ARTICLES
https://doi.org/10.1038/s41551-019-0487-z

nature biomedical engineering

There are amendments to this paper

Detection of anaemia from retinal fundus images via deep learning

Akinori Mitani[1]*, Abigail Huang[1], Subhashini Venugopalan[2], Greg S. Corrado[1], Lily Peng[1], Dale R. Webster[1], Naama Hammel[1], Yun Liu[1,3] and Avinash V. Varadarajan[1,3]

Owing to the invasiveness of diagnostic tests for anaemia and the costs associated with screening for it, the condition is often undetected. Here, we show that anaemia can be detected via machine-learning algorithms trained using retinal fundus images, study participant metadata (including race or ethnicity, age, sex and blood pressure) or the combination of both data types (images and study participant metadata). In a validation dataset of 11,388 study participants from the UK Biobank, the metadata-only, fundus-image-only and combined models predicted haemoglobin concentration (in g dl^{-1}) with mean absolute error values of 0.73 (95% confidence interval: 0.72–0.74), 0.67 (0.66–0.68) and 0.63 (0.62–0.64), respectively, and with areas under the receiver operating characteristic curve (AUC) values of 0.74 (0.71–0.76), 0.87 (0.85–0.89) and 0.88 (0.86–0.89), respectively. For 539 study participants with self-reported diabetes, the combined model predicted haemoglobin concentration with a mean absolute error of 0.73 (0.68–0.78) and anaemia an AUC of 0.89 (0.85–0.93). Automated anaemia screening on the basis of fundus images could particularly aid patients with diabetes undergoing regular retinal imaging and for whom anaemia can increase morbidity and mortality risks.

Anaemia is a public health problem that affects an estimated 1.62 billion people[1]. In 2011, 29% of non-pregnant women worldwide were affected by anaemia[2]. As a major contributor to the global burden of disease, anaemia has far-reaching consequences for work and productivity and quality of life[1,3,4]. As anaemia is usually correctable[5], timely detection and intervention are key. The most reliable indicator of anaemia is haemoglobin concentration (Hb)[1], which is traditionally measured using a venous or capillary blood sample. However, these procedures are invasive and painful, can cause infection in patients and healthcare workers and generate biohazardous waste[6]. Thus, there is a clear need for a non-invasive procedure.

Several non-invasive methods of estimating Hb are now available. Traditionally, subjective assessment of pallor of the conjunctiva, nail beds, tongue and palms have been used as clinical signs indicating the presence of severe anaemia, with a wide range of estimated sensitivities and specificities[7,8]. Recently, it was reported that Hb can be estimated with high accuracy using automated algorithms to analyse the colour of fingernail beds from digital photographs taken by smartphones[9]. However, the algorithms were based on manually selected regions of interest on the fingernails, and their robustness and real-world performance remains to be assessed. Another two methods, occlusion spectroscopy and pulse CO-oximetry, use spectrophotometric sensors that non-invasively assess Hb by measuring light transmission through the tissue[10,11]. These non-invasive methods are less accurate than the gold standard of venous blood laboratory analysis[12] and represent trade-offs between invasiveness, time, cost and accuracy[9].

Interestingly, anaemia of sufficient severity has been known to manifest characteristic signs in the fundus of the eye[13], and 20% of patients with anaemia are reported to develop extravascular lesions, with the severity of anaemia related to venous tortuosity[14]. Retinopathy is observed in 28.3% of patients with anaemia and/or thrombocytopaenia, and low Hb was associated with the presence of retinopathy[14]. However, the low prevalence of retinopathy among patients with anaemia limits its potential sensitivity as a stand-alone diagnostic feature, and fundus photographs have not been used to either detect anaemia or quantify more precise Hb levels.

In this work, we explore the hypothesis that Hb can be quantified using non-invasive fundus photographs and deep learning. Deep learning has been previously shown to be highly effective in extracting information from images[15]. In ophthalmology, deep-learning algorithms can detect eye conditions such as diabetic retinopathy, age-related macular degeneration and glaucoma with accuracy comparable to human experts[16–21]. Additionally, some previously unknown information can be extracted from fundus images, such as refractive error[22], age, sex and cardiovascular risk[23]. Extending this, we show that deep learning can be leveraged to quantify Hb and detect anaemia.

Results

This study was conducted using data from the UK Biobank, which is a population-based prospective study[24]. In total, 114,205 fundus images from 57,163 study participants were included in this study. In the validation set, the median age of the study participants was 57.9 years (interquartile range of 50.0–63.8 years), 54.9% were female and 91.1% were white (Table 1). Among all the study participants, 3.7% had anaemia, with Hb ranging from 6.4 g dl^{-1} to 19.6 g dl^{-1}. Additional study participant demographics are summarized in Table 1.

Results of many blood tests, including Hb, are correlated with the metadata of the study participants, such as demographics. Some of these metadata, such as age and sex, have previously been shown to be predictable using fundus images[23]. Therefore, to ensure that our predictors were not solely predicting Hb measurements via age and sex, we first developed baseline linear regression models using these metadata (metadata-only model). Next, we developed deep-learning models based on the Inception-v4 architecture[25] (Methods)

[1]Google Health, Google, Mountain View, CA, USA. [2]Google Research, Google, Mountain View, CA, USA. [3]These authors contributed equally: A. V. Varadarajan and Y. Liu. *e-mail: amitani@google.com

Table 1 | Basic characteristics of the development datasets and the validation dataset

	Development datasets		Validation dataset
	Training dataset	Tuning dataset	
Total no. of images	80,006	11,457	22,742
No. of participants	40,041	5,734	11,388
Age (years)[a]	57.9 (50.0–63.7)	58.0 (49.9–63.7)	57.9 (50.0–63.8)
Females (%)	21,944 (54.8%)	3,152 (55.0%)	6,255 (54.9%)
Race/ethnicity (%)			
Black	468 (1.2%)	74 (1.3%)	145 (1.3%)
Asian	1,330 (3.3%)	192 (3.3%)	404 (3.5%)
White	36,606 (91.4%)	5,247 (91.5%)	10,369 (91.1%)
Other	1,637 (4.1%)	221 (3.9%)	470 (4.1%)
Current smoker (%)	3,794 (9.5%)	544 (9.5%)	1,120 (9.8%)
Body mass index (kg m^{-2})[a]	26.6 (24.0–29.7)	26.7 (24.1–29.9)	26.6 (24.1–29.7)
Height (cm)[a]	168 (162–176)	168 (162–176)	168 (162–175)
Weight (kg)[a]	76.3 (66.2–87.5)	76.7 (66.4–87.6)	76.3 (66.4–87.8)
Heart rate (b.p.m.)[a]	67.5 (61.0–74.5)	67.5 (60.5–75.0)	67.0 (61.0–74.0)
Systolic blood pressure (mmHg)[a]	135 (124–148)	135 (124–148)	136 (124–148)
Diastolic blood pressure (mmHg)[a]	82 (75–88)	82 (75–89)	82 (75–88)
Hb (g dl^{-1})[a]	14.3 (13.4–15.2)	14.3 (13.4–15.1)	14.2 (13.4–15.1)
Distribution of anaemia levels[b]			
None (12+ (F), 13+ (M))	38,628 (96.5%)	5,539 (96.6%)	10,949 (96.1%)
Mild (11–12 (F), 11–13 (M))	1,134 (2.8%)	164 (2.9%)	347 (3.0%)
Moderate (8–11 (Both))	267 (0.7%)	31 (0.5%)	90 (0.8%)
Severe (0–8 (Both))	12 (0.0%)	0 (0.0%)	2 (0.0%)

[a]Results are presented as median values (interquartile range). [b]The anaemia categories are based on those from the WHO (please see the "Definitions of anaemia" section in the Methods), and the ranges are shown in g dl^{-1}. b.p.m., beats per minute; F, female; M, male.

using fundus images (fundus-only model). Last, we hypothesized that a model utilizing both types of input data may be even more accurate and thus we developed combined models that use both metadata and fundus images (combined model).

First, we compared the performance of metadata-only, fundus-only and combined models trained to predict Hb, haematocrit (HCT) and red blood cell count (RBC), which are all correlated with each other and related to anaemia (Fig. 1). The mean absolute error (MAE) for predicting Hb when using the metadata-only model was 0.73 g dl^{-1} (95% confidence interval (CI): 0.72–0.74 g dl^{-1}). The MAE was 0.67 g dl^{-1} (95% CI: 0.66–0.68 g dl^{-1}) for the fundus-only model and 0.63 g dl^{-1} (95% CI: 0.62–0.64 g dl^{-1}) for the combined model. The performance of predicting HCT and RBC followed a similar trend across the three models (Supplementary Figs. 1 and 2). In addition, the performance was assessed in three age groups (between 40 and 49 years, 50 and 59 years and 60 and 69 years) (Supplementary Fig. 3). Study participants below 40 years old and above 70 years old were excluded from this analysis. The fundus-only model and the combined model did not show any significant difference in performance between age groups. Thus, the combined model predicted Hb, HCT and RBC more accurately than either the fundus-only or metadata-only models, which indicates that both metadata and fundus images were important for the accurate prediction, and the performance was consistent across age groups. We also examined the performance of the models when only including the 539 study participants who had self-reported diabetes, and we found a slightly larger MAE (for example, 0.73 g dl^{-1} (95% CI: 0.68–78 g dl^{-1}) for the combined model; Supplementary Fig. 4).

We next analysed the predictions using a Bland–Altman plot[26] (Fig. 1), and observed a negative slope in the linear fit, which indicates a proportional bias. That is, study participants in the lower range were overestimated, and participants in the higher range were underestimated. We hypothesized that the proportional bias could be reduced if the model outputs were calibrated to have the same variance as that of the ground-truth measurements (Methods). This calibration reduced the proportional bias (slope of linear fit) from −0.31 (95% CI: −0.33, −0.29) to −0.01 (95% CI: −0.03, 0.00) for the combined model, while slightly increasing the MAE (0.63 (95% CI: 0.62–0.64) to 0.67 (95% CI: 0.66–0.68)) (Supplementary Fig. 5).

Additionally, we investigated whether errors made by the model were subject-specific. In the validation set, 342 study participants had two visits with both fundus images and a Hb measurement. We applied the combined model to the two visits and found that the residual error was correlated between multiple visits over time by the same patient (Pearson's correlation coefficient $r = 0.38$ (95% CI: 0.18–0.65); Supplementary Fig. 6).

Next, we examined whether fundus images could be used to predict anaemia by developing a deep-learning-based classification model to directly predict whether a patient is anaemic. Using the Hb cut-off values for anaemia published by the WHO (World Health Organization), we trained each model to perform three binary classification tasks (anaemia: normal versus mild, moderate or severe; moderate anaemia: normal or mild versus moderate or severe; and, to enable comparison with results from a previous publication[9], 'approximate anaemia' (Methods)). For all tasks, the combined model and the fundus-only model performed better than the metadata-only model (Fig. 2; Table 2). The AUC for detecting anaemia was 0.73 (95% CI: 0.71–0.76) for the metadata-only model, 0.87 (95% CI: 0.85–0.89) for the fundus-only model and 0.88 (95% CI: 0.86–0.89) for the combined model (Fig. 2a). The AUC

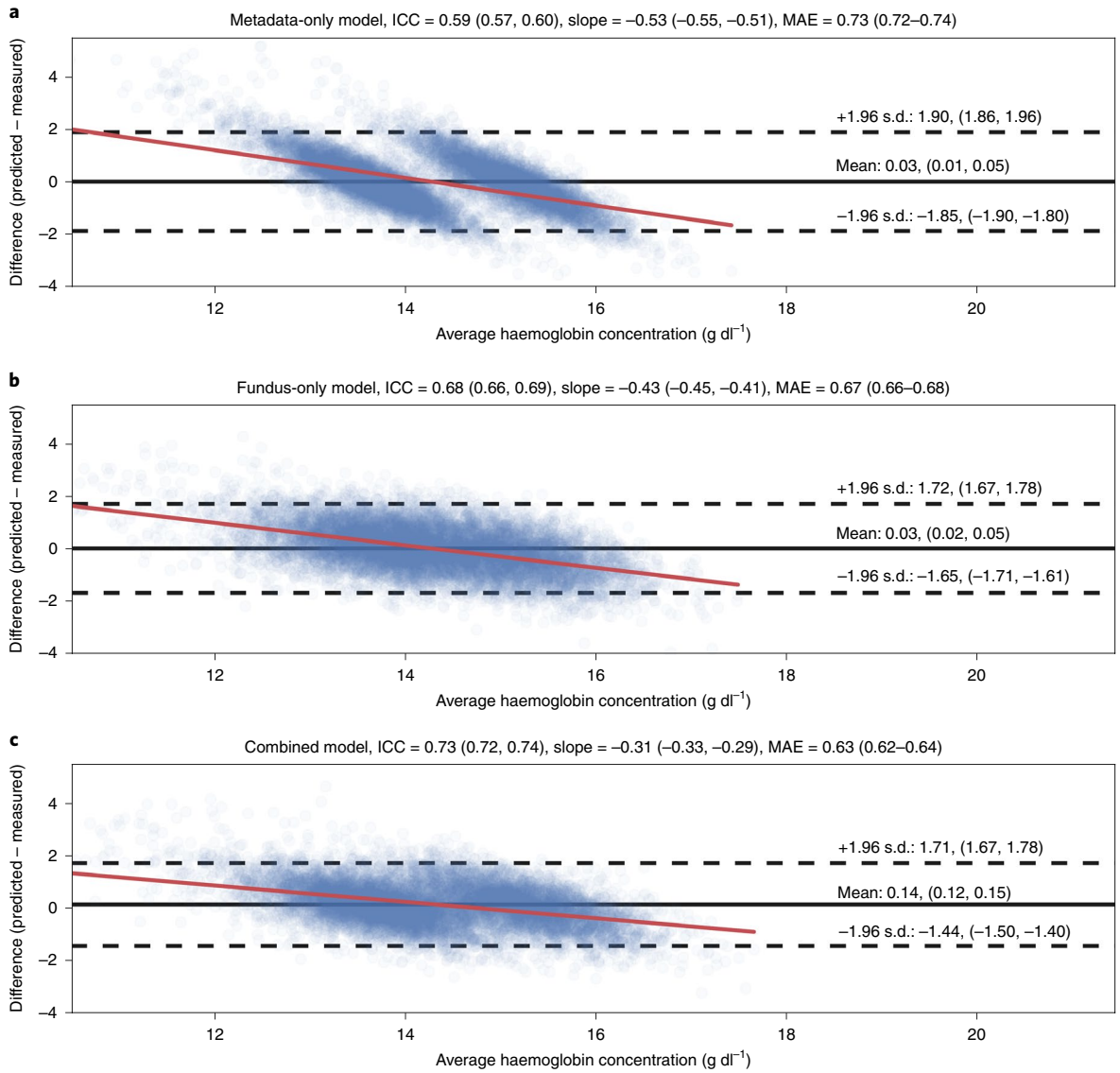

Fig. 1 | Bland–Altman plot for predicted and measured Hb. a, Each blue dot represents the difference between the measured Hb and the predicted value of a participant using the metadata-only model against the average of the two. The black unbroken line represents the mean of the difference, and black broken lines represent 95% limits of agreement (mean ± 1.96× s.d.). The red line represents a linear fit. The title text shows the intraclass correlation coefficient (ICC), the slope of the linear fit and the MAE. The numbers are shown with 95% CIs (n = 11,388). **b**, Same as **a**, but for the fundus-only model. **c**, Same as **a**, but for the combined model (leveraging both metadata and fundus images). The 95% CI of the limits of agreement did not overlap between that of the metadata-only model and the combined model; for example, 1.67, 1.78 versus 1.86, 1.96.

for detecting moderate anaemia was 0.79 (95% CI: 0.74–0.84) for the metadata-only model, 0.95 (95% CI: 0.93–0.97) for the fundus-only model and 0.95 (95% CI: 0.93–0.97) for the combined model (Fig. 2a). At 80% specificity, the sensitivity for detecting moderate anaemia was 64.1% (95% CI: 54.1–73.8) for the metadata-only model, 94.6% (95% CI: 91.0–99.0) for the fundus-only model and 93.5% (95% CI: 88.1–97.8) for the combined model (Table 2). We also examined classification task performance using the predicted Hb of the regression models. The AUC represents the probability that the predicted value for a study participant from one category (for example, anaemia) is lower than the predicted value for a second category (for example, not anaemia), and thus the AUC of a regression model shows the utility of the predicted value in making binary decisions by applying a threshold. Metadata-only, fundus-only and combined regression models had similar AUC values to the corresponding classification models (Fig. 2b). To further examine the classification performance of the regression models, the positive predictive values of each model with varying thresholds were plotted (Supplementary Fig. 7). The results showed that the study participants with lower predicted Hb were more likely to have anaemia, and the fundus-only model and the combined model achieved higher positive predictive values than the metadata-only model. We also examined the performance of the models in a subgroup with self-reported diabetes (n = 539), and confirmed that the models had comparable performance. For example, the AUC for detecting anaemia of the combined model was 0.89 (95% CI: 0.85–0.93; Supplementary Fig. 8). These results show that both the fundus-only model and the combined model successfully extracted information about anaemia from fundus images, and it supports the hypothesis that deep-learning models can help detect anaemia using fundus images.

We further investigated the importance of different anatomical features to the prediction of anaemia by perturbing the images

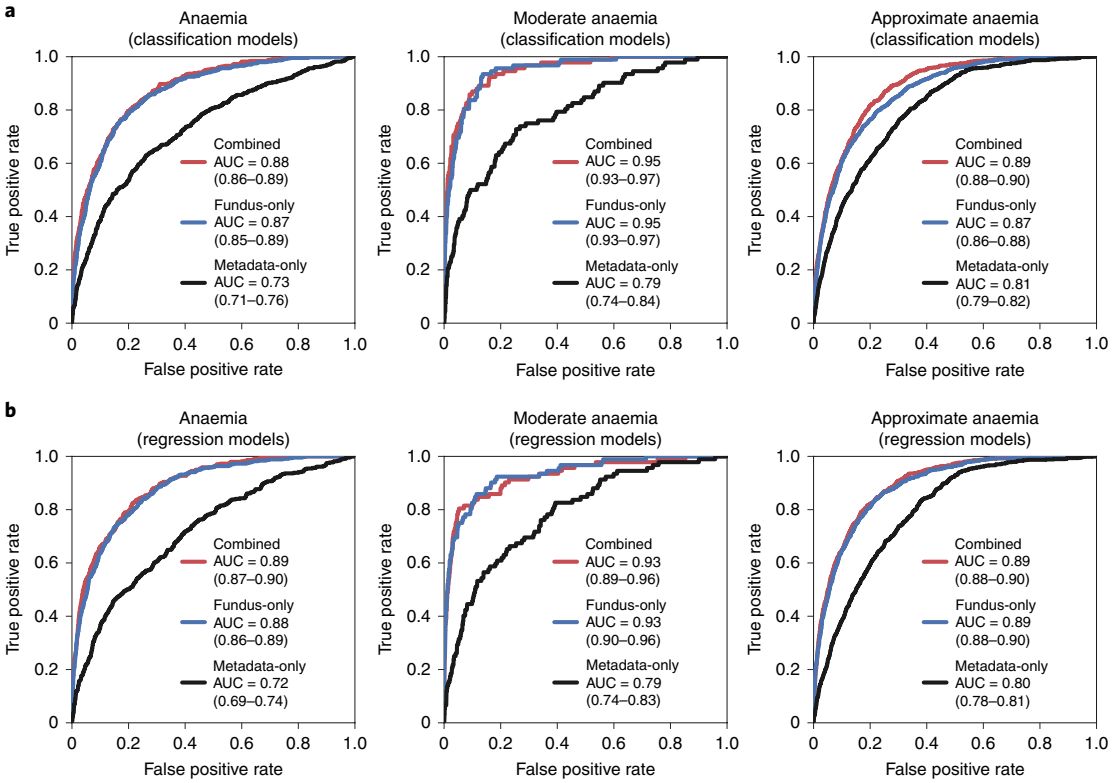

Fig. 2 | Prediction of anaemia classifications. a, ROC curves for detecting anaemia (left), moderate anaemia (middle) and approximate anaemia (right; see Methods) using the metadata-only model, the fundus-only model and the combined model. AUC values for models are shown with 95% CIs ($n = 11{,}388$). **b**, Same as **a**, but using the predicted Hb from the regression models instead of the outputs of the classification models.

Table 2 | Sensitivity at various levels of specificity

	Specificity		
	70%	**80%**	**90%**
Anaemia			
Metadata-only	65.1% (60.8–69.5)	53.5% (49.0–58.7)	38.5% (33.6–43.3)
Fundus-only	86.8% (83.6–90.0)	78.6% (74.9–82.5)	60.4% (55.0–65.1)
Combined	**87.5%** (84.7–90.9)	**79.5%** (75.6–83.4)	**62.2%** (57.2–66.7)
Moderate anaemia			
Metadata-only	73.9% (65.9–83.7)	64.1% (54.1–73.8)	48.9% (39.8–60.4)
Fundus-only	**95.7%** (92.7–100.0)	**94.6%** (91.0–99.0)	83.7% (75.8–91.0)
Combined	**95.7%** (92.5–100.0)	93.5% (88.1–97.8)	**85.9%** (79.0–93.3)
Approximate anaemia			
Metadata-only	75.7% (72.2–78.8)	61.3% (57.9–65.1)	43.9% (39.6–47.6)
Fundus-only	85.2% (82.8–87.8)	76.3% (73.2–79.3)	60.5% (56.3–64.0)
Combined	**89.7%** (87.6–91.9)	**81.6%** (78.5–84.4)	**61.1%** (57.3–64.9)

Sensitivity is presented with 95% CIs. Bold text indicates the highest sensitivity among the models at each specificity and condition.

during both model training and validation (Figs. 3 and 4). Most notably, when the upper and lower parts of the images were masked, the performance started to decline only after about 80% of each image was masked (Fig. 3a,b, left). Masking the horizontal stripes through the middle of the images decreased the performance after about 20% of each image was masked (Fig. 3c,d, left). When either the circular central core or the outer rim of the image was masked, the performance started to decline after about 40% was masked (Fig. 3b,d, right). Masking using a central horizontal stripe caused the biggest drop in AUC (about 3%) when 10% of the image was masked. In particular, Fig. 3a,b illustrates that including the disc and the macula horizontally increased the AUC more than only including the central circular part around the macula. By contrast, Fig. 3c,d shows that masking both the disc and the macular horizontally decreased the AUC more than just masking the macula. We also examined the effect of removing high-frequency image information via Gaussian blurs (Fig. 4), and found that applying a Gaussian blur with $\sigma = 8$ pixels decreased the AUC for predicting moderate anaemia from 0.92 to 0.83 (Fig. 4c). Notably, the models performed better than chance even with severe perturbations (for example, the AUC for predicting anaemia was 0.63 after applying a Gaussian blur with $\sigma = 32$ pixels). We hypothesized that the models could make use of the colour distribution (for example, corresponding to the general pallor of the retina with severe anaemia). To test this, we removed spatial information by randomly scrambling the order of the image pixels during training and evaluation, and the AUC of the model (0.60, average of 3 runs) remained better than chance for predicting anaemia, which supports the hypothesis.

Results from applying multiple model explanation techniques (GradCAM[27], Smooth Integrated Gradients[28,29] and Guided-

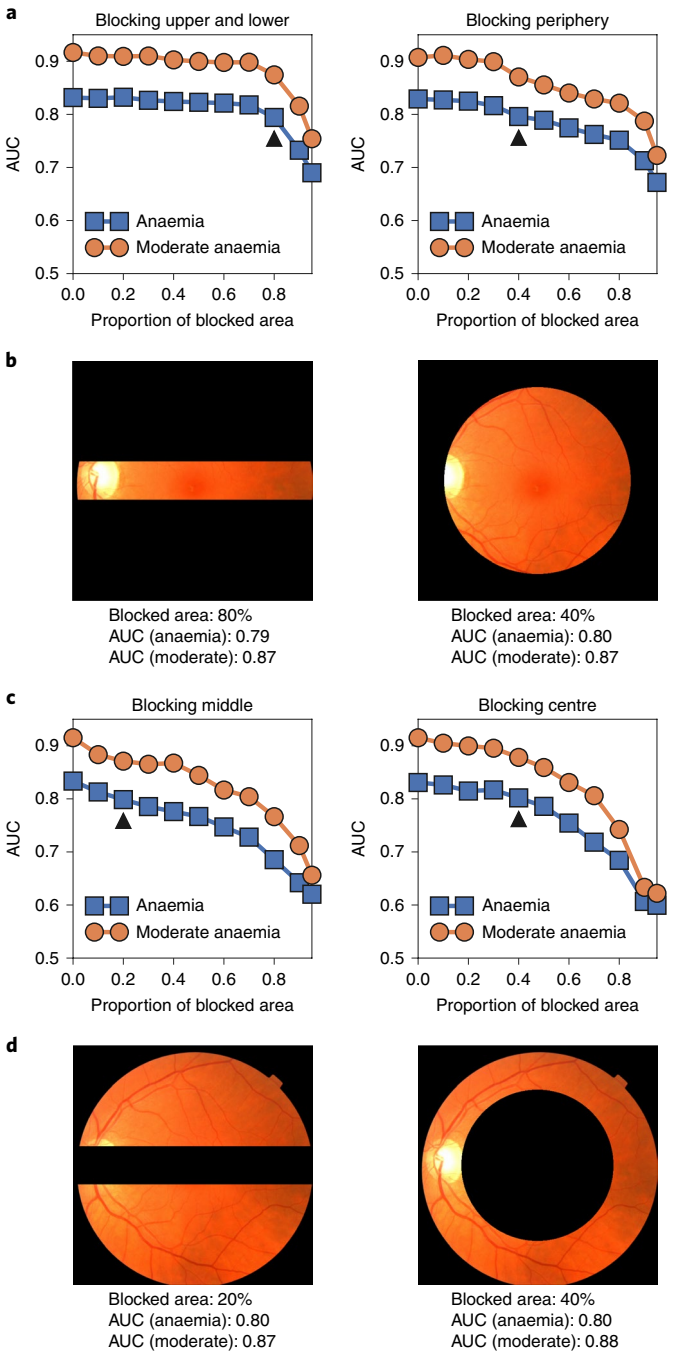

Fig. 3 | Effects of masking parts of the image on the prediction of anaemia and moderate anaemia. The masking was applied during both training and validation. **a**, Left: masking the upper and lower parts of the images. Right: masking the outer rim of the images. The arrowheads correspond to the respective examples shown in **b**. **b**, Examples of the masked images used for **a**. **c**, Left: masking a horizontal stripe through the middle of the images. Right: masking a central core of the images. The arrowheads correspond to the respective examples shown in **d**. **d**, Examples of the masked images used for **c**. For **a** and **c**, all arrowheads on the curves were manually chosen to represent medium performance (anaemia AUC near 0.80). The *x* axes end at 95% occlusion by area, which is equivalent to retaining a square crop that has 22% of the height and width ($0.22 \times 0.22 = 0.048$).

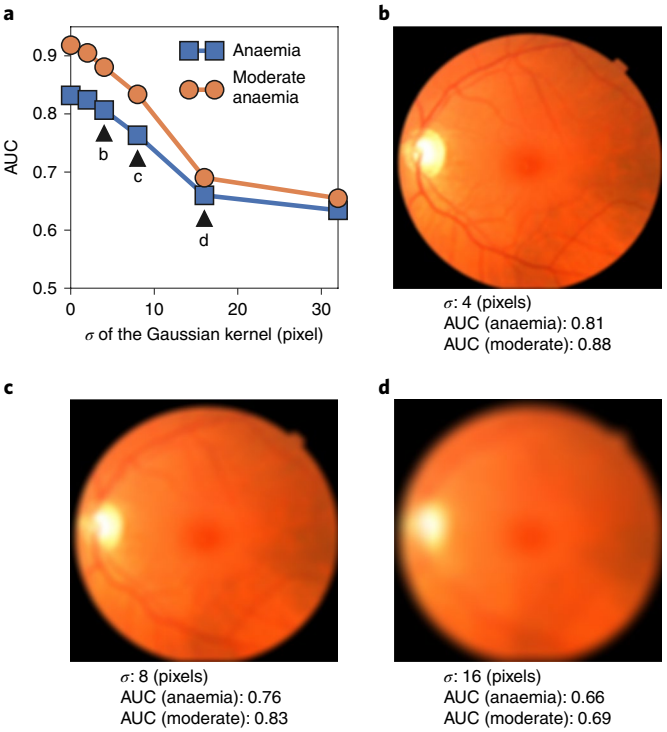

Fig. 4 | Effects of removing high-frequency information using Gaussian blur on the prediction of anaemia and moderate anaemia. The ablation was applied during both training and validation. **a**, Model performance as a function of a Gaussian blur amount. Arrowheads correspond to examples shown in **b**–**d**. **b**–**d**, Example images with varying blur amounts. Arrowheads were chosen manually to represent medium performance (anaemia AUC near 0.80 (**b**)) low performance (anaemia AUC near 0.75 (**c**)) and visually based on 'hinges' in the curve (**d**). σ, s.d. of the Gaussian kernel used for a blur, which was applied after images were resized to 587×587 pixels.

backprop[30]) to the fundus-only model are presented in Fig. 5. The saliency maps from the three explanation techniques suggest that the model tends to focus on the spatial features around the optic disc, sometimes extending along the entire length of the blood vessel in the image.

Last, we examined whether other components of the complete blood count (CBC) could be predicted from fundus images. Anaemia is diagnosed on the basis of blood Hb measurements, which is often measured as a part of the CBC. In addition to Hb, HCT and RBC, other measurements such as the mean corpuscular volume (MCV) also help diagnose anaemia and identify the subtype. Since the different components of CBC are measured on separate scales, we compared the performance of the model across tests with the R^2 coefficient of determination. The combined model was not able to predict MCV with high accuracy, with a low R^2 of 0.12. By contrast, the model predicted the three anaemia-related measurements (Hb, HCT and RBC) most accurately, with R^2 values of 0.52, 0.49, and 0.36, respectively (Supplementary Table 1).

Discussion

This study shows that a deep-learning-based approach that leverages retinal fundus images and metadata can both detect anaemia and quantify Hb measurements, potentially enabling automated anaemia screening using fundus images.

To help put the accuracy of our and other non-invasive anaemia detection methods in context, it can be useful to consider the variability of the ground truth itself[31,32]. Consistent with other studies, the ground truth in this study was the Hb measured using laboratory haematology analysers[33]. The s.d. of the difference between haematology analysers and the haemoglobincyanide method (HiCN, the gold standard for Hb measurement for

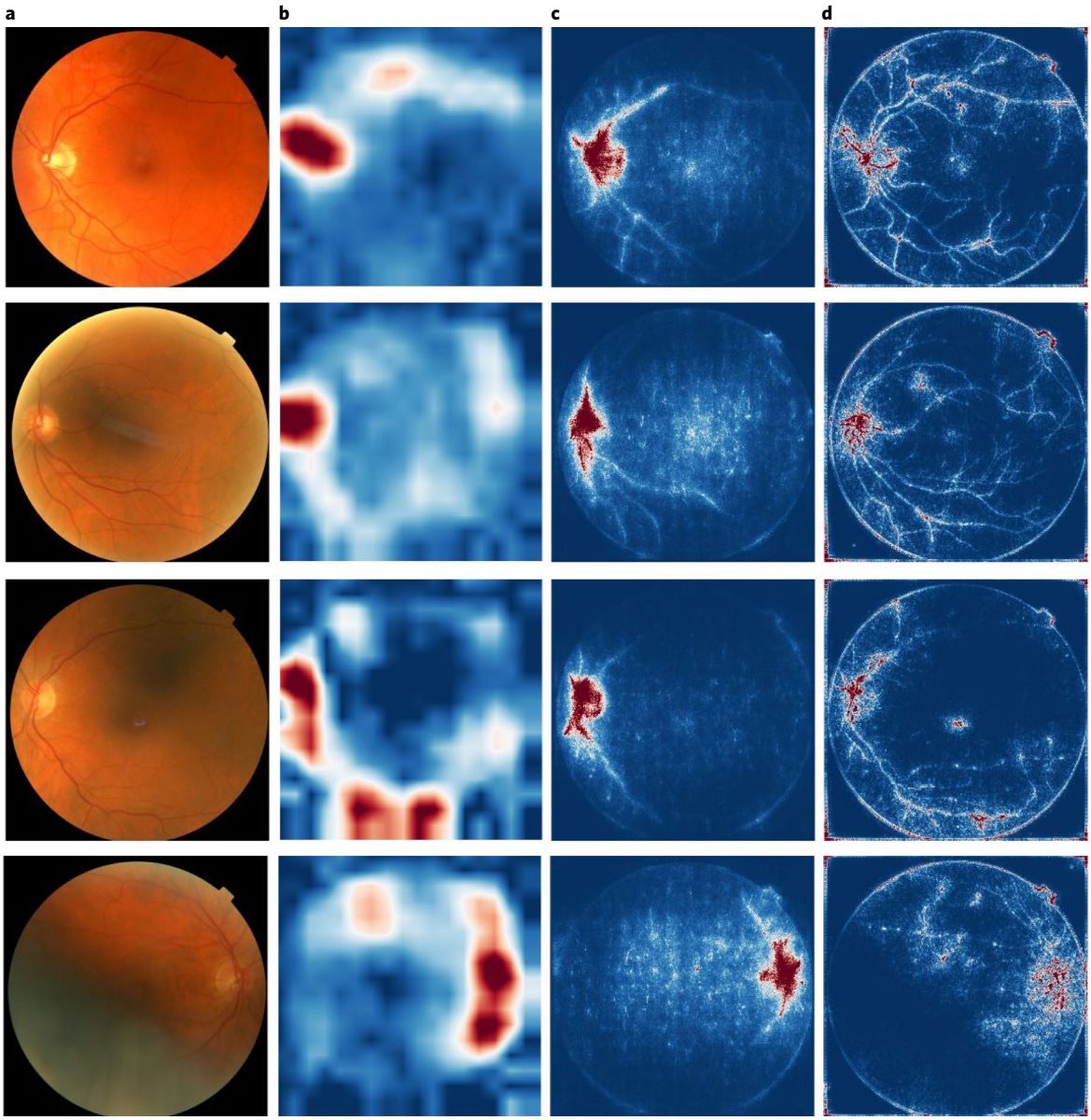

Fig. 5 | Examples applying different explanation techniques to generate saliency maps highlighting the regions the model focuses on when predicting anaemia. a, The original fundus images. Explanation techniques were applied to each image to generate saliency maps in the same row. Saliency maps from GradCAM[27] (**b**), Smooth Integrated Gradients[28,29] (**c**) and Guided-backprop[30] (**d**). In each saliency map, red and white areas represent regions the model is positively influenced by when predicting anaemia. Red indicates a stronger contribution than white, and blue regions have little to no contribution towards the prediction. The model correctly predicts moderate anaemia in each of the four cases.

research) was 0.18 g dl^{-1} (ref. [34]), which is equivalent to a MAE of 0.14 g dl^{-1}. Thus, a portion of the MAE of our model (0.63 g dl^{-1}) may be attributable to variability in the laboratory Hb measurement. The accuracy of our approach is also comparable to invasive (the pooled s.d. of difference was 0.64 g dl^{-1} (ref. [35]), which is equivalent to a MAE of 0.51 g dl^{-1}) and non-invasive point-of-care devices (the pooled s.d. was 1.4 g dl^{-1} (refs. [35,36]), which is equivalent to a MAE of 1.1 g dl^{-1}), and a non-invasive smartphone-based application (the 95% limits of agreement was 2.4 g dl^{-1} (ref. [9]), equivalent to a MAE of 0.96 g dl^{-1}) (Table 3).

There are several potential applications of our approach to anaemia detection. First, because fundus photographs are routinely captured as part of teleretinal screening for diabetic retinopathy[37–39], the ability to predict Hb from these photographs may enable seamless anaemia screening in the patient population with diabetes, with minimal additional cost as an add-on to diabetic retinopathy screening. Additionally, we confirmed that the models perform similarly in detecting anaemia in a subgroup with self-reported diabetes (Supplementary Figs. 4 and 8). This is important because anaemia is twice as common in patients with diabetes, and although anaemia in this patient population is frequently caused by nephropathy, up to 23% of those with anaemia remain unrecognized in patients with diabetes, a proportion that is higher than the general population both with and without chronic kidney disease[40–43]. Anaemia in patients with diabetes have various aetiologies[44,45], and when anaemia is detected, laboratory workup (for example, serum ferritin, vitamin B12 and folate levels) is indicated to isolate the cause, and treatment may be considered based on the aetiology[45]. Ultimately, the correction of anaemia improves quality of life and may reduce diabetic complications[41].

Table 3 | Accuracy of different methods to measure Hb

	Mean of differences (g dl^{-1})	95% limits of agreement (g dl^{-1})		(Estimated) MAE[a] (g dl^{-1})
		Lower limit	Upper limit	
Haematology analyser[34]	0.26	−0.11[b]	0.63[b]	0.14[a]
Invasive point-of-care devices[35]	−0.03 (−0.30 to 0.23)	−1.3 (ref. [35])	1.4 (ref. [35])	0.51[a]
Non-invasive point-of-care devices[35,36]	0.08 (−0.04 to 0.20)[35]	−3.0 (ref. [35])	2.9 (ref. [35])	1.2 (ref. [35])[a]
	0.10 (ref. [36])	−2.6 (ref. [36])	2.8 (ref. [36])	1.1 (ref. [36])[a]
Smartphone-based method[9]	0.2	−2.2[b]	2.6[b]	0.96[a]
Our method (combined model)	0.14 (0.12 to 0.15)	−1.85 (−1.90 to −1.80)	1.90 (1.86 to 1.96)	0.63 (0.62 to 0.64)

[a]MAE was estimated assuming that the error has a Gaussian distribution. Haematology analyser was compared with the haemoglobincyanide method. Other methods were compared with the haematology analyser. Numbers in parentheses indicate 95% CIs. [b]Calculated from mean of difference and 2× s.d.

In addition, from an ophthalmic standpoint, anaemia is an independent risk factor for developing high-risk proliferative diabetic retinopathy[46]. Identifying this risk factor can potentially be used to individualize the follow-up intervals of diabetic retinopathy screening to improve the effectiveness of the programme[47]. Thus, screening for anaemia as an add-on is beneficial to patients with diabetes enrolled in a teleretinal-screening programme.

Teleretinal screening is also increasingly performed in non-specialist settings, such as primary care and retail stores, where ophthalmologists are not on-site. In fact, teleretinal screening has been used in diabetic retinopathy screening programmes in the United States (predominantly serving minority populations)[48] and represents the vast majority of diabetic eye screening in the United Kingdom[49], and, increasingly, the standard of care globally[50,51]. Together with the development of fully automated diabetic retinopathy screening tools based on artificial intelligence[52] and low-cost mobile devices with image qualities approaching that of standard 'table-top' fundus cameras[53], the prevalence of fundus photographs is likely to increase in the future, leading in principle to an even broader application of our proposed method.

Besides applications as an add-on to diabetic retinopathy screening, there have been other published works in this area, such as predicting cardiovascular disease from fundus photographs[23]. As such, our work adds to the body of literature in this area, potentially enabling eventual opportunities for low-cost, non-invasive screening for multiple diseases in the general population and is not limited to patients with diabetes, as ocular imaging becomes more widely available.

Finally, our method can be valuable for clinical research because new information such as Hb levels can be inferred from previously collected patient data from existing research or clinical care. Of particular value is the potential to study the association of anaemia with ocular disease, for which fundus images are frequently already available. For example, there has been interest in assessing anaemia as a risk factor for glaucoma[54]. Our method can be retrospectively applied to the existing clinical data to conduct exploratory analysis of the association of ocular findings or diseases with anaemia, without the need to prospectively enrol patients and perform invasive Hb measurements.

One potential caveat of our algorithm is its proportional bias, albeit resolvable via calibration. Generally, a difference in variance between two methods leads to proportional bias, and regression models by definition have a residual error term, resulting in a smaller output variance than the variance of the ground-truth data. Therefore, our calibration 'expands' the model prediction around the mean such that the predictions have the same variance as the ground-truth data. During calibration, the ratio between the variance of the model outputs and that of the measurements was estimated using the tuning set, and the calibration based on this ratio reduced the proportional bias in the validation set, but at the cost of a slightly higher MAE. Notwithstanding this proportional bias and need for calibration, the receiver operating characteristic (ROC) analysis using the regression model outputs showed that the predicted values are adequately rank-ordered.

Another interesting observation here is whether users of this technology should focus on the classification versus regression model and how to use it. Although we may expect from a statistical machine-learning perspective that optimization of the specific classification label (for example, anaemia versus no anaemia) would produce the most accurate results in classification tasks, our results did not demonstrate a clear superiority of the classification models. Given the clear advantage of the regression models to also predict the actual Hb (enabling the use of adjusted anaemia cut-off values as desired), we recommend using the regression models. In this regard, the proposed calibration step will remove proportional bias, leading to more accurate Hb predictions at both extremes. However, clinicians should still note that the Hb prediction accuracy is lower than invasive measurements and it may not be appropriate to either rule in or rule out anaemia based on the predictions by the model using the same threshold. For example, when the WHO-based threshold (12 g dl^{-1} for female and 13 g dl^{-1} for male) was used for predictions using the combined model, the specificity of detecting anaemia was 0.999, but the sensitivity was low at 0.073. By increasing the threshold by 1.5 g dl^{-1} to 13.5 g dl^{-1} for female and 14.5 g dl^{-1} for male, the specificity was only decreased to 0.80 and the sensitivity was increased to 0.79. It would be more useful to suspect anaemia in patients below this threshold and conduct follow-up venipuncture testing and/or blood tests if necessary.

Beyond the initial diagnosis, subsequent Hb measurements are used to track progression of anaemia and response to treatment. In addition, even without a diagnosis of anaemia, rapid decreases in Hb may indicate the existence or onset of an underlying disease. Thus, in addition to the Hb at one visit, the difference in Hb across visits is clinically important. Interestingly, our prediction error between multiple visits of the same study participant was correlated, showing that some component of the error is subject-specific. When comparing the differences in predictions across multiple visits, the subject-specific components cancelled out. Thus, our approach may provide additional value in monitoring the trend in Hb. This premise will need to be further assessed in future work, such as assessing the time delay between true Hb changes and changes in Hb predicted by the algorithm using fundus images; that is, whether the prediction of the model reflects instantaneous Hb or the average Hb over a certain time window. Understanding how the algorithm works would help answer this question. If, on the one hand, the algorithm is quantifying the degree of pallor in the fundus, we would expect minimal to no time delays in the order of minutes or hours. On the other hand, if the algorithm is examining features in microvasculature that develop over time, the time delay may be

weeks or months. Investigating how the algorithm reflects recent Hb changes and validating it with multiple measurements over time will be an important step towards determining clinical use cases of the algorithm.

To understand the underlying mechanisms of the model, we examined how the performance was affected by applying image ablation. First, we hypothesized that if the algorithm was based on the degree of pallor of the fundus as a whole, applying a Gaussian blur would have little effect on the performance. However, applying a Gaussian blur decreased the model performance, which indicates that the model was dependent on the fine spatial features of the fundus images. In addition, we applied various masking methods to examine which portions of the fundus images are relevant to the model performance. Masking the upper and lower parts had little to no effect on the model performance, which shows that the information contained in those areas was redundant. Masking the central core, which includes the macula, had less effect than masking the horizontal stripe through the middle part, which includes both the macula and disc. These results suggest that fine spatial features around the optic disc are crucial. Supporting this hypothesis, heatmaps created by GradCAM[27], Smooth Integrated Gradients[28,29] and Guided-backprop[30] highlighted the optic disc (Fig. 5), and Integrated Gradients[28,29] and Guided-backprop[30] particularly highlighted the blood vessels (Fig. 5c,d). Attribution to the camera notch and the four corners, in the case of Guided-backprop[30], is an artefact of the attribution method, and we also empirically show in Fig. 3g that excluding the peripheral rim of the image had little to no effect on model performance.

Another aspect of understanding the algorithm is whether it was detecting anaemia or the underlying pathophysiology specific to each subtype of anaemia. Anaemia has multiple subtypes, and each has a different underlying aetiology and requires different management. Other components of the CBC, in particular the MCV (average volume of red blood cells), are used to differentiate between the subtypes. For example, while iron-deficiency anaemia typically presents with normal MCV, vitamin B12 or folate deficiency typically presents with elevated MCV, and anaemia of chronic disease and thalassaemia presents with decreased MCV. Thus, we hypothesized that if fundus images contained information about subtype-specific pathophysiology, the algorithm would be able to predict CBC results beyond Hb, HCT and RBC. However, the results did not support this hypothesis, which indicates that the algorithm may be responding to features associated with the lack of haemoglobin itself. These results also illustrate that the algorithms may be useful for screening, but not for diagnosis. Patients would require referral and follow-up examinations, such as blood tests, before treatment. In addition, we did not distinguish between different anaemia subtypes, and the algorithm performance should be further validated using patients with various subtypes of anaemia across different demographic groups.

When developing machine-learning algorithms, it is crucial to include a broad range of examples in the training set so that the developed algorithm generalizes well in various settings. One of the limitations of our study is that we used a dataset from a single source. For example, ethnicity is highly biased towards whites in the UK Biobank dataset. In a community setting, we may see more patients with cataracts, and images may have more artefacts. In addition, this dataset consisted of mostly healthy people, and the number of patients with anaemia, especially those with severe anaemia, was limited. This could have contributed to the proportional bias in the regression tasks without calibration. Training and validating on multiple diverse datasets will be important for creating a generalizable algorithm.

To conclude, we showed that anaemia and Hb can be predicted from fundus images. Further research is warranted to examine whether the approach is useful for scalable screening of anaemia.

Methods

Study participants. The dataset for this study consisted of fundus images obtained from the UK Biobank[24], an observational study that recruited 500,000 participants, aged 40–69 years, across the United Kingdom between 2006 and 2010. The study was reviewed and approved by the North West Multi-Centre Research Ethics Committee. Each participant provided consent and underwent a series of health measurements and questionnaires. Age, race/ethnicity, sex and current smoking status were self-reported by the participants via questionnaires. Each participant also provided blood, urine and saliva samples. Approximately 70,000 study participants also subsequently underwent ophthalmological examinations with paired retinal fundus and optical coherence tomography imaging. Only retinal fundus images were included in this study. About 12% of the study participants were excluded due to poor image quality, as previously described[22]. Only study participants with at least one fundus image paired with a Hb measurement were included in this study ($n=57{,}163$). For CBC analysis, only the study participants who had all the CBC components measured were included ($n=53{,}473$). If a study participant had multiple visits with paired fundus images and Hb measurements, only the first visit was included, except in the multiple-visit analysis described. We randomly divided this dataset into a development set to develop our models (80%, further divided into 70% for training and 10% for tuning) and a validation set to assess the performance of our model (20%) after stratifying study participants by their sex and age. The validation set was not accessed during model development. The tuning set was used during model development for tuning hyper parameters such as learning rate and criterion for early stopping, with the training set used for training the parameters of the neural networks.

Definitions of anaemia. Using guidelines from the WHO[55], we used the following three sets of cut-off values based on Hb measurements: 12 g dl^{-1} for women and 13 g dl^{-1} for men (anaemia), and 11 g dl^{-1} (moderate anaemia). In addition, we assessed our results using a previously described sex-neutral average anaemia cut-off at 12.5 g dl^{-1} (ref. [9]) for both men and women, which we call 'approximate anaemia'.

Categories of predictive models. In this study, we made two different types of predictions: continuous values (for example, Hb or HCT; henceforth termed 'regression tasks') and categorical values (for example, the presence or absence of anaemia; henceforth termed 'classification tasks'). Although a single model can in principle be trained for both regression tasks and multiple classification tasks, separate models were trained for regression tasks and classification tasks to keep the loss functions on a consistent scale. For each of these tasks, we compared the ability of three different categories of prediction models, each with a different set of input data. As a baseline, we used linear regression for regression tasks and logistic regression for classification tasks. These linear and logistic regression models used only demographic and clinical information ('metadata', which are race/ethnicity, age, sex, current smoking status, systolic and diastolic blood pressure, pulse rate, height, weight and body mass index). We will refer to these as metadata models. Our second type of model used a deep convolutional neural network (details in the next section) with fundus images as input (fundus-only models). Our third and last type of model used both metadata and fundus images as input; the metadata was concatenated with the output of the Inception-v4 architecture[25] before the fully connected layer (combined models). Specifically, the fundus images were used as an input to a deep convolutional neural network (same structure as the fundus-only model), and the outputs of the convolutional neural network and metadata were provided to the combined model before the final layer (that is, a 'late fusion' model). The fully connected layer of the fundus-only models and the combined models had one output for each regression task and multiple outputs for each classification task (each output corresponds to each class).

Development of the deep-learning algorithms. Fundus images were preprocessed as previously described[23], while the categorical input metadata were represented as one-hot vectors (four classes for race/ethnicity and two classes each for sex and smoking status), and the continuous input metadata (age, blood pressures, pulse rate, height, weight and body mass index) and continuous output values (for example, Hb) were standardized to have zero mean and unit variance. Using these data, a deep convolutional neural network using the Inception-v4 architecture[25] was developed and trained in TensorFlow[56]. The Inception-v4 network was initialized using parameters from a network pretrained to classify objects in the ImageNet dataset[57], and the weights on the auxiliary connections from metadata were randomly initialized. Mean squared error was used as a loss function for regression tasks, and cross entropy was used for classification tasks. The models were trained with mini-batch stochastic gradient descent with momentum[58] with linear warm up[59] using Google Tensor Processing Unit (TPU) accelerators[60] (see Supplementary Methods). The learning rate was chosen to minimize the error in the tuning dataset. Since our network had a large number of parameters (43 million), to prevent overfitting, training was terminated before convergence using early stopping[61] based on the performance on the tuning dataset. An ensemble of ten networks[62] was trained on the same development set, and the outputs were averaged to yield the final prediction. For each study participant, the final prediction was the average across both eyes.

Evaluating the algorithms. To evaluate the model performance for continuous predictions, we used the MAE, 95% limits of agreement and R^2 values. For binary classification, we used the AUC and sensitivity at various levels of specificity. When plotting the ROC curves for the anaemia classification task using predicted Hb by regression models (Fig. 3d), the threshold for females was kept 1 g dl^{-1} lower than the threshold for males, as consistent with the WHO's definition. To obtain 95% CIs for these performance metrics, we used the non-parametric bootstrap procedure with 2,000 samples and report the 2.5 and 97.5 percentiles.

Age group analysis. To assess the effect of age on the model performance, the study participants in the validation set were stratified into three age groups (40–49 years, 50–59 years and 60–69 years). Participants aged below 40 years and those above 70 years were excluded due to the small numbers of participants in those ranges. Pairwise comparisons between age groups were conducted for each model using the bootstrap method, and *P* values were corrected for multiple comparisons using the Holm–Bonferroni method.

Ablation analysis. While applying ablation during both training and validation, we trained the fundus-only model for classification tasks and assessed the performance (AUC for predicting anaemia and moderate anaemia) of the model without ensembling or averaging across eyes. For each ablation method (for example, masking 20% of the fundus at the centre), three networks were trained, and the performance metrics were averaged across the three networks.

Model explanation. We used visual explanation tools to understand which regions in the fundus image our deep-learning model is most influenced by when predicting anaemia or moderate anaemia. We present the saliency maps obtained using GradCAM[27], Smooth Integrated Gradients[28,29] and Guided-backprop[30] in Fig. 5. All three methods present different ways of attributing the contributions of parts of the network to regions in the image. These can be presented as coloured heatmaps to give a visual indicator of the importance of any given region on the image. We applied GradCAM[27] to the final convolutional layer, and the effective resolution of the heatmap was equivalent to the resolution of the final layer. Integrated Gradients[28] measures the contribution of each pixel in fundus images and we additionally used the Smooth-Grad[29] variant to generate more refined heatmaps. Guided-backprop[30] inverts the data flow in a neural network by keeping track of only the positive gradients and discarding the negative gradients in each part of the network as it traces back to a specific location on the image to highlight features that positively contributed to the prediction.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data supporting the findings of this study are available, with restrictions, from the UK Biobank[24].

Code availability

The machine-learning models were developed by using standard model libraries and scripts in TensorFlow[26]. Custom code was specific to our computing infrastructure and mainly used for data input/output and parallelization across computers.

Received: 9 April 2019; Accepted: 11 November 2019;
Published online: 23 December 2019

References

1. McLean, E., Cogswell, M., Egli, I., Wojdyla, D. & de Benoist, B. Worldwide prevalence of anaemia, WHO Vitamin and Mineral Nutrition Information System, 1993–2005. *Public Health Nutr.* **12**, 444–454 (2008).
2. Stevens, G. A. et al. Global, regional, and national trends in haemoglobin concentration and prevalence of total and severe anaemia in children and pregnant and non-pregnant women for 1995–2011: a systematic analysis of population-representative data. *Lancet Glob. Health* **1**, e16–e25 (2013).
3. Stoltzfus, R. J. Iron-deficiency anemia: reexamining the nature and magnitude of the public health problem. Summary: implications for research and programs. *J. Nutr.* **131**, 697S–701S (2001).
4. Milman, N. Anemia—still a major health problem in many parts of the world! *Ann. Hematol.* **90**, 369–377 (2011).
5. Smith, R. E. Jr. The clinical and economic burden of anemia. *Am. J. Manag. Care* **16** (Suppl.), S59–S66 (2010).
6. Shah, N., Osea, E. A. & Martinez, G. J. Accuracy of noninvasive hemoglobin and invasive point-of-care hemoglobin testing compared with a laboratory analyzer. *Int. J. Lab. Hematol.* **36**, 56–61 (2014).
7. Kalantri, A., Karambelkar, M., Joshi, R., Kalantri, S. & Jajoo, U. Accuracy and reliability of pallor for detecting anaemia: a hospital-based diagnostic accuracy study. *PLoS ONE* **5**, e8545 (2010).
8. Kasper, D. L. et al. *Harrison's Principles of Internal Medicine* (McGraw Hill Professional, 2006).
9. Mannino, R. G. et al. Smartphone app for non-invasive detection of anemia using only patient-sourced photos. *Nat. Commun.* **9**, 4924 (2018).
10. Barker, S. J. & Badal, J. J. The measurement of dyshemoglobins and total hemoglobin by pulse oximetry. *Curr. Opin. Anaesthesiol.* **21**, 805–810 (2008).
11. Pinto, M. et al. The new noninvasive occlusion spectroscopy hemoglobin measurement method: a reliable and easy anemia screening test for blood donors. *Transfusion* **53**, 766–769 (2013).
12. Wittenmeier, E. et al. Comparison of the gold standard of hemoglobin measurement with the clinical standard (BGA) and noninvasive hemoglobin measurement (SpHb) in small children: a prospective diagnostic observational study. *Paediatr. Anaesth.* **25**, 1046–1053 (2015).
13. Posey, W. M. C. The ocular manifestations of anemia. *JAMA* **XXIX**, 169–171 (1897).
14. Aisen, M. L., Bacon, B. R., Goodman, A. M. & Chester, E. M. Retinal abnormalities associated with anemia. *Arch. Ophthalmol.* **101**, 1049–1052 (1983).
15. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
16. Gulshan, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**, 2402–2410 (2016).
17. Krause, J. et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology* **125**, 1264–1272 (2018).
18. Ting, D. S. W. et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA* **318**, 2211–2223 (2017).
19. Liu, S. et al. A deep learning-based algorithm identifies glaucomatous discs using monoscopic fundus photographs. *Ophthalmol. Glaucoma* **1**, 15–22 (2018).
20. Christopher, M. et al. Performance of deep learning architectures and transfer learning for detecting glaucomatous optic neuropathy in fundus photographs. *Sci. Rep.* **8**, 16685 (2018).
21. Li, Z. et al. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. *Ophthalmology* **125**, 1199–1206 (2018).
22. Varadarajan, A. V. et al. Deep learning for predicting refractive error from retinal fundus images. *Invest. Ophthalmol. Vis. Sci.* **59**, 2861–2868 (2018).
23. Poplin, R. et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.* **2**, 158–164 (2018).
24. Sudlow, C. et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
25. Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. A. Inception-v4, inception-ResNet and the impact of residual connections on learning. In *Proc. of the 31st AAAI Conference on Artificial Intelligence* 4278–4284 (AAAI Press, 2017).
26. Bland, J. M. & Altman, D. G. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **1**, 307–310 (1986).
27. Selvaraju, R. R. et al. Grad-CAM: visual explanations from deep networks via gradient-based localization.*International Journal of Computer Vision* https://doi.org/10.1007/s11263-019-01228-7 (2019).
28. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. In *Proc. of the 34th International Conference on Machine Learning* 3319–3328 (Microtome Publishing, 2017).
29. Smilkov, D., Thorat, N., Kim, B., Viégas, F. & Wattenberg, M. SmoothGrad: removing noise by adding noise. Preprint at https://arxiv.org/abs/1706.03825 (2017).
30. Springenberg, J. T., Dosovitskiy, A., Brox, T. & Riedmiller, M. Striving for simplicity: the all convolutional net. Preprint at https://arxiv.org/abs/1412.6806 (2014).
31. Bland, J. M. & Altman, D. G. Measuring agreement in method comparison studies. *Stat. Methods Med. Res.* **8**, 135–160 (1999).
32. Barker, S. J., Shander, A. & Ramsay, M. A. Continuous noninvasive hemoglobin monitoring: a measured response to a critical review. *Anesth. Analg.* **122**, 565–572 (2016).
33. Elliott, P. & Peakman, T. C. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *Int. J. Epidemiol.* **37**, 234–244 (2008).
34. Gehring, H. et al. Accuracy of point-of-care-testing (POCT) for determining hemoglobin concentrations. *Acta Anaesthesiol. Scand.* **46**, 980–986 (2002).
35. Hiscock, R., Kumar, D. & Simmons, S. W. Systematic review and meta-analysis of method comparison studies of Masimo pulse co-oximeters (Radical-7™ or Pronto-7™) and HemoCue® absorption spectrometers (B-Hemoglobin or 201+) with laboratory haemoglobin estimation. *Anaesth. Intensive Care* **43**, 341–350 (2015).
36. Kim, S.-H. et al. Accuracy of continuous noninvasive hemoglobin monitoring: a systematic review and meta-analysis. *Anesth. Analg.* **119**, 332–346 (2014).

37. Tsan, G. L. et al. Assessment of diabetic teleretinal imaging program at the Portland Department of Veterans Affairs Medical Center. *J. Rehabil. Res. Dev.* **52**, 193–200 (2015).
38. Conlin, P. R. et al. Nonmydriatic teleretinal imaging improves adherence to annual eye examinations in patients with diabetes. *J. Rehabil. Res. Dev.* **43**, 733–740 (2006).
39. Garg, S., Jani, P. D., Kshirsagar, A. V., King, B. & Chaum, E. Telemedicine and retinal imaging for improving diabetic retinopathy evaluation. *Arch. Intern. Med.* **172**, 1677–1678 (2012).
40. Jones, S. C. et al. Prevalence and nature of anaemia in a prospective, population-based sample of people with diabetes: Teesside anaemia in diabetes (TAD) study. *Diabet. Med.* **27**, 655–659 (2010).
41. Thomas, M. C., MacIsaac, R. J., Tsalamandris, C., Power, D. & Jerums, G. Unrecognized anemia in patients with diabetes: a cross-sectional survey. *Diabetes Care* **26**, 1164–1169 (2003).
42. Wright, J. A., Oddy, M. J. & Richards, T. Presence and characterisation of anaemia in diabetic foot ulceration. *Anemia* **2014**, 104214 (2014).
43. AlDallal, S. M. & Jena, N. Prevalence of anemia in type 2 diabetic patients. *J. Hematol.* **7**, 57–61 (2018).
44. Mehdi, U. & Toto, R. D. Anemia, diabetes, and chronic kidney disease. *Diabetes Care* **32**, 1320–1326 (2009).
45. Kliger, A. S. et al. KDOQI US commentary on the 2012 KDIGO clinical practice guideline for anemia in CKD. *Am. J. Kidney Dis.* **62**, 849–859 (2013).
46. Davis, M. D. et al. Risk factors for high-risk proliferative diabetic retinopathy and severe visual loss: early treatment diabetic retinopathy study report #18. *Invest. Ophthalmol. Vis. Sci.* **39**, 233–252 (1998).
47. Taylor-Phillips, S. et al. Extending the diabetic retinopathy screening interval beyond 1 year: systematic review. *Br. J. Ophthalmol.* **100**, 105–114 (2016).
48. Owsley, C. et al. Diabetes eye screening in urban settings serving minority populations: detection of diabetic retinopathy and other ocular findings using telemedicine. *JAMA Ophthalmol.* **133**, 174–181 (2015).
49. Scanlon, P. H. The English national screening programme for diabetic retinopathy 2003–2016. *Acta Diabetol.* **54**, 515–525 (2017).
50. Das, T. & Pappuru, R. R. Telemedicine in diabetic retinopathy: access to rural India. *Indian J. Ophthalmol.* **64**, 84–86 (2016).
51. American Diabetes Association. Standards of medical care in diabetes—2018 Abridged for primary care providers. *Clin. Diabetes* **36**, 14–37 (2018).
52. Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N. & Folk, J. C. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit. Med.* **1**, 39 (2018).
53. Tran, K., Mendel, T. A., Holbrook, K. L. & Yates, P. A. Construction of an inexpensive, hand-held fundus camera through modification of a consumer 'point-and-shoot' camera. *Invest. Opthalmol. Vis. Sci.* **53**, 7600–7607 (2012).
54. Firat, P. G., Demirel, E. E., Dikci, S., Kuku, I. & Genc, O. Evaluation of iron deficiency anemia frequency as a risk factor in glaucoma. *Anemia* **2018**, 1456323 (2018).
55. WHO. Haemoglobin concentrations for the diagnosis of anaemia and assessment of severity. *WHO* https://www.who.int/vmnis/indicators/haemoglobin.pdf (2011).
56. Abadi, M. et al. TensorFlow: a system for large-scale machine learning. In *Proc. of the 12th USENIX Symposium on Operating Systems Design and Implementation* 265–283 (USENIX Association, 2016).
57. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Proc. of the 25th Conference on Advances in Neural Information Processing Systems* 1097–1105 (Curran Associates, 2012).
58. Sutskever, I., Martens, J., Dahl, G. & Hinton, G. On the importance of initialization and momentum in deep learning. In *Proc. of the 30th International Conference on Machine Learning* 1139–1147 (Microtome Publishing, 2013).
59. Priya, G. et al. Accurate, Large Minibatch SGD: training ImageNet in 1 hour. Preprint at https://arxiv.org/abs/1706.02677 (2017).
60. Jouppi, N. P. et al. In-datacenter performance analysis of a tensor processing unit. In *Proc. of the 44th Annual International Symposium on Computer Architecture* 1–12 (ACM New York, 2017).
61. Caruana, R., Lawrence, S. & Giles, L. Overftting in neural nets: backpropagation, conjugate gradient, and early stopping. In *Proc. of the 13th Conference on Advances in Neural Information Processing Systems* 381–387 (MIT Press, 2001).
62. Opitz, D. & Maclin, R. Popular ensemble methods: an empirical study. *J. Artif. Intell. Res.* **11**, 169–198 (1999).

Acknowledgements

This research was conducted using the UK Biobank Resource under application number 17643. The work of A.H. was done via Advanced Clinical, San Francisco, USA. The authors thank C. Angermueller from Google Research for his engineering contributions and A. Zaidi, A. Narayanaswamy, C. Chen, J. Krause and R. Sayres from Google Research for their advice and assistance with reviewing the manuscript.

Author contributions

A.M., G.S.C., L.P., D.R.W., N.H. and A.V.V. designed the research. A.M., L.P. and A.V.V. acquired data from the UK Biobank. A.M. executed the research and analysed the data. S.V. conducted the model explanation analysis. A.M., Y.L. and A.V.V. interpreted the results. A.M., A.H., Y.L. and N.H. prepared the manuscript. All authors contributed to manuscript revision and approved the submitted version.

Competing interests

The authors are employees of Google and own Alphabet stock or are working at Google. A.M., A.V.V., L.P. and D.R.W. are inventors on a patent applied by Google related to this work (current status: pending).

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41551-019-0487-z.

Correspondence and requests for materials should be addressed to A.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Akinori Mitani

Last updated by author(s): Nov 4, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software specific to this study was used for data collection.
Data analysis	TensorFlow and python libraries.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data supporting the findings of this study are available, with restrictions, from the UK Biobank.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	On the basis of our previous experience and of published literature, we know that deep learning requires on the order of tens of thousands or of hundreds of thousands of examples. Therefore, we included as much available data as possible.
Data exclusions	We excluded any images that were of poor quality or that had missing data. These were pre-established exclusions.
Replication	The model was developed by using the training dataset, and the hyper-parameters were tuned on the basis of the performance on the tuning dataset. The validation dataset was not accessed during model development, and the reported performance is based on the validation dataset.
Randomization	Samples were randomly allocated to the training, tuning and validation datasets after stratifying for age and gender.
Blinding	This is a retrospective study. Splits for validation were random and automatically generated. No blinding was necessary.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Retinal fundus images were obtained from the general adult population in the UK.
Recruitment	Participants were recruited at 22 recruitment centers in the UK. A wide range of backgrounds are represented.
Ethics oversight	The North West Multi-Centre Research Ethics Committee

Note that full information on the approval of the study protocol must also be provided in the manuscript.