AI-Driven Insights from Ocular and Medical Imaging

A series of research initiatives demonstrate that deep learning models can extract a remarkable depth of novel systemic health information from ocular and other medical images, significantly extending diagnostic capabilities beyond traditional human interpretation. These AI systems can accurately predict a wide range of conditions and biomarkers—including refractive error, anemia, and indicators of kidney disease—from retinal fundus and external eye photographs.

A critical theme across this body of work is the advancement of explainability techniques to build trust and drive scientific discovery. Initial methods like attention maps successfully identified *where* models focused (e.g., the fovea for refractive error), but recent breakthroughs using generative AI (the "StylEx" framework) now visualize *what* specific visual features are being altered, creating counterfactual images that illustrate these changes.

This advanced explainability has yielded profound insights. The models have not only rediscovered known clinical signs, validating their efficacy, but have also uncovered how AI systems learn unintended correlations from dataset biases and socio-cultural confounders (e.g., associating eyeliner with low hemoglobin due to a shared correlation with sex). Most significantly, this AI-driven approach establishes a new paradigm for biomedical research, generating novel, testable hypotheses about disease pathophysiology. However, the research strongly concludes that interpreting these complex findings requires a collaborative, interdisciplinary approach, uniting clinical, technical, and socio-technical expertise to correctly distinguish between true biological signals, data artifacts, and the influence of social determinants on health data.

1. Expanding Diagnostic Horizons with Deep Learning on Ocular Images

Deep learning has unlocked the ability to predict a variety of systemic health factors from images of the eye, a capability not previously thought possible. These models identify subtle patterns in retinal fundus and external eye photographs to infer conditions related to vision, blood health, and kidney and liver function.

1.1 Predicting Refractive Error from Retinal Fundus Images

Uncorrected refractive error is a leading cause of visual impairment globally. A landmark study demonstrated that a deep learning model can predict this condition directly from retinal fundus images.

- **Objective:** To train and validate a deep learning algorithm to predict spherical equivalent, spherical power, and cylindrical power from fundus photographs.
- **Methodology:** The model, which combines a ResNet architecture with a "soft-attention" layer, was trained on a total of 226,870 images from two large datasets: the UK Biobank and the Age-Related Eye Disease Study (AREDS).

- **Performance:** The algorithm demonstrated high accuracy, particularly on the UK Biobank validation set, significantly outperforming a baseline statistical model.
    - **Spherical Equivalent:** The model achieved a Mean Absolute Error (MAE) of 0.56 diopters (D), compared to a baseline MAE of 1.81 D. The model's predictions were within 1D of the actual value 86% of the time.
    - **Refractive Components:** The model was highly accurate for spherical power (MAE of 0.63 D) but, as expected, was not accurate for cylindrical power (astigmatism), which is primarily related to corneal and lens shape not visible in a fundus image.
- **Explainability:** Attention maps, which visualize the regions most influential to the model's prediction, consistently highlighted the foveal region of the retina as a critical area for prediction, regardless of the type of refractive error. This finding suggests a new, non-biased avenue for research into the pathophysiology of myopia and hypermetropia.

1.2 Detecting Anemia from Retinal Fundus Images

Anemia affects over 1.6 billion people worldwide, but diagnosis typically requires an invasive blood test. Research has shown that deep learning can non-invasively quantify hemoglobin (Hb) concentration and detect anemia from fundus images, often captured during routine diabetic eye screenings.

- **Objective:** To develop a deep learning algorithm to predict Hb levels using fundus images, demographic metadata, or a combination of both.
- **Methodology:** Models were developed using data from 57,163 subjects in the UK Biobank. Three model types were compared: metadata-only, fundus-only, and a combined model using both inputs.
- **Performance:** The combined model, leveraging both the fundus image and metadata, performed best, indicating that the images contain predictive information independent of demographic data.

Model Type	Hemoglobin MAE (g/dL)	Anemia Detection AUC
Metadata-only	0.73 [0.72-0.74]	0.74 [0.71-0.76]
Fundus-only	0.67 [0.66-0.68]	0.87 [0.85-0.89]
Combined	**0.63 [0.62-0.64]**	**0.88 [0.86-0.89]**

- **Explainability:** Model explanation techniques (GradCAM, Smooth Integrated Gradients) and image ablation studies suggested that the model focuses on the optic disc and the surrounding blood vessels to make its predictions. The model's accuracy decreased when fine spatial features were blurred, indicating its reliance on more than just the general "pallor" of the retina.

1.3 Identifying Systemic Biomarkers from External Eye Photographs

Building on the success with fundus images, a subsequent study investigated whether photographs of the external eye—which are easier to capture—could also reveal signs of systemic disease.

- **Objective:** To develop a deep learning system (DLS) to predict nine pre-specified biomarkers related to kidney, liver, blood, and endocrine function from external eye photos.
- **Methodology:** A DLS was developed using 123,130 images from patients in Los Angeles County and evaluated on three independent validation sets from Los Angeles and the Atlanta area, encompassing diverse patient populations.
- **Performance:** The DLS demonstrated a statistically significant ability to outperform baseline models that used only clinicodemographic data (e.g., age, sex, race) for several key biomarkers. Performance was particularly robust for markers of kidney disease and anemia across all three diverse validation sets.
    - **Severely Increased Albuminuria (ACR ≥ 300 mg/g):** AUC improvement over baseline ranged from 8.7% to 13.2%.
    - **Moderate Anemia (Hgb < 11.0 g/dL):** AUC improvement over baseline ranged from 7.3% to 19.9%.
    - **Other Successes:** The DLS also significantly outperformed the baseline in detecting abnormalities in eGFR, AST (liver), platelets, and WBC in the validation set most similar to the development data.
- **Explainability:** Ablation experiments, where parts of the image like the pupil or iris were masked, suggested that predictive signals are distributed across the eye and that color information is important for many predictions.

2. The Evolution of Explainability in Medical AI

A central challenge in deploying medical AI is understanding *how* it arrives at a conclusion. The research reflects a significant evolution in explainability techniques, moving from simply identifying areas of interest to visualizing the specific, fine-grained features the models have learned.

2.1 Early Approaches: Localization and Saliency

Initial explainability efforts focused on localizing the most important pixels or regions in an image for a given prediction.

- **Attention Maps:** Used in the refractive error study, these heatmaps highlight regions the model "attends" to, successfully identifying the fovea as a key area.
- **Ablation and Saliency Maps:** Employed in the anemia and external eye studies, these methods involve masking parts of an image to see how

performance is affected or using techniques like GradCAM to generate heatmaps. These confirmed the importance of the optic disc and conjunctiva.
- **Limitation:** While these methods effectively answer "where" the model is looking, they do not explain "what" specific features (e.g., texture, shape, color change) within that region are driving the prediction.

2.2 A New Frontier: Generative AI for Counterfactual Explanations
To overcome the limitations of saliency maps, a new framework was developed using generative AI to produce intuitive visual explanations.

- **The StylEx Framework:** This novel, four-stage workflow uses a StyleGAN-based image generator (StylEx) guided by a pre-trained classifier.
    1. **Train Classifier:** A high-performing classifier is trained for a specific task.
    2. **Train StylEx:** A generative model is trained to reconstruct images while ensuring the classifier makes the same prediction on the reconstructed image, forcing the generator to learn the features most relevant to the classifier.
    3. **Extract Attributes:** The system automatically identifies independent visual attributes in the generator's latent space that most affect the classifier's output.
    4. **Expert Evaluation:** The system generates counterfactual visualizations—images showing what the input would look like with a specific attribute increased or decreased—for review by an interdisciplinary panel of experts.
- **Core Advantage:** This method moves beyond "where" to explain "what." By isolating and visualizing discrete attributes (e.g., vessel dilation, lens opacity), it provides a powerful tool for generating and testing hypotheses about what the AI model has learned.

3. Key Insights from Generative AI Explanations

Applying the StylEx framework across eight prediction tasks and three imaging modalities (fundus photos, external eye photos, chest radiographs) yielded a rich set of findings, demonstrating the method's power to validate, debug, and inspire.
3.1 Rediscovery and Validation of Known Clinical Signs
The framework successfully served as a positive control by automatically identifying and visualizing well-established clinical signs, confirming that the underlying classifiers learned medically relevant features.

- **Cataract Presence (External Eye):** StylEx identified attributes corresponding to the development of **cortical cataract spokes** and a **dimmer red reflex**, both classic signs of cataracts.

- **Abnormal Chest X-ray:** The model learned to identify **left ventricular enlargement** (cardiomegaly) as a key attribute for abnormality.
- **Hypertension (Fundus):** For predicting elevated systolic blood pressure, the model identified **retinal arteriolar narrowing**, a known sign of hypertensive retinopathy.
- **Anemia (External Eye):** The model found that **decreased conjunctival vessel prominence**, consistent with conjunctival pallor, was associated with low hemoglobin.

3.2 Uncovering Confounders and Dataset Biases

A critical finding was that models learn correlations beyond pathophysiology, picking up on signals related to data collection protocols, patient demographics, and socio-cultural factors. Interdisciplinary expert review was essential to identify these confounders.

- **Low Hemoglobin & Eyeliner:** StylEx showed that increased eyeliner was associated with a higher probability of low hemoglobin. The expert panel hypothesized this is not a biological signal but a confounder, as eyeliner use is more common in females, who also have a higher prevalence of anemia.
- **Abnormal CXR & Image Exposure:** An attribute for increased image over-exposure (darker images) correlated with abnormality. The panel concluded this was likely due to portable (AP) X-rays being used more frequently for sicker, less mobile inpatients, and these images often have different exposure characteristics than standard (PA) X-rays.
- **Race Prediction from CXR:** The model associated **increased skeletal conspicuity** with predicting Black race, potentially related to known population differences in bone mineral density. The expert panel stressed that since race is a social construct, this finding should not be interpreted as a biological determinant and requires further investigation into unmeasured environmental or structural factors.

3.3 Generating Novel Scientific Hypotheses

The most exciting application of the framework is its ability to reveal previously unknown or poorly understood correlations, generating plausible hypotheses for future scientific investigation.

- **High HbA1c & Eyelid Margin Pallor (External Eye):** The model associated increased pallor of the eyelid margin with poorly controlled diabetes (HbA1c $\geq$ 9%). The expert panel hypothesized this could be a subtle manifestation of **meibomian gland disease**, which is more prevalent in individuals with diabetes.
- **Sex Prediction & Choroidal Vasculature (Fundus):** StylEx found that greater visibility of the choroidal vasculature was associated with male sex. This was surprising, as some prior literature suggests the opposite. The

panel hypothesized this could be driven by dataset-specific distributions of factors like myopia or fundus pigmentation between sexes.

4. Implications and Future Directions

This collective body of research has profound implications for the future of medical diagnostics and biomedical discovery, highlighting both the immense potential of AI and the critical need for responsible and rigorous interpretation.

4.1 A New Paradigm for Biomedical Research

The work establishes a powerful new methodology for scientific inquiry: using deep learning to first predict a phenotype of interest and then deploying advanced explainability tools like StylEx to localize and visualize the most predictive features. This approach can rapidly generate novel, non-obvious, and testable hypotheses from vast existing datasets, potentially accelerating our understanding of disease mechanisms.

4.2 The Critical Role of Interdisciplinary Expert Review

A recurring and crucial conclusion is that interpreting the findings of complex AI models cannot be done in a vacuum. The StylEx study, in particular, demonstrates that an interdisciplinary panel—including clinicians, ML engineers, social scientists, and socio-technical experts—is essential. This collaboration is necessary to:

- Distinguish true pathophysiological signals from confounders.
- Contextualize findings within the social and structural determinants of health.
- Prevent the misinterpretation of spurious correlations (e.g., social factors) as biological causality.

4.3 Potential Applications and Limitations

The demonstrated capabilities open doors for significant clinical and research applications while also requiring careful consideration of their limitations.

- **Potential Applications:**
  - **Opportunistic Screening:** AI tools could analyze images from existing screening programs (e.g., using diabetic retinopathy photos to also screen for anemia or cardiovascular risk) at minimal additional cost.
  - **Epidemiological Research:** These algorithms can be applied retrospectively to large image datasets to study associations between diseases without needing new, invasive tests.
- **Limitations and Caveats:**

- **Generalizability:** The models require further validation on more diverse populations, different ethnicities, and images captured with various devices (e.g., smartphones vs. specialized fundus cameras).
- **Research, Not Products:** The information presented is from research studies and does not reflect commercially available products; future availability is not guaranteed.
- **Causality:** The methods identify strong correlations and generate hypotheses but do not establish causality. Further prospective studies are needed to validate the clinical utility and impact of these findings.