Title: Synthetic Data for Equitable Artificial Intelligence in Rapid Diagnostic Test Interpretation

Authors: Elliott Rogers[1,2], Valérian Turbé[1], Dickman Gareta[3,4,5,6], Carina Herbst[3], Rachel A. McKendry[1], Kobus Herbst[3,7], Maryam Shahmanesh[2,3*]

***Corresponding author email: m.shahmanesh@ucl.ac.uk**

Affiliations:

[1] London Centre for Nanotechnology and Division of Medicine, University College London, London, UK

[2] Institute of Global Health, University College London, London, UK

[3] Africa Health Research Institute, Durban, South Africa

[4] Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland

[5] Graduate School for Health Sciences, University of Bern, Bern, Switzerland

[6] School of Laboratory Medicine and Medical Sciences, University of KwaZulu Natal, Durban, South Africa

[7] DSI-SAMRC South Africa Population Research Infrastructure Network, Durban, South Africa

**Corresponding author*

Abstract

Rapid diagnostic tests (RDTs) support affordable disease diagnosis. Machine learning (ML) can improve RDT interpretation but often relies on large, proprietary, and costly real-world image libraries. We present SynSight – a ML-enabled RDT segmentation and classification pipeline trained on synthetic data. Validated on HIV (98% sensitivity, 99% specificity) and COVID-19 RDTs (up to 99% accuracy), SynSight enables rapid ML training without real-world images, keeping pace with new RDT development.

Main text

Lateral flow tests are indispensable rapid diagnostic tests (RDT) valued for delivering results within minutes and meeting the REASSURED criteria[1]. Simple RDTs, such as COVID-19 RDTs, display a red test line to indicate antigen detection, while multiplex RDTs detect multiple targets simultaneously, such as Immunoglobulin G (IgG) and IgM antibodies. RDT use for managing acute and chronic diseases is ever increasing[1]. Over 3 billion COVID-19 RDTs were performed by late-2022[2]. In 2024, GAVI initiated the largest global deployment of 1.2 million cholera RDTs[3]. HIV RDT use is projected to reach 206 million by 2027[4]. However, RDT performance – reported through sensitivity, specificity, and accuracy – can be undermined by subjective human interpretation, leading to result misclassifications[5-7].

Machine learning (ML) is implemented in many point-of-care testing use cases[8], supports interpretation of RDT results, including for HIV[7,9], COVID-19[9-13], influenza[14], and malaria[14-17], and can be deployed via increasingly available mobile devices[18]. However, training ML models often relies on large, proprietary real-world image libraries – typically clinically sensitive, difficult to access, and expensive – or pre-existing models. Real-world image libraries are often privately owned, only available as example images, or restricted to protect patient's data[7,11,12]. Where large real-world image libraries are available, data sharing agreements can present barriers, delays and costs, and incorrect ground truth labelling can complicate ML training and evaluation[19]. This hinders the development and deployment of convolutional neural networks (CNNs).

In 2021, our previous study (*Turbé et al.*) introduced ClassifyAI, a CNN trained on over 6,000 operationally acquired HIV RDT images collected by sixty trained workers over two years which were

manually cropped to the region of interest (ROI) and manually labelled[7]. In 2023, a state-of-the-art few-shot learning model, AutoAdaptPOC, reduced the number real-world training images to 1/50th of those previously needed; however, it still relied on a base model trained on hundreds of real-world images[9].

Synthetic data offers a promising alternative to real-world data and is used in healthcare to overcome challenges around data paucity, class imbalance, model generalisability, cost and privacy[20,21]. Broadly, synthetic images can be created by either training a model on real-world data to create similar images (eg. Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs)), with accompanying privacy concerns[21], or by using virtual environments (eg. mechanistic simulators) to generate labelled datasets using no real-world data[22]. Mechanistic simulators have been highlighted for their potential in computer vision for diagnostics[22].

Herein, we present SynSight, a two-stage ML-enabled RDT processing pipeline consisting of ROI segmentation models (Fig. 1a) and result classification models (Fig. 1b) trained on mechanistically generated synthetic data and publicly available models (MobileNetV2[23] and You-Only-Look-Once (YOLO) v11[24]). A single real-world image can be processed and classified using models trained on synthetic data (Fig. 1c). We use the "train on synthetic, test on real" (TSTR) versus "train on real, test on real" (TRTR) evaluative framework[22,25] to compare the performance of published models trained on hundreds to thousands of real-world images (TRTR) to our models trained on synthetic data (TSTR). This framework evaluates a model's generalisability to real-world settings, a key challenge of synthetic data. We validate SynSight on four RDTs, two disease targets (HIV and COVID-19) and both antigen and antibody RDTs, using previously published datasets including result classifications as ground truths (Fig. 2 and 3).

To start, we mechanistically generated diverse synthetic training image libraries in minutes, significantly faster than the weeks to years required to collect real-world data (Table 1)[7,10]. A mechanistic simulator was used instead of deep generative models to address challenges of sim-to-real generalisation[26]. All features of the virtual environment were controlled, such as lighting (e.g. intensity) and camera parameters (e.g. angle).

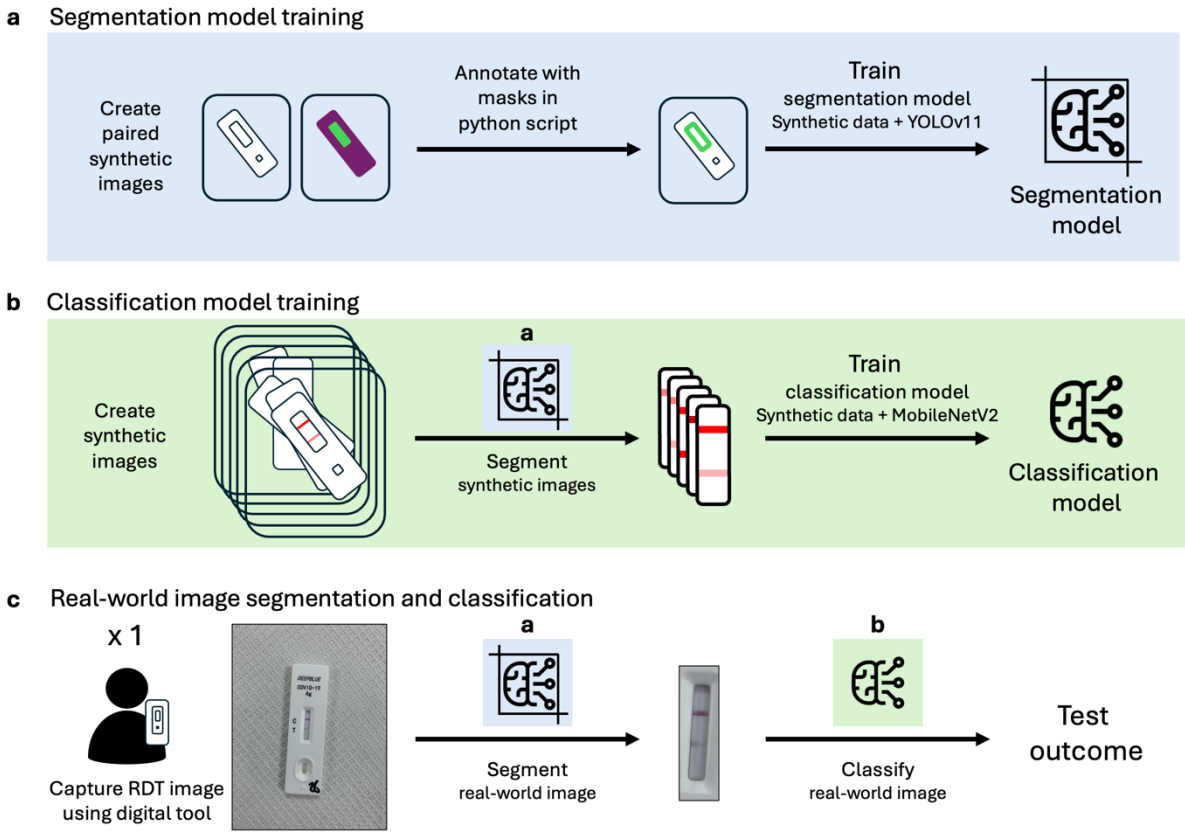


Fig. 1: Schematic of (a) segmentation and (b) classification model training on synthetic data and (c) processing a real-world image. a) Segmentation model training is facilitated by automating the ROI annotation of a synthetic image library by using python scripts implementing masks and paired synthetic images. The resulting annotated synthetic image library is used with You-Only-Look-Once v11 (YOLOv11) state-of-the-art segmentation model to train a ROI segmentation model. **b)** Synthetic images of full RDTs are segmented using the model trained in (a). Class balanced datasets of varying results can be created, including for strong positives, weak positives, negatives, invalid results, and results underrepresented in real-world image libraries. These images can be used with MobileNetV2 to train classification models. **c)** Single real-world RDT results can be classified by implementing the (a) segmentation model and (b) classification model trained on synthetic data. This allows image segmentation and classification for RDTs with no existing large real-world image libraires.

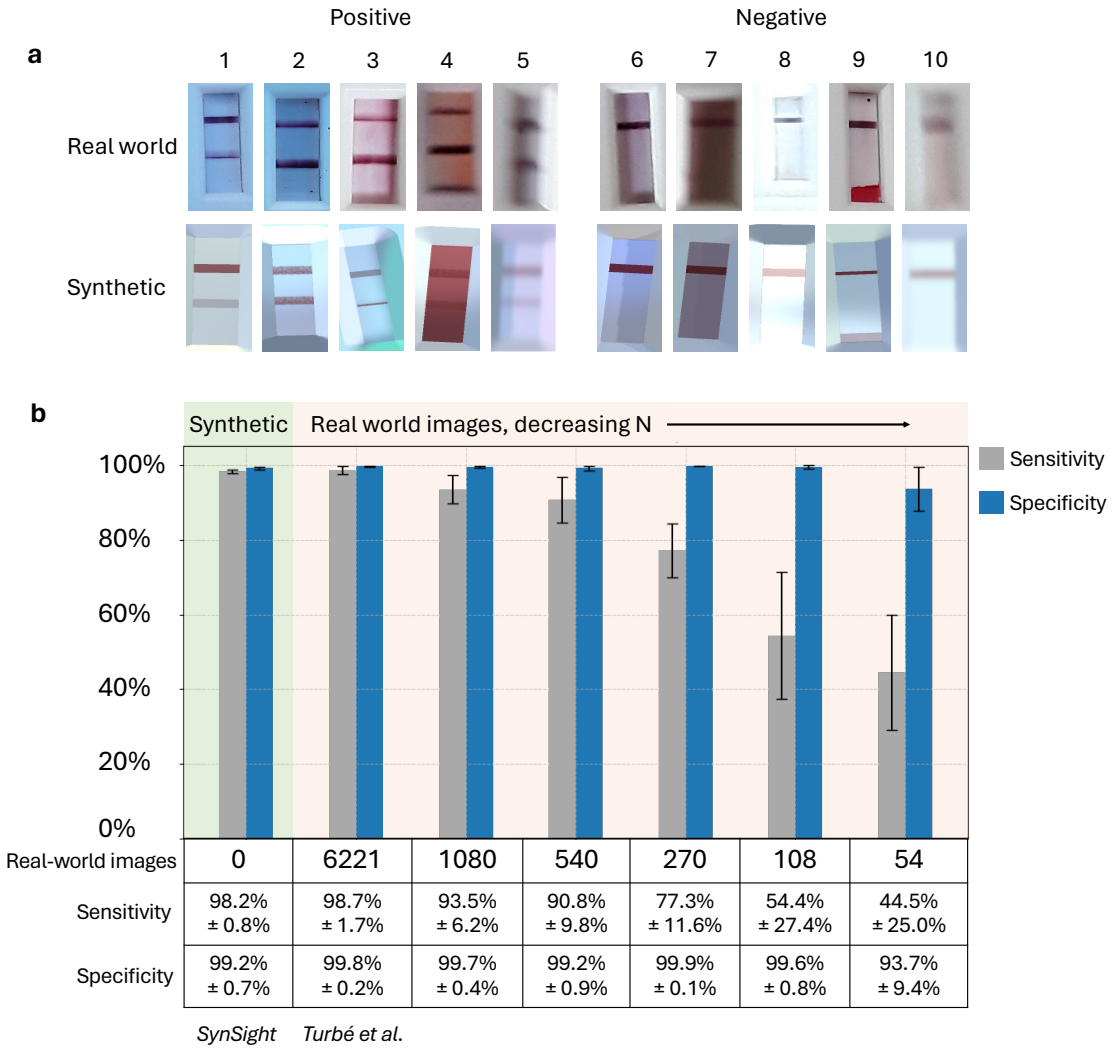


Fig. 2. Comparison of (a) real-world HIV RDT images and synthetic images, and (b) models trained on synthetic images (SynSight, this work), 6,221 real-world images (*Turbé et al.*, 2021) and decreasing real-world image training libraries. a) Synthetic images showing varying test result, line intensity, background strip colour, blur, bleed, shadows compared to real-world HIV RDTs (*Turbé et al*[7]). **b)** Plot showing the average sensitivity and specificity of HIV RDT classification using our classification model trained on synthetic data (TSTR, synthetic training n=8,000, real-world training n=0, test n = 5,685; green shading) compared to our previous model trained on thousands of real-world images (TRTR, real-world training n=6,221)[7] and decreasing number of real-world images (TRTR, real-world training n=1080 to n=54; orange shading). The threshold for the synthetic model was calculated using 18% of the real-world dataset, as per our previous publication[7]. The TRTR model was trained using 10 fold cross-validation, with sensitivity and specificity obtained using a hold-out dataset (TRTR, test n=693)[7].

a

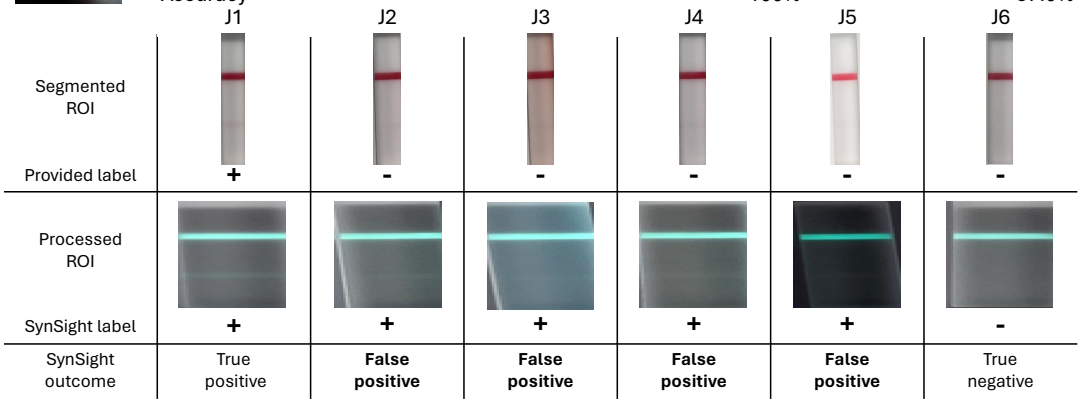
HIV	ClassifyAI	SynSight (this work)
Real-world training images	5,528	0
Real-world evaluation images	693	5,685
Sensitivity	98.7%	98.3%
Specificity	99.8%	99.2%
Accuracy	Not reported	99.2%

b

COVID-19 (DeepBlue)	AutoAdaptPOC	SynSight (this work)
Real-world training images	EcoTest base model and 14 adaption	0
Real-world evaluation images	184	283
Sensitivity	97.0%	100%
Specificity	100%	99.0%
Accuracy	98.9%	99.3%

c

COVID-19 (Jinwofu)	AutoAdaptPOC	SynSight (this work)
Real-world training images	EcoTest base model and 10 adaption	0
Real-world evaluation images	104	200
Sensitivity	100%	98.1%
Specificity	100%	95.6%
Accuracy	100%	97.0%



d

COVID-19 (EcoTest)	AutoAdaptPOC (base model)	SynSight (this work)
Real-world training images	383 train and 254 validation	0
Real-world evaluation images	102	610
Sensitivity	98%	92.0%
Specificity	100%	99.5%
Accuracy	100%*	96.6%

*As reported in previous publication

Fig. 3. Comparison of the performance of SynSight, our machine learning models trained on synthetic data and no real-world training images, to published model performance of one HIV antibody (rapid diagnostic test) RDT[7] and two COVID-19 antibody and one COVID-19 antigen RDTs[9]. SynSight (0 real-world training images) matches the HIV RDT sensitivity (98.3%) and specificity (99.2%) compared to the benchmark model, ClassifyAI (5,528 real-world training images, 98.7% sensitivity, 99.8% specificity), and for the COVID-19 RDTs, meets the performance on the DeepBlue COVID-19 RDT (SynSight: 100% sensitivity, 99.0% specificity, 99.3% accuracy; comparative AutoAdapt benchmark model: EcoTest base model and 14 adaption training images, 97.0% sensitivity, 100% specificity, 98.9% accuracy), the Jinwofu RDT (SynSight: 98.1% sensitivity, 95.6% specificity, 97.0% accuracy; AutoAdapt: EcoTest base model and 10 adaption training images: 100% sensitivity, 100%

specificity, 100% accuracy), and the EcoTest RDT (SynSight: 92.0% sensitivity, 99.5% specificity, 96.6% accuracy; AutoAdapt: EcoTest base model (383 training and 254 validation training images), 98% sensitivity, 100% specificity, 100% accuracy). Segmented Regions of Interest (ROIs) and processed ROIs are displayed for all false positives predicted by SynSight (J2-5), alongside a true positive (J1) and true negative (J6), based on provided ground truth labels. Test lines visible in J2-4 suggest the provided ground truth labelling is incorrect. Due to synthetic training, for the HIV RDT, SynSight can evaluate an image library over eight times larger than ClassfyAI, and 1,093 COVID-19 RDTs compared to 390 RDTs of AutoAdaptPOC.

Table 1. Comparison of machine learning (ML) model architecture and real-world training images used for three models: ClassifyAI (our previous work), AutoAdaptPOC (2023), and SynSight (this work). ClassifyAI (2021), trained on data collected over years, features manual segmentation of regions of interest (ROI) and a classification model built on MobileNetV2 fine-tuned on 5,528 real-world images. AutoAdaptPOC (2023), trained on real-world data collected over weeks, features automated segmentation using a Mask R-CNN, an instance segmentation model, pre-trained on 60 real-world images and 20-75 images to adapt the model to each new RDT. AutoAdaptPOC classifies images using a pretrained ResNet18 model trained on 383 real-world images and adapted using supervised contrastive learning on 10-14 labelled images. SynSight is a ML-enabled RDT segmentation and classification pipeline trained on synthetic data generated in minutes. Paired synthetic images are generated and automatically annotated using python scripts to produce a labelled ROI dataset. The annotated synthetic image library is then used to train the pre-trained You-Only-Look-Once (YOLO) v11 nano segmentation model, a state-of-the-art open-source segmentation model. SynSight classifies images using a classification model built on MobileNetV2 pretrained on ImageNet and fine-tuned on synthetic images and no real-world images.

		2021 ClassifyAI (previous work)	2023 AutoAdaptPOC	2025 SynSight (this work)
	Time to model	Years	Weeks - months	Minutes - days
Segmentation	Model architecture	Manual	Mask R-CNN	YOLOv11
	Real-world training images		60 (pretrain), 20-75 per new kit	0
Classification	Model architecture	MobileNetV2	ResNet-18 + supervised contrastive few-shot learning	MobileNetV2
	Real-world training images	5,528	383 - 397	0

Next, we trained segmentation models for the ROI of RDTs, a key precursor to ML result classification[7,9,27]. We leveraged mechanistic image generation to create paired images that allowed automated ROI annotation using python scripts (Fig. 1a). The segmentation model demonstrated high performance for bounding box prediction on synthetic data (F1 = 1.00 at confidence 0.740, mAP@0.5 = 0.995, Supplementary Figure 4) and segmented 99.9% (1093/1094) of real-world COVID-19 RDTs, confirming both strong performance and sim-to-real generalisation. This approach could be applied to other healthcare segmentation use cases of objects not represented in training libraries. We also developed an RDT segmentation model presented through Augmented Reality which could guide smartphone data capture (Supplementary Figure 1).

For the classification model, synthetic image generation was informed by our previous studies and the literature[6,7,10]. Synthetic libraries can be created to improve model performance on challenging images by increasing the frequency of weak positives[10] (Fig 2a. 1 and 5), underrepresented edge cases (eg. blood in the ROI[10], Figure 2a. 9), and rare images (eg. no control line[7]). Unlike real-world libraries which can suffer class-imbalance (eg. 1:8 positive to negative class-imbalance for ClassifyAI[7]), synthetic libraries can be class-balanced.

For the HIV RDT, we trained a classification model on 8,000 synthetic images and achieved a TSTR performance of 98.2% (±0.8%) sensitivity, 99.2% (±0.7%) specificity, and 99.2% (±0.6 %) accuracy (test n=5,685). This was comparable to our previous model, ClassifyAI, trained on 6,221 real-world images (TRTR performance of 98.7%±1.8% sensitivity, 99.8%±0.2% specificity, accuracy not reported, test n=693)[7] (Fig. 2). We also evaluated SynSight against ClassifyAI when trained on fewer real-world images, with sensitivity of ClassifyAI dropping to 77.3% (±11.6%), over 20 percentage points lower than SynSight (Fig. 2b), when trained on 270 real-world images. This highlights the strength of the synthetic approach when limited real-world image libraries are available.

Next, we applied the full SynSight pipeline, implementing both the segmentation and classification models, to two antigen and one antibody COVID-19 RDT. Training and best threshold selection only used synthetic images, allowing classification of full image libraries instead of limited test sets. Classification accuracy of two antigen RDTs was 99.3% (test n=293) (DeepBlue, TRTR:98.9%, test n=184) and 97.0% (test n=200) (Jinwofu, TRTR:100%, test n=104), meeting the performance of the few-shot learning approach trained on approximately 400 real-world images[9]. For the antibody RDT, the base-model training set used in the few-shot learning approach, SynSight achieved 96.6% accuracy (test n= 610) (EcoTest, TRTR:100%, test n=102) (Figure 3)[9].

SynSight outperformed the benchmark model on the DeepBlue RDT but underperformed on the Jinwofu RDT. We reviewed SynSight's false positives on Jinwofu (n=4) (Fig. 3c J2-5) alongside a true positive (J1) and negative (J6). Ground truth labels in real-world image libraries may contain inaccuracies and so disagreements between ML and human interpretation should be re-evaluated. False negatives (ML negative, ground truth positive) can be difficult to verify as weak test lines may not be captured in images, a limitation of ML RDT analysis. False positives (ML positive, ground truth negative), however, are easier to verify as a test line indicates incorrect ground truth labelling. A test line is visible in three of the processed ROI results (Fig. 3c J2-4) suggesting the provided ground truth labels are incorrect. This underscores a challenge of working with real-world image libraries for both training (if using real-world images) and evaluating models.

Our synthetic approach marks the next step in training RDT ML models, reducing real-world training images from thousands (ClassifyAI[7]), to hundreds (AutoAdaptPOC[9]), to none (SynSight, this work). SynSight is trained on synthetic data created on free and widely available 3D modelling software (Unity) and on open-source models (MobileNetV2, YOLOv11) instead of closed or proprietary base models. SynSight offers an accessible, scalable, privacy preserving RDT processing pipeline which can be applied to existing and novel RDTs where real-world image libraries are unavailable. SynSight does not require data sharing agreements or extensive resources for data capture or model training (conducted on both CPU and GPU), facilitating more equitable AI deployment. Real-world images remain important to ensure the validation of ML models.

Limitations may arise when image generation does not capture the variability of real-world images, leading to poor performance, bias, and confabulation – plausible but incorrect predictions based on patterns in synthetic images. For example, our model incorrectly classified 1/283 DeepBlue RDTs with excessive streaking (where nanoparticles are deposited throughout the ROI) and 2/610 EcoTest RDTs with marks near test line locations as false positives (Supplementary Figure 7). Precursor qualitative research is key for developing representative synthetic image libraries. For future synthetic image libraries, experts, including individuals creating and implementing RDTs, should be consulted to identify challenges in result classification. Further, clear guidance and standardised image capture protocols should be developed when collecting real-world images.

In conclusion, synthetic data and digital tools can support the use of RDTs, improving interpretation by reducing false positives and negatives. Developing accurate ML models quickly is crucial as novel RDTs are continuously developed for emerging and re-emerging diseases. Synthetic data removes privacy concerns of training on sensitive, clinical data. Our SynSight pipeline trained on synthetic data

segments and classifies four RDTs, including antigen and multiplex antibody RDTs for two different diseases, with similar performance as models trained on hundreds and thousands of real-world images. The high accuracy of SynSight is marked when the real-world training dataset is small. Such digital tools offer scalable expert-level RDT interpretation and quality control, and are increasingly important as novel, multiplex RDT use and development continues.

Methods

Ethics

The work presented are secondary analyses of existing real-world image libraries of HIV and COVID-19 RDTs. Ethical approval for the Africa Health Research Institute's (AHRI) HIV image library was collected as part of the demographic surveillance study, with ethics granted by the Biomedical Research Ethics Committee of the University of KwaZulu-Natal, South Africa (no. BE435/17)[7,28]. The ethics for the collection of the COVID-19 RDT image library are outlined in a previous publication[9].

Real-world data

Images of HIV RDTs were collected by the AHRI. Image capture of Advanced Quality One Step Anti-HIV (1&2) Test (InTec Products, Inc.)) tests was embedded into trained staff workflows in Kwazulu-Natal, South Africa, using minimal hardware (Samsung SM-P585 tablet and customised imaging tray), as previously described[7]. The two antigen COVID-19 RDTs were the Anhui DeepBlue SARS-CoV-2 Antigen Test and the Jinwofu SARS-CoV-2 Antigen Rapid Test. The antibody COVID-19 RDT was the AssureTech EcoTest COVID-19 IgG/IgM Antibody Test kit. These COVID-19 RDTs were conducted and imaged by researchers at Columbia University using nasopharyngeal swabs from Mayo Clinic Hospital patients (Mayo Clinic IRB 20-010688)[9]. 25 duplicate RDT images and 1 test labelled 'handmarked', with a test line clearly marked on the RDT, were identified for the EcoTest RDT and excluded from analyses.

Synthetic data generation

Synthetic images of RDTs were mechanistically generated using 3D modelling software, Unity. This allowed control of all parameters of each image using C# scripts and features in Unity. Controlled parameters are detailed below and can be grouped into test specific and condition specific parameters.

Test specific parameters included test size, test result, control and test line sizes, opacities, appearances (solid and particle) and positions, test strip colour, simulations of liquid progressing into the ROI, appearance of randomised alphanumeric text above and to the right of the test window, appearance of outcome labelling text to the left-hand side of the test window, and appearance of symbols below the test window. A thin or thick solid test line was used to replicate homogeneous control and test line results, with intensity set between 0.5 to 100% opacity, while multiple overlayed particle systems fixed to two dimensions of varying sizes, number and area were used to replicate

non-homogenous real-world test lines. Condition specific parameters included lighting, camera yaw, position of the camera by translating randomly in three dimensions, camera layer culling masks, camera field of view, camera aspect ratio, and output image size. Most parameters were governed by probabilistic mechanisms to create a wide variety of image properties. This probabilistic variation across multiple parameters results in a virtually infinite number of unique images. The image creation parameters were configured differently depending on the desired training dataset.

Images for training the segmentation model were created in pairs. In one image, the ROI (test window) can be assigned a clear block colour receiving no shadows (green shading), and the full cassette assigned a different clear block colour receiving no shadows (purple shading) (Fig. 1a, Supplementary Figure 2 and 3). A python script annotated the location of the ROI in each image pair. 4871 training images and annotated ROIs were used for training the segmentation model. 4000 positive and 4000 negative images were generated to train the HIV RDT image classification model (Supplementary Figure 5). 1000 positive and 1000 negative images were generated to train the COVID-19 RDT image classification model. All segmentation and classification synthetic images are available via the Data availability statement

The Supplementary Information and material available through the Code availability statement and Data availability statement provides all information required to produce synthetic images for a new RDT, including the 3D assets (and video outlining the creation of 3D assets), C# code and a detailed text guide.

Image processing

Image processing consisted of several transformations aimed at standardising and improving data before training. Training images for the segmentation model were processed using various data augmentation techniques, including scaling images, mosaicking, random erasing, adjusting Hue, Saturation and Value (HSV), rotations, translations, and other data augmentation techniques (see Code availability statement for full training script). These were applied randomly through a predefined augmentation policy (RandAugment, https://docs.ultralytics.com/modes/train/).

For the HIV classification model, training images were subject to four image processing steps: resizing to 224 by 224 pixels preparing the image for the model input; blurring by applying a random intensity blur filter to soften image details; inverting by subtracting the value of each pixel from 1; and, finally, normalising by subtracting the mean pixel value and dividing by the standard deviation of

all pixels in the image, closely following previous image processing to allow comparision[7]. Real-world images were subject to the same image processing excluding blurring.

For the COVID-19 classification model, training images had slightly different processing steps to reduce computational demand. Images were first segmented to the ROI using a segmentation model, resized to 224 by 224, normalised by dividing each pixel value by 255, blurred and inverted. For the real-world images, images were normalised, rescaled to emphasise pixel intensities above a threshold of 0.1, suppressing background noise and enhancing contrast in the test region, and inverted.

Image labels

All synthetic images were labelled according to the virtual camera capturing the image from the mechanistic simulation environment. The C# script for generating image parameters, layer assignment and camera culling masks were checked before image generation to ensure correct labelling. Parameter extremes, such as highest light intensity and lowest test line intensity, were produced and visually checked to ensure the test outcome label was accurate.

The HIV RDT real-world image library was labelled by a panel of three RDT experts[7]. The COVID-19 RDT real-world image library was labelled as per previous publication[9]. All real-world RDT images were classified by ML into two outcome categories (positive/negative).

Machine learning: training

For both classification models, a base model of MobileNetV2 was used with weights pre-trained on the ImageNet dataset, excluding the top layer. The model was prepared to receive images of 224 by 224 pixels with three channels corresponding to Red, Green, and Blue (RGB). After the base model, the output undergoes Global Average Pooling reducing the spatial dimensions to a single vector per image. This vector is processed through a fully connected layer consisting of 256 neurons with rectified linear unit (ReLU) activation. For the HIV classification model, the final dense layer is a single neuron with a sigmoid activation function outputting a probability value. This is the same as previously described[7] with two changes: the number of epochs was reduced from 50 to 10 in order to output high performance with faster training, and image preprocessing was updated to include the aforementioned blurring and inversing steps. Models were trained using GPU for comparison to previous work and training speed. For the COVID-19 classification model, an additional dropout layer with a rate of 0.5 was added to prevent overfitting and allow the model to generalise well to new

data. Training was conducted on a CPU demonstrating the technological accessibility of model development.

For the segmentation model, a nano YOLOv11 segmentation (yolo11n-seg.pt) base model was used, pre-trained on Ultralytics' library[24]. A custom synthetic dataset defined in a YAML file, including synthetic images and annotations, was used for further training. The model was trained for 5 epochs on images resized to 640 by 640 pixels. Mixed precision training was used to balance memory usage and efficiency. All training scripts are either previously published[7] or available via the Code availability statement.

The ClassifyAI and AutoAdaptPOC models were trained as previously described[7,9]. Briefly, ClassifyAI is a classification model built on MobileNetV2 pretrained on ImageNet and fine-tuned on 5,528 real-world images[7]. AutoAdaptPOC features automated segmentation using Mask R-CNN, an instance segmentation model pre-trained on the ImageNet1K dataset, combined with RDT kit and ROI dimensions, along with 60 real-world images to pre-train a segmentation model and 20-75 images to adapt to each new RDT kit. AutoAdaptPOC classifies images using a ResNet18 model pretrained via self-supervised learning on the ImageNet1K dataset and 383 real-world images and adapted using supervised contrastive learning on 10-14 labelled images[9].

Machine learning: best threshold

For the HIV RDT models, following training of the classification model on synthetic data, a best threshold was calculated by using 18% of the full dataset, the same proportion of data for validation as the benchmark ClassifyAI model trained on real-world data[7]. First, receiver operating characteristic curves (ROC) were constructed for the validation dataset, providing false positive rate (FPR (1 – specificity)) and true positive rate (TPR (sensitivity)) across many thresholds. The best threshold was determined by selecting the threshold which maximised the sum of TPR and 1 – FPR, maximising true positives and minimising false positives. The ratio of classes in the validation set was matched to the 1:8 positive to negative ratio in the full HIV RDT image library, as conducted for ClassifyAI. For the models trained on increasing real-world data (Fig. 2b), a 1:8 positive to negative ratio was also maintained in training datasets. For the first training dataset (N = 54, positive n = 5), the 18% validation dataset had 1 positive image and 8 negative images. For each for the increasing real-world image datasets this ratio is maintained.

For the COVID-19 RDTs, no real-world RDT images were used during training, validation, or best threshold calculation. Instead, a separate synthetic image library which the classification model was not trained on was used to calculate the best threshold by maximising the sum of TPR and 1 – FPR, as detailed above. This best threshold was then applied to all three COVID-19 RDTs.

Machine learning: evaluation

For the HIV RDTs, the remaining real-world image libraries were classified based on the best threshold, and sensitivity, specificity and accuracy obtained. This process was repeated for 10 random validation subsets of each dataset, as previously described[7]. For the COVID-19 RDTs, full real-world image libraries were classified based on the synthetically calculated best threshold, and sensitivity, specificity and accuracy obtained.

Data availability

All synthetic images are available for immediate access through the following link: https://doi.org/10.5522/04/28351934. Source data for Fig. 2 is available in Supplementary Table 1 and Supplementary Table 2. The real-world image library used for evaluating the performance of HIV RDT classification models is available through the AHRI repository, as previously published: https://data.ahri.org/index.php/catalog/923#metadata-data_access. Source data for Fig. 3 is available in Supplementary Figure 6. The real-world image library used for evaluating the performance of the COVID-19 RDT segmentation and classification models is available through the following link, as previously published: https://figshare.com/projects/AutoAdapt/188457.

Code availability

Custom code used in training and analysing the machine learning models presented in this study is available on GitHub at https://github.com/ElliottRogers/SynSight. The code was developed for Python 3.9. A detailed description of the code, including usage instructions and dependencies, is provided in several README files within the repository. This includes the core machine learning models used in this study (HIV classification, COVID segmentation, and COVID classification). In addition, previously published, free and easy to access code used to train the ClassifyAI model and the classification models trained on increasing real-world HIV data (Fig. 2b) is available here: https://xip.uclb.com/product/classify_ai.

The code was primarily tested on macOS 14.6.1. Other operating systems may be compatible but were not extensively tested. The following hardware was used for training: HIV Classification, Apple M2 (GPU - Metal); COVID Segmentation, Apple M2 (GPU - Metal); COVID Classification: CPU (tested on Google Colab GPU).

Clinical trial number: not applicable.

Funding

We acknowledge funding for this research: Medical Research Council for the PhD studentship (MR/W006774/1 to E.R); EPSRC Digital Health Hub for AMR (EP/X031276/1 to R.A.M); MS is an NIHR Research Professor (NIHR 301634); AHRI is supported by core funding from the Wellcome Trust

(082384/Z/07/Z and 201433/Z/16/Z to KH and MS). For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Acknowledgements

We would like to thank the authors, institutions and participants who contributed to collecting and providing access to real-world image libraries, as outlined in *Turbé et al.* and *Arumugam et al.*

Author Contributions

ER, RAM, MS, VT, KH contributed to the conceptualisation of the manuscript. RAM and MS supervised the study. All authors – ER, VT, DG, CH, RAM, KH, MS - read and commented on early drafts, and approved the final manuscript. KH was principal investigator of the original study that produced the real-world HIV RDT images.

Competing Interests

None

References

1. Land, K.J., Boeras, D.I., Chen, X.-S., Ramsay, A.R. & Peeling, R.W. REASSURED diagnostics to inform disease control strategies, strengthen health systems and improve patient outcomes. *Nature microbiology* **4**, 46-54 (2019).
2. Organization, W.H. The ACT-accelerator: Two years of impact. *URL: https://www. who. int/publications/m/item/the-act-accelerator--two-years-of-impact (abgerufen am 11.08. 2022)* (2022).
3. WHO. Global deployment of rapid diagnostic tests to boost fight against cholera. (2024).
4. Consortium, E.I. HIV rapid diagnostic test market landscape. (2023).
5. Gaydos, C.A.*, et al.* Will patients "opt in" to perform their own rapid HIV test in the emergency department? *Annals of emergency medicine* **58**, S74-S78 (2011).
6. Mukadi, P.*, et al.* SMS photograph-based external quality assessment of reading and interpretation of malaria rapid diagnostic tests in the Democratic Republic of the Congo. *Malaria Journal* **14**, 1-12 (2015).
7. Turbé, V.*, et al.* Deep learning of HIV field-based rapid tests. *Nature medicine* **27**, 1165-1170 (2021).
8. Han, G.-R.*, et al.* Machine learning in point-of-care testing: innovations, challenges, and opportunities. *Nature Communications* **16**, 3165 (2025).
9. Arumugam, S.*, et al.* Rapidly adaptable automated interpretation of point-of-care COVID-19 diagnostics. *Communications Medicine* **3**, 91 (2023).
10. Wong, N.C.*, et al.* Machine learning to support visual auditing of home-based lateral flow immunoassay self-test results for SARS-CoV-2 antibodies. *Communications medicine* **2**, 78 (2022).
11. Beggs, A.D.*, et al.* Machine learning for determining lateral flow device results for testing of SARS-CoV-2 infection in asymptomatic populations. *Cell Reports Medicine* **3**(2022).
12. Lee, S.*, et al.* Sample-to-answer platform for the clinical evaluation of COVID-19 using a deep learning-assisted smartphone-based assay. *Nature communications* **14**, 2361 (2023).
13. Mendels, D.-A.*, et al.* Using artificial intelligence to improve COVID-19 rapid diagnostic test result interpretation. *Proceedings of the National Academy of Sciences* **118**, e2019893118 (2021).
14. Park, C.*, et al.* The design and evaluation of a mobile system for rapid diagnostic test interpretation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **5**, 1-26 (2021).
15. Park, C.*, et al.* Supporting smartphone-based image capture of rapid diagnostic tests in low-resource settings. in *Proceedings of the 2020 International Conference on Information and Communication Technologies and Development* 1-11 (2020).
16. Bahati, F.*, et al.* Artificial intelligence (AI)-driven rapid diagnostic test interpretation in a Connected Diagnostics (ConnDx) system for dynamic malaria surveillance. *BMC Digital Health* **3**, 62 (2025).
17. Frade, S.*, et al.* Malaria RDT (mRDT) interpretation accuracy by frontline health workers compared to AI in Kano state, Nigeria. *BMC Digital Health* **3**, 50 (2025).

18. Wood, C.S.*, et al.* Taking connected mobile-health diagnostics of infectious diseases to the field. *Nature* **566**, 467-474 (2019).
19. Roche, S.D.*, et al.* Measuring the performance of computer vision artificial intelligence to interpret images of HIV self-testing results. *Frontiers in Public Health* **12**, 1334881 (2024).
20. Chen, R.J., Lu, M.Y., Chen, T.Y., Williamson, D.F. & Mahmood, F. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering* **5**, 493-497 (2021).
21. Giuffrè, M. & Shung, D.L. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *NPJ digital medicine* **6**, 186 (2023).
22. van Breugel, B., Liu, T., Oglic, D. & van der Schaar, M. Synthetic data in biomedicine via generative artificial intelligence. *Nature Reviews Bioengineering*, 1-14 (2024).
23. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 4510-4520 (2018).
24. Glenn Jocher, A.C., and Jing Qiu. Ultralytics YOLO. Vol. 2023 (Ultralytics, https://github.com/ultralytics/ultralytics, 2023).
25. Esteban, C., Hyland, S.L. & Rätsch, G. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633* (2017).
26. Gibney, E. AI models fed AI-generated data quickly spew nonsense. *Nature* **632**, 18-19 (2024).
27. Lee, S.*, et al.* Rapid deep learning-assisted predictive diagnostics for point-of-care testing. *Nature Communications* **15**, 1695 (2024).
28. Gareta, D.*, et al.* Cohort profile update: Africa Centre Demographic Information System (ACDIS) and population-based HIV survey. *International journal of epidemiology* **50**, 33-34 (2021).