AI-Driven Pipelines for Science: From Easy To Hard

Antoine Sanchez [1] **Viktor Schlegel** [2 3] **Anil A Bharath** [2 1] **Nathan CK Wong** [4] **Christina Atchison** [1]
Christina Vagena-Pantoula [1] **Sulaimaan Lim** [1] **Julien Vermot** [1]

1. Introduction

The term "pipeline" is used in many contexts for processing large datasets [1], [2]. We focus on the use of pipelines for science. We discuss two pipelines that leverage machine learning, and suggest design principles that are important to the analysis of visual data from scientific experiments at scale. Our two use cases span two extremes of measurement complexity. Both cases present unique challenges, and these are briefly discussed, along with elements of data pipeline design for visual data in science.

2. Overview of Pipelines

2.1 ALFA

During the SARS-2-CoV pandemic, the UK's REACT-2 antibody prevalence study was run, consisting of a series of non-overlapping cross-sectional population surveys of the prevalence of SARS-CoV-2 antibodies in the general population of England [3]. The study used lateral flow immunoassays (LFIAs) to detect antibodies in blood samples. Participants were sent test kits and instructions with results being self-reported, alongside a photograph taken and uploaded without a dedicated App. A data processing pipeline was built to read lateral flow test results from more than 500,000+ user-posted images [4].

2.2 HARMLESS

Zebrafish are widely used as models for developmental biology due to their optical transparency and rapid rate of development. The zebrafish is ideal for studying some forms of severe cardiac disorder [5], since oxygen diffusion allows the development of even non-functioning cardiac structures. The role of specific genes, and even mechanical factors [6] influencing cardiac morphogenesis can be studied. A pipeline for dealing with the huge variability in imaging protocols, spatial deformations and strong spatial anisotropy was developed to allow data to be pooled, compared, mapped into common frames of reference and analysed. A number of deep networks were trained to support the different functions of the pipeline.

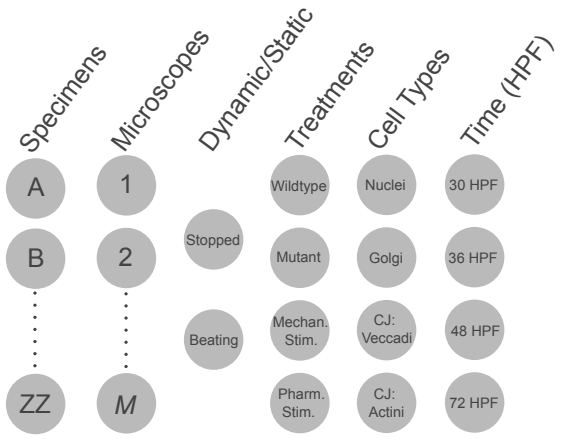

Fig. 1: The combinatorial space of the imaging experiments encompasses different biological specimens (individuals), microscopes, contrast mechanisms, instruments, cell types and time points; static and beating hearts: all must be brought into one common space for joint analysis and hypothesis testing.

3. Components of the Pipelines

Both pipelines contain modules, trained by machine learning, to perform specific tasks. ALFA uses a trained CNN to label three key areas of the LFIA device; these were used to quickly detect errors in segmentation using (non-ML based) computer vision libraries. HARMLESS required a variety of data-driven deep networks to segment different components of the zebrafish heart, and to do so from different contrast mechanisms.

Both pipelines made of use of geometric measures to constrain solutions or to provide common coordinate systems over which ensemble statistics could be generated. Projection-based dimensionality reduction was used to drop from 3D space to local 2D space, or from 2D space to 1D, and we found this particularly helpful in constructing interpretable renderings of data.

For the case of HARMLESS, projection onto reference heart manifold was used to visualise and analyse properties of cell junction connections, and to probe differences in cell junction organisation under treatment or mutation conditions. This work is unpublished, but an example of the visualisations produced from the pipeline is shown in Figure 4.

3.1 Findings from ALFA

We found a Cohen's kappa of 0.797 (95% CI: 0.794–0.799) between participants' and ALFA readout of test validity and IgG status. Disagreements

[1]*Imperial College London, Department of Bioengineering, United Kingdom* [2]*Imperial College London, Imperial Global Singapore* [3]*University of Manchester, Department of Computer Science, United Kingdom* [4]*Kings College London*. Correspondence to: Antoine Sanchez a.sanchez20@imperial.ac.uk.

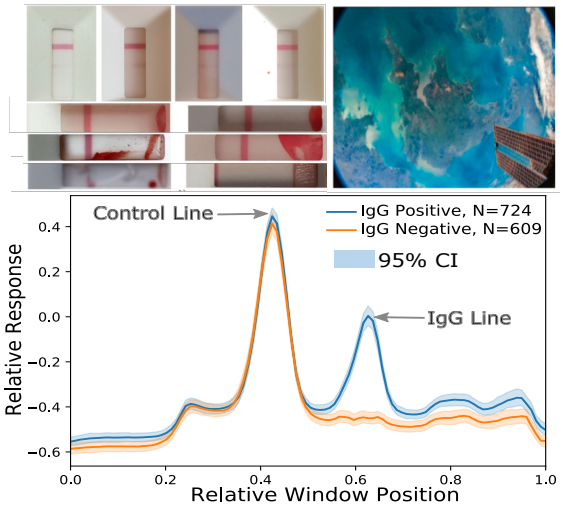

Fig. 2: (Top) Participants' phone-recorded submissions; weak responses, broken devices, and bizarrely erroneous submissions were found, often by anomaly detection; (Bottom) ML-based segmentation supported analytics, and anomaly detection. Ensemble-averaged profiles from lateral flow tests led us to discover reading errors made by participants [4].

occurred primarily when the algorithm reported an IgG positive test and study participants submitted a negative result. Using plots similar to the profiles of Figure 2, we also found evidence of user-submitted false positives in one of the antibody measures on the multiplexed lateral flow device, with implications for the design/quality assurance of future devices.

3.2 Findings from HARMLESS

- Using the data produced by the HARMLESS platform, we found that specific membrane proteins play a role in zebrafish cardiac shape. We found differences in volume changes in certain structures of the heart with certain mutant embryo cells; looping ratio was also found altered, leading to later morphogenesis effects.
- We found evidence for the role of passive mechanical forces; this has implications for the interaction between endocardium and myocardium in the developing heart.

4. Common Elements of Good Pipeline Design

We found some following elements were common to successful development and operation of both pipelines:

- Bootstrapping was helpful to iteratively increase labelling, so that components (such as segmentation networks) based on data driven learning achieved satisfactory performance.
- Geometric heuristics are useful to constrain solutions, or to provide common coordinate systems over which data can be analysed.

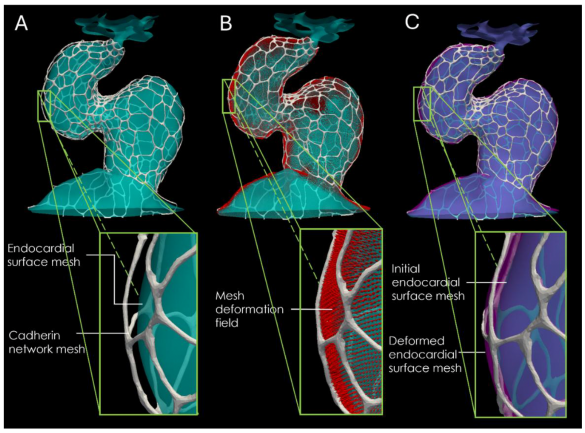

Fig. 3: The HARMLESS pipeline supports the integration of several imaging channels, and the analysis of the data in a common space; mapping of data from one space or channel to another is vital to optimally combine information over diverse experiments.

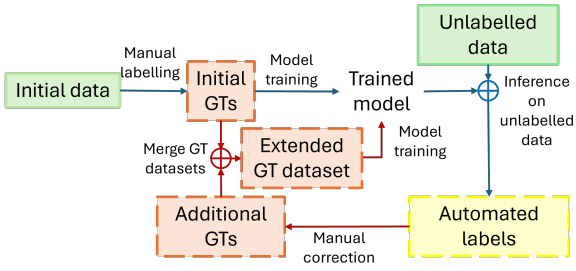

Fig. 4: Because of the high combinatorial space of imaging data, conditions and specimens handled by HARMLESS, components used the data pipeline are bootstrapped to the point of usability, then re-deployed to collect and label new data. A similar strategy was used for the ALFA pipeline.

- Maintaining logs of pipeline processes supports restarting the pipeline mid-process when errors occur due to equipment/network failure, exceptions being thrown due to rare data events (outliers), and to track the provenance of data.
- Care was necessary – in both pipelines – to ensure that components using data-driven machine learning did not inadvertently introduce biases into downstream statistical analyses.
- Independent algorithms for anomaly detection can be useful when datasets become too large even for checks by random sampling; anomalous data can then be scrutinised, and adjustments made, or data rejected.

Acknowledgments

AAB wishes to acknowledge the support of the NRF CREATE Programme. REACT was supported as a partnership between Ipsos MORI and the Department of Health and Social Care. The HARMLESS project was supported by the Wellcome Trust and the British Heart Foundation.

References

[1] Amar Shukla, Rajeev Tiwari, and Shamik Tiwari. Review on alzheimer disease detection methods: Automatic pipelines and machine learning techniques. *Sci*, 5(1):13, 2023.

[2] Alexandra Posoldova. Machine learning pipelines: From research to production. *IEEE Potentials*, 39(6):38–42, 2020.

[3] Steven Riley, Christina Atchison, Deborah Ashby, Christl A Donnelly, Wendy Barclay, Graham S Cooke, Helen Ward, Ara Darzi, Paul Elliott, REACT Study Group, et al. Real-time assessment of community transmission (react) of sars-cov-2 virus: study protocol. *Wellcome Open Research*, 5:200, 2021.

[4] Nathan CK *et al* Wong. Machine learning to support visual auditing of home-based lateral flow immunoassay self-test results for sars-cov-2 antibodies. *Communications medicine*, 2(1):78, 2022.

[5] Despina Bournele and Dimitris Beis. Zebrafish models of cardiovascular disease. *Heart Failure Reviews*, 21:803–813, 2016.

[6] Hajime Fukui, Renee Wei-Yan Chow, Choon Hwai Yap, and Julien Vermot. Rhythmic forces shaping the zebrafish cardiac system. *Trends in Cell Biology*, 2024.