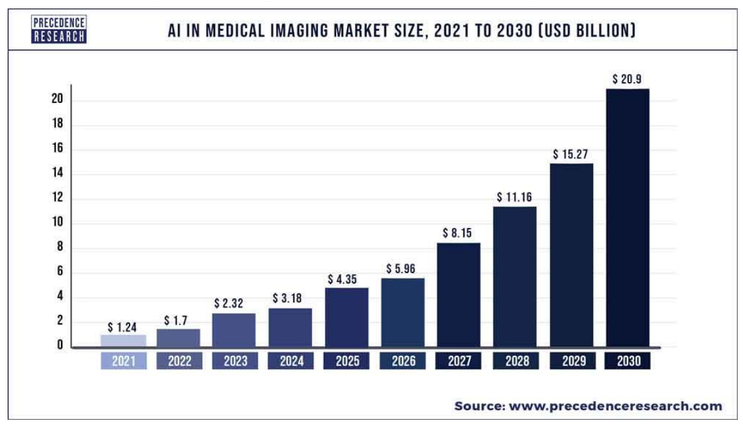
USING COMPUTER VISION TO DISENTANGLE FEATURES ENABLING AI TO LEARN SELF-REPORTED RACE AND ETHNICITY FROM MEDICAL IMAGES

INTRODUCTION

The Danger of Algorithmic Bias in Healthcare

A 2019 clinical Artificial Intelligence (AI) applied to **~200 million Americans** underdiagnosed patients of color by 50%.

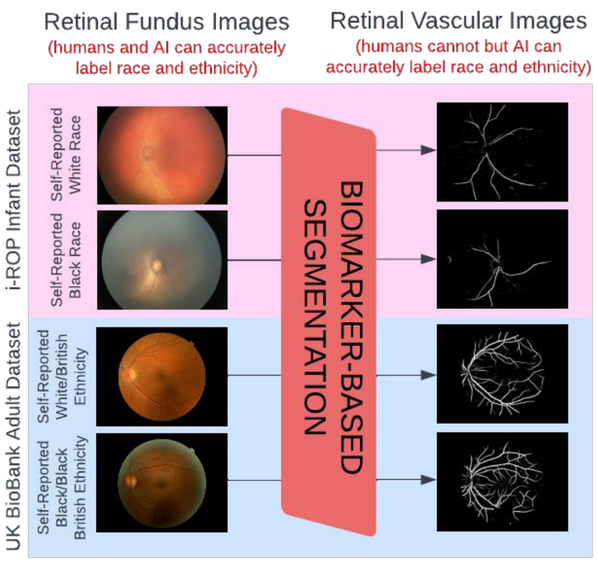


The global market for AI in Medical Imaging is estimated to grow **>10 times by 2033.**

AI-powered medical imaging tools are expanding and exacerbating inequity in clinical care for Black and Brown patients, and other vulnerable communities.

AI Can Learn Self-Reported Race And Ethnicity

In 2021, AI models were trained to recognize patients' self-reported race and ethnicity from medical images**, even when there are no indications of race or ethnicity visible to human experts. This has stumped experts worldwide.**



AI can learn self-reported race and ethnicity from RVMs with AU-ROCs from 92.0 to 95.0.

AI *might* be learning false correlations between race or ethnicity and disease, but we're not sure because the hidden features that signal AI to race and ethnicity are **UNKNOWN.**

We need to discover how AI learns self-reported race and ethnicity when humans cannot.

RESEARCH GOAL

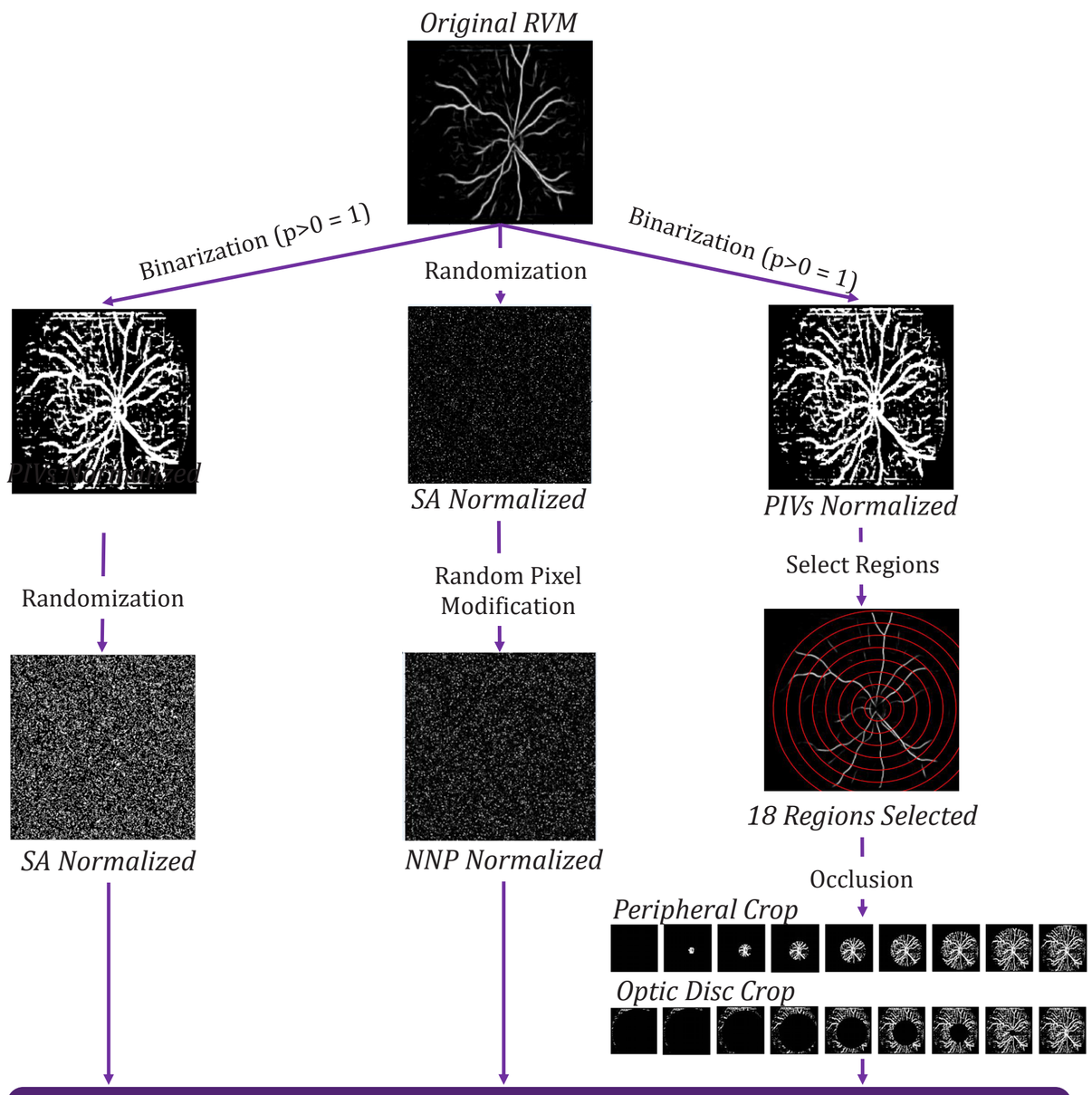
Discover the hidden signals in retinal images that enable algorithms to learn self-reported race and ethnicity.

METHODS

Discovering Hidden Signals Using CNNs

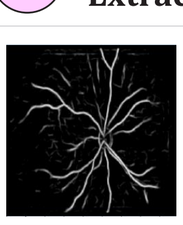
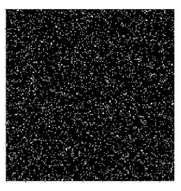
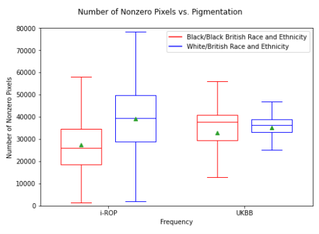
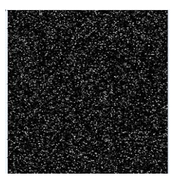
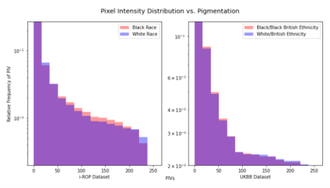
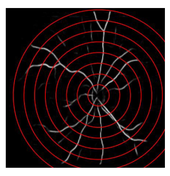
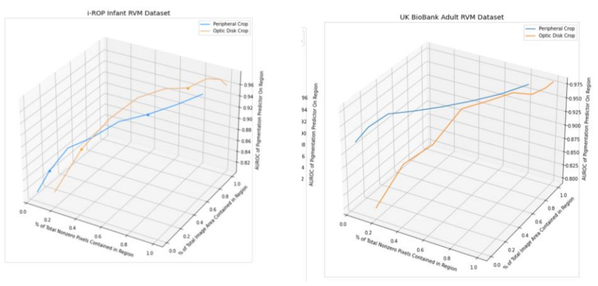
My goal was threefold: (1) to identify key image features that could be hidden signals, (2) to extract each feature from RVMs, and (3) to train an AI model to learn race and ethnicity from the isolated feature to assess its individual significance.

After conducting 100+ experiments to identify key image features, I designed a novel approach to deconstruct an RVM into three key features: **Number of Nonzero Pixels (NNP), Pixel Intensity Values (PIVs), and Spatial Arrangement (SA).**



If the model trained on an isolated feature performs better than random in detecting self-reported race and ethnicity on a modified test set, we infer that the specific RVM feature is a hidden signal!

RESULTS

	A RVM Features Extracted	B CNN Performance (AU-ROC, p-value vs. random)	C Why Can a CNN Predict Self-Reported Race So Well?	D Hidden Signals Discovered by CNNs?
Original RVM		**i-ROP:** 95.0, $3.94 \cdot 10^{-8}$ **UKBB:** 92.0, $1.70 \cdot 10^{-10}$	"The members of our research team **could not come anywhere close** to identifying a good proxy for self-reported race." **- Dr. Marzyeh Ghassemi, MIT Professor**	✗
NNP* Isolated		**i-ROP:** 79.7, $1.65 \cdot 10^{-3}$ **UKBB:** 70.0, $1.14 \cdot 10^{-7}$	 Number of Nonzero Pixels **i-ROP:** $p = 2.45 \cdot 10^{-160}$ **UKBB:** $p = 2.14 \cdot 10^{-24}$	There are fewer nonzero pixels in Black RVMs vs. White RVMs ✓
PIVs* Isolated		**i-ROP:** 79.6, $3.01 \cdot 10^{-5}$ **UKBB:** 58.1, $2.55 \cdot 10^{-2}$	 Pixel Intensity Distributions **i-ROP:** z = 14.3 **UKBB:** z = 43.0	Black RVMs have more pixels of medium intensity, White RVMs have more pixels of high intensity ✓
SA* Extracted		**i-ROP:** 82.0 – 97.5, $3.72 \cdot 10^{-13}$ – $1.16 \cdot 10^{-6}$ **UKBB:** 67.0 – 97.0, $5.24 \cdot 10^{-12}$ – $6.25 \cdot 10^{-6}$		Retinal vessels in center regions are very highly weighted in learning self-reported race ✓ Model artifacts in exterior regions are very highly weighted in learning self-reported race ✓

* NNP: Number of Nonzero Pixels
PIVs: Pixel Intensity Values
SA: Spatial Information

Hidden Signals

1. Black RVMs contain **more fragmented veins** than White RVMs
2. **Fewer large veins are present** in Black RVMs vs. White RVMs in the central regions near the optic disc.
3. The peripheral regions in Black RVMs contain **more choroidal vessels/capillaries** than White RVMs.

All images created by student researcher unless stated otherwise.