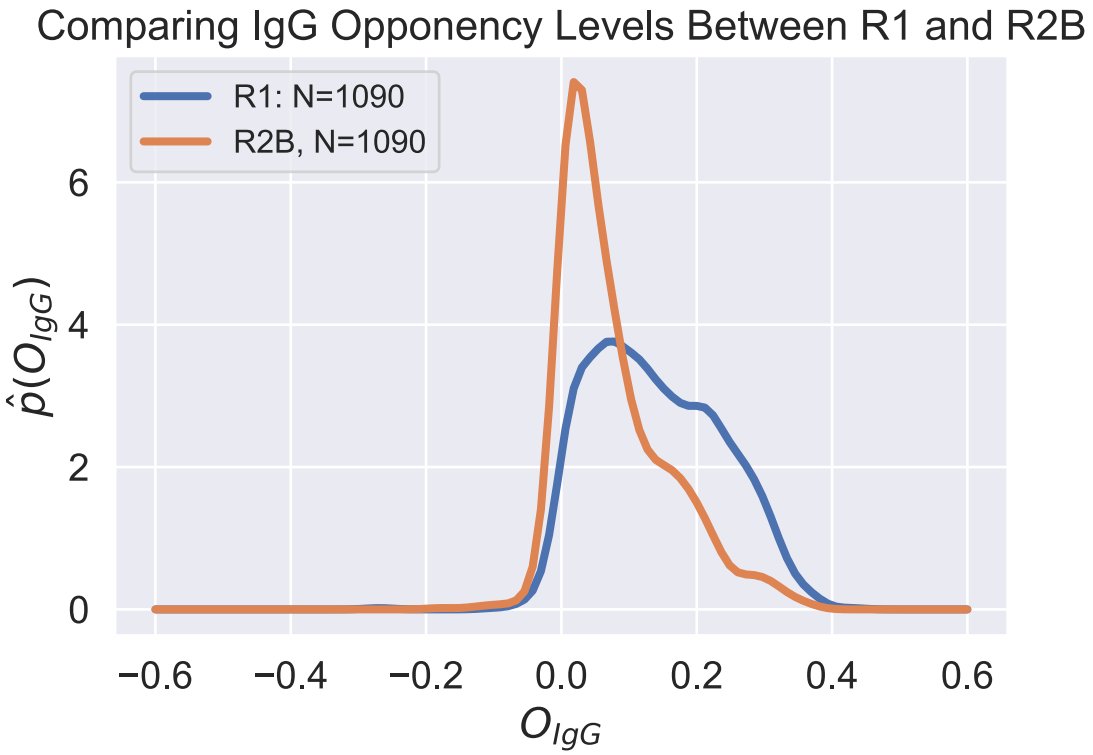


**Figure 19: Unpaired Comparison of IgG responses taken one month apart.** By extracting amplitudes in locations around the IgG peaks in aligned signals, we can compare the amplitudes of colour-opponent signal levels through repeated home-testing in participants, taken a minimum of 1 month apart. This suggests a strong shift in the distribution of amplitudes - represented on the horizontal axis - downwards, toward 0. Pairwise comparisons are performed in the next section.

Figure 19 suggests a significant alteration in the distributions of amplitudes of opponency signals taken 1 month apart; though normalising to peak values, and the close agreements of control line amplitudes give some confidence that amplitudes are indeed comparable, caution should still be exercised in interpreting such results. Subsequent pairwise comparisons, and a separate study comparing responses in the same round, and in sub-cohorts with different times of infection lend support to the validity of an approximate quantitative relationship.

### 9.3.2 Paired Testing

We can also look at paired responses, matching each positive-reporting individual with their subsequent self-reported home-test results. Due to the large number of comparisons, we chose to randomly draw 9 samples, each containing 20 individuals to provide a more accessible illustration of the general trend

of IgG change from first to second readings.

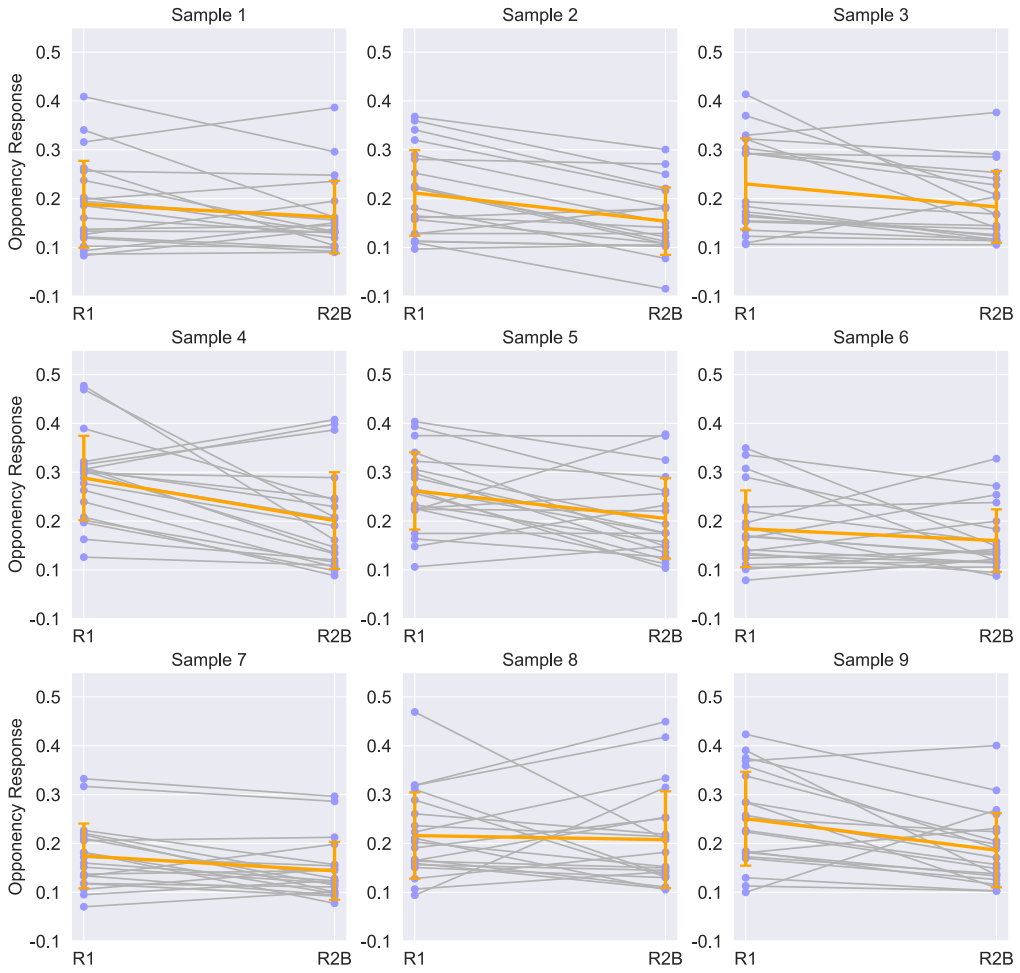


**Figure 20: Paired Comparison of IgG responses.** By extracting the amplitudes of projection signature in locations around the IgG peaks in aligned signals, we can compare the opponency channel projection signature levels between repeated home antibody testing. Here, participants repeat the test a minimum of 1 month apart. We used a bootstrap-type of approach to explore, in an intuitive way, the consistency of these changes by sampling 9 groups of 20 individuals at random from 1,090 individuals for whom we have paired responses that are valid, and for which alignment of the control peaks and signals are within tolerance limits (see text for details). Though there are some paired readings showing increases, the general trend of a decrease is present even when taking averages across these small “minibatches” of randomly selected participants. Orange lines represent the average response change over each of the randomly selected batches of 20 participants, with error bars indicating 1 standard deviation. The statistical significance of this change is reported for all participants in the text.

Figure 20 illustrates a general trend in which a decrease in the amplitude of IgG signals is the predominant trend between R1 and R2B. There are some instances of *increases*, but the infrequency of these occurrences and lack of formal ground truth assays makes it difficult to state for certain that these represent genuinely increased responses or not.

We performed a two-sided, pairwise Mann-Whitney-Wilcoxon test (`scipy.stats`), against the null hypothesis of there being no difference in the opponency amplitudes corresponding to the IgG line; this yielded a rejection of the null hypothesis at a significance level of $p < 1e^{-5}$. Indeed, in our most conservative exploration of such pairwise comparisons, using a separate cohort of positive self-reporting additional subjects for which the presence of simultaneous IgM lines suggested more recent infections, we obtained our most conservative $p$ value of 0.0051. We explore this sub-group further in the next Section.

### 9.3.3 Early vs Recent Infections

Though several factors – quantities of blood and timing of taking a photograph of the test result – may change the response level, we have found strong evidence that the strength of the projection signature of the opponency channel, $O$, over a cohort can provide an indication of the strength of response. This makes use of ensemble-averaging of the projection signatures, and a pair of studies from REACT-2, R1 and R2B, for which the same participants were tested twice at an interval of around 1 month.

We hypothesized that, after removing signals corresponding to incorrect interpretation of IgM positive results induced by blood leakage, participants who displayed both IgG and IgM positive responses in R1 would have been recently infected, whilst those who were IgG positive only in R1 were infected earlier. Taking participants that reported being positive in **both** rounds, we compared the strength of response around the IgG peaks for these two groups extracted from the participants' submitted data of their R2B tests. The probability density functions for amplitudes extracted from the respective opponency signals taken from R2B images are plotted in Figure 21. The two distributions appear different, and provide evidence of waning of the immune response amongst those participants that were infected earlier than significantly before June, R1 and those that were infected closer to the time at which R1 was taken. The shift is in a direction consistent with observations of waning estimated from numbers of participant-read responses, but these shifts were also found to be *independent* of the candidate's positive self-reading status for R2B.

A two-sided, two-sample Kolmogorov-Smirnoff test (*SciPy* 1.16, `ks_2samp`), applied directly to the samples, yields a significance level of $p < 0.03$, suggesting that we can reject the null hypothesis that the IgG opponency samples taken from R2B under two different reported states of R1 show no shift. The implication is that participants who were more recently infected - suggested by the continued presence of the IgM line *in addition to the IgG line* in R1 returns - display slightly different opponency readings in R2B; the trend is toward an increase. Though a small effect, there is some confidence in

this observation. Because the comparisons are made from measures taken within the same round of study; possible batch - or timing - related effects, which are possible confounding factors for the R1/R2B comparison in Sections 9.3.1 and 9.3.2 cannot be in play. Though the data plotted in Figure 21 come from samples acquired during R2B, the *conditions* being compared within the two curves are separated not by data from R2B, but by those from R1.

A few final remarks on these waning signs are in order:

- we have seen in Section 9.3.2 that the direction of the effect is toward a decrease in opponency signal over time, and the curves in Figure 21 suggest that this decrease is smaller between two tests in subjects who have been recently infected.
- the direction of the difference shown in Figure 21 is likely to be quite dependent on the timing between tests, and the time of infection between the two rounds, and the rate of development of IgG response, which is known to vary between individuals.

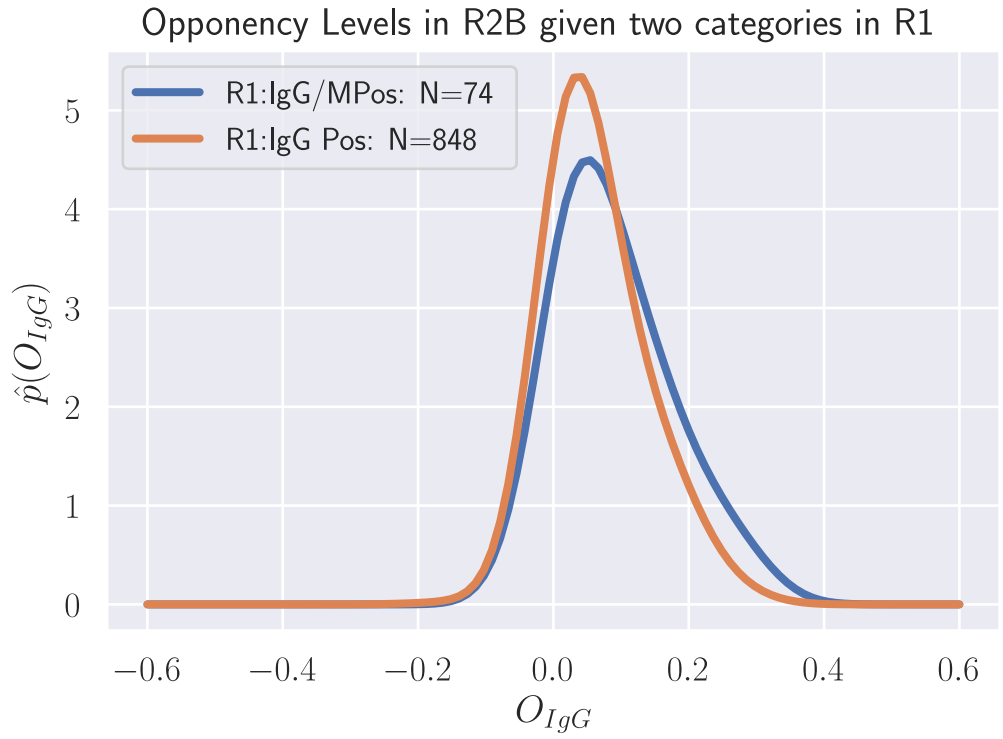


**Figure 21: Illustration of Probable IgG Response Waning.** The projection signatures described in Section 7.2.2 show changes that are likely to be due to immune response associated with waning. Earlier studies inferred waning by looking at the changes in the proportion of participants reporting a positive response. But these show that an effect is observable even for those participants that self-reported as positive in R2B, having also self-reporting as positive in R1. Those that were more recently infected (blue trace) have a higher amplitude of projection signature response - a slight shift to the right - than those that were infected earlier (orange trace); the evidence for time of infection was simply distinguished by the co-presence/absence of IgM response in the earlier (R1) test result.

## 9.4 A source of false-positives

Instances of blood leakage are hopefully rare, and this might be expected to spoil some results, and perhaps in an unbiased way. However, during the process of generating class-conditional projection signatures, we observed a strong difference in profiles, correlated with user response (see Figure 22). The odd shape of this curve showed a strong opponency response at the very end of the trace. These were not present in other conditional plots, i.e. either the IgG positive, or the negative plots. It turned out that there was a causal relationship between the appearance of the elevated end section of the trace (close to the $x_\theta = 1$ end) and the self-reported IgM readings.

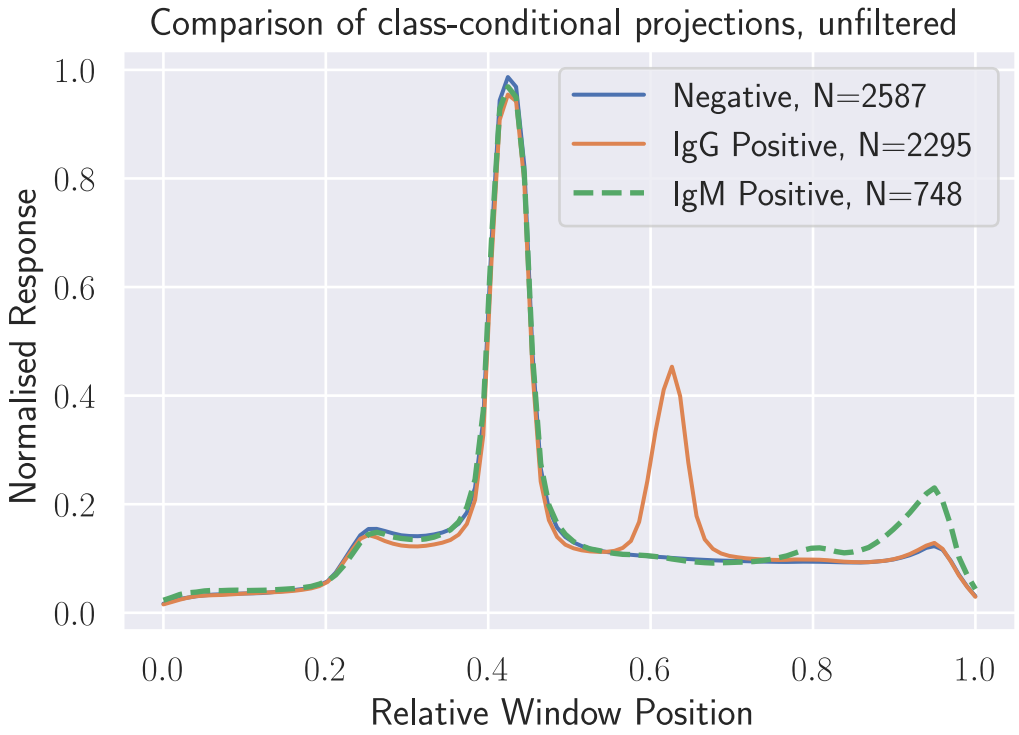


**Figure 22: Illustration of opponency projection signatures.** Opponency Signals, *before* filtering for anomalies. Note that there appears to be a very large peak beyond the location of the IgM region; this figure should be compared with Figure 13, which also provides indications of the locations of IgM and IgG lines.

Closer examination revealed that the appearance of the elevated opponency signal near the $x=0.95 location was overwhelmingly due to blood leakage, misinterpreted by many participants as positive IgM readings. Because this effect might be dependent on manufacturer or batch, ensemble projection responses allows the quality of devices and their usage to be quickly assessed at scale, with respect to the responses submitted by participants.

We subsequently used simple checks on the signals to detect and remove these traces. Many of the images flagged in this process contained examples, shown in Figure 24, of blood leakage or represented damaged devices. By eliminating these signals, it is possible to significantly reduce these artefacts from the ensemble-averaged projection signatures.

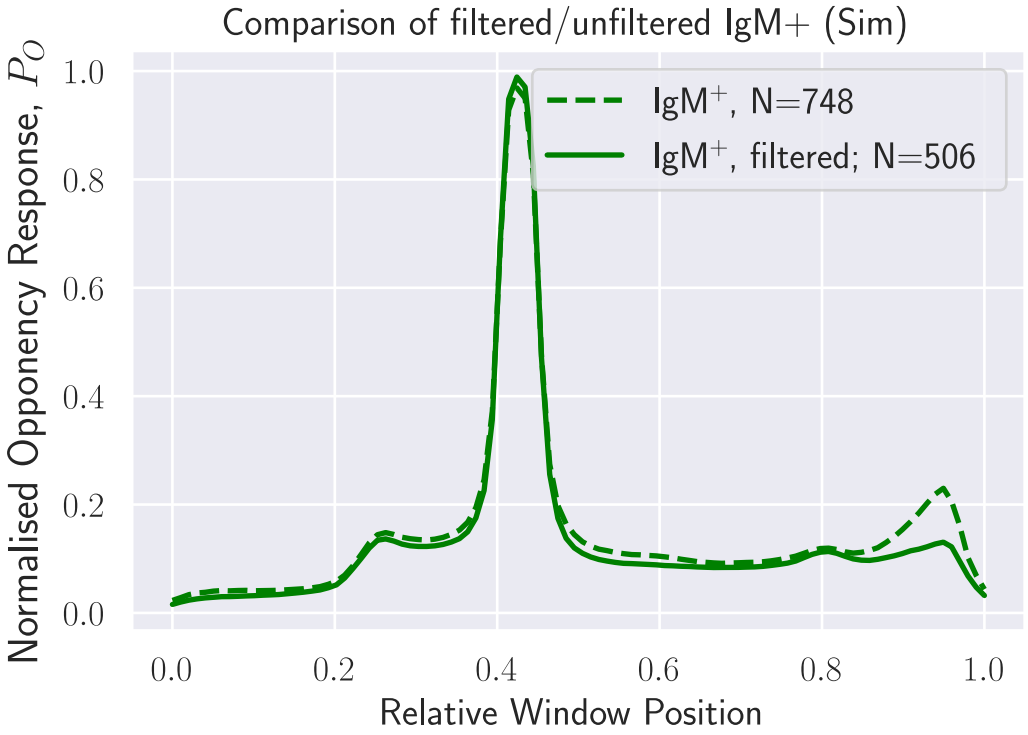


**Figure 23: The effect of misinterpretation due to leakage** Opponency Signals, *before* and *after* filtering for anomalies. Note the symmetry of the pattern at either end of the read-out window, i.e. there are small 'bumps' at horizontal positions of 0.25 and 0.95 that represent the top and bottom of the recess within the read-out window. In this figure, filtering was performed using a similarity measure between the mean signal and each of the class conditional signals.

The images associated with these signals should be flagged, and ideally removed; their interpretation might yield unpredictable outputs in a classifier that has been trained only on images that are considered to be good examples. Subsequent training with a large enough class of these anomalous images would allow them to be classified as invalid by a classification network, and we are currently collecting and curating these, with the ultimate aim being to use them in subsequent training of a multi-class CNN that provides indications of the nature of the anomaly/flaw in the read-out window.

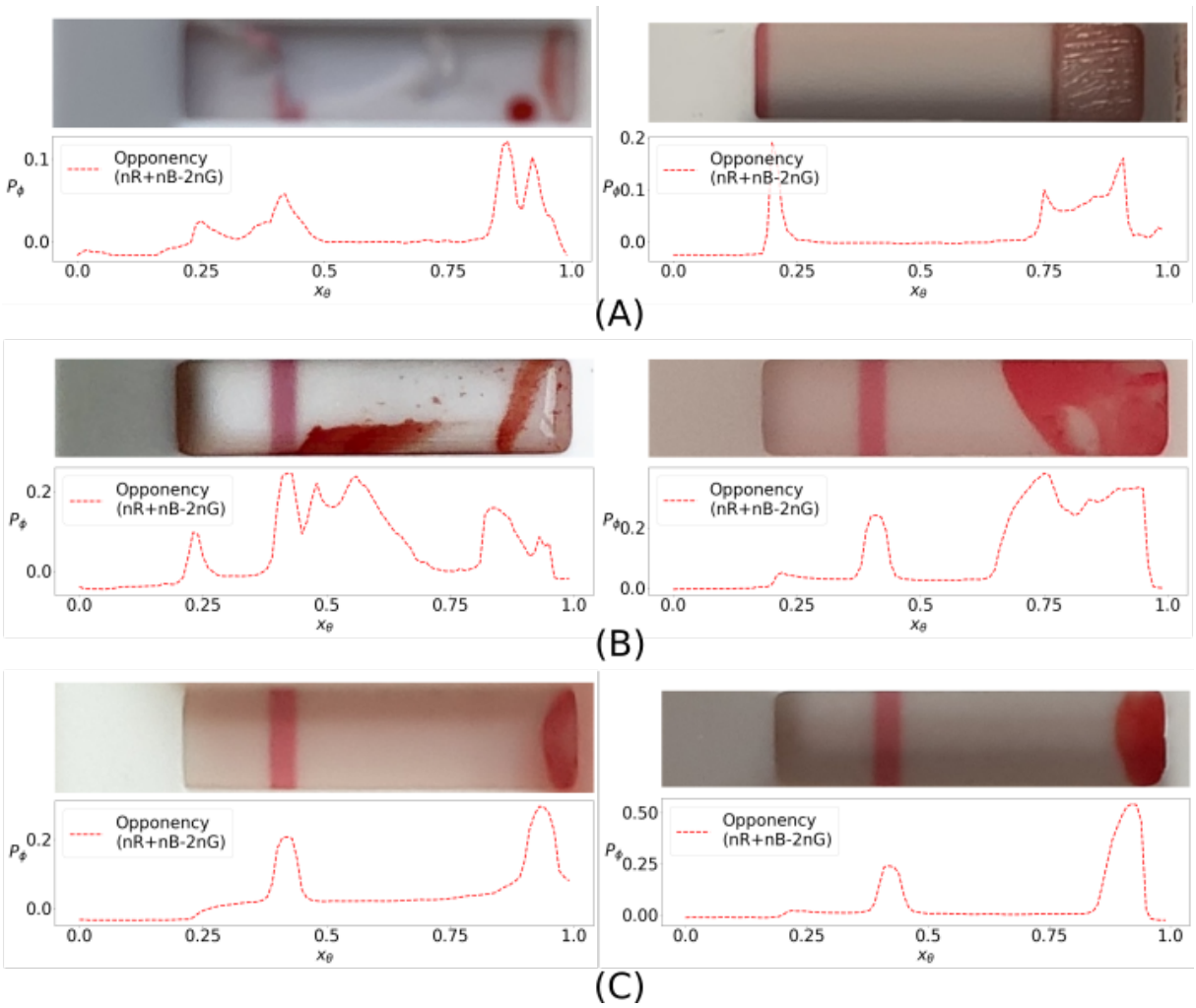


**Figure 24: Examples of broken devices and blood-leakage into the read-out.** A: Examples of broken devices where the read-out slip has been damaged. B: Examples of "heavy" blood-leakage. C: Examples of "light" blood-leakage

# 10 Discussion

## 10.1 Utilisation in a QA Setting

To gain some insight into the utility of the ALFA pipeline, let us consider the following scenario: we assume that the latent accuracy of the LFIA test is 1 (100%) for 'IgG' seroprevalence, but that 0.5% of participants misread negative test results as positive. Assume that the false positive rate of the automatic reading system is 5%. Then, the probability of both algorithm and participant reporting a positive result for a seronegative sample (false positive rate) is of the order of 0.05 x 0.005 = 0.00025.

The assumptions are that the test validity (control line) and readability errors (image is not blurred or too small in size) are negligible, which is easily enforced in a pipeline by checking region sizes and measuring the width of the control line. Therefore, over a 100,000 cohort size, the expected number of false positives arising from participant-submitted results reduces from 500 down to 25.

To utilise the pipeline in a conservative manner, we would require both participant's and ALFA to read a sample as positive before deeming it seropositive and expert reviews in cases of discrepancy. If ALFA is used in this manner, the significance is reduced risk for home-based surveys to overestimate the seroprevalence in the population. However, this is a theoretical benefit; and the operational characteristics of a live system that performs such a review will depend on the engineering details of any particular pipeline and the degree of training used to construct the data-driven elements.

An alternative argument could be made that by solely relying on the ALFA pipeline, there would be a reduced risk of underestimating the population seroprevalence. As seen, there appears to be a trend where participants are miss-reporting weak-positives as negatives and manual inspection of each sample is impossible. However, as mentioned earlier, this too would be reliant on external factors, in particular the degree of training for the algorithms and quality of the dataset.

## 10.2 Generalisation To New Immunoassay Devices

The modular nature of the ALFA pipeline allows a step-by-step approach to engineering a modified version that can support large-scale monitoring; this requires taking a similar approach to bootstrapping through progressively improved versions of the network.

1. If the device shape is substantially different from the Fortress test, hand label data according to the classes of regions in the device that are visually distinctive, including the device itself. Train the segmentation network.
2. Use the region descriptions (taking the form of simple geometric primitives) and modify the bounds-checking to reject incorrectly segmented samples. For correctly segmented samples, extract the read-out windows.
3. Label the read-out window according to expert-interpreted readings, or study participants, if confident in self-read interpretations. Use multi-class labels if devices contain multiplexed assays.
4. Train one or more classifiers to perform read-out, and make further adjustments to the pipeline to remove potential sources of potentially erroneous reading; this relies on the assumption that such errors in usage are not correlated with the diagnostic state of the participant, a reasonable assumption in many cases.

### 10.3 App-Based Reading

The decision not to use an App to acquire, or even guide, data acquisition has advantages and drawbacks. App development and support costs across multiple device versions and operating system platforms are perhaps the most serious. Single-platform Apps are typically quoted as starting at around £40,000, but this neglects the nature of support and infrastructure costs, nor does it consider the percentage of platform variations covered: prices are often dramatically understated. Hardware-dependent variations in user cameras, phones and tablets will also significantly increase support costs, particularly for Apps with some basic AI to guide the user toward taking a good image of the LFIA. Apps that use image acquisition, where images are acquired from many users in a short time frame, will also require higher back-end costs to meet surge usage.

## 11 Future Work

Variations in device layout need to be better handled, circumventing the need for changes to training, either for the segmentation network or for the result window interpretation. A solution could be a closed-loop system in which CAD/design drawings of the physical device are used directly as part of a semi-supervised technique to create automated analytics; this requires good rendering tools to create synthetic training data. This method is more straightforward than trying to provide invariance through hand-engineered algorithms. Additional benefits include a reliable ground truth and potential augmentations to reproduce variation in lighting, camera angle and 'blood-leakage' into the result window.

Detecting infection based on the IgM reading was not built into the detection function of the pipeline (though clear changes in mean signal level are observable in the region corresponding to. the IgM line), as IgM readings were deemed significantly less relevant to the aims of the REACT-2 studies. Much of the training data leveraged expert readings obtained in the normal course of the REACT-2 programme, where the status of the IgG line was used at scale to inform public health decisions.

New networks will be designed to replace dhSegment and the 2D CNN, which are parameter-heavy, and therefore compute heavy. Given the quantity of verified data, we can now use the pipeline outputs in Neural Architecture Search (NAS) (Elsken et al. 2019) to create optimised networks for all sub-tasks. Another logical step would be to leverage the outputs from the existing pipeline, not just the diagnostic readings, to create a single network architecture that will analyse the REACT-2 test results and reproduce multiple outputs that support interpretability. For example, the projection signatures, a sketch of regions of interests, and intermediate reports supporting a trained model's decisions; for further discussions around this particular design requirement, see (Carvalho, Pereira, and Cardoso 2019).

## 12 Bibliography

Abadi, Martın, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. “Tensorflow: A System for Large-Scale Machine Learning.” In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 265–83.

Adams, Emily, Mark Ainsworth, Rekha Anand, Monique I Andersson, Kathryn Auckland, J Kenneth Baillie, Eleanor Barnes, et al. 2020. “Evaluation of Antibody Testing for SARS-Cov-2 Using ELISA and Lateral Flow Immunoassays.” *Wellcome Open Research*.

Atchison, Christina, Philippa Pristerà, Emily Cooper, Vasiliki Papageorgiou, Rozlyn Redd, Maria Piggin, Barnaby Flower, et al. 2020. “Usability and Acceptability of Home-based Self-testing for Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Antibodies for Population Surveillance.” *Clinical Infectious Diseases*, August. https://doi.org/10.1093/cid/ciaa1178.

Buda, Mateusz, Atsuto Maki, and Maciej A Mazurowski. 2018. “A Systematic Study of the Class Imbalance Problem in Convolutional Neural Networks.” *Neural Networks* 106: 249–59.

Carvalho, Diogo V, Eduardo M Pereira, and Jaime S Cardoso. 2019. “Machine Learning Interpretability: A Survey on Methods and Metrics.” *Electronics* 8 (8): 832.

De Fauw, Jeffrey, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, et al. 2018. “Clinically Applicable Deep Learning for Diagnosis and Referral in Retinal Disease.” *Nature Medicine* 24 (9): 1342–50.

Deng, J., W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. 2009. “ImageNet: A Large-Scale Hierarchical Image Database.” In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–55. https://doi.org/10.1109/CVPR.2009.5206848.

Elsken, Thomas, Jan Hendrik Metzen, Frank Hutter, and others. 2019. “Neural Architecture Search: A Survey.” *J. Mach. Learn. Res.* 20 (55): 1–21.

Flower, Barnaby, Jonathan C Brown, Bryony Simmons, Maya Moshe, Rebecca Frise, Rebecca Penn, Ruthiran Kugathasan, et al. 2020. “Clinical and Laboratory Evaluation of SARS-CoV-2 Lateral Flow Assays for Use in a National COVID-19 Seroprevalence Survey.” *Thorax* 75 (12): 1082–88. https://doi.org/10.1136/thoraxjnl-2020-215732.

GOV.UK. 2021. https://coronavirus.data.gov.uk/.

Group, Visual Geometry. n.d. https://www.robots.ox.ac.uk/~vgg/software/via/.

Kartikeyan, B, A Sarkar, and KL Majumder. 1998. “A Segmentation Approach to Classification of Remote Sensing Imagery.” *International Journal of Remote Sensing* 19 (9): 1695–1709.

Leevy, Joffrey L, Taghi M Khoshgoftaar, Richard A Bauder, and Naeem Seliya. 2018. “A Survey on Addressing High-Class Imbalance in Big Data.” *Journal of Big Data* 5 (1): 1–30.

Montesinos, Isabel, Damien Gruson, Benoit Kabamba, Hafid Dahma, Sigi Van den Wijngaert, Soleimani Reza, Vincenzo Carbone, et al. 2020. “Evaluation of Two Automated and Three Rapid Lateral Flow Immunoassays for the Detection of Anti-SARS-CoV-2 Antibodies.” *Journal of Clinical Virology* 128: 104413. https://doi.org/https://doi.org/10.1016/j.jcv.2020.104413.

Murmann, Lukas, Michael Gharbi, Miika Aittala, and Fredo Durand. 2019. “A Dataset of Multi-Illumination Images in the Wild.” In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4080–89.

Olah, Chris, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. 2018. “The Building Blocks of Interpretability.” *Distill* 3 (3): e10.

Oliveira, Sofia Ares, Benoit Seguin, and Frederic Kaplan. 2018. “dhSegment: A Generic Deep-Learning Approach for Document Segmentation.” In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 7–12. IEEE.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, et al. 2019. “PyTorch: An Imperative Style, High-Performance Deep Learning Library.” In *Advances in Neural Information Processing Systems 32*, 8024–35. Curran Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Payne, Andrew, and Sameer Singh. 2005. “A Benchmark for Indoor/Outdoor Scene Classification.” In *International Conference on Pattern Recognition and Image Analysis*, 711–18. Springer.

Riley, Steven, Christina Atchison, Deborah Ashby, Christl A Donnelly, Wendy Barclay, Graham Cooke, Helen Ward, et al. 2020. “REal-Time Assessment of Community Transmission (REACT) of SARS-CoV-2 Virus: Study Protocol.” *Wellcome Open Research* 5 (200): 200.

Sandler, M., A. Howard, M. Zhu, A. Zhmoginov, and L. Chen. 2018. “MobileNetV2: Inverted Residuals and Linear Bottlenecks.” In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4510–20. https://doi.org/10.1109/CVPR.2018.00474.

Tkalcic, Marko, and Jurij F Tasic. 2003. *Colour Spaces: Perceptual, Historical and Applicational Background*. Vol. 1. IEEE.

Turbé, Valérian, Carina Herbst, Thobeka Mngomezulu, Sepehr Meshkinfamfard, Nondumiso Dlamini, Thembani Mhlongo, Theresa Smit, et al. 2021. “Deep learning of HIV field-based rapid tests.” *Nature Medicine*. https://doi.org/10.1038/s41591-021-01384-9.

Ward, Helen, Graham Cooke, Christina Atchison, Matthew Whitaker, Joshua Elliott, Maya Moshe, Jonathan C Brown, et al. 2020. “Declining Prevalence of Antibody Positivity to SARS-CoV-2: A

Community Study of 365,000 Adults." *medRxiv*. https://doi.org/10.1101/2020.10.26.20219725.

# 13 Appendix A1 – Metrics

## 13.1 Measurement of Segmentation Performance

We use the following definition for the measure of quality of segmentation: given pixel sets representing the output labelling of a set of pixels corresponding class $\mathcal{C}_k$, and a ground truth set of pixels for that class, $\mathcal{G}_k$:

$$d_k = \frac{2|\mathcal{C}_k \bigcup \mathcal{G}_k|}{\mathcal{C}_k| + |\mathcal{G}_k|} \tag{2}$$

## 13.2 Sensitivity, Specificity and Accuracy

$$Specificity = \frac{True\ negative}{True\ negative\ +\ False\ positive} \tag{3}$$

$$Sensitivity = \frac{True\ positive}{True\ positive\ +\ False\ negative} \tag{4}$$

$$Accuracy = \frac{True\ positive\ +\ True\ negative}{Total\ no.\ samples} \tag{5}$$

## 13.3 Cohen's kappa

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{6}$$

where $p_o$ is the relative observed agreement between method/participants and ground truth, and $p_e$ is the hypothetical probability of chance agreement (Eq~(7), For $c$ categories, $N$ observations, and $n_{ci}$ the number of times rater $i$ predicted category $c$).

$$p_e = \frac{1}{N^2} \sum_c n_{c1} n_{c2} \tag{7}$$

**Table 6:** **Cohen's Kappa interpretation of values.**

| Cohen's kappa | Interpretation of quality of agreement |
|---|---|
| kappa < 0 | Poor agreement |
| 0.00 <= kappa <= 0.20 | Slight agreement |
| 0.20 < kappa <= 0.40 | Fair agreement |
| 0.40 < kappa <= 0.60 | Moderate agreement |
| 0.60 < kappa <= 0.80 | Substantial agreeement |
| 0.80 < kappa | Almost perfect agreement |

## 14 Appendix A2 – Geometric Priors

For simplicity, we tabulate the relationships that are used to identify failed segmentation. Probabilistic versions of these are straightforward to write (e.g. replacing firm rules by likelihood functions) so that confidence can be specified, rather than rigid decisions on acceptance or rejection.

**Table 7:** **Geometric priors used to filter out-of-distribution images from good quality images of Fortress COVID-19 LFIA.**

| Geometric-Priors | Property | Pass Range |
|---|---|---|
| Result window / Blood well | Long-length | 0.45 < r < 0.85 |
| Result window / Blood well | Short-length | 0.75 < r < 1.2 |
| Blood well / LFIA (device) | Short-length | 0.3 < r < 0.65 |
| Result window / LFIA (device) | Short-length | 0.3 < r < 0.65 |
| Blood well / Result window | Area | 0.25 < r |
| Blood well / Result window | Perimeter | 0.55 < r |

## 15 Appendix A3 – Codebase and usage

0. Documentation is in progress
1. Repo contains approximately 1.5 million lines of code. Not all written by us!
2. We are working to reduce code complexity on a best-efforts basis due to reaching the end of funding; an MSc student will tackle some of the optimisations over the summer with the help of volunteers from the BICI lab.

# 16 Appendix A4 – 1D Convolutional neural network architectures

**Table 8: The five different 1D CNN architectures.** 'conv' are 1D convolutions (no. of input channels, no. of output channels, filter length) and 'FC' are full connected layers (no. of inputs, no. of outputs. 'batchnorm' are 1D batch nomalisation and 'MaxPool' are 1D max pooling functions with kernel size to 2.)

| **Model Architectures** | | | | |
|---|---|---|---|---|
| **1** | **2** | **3** | **4** | **5** |
| Input (10) | | | | |
| conv(10,20,5) | conv(10,30,5) | conv(10,30,5) | conv(10,40,13) | conv(10,30,13) |
| batchnorm + MaxPool | | | | |
| conv(20,20,5) | conv(30,20,5) | conv(30,30,5) | conv(40,40,9) | conv(30,30,9) |
| batchnorm + MaxPool | | | | |
| FC(440,170) | FC(440,170) | conv(30,20,5) | conv(40,30,5) | conv(30,20,5) |
| FC(170,70) | FC(170,70) | batchnorm + MaxPool | | |
| FC(70,1) | FC(70,1) | FC(180,170) | conv(30,20,3) | FC(140,170) |
| Sigmoid | Sigmoid | FC(170,70) | batchnorm | FC(170,70) |
| - | - | FC(70,1) | FC(100,170) | FC(70,1) |
| - | - | Sigmoid | FC(170,70) | Sigmoid |
| - | - | - | FC(70,1) | - |
| - | - | - | Sigmoid | - |