Sex Prediction from Polish Ethnicity Fundus Images using Foundation Model

Paweł Niedziółka[1 [0009−0006−8189−5273]](✉), Paweł Zyblewski[1 [0000−0002−4224−6709]], Andrzej Grzybowski[2 [0000−0002−3724−2391]], Michał Woźniak[1 [0000−0003−0146−4205]], Łukasz Lisowski[3 [0000−0002−5202−1968]], Marlena Dubatówka[4 [0000−0001−5977−0745]], and Karol Kamiński[4 [0000−0002−9465−2581]]

[1] Wrocław University of Science and Technology, Faculty of Information and Communication Technology, Wrocław, Poland
{pawel.niedziolka,pawel.zyblewski,michal.wozniak}@pwr.edu.pl
[2] University of Warmia and Mazury, Olsztyn, Poland
ae.grzybowski@gmail.com
[3] Medical University of Bialystok, Department of Ophthalmology, Bialystok, Poland
lukasz.lisowski@umb.edu.pl
[4] Medical University of Bialystok, Center for Population Research, Bialystok, Poland
{marlena.dubatowka,karol.kaminski}@umb.edu.pl

Abstract. The paper aims to use fundus images to determine the patient's sex . Admittedly, such a possibility has been confirmed by previous studies, but the paper will consider how transfer learning from the RET-Found foundation model for retinal images dedicated to the diagnosis of diabetic retinopathy can be used to build such a system and whether the quality of the transfer is sufficient when analyzing data from a different population than the one used to build the fundus model. The study presented here is based on a non-public dataset of fundus images collected from individuals of Polish ancestry. Our study shows that transfer learning, a proven tool for applying and achieving better results when reusing models tuned for different tasks, has some limitations when applied to medical data. In addition, we address emerging concerns about data leakage in medical imaging, as such data may contain overt patient metadata and hidden patterns invisible to humans, which may encode sensitive information that is impossible to capture even by medical experts aided by Explainable artificial intelligence (XAI) techniques.

Keywords: Original dataset exploration · Sex prediction · Transfer learning · Fundus imaging

1 Introduction

Medical imaging plays a crucial role in disease detection, with advances in diagnostic imaging technology enabling the early identification of various medical conditions. The screening used for this purpose aims to find irregular features

in the population and conduct a deeper analysis [14]. Various tests and population characteristics are employed to assess the benefits and limitations of such screenings [16]. In the case of diabetic retinopathy, the key is to minimize the likelihood of visual impairment in people with diabetes through rapid detection and appropriate treatment [19].

Analyzing medical data, it is essential to consider the origin of the population. Different ethnic or geographic groups may show significant differences in the distribution of biological traits [7]. This is why predictive models, based on machine learning, can be subject to the selection bias, where the data on which they were trained are favored at the expense of the overall ability to generalize the results [12].

Lack of representativeness can reduce the model's effectiveness in different populations and increase the risk of misdiagnoses [6]. Therefore, accounting for population diversity and performing a thorough dataset characterization is essential when designing and evaluating machine learning models.

The main contributions of this study are as follows:

- One of the first explorations of a non-public fundus image dataset that contains Polish participants,
- Adaptation of convolutional and transformer-based deep learning models for the sex classification task
- Use of the *GradCam* XAI approach in an attempt to interpret the results obtained and spark future discussions in the ophthalmology community

2 Related Works

This section briefly introduces this article's two most important research areas, i.e., employing deep learning for fundus image classification and transferability estimation.

2.1 Deep Learning for Fundus Image Classification

Deep learning has recently revolutionized automated Diabetic Retinopathy (DR) detection and classification [2], especially with the deployment of Convolutional Neural Networks (CNN) and emerging transformer-based architectures. CNNs such as InceptionV3 and ResNet have shown themselves to be very sensitive and specific to detecting retinal pathologies directly from fundus images, outperforming prior handcrafted methods for features and delivering performance close to that of a human expert in screening applications [11, 25]. This architecture can be applied to medical imaging studies as it can also effectively capture low-level and high-level features [5].

Transformer-based models have been proposed to address the limitations of the capacity of CNNs to compute long-range dependencies and contextual retinal features. The research of DR has proved that ViTs [28] based models can help learn richer spatial hierarchies and attention-based representations in the

classification task. Notably, RETFound [33], trained as a foundation model on millions of unlabeled retinal images using self-supervised learning, can transfer to other downstream ophthalmic tasks, including DR detection.

Explainable AI methods, including Grad-CAM, LIME, and attention maps, have been incorporated into DR diagnostic pipelines to produce visual explanations for model predictions, which is a necessary step towards clinical acceptance and regulatory approval [3, 32]. These interpretability aids help clinicians confirm that the model is attending to clinically relevant features like microaneurysms, exudates, and hemorrhages.

2.2 Transferability estimation

The canonical machine learning methods assume that training and testing data are i.d.d. (independent and identically distributed). This assumption may be violated when constructing pattern recognition systems to solve real-world tasks. For example, when classifying biosignals such as EEG, considerable variances might be observed between individuals, resulting in poor generalization capacity [27]. This can be addressed by collecting additional fully or partially labeled data, which is subsequently used to train the model. However, such a solution usually comes with a significant time and monetary cost [20].

The alternative solution is transfer learning [18], which aims to boost the model's generalization capacity for a target domain, which generally contains a small number of labeled samples, by applying knowledge learned during training on an associated source domain. The effectiveness of transfer learning typically relies on three assumptions [31]: **(i)** *Task correlation*, which requires that the learning task in the source and target domains are similar, **(ii)** *Domain correlation*, which requires that the data distributions in the source and target domains be similar, and **(iii)** *Ideal joint error*, which requires that an appropriate hypothesis be applied to both domains. Negative knowledge transfer may result from not meeting any of the above assumptions [31], which could hinder the pre-trained model's capacity to generalize to the target dataset.

Retraining or fine-tuning the previously learned model is one of the most fundamental but effective transfer learning strategies [18]. The development of metrics for matching the right pre-trained model to the target problem at hand has become essential due to the enormous popularity of transfer learning and the substantial variability in the benefits derived from its use, depending on the network architecture, the source dataset used for pre-training, or even the layer from which the representation is obtained. According to the inherent computational complexity, available transferability estimation techniques can be classified as either **(i)** computation-intensive or **(ii)** computation-efficient [13].

Methods from the first category can typically be used to estimate the knowledge transferability of unsupervised pre-trained models and determine the network layer from which the transfer should occur. However, their computational complexity corresponds with executing fine-tuning on a target dataset, making it impossible to use them to assess transferability before fine-tuning [13]. Dwivedi

and Roig presented *Representation Similarity Analysis* (RSA), which uses correlations between pre-trained models on different tasks to compute a similarity score [10]. Song et al. introduced *DEeP Attribution gRAph* (DEPARA) [23], which creates a similarity graph between representations of samples from specific datasets derived from pre-trained deep neural networks. Zamir et al. presented *Taskonomy* [30], which analyzes the source and target task linkage by performing fine-tuning on a model pre-trained on the target dataset and estimating the loss, whereas Achille et al. [1] employ *Fischer Information Matrix*.

In contrast to its counterpart, computationally efficient transferability metrics do not require model training on the target dataset [13]. Cui et al. calculated the *Earth Mover's Distance* between source and target data features [8], which requires access to the source dataset. You et al. used the *Logarithm of Maximum Evidence* (LogME) to evaluate pre-trained models for transfer learning [29] and discovered a link between high LogME and increased generalization ability. Tran et al. assessed transferability by calculating *Negative Conditional Entropy* (NCE) between source and target labels [26]. The computationally efficient category also contains metrics that do not require access to source data but cannot be used to choose layers. Nguen et al. presented the *Log Expected Empirical Prediction* (LEEP) score, working similarly to NCE, which replaces source dataset labels with pseudo-labels derived from a pre-trained model [17]. Bao et al. presented an *H-score* to measure knowledge transfer between source and destination datasets using statistical and information-theoretic concepts, although, like LogME, they only considered the network's last layer [4]. Huang et al. developed *TransRate*, which can be expressed as the mutual information between embeddings of target samples and their labels [13]. This only requires access to target data and a single pass through the pre-trained model.

3 Bialystok PLUS prospective cohort study

The dataset used in this study originates from the non-public repository of the *Bialystok PLUS* prospective cohort study[5]. It includes high-resolution fundus images – as shown in Fig. 1 – collected from the left and right eyes of individual participants. Multiple photo acquisitions per participant per eye provide variability in quality and capture conditions. All sensitive, identifiable information was removed following ethical standards and data protection regulations. Aside from imaging data, the dataset is complemented by a broad range of clinical and demographic data. This includes, e.g., age, sex , and various physiological or diagnostic markers, which are listed in Tab. 1. Due to its proprietary nature, this dataset is not available for public distribution.

However, this paper focuses on sex prediction using different convolutional models and fundus images as a base for prediction. To support that research in the dataset, 18,472 fundus images were collected from 2,854 individual participants, including 1,271 women and 1,583 men of different age groups. The age distribution by sex is visible in Figure 2.

[5] www.bialystok.plus

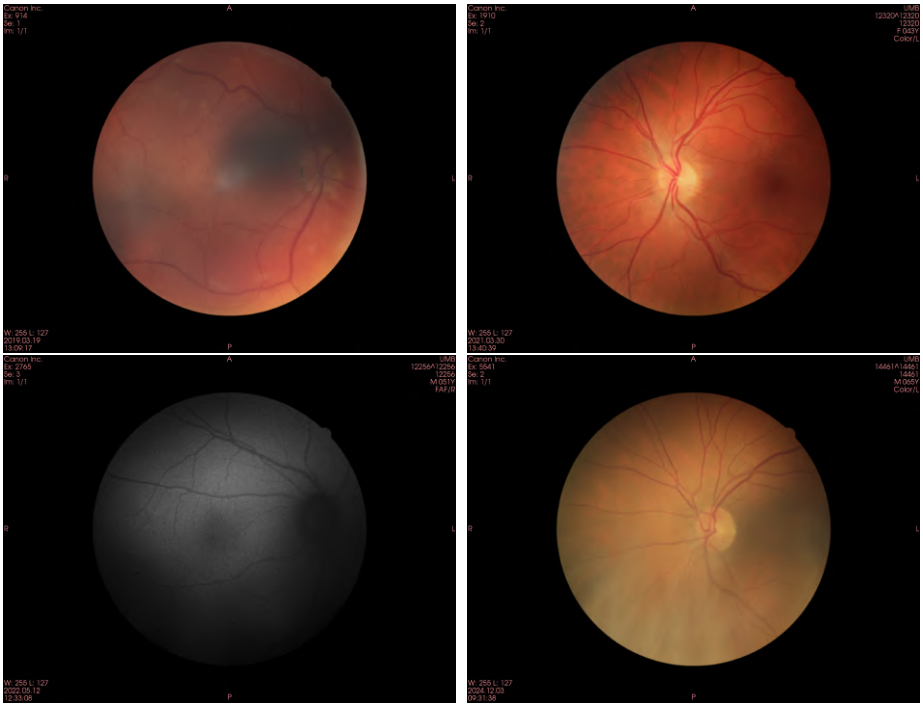


Fig. 1: Sample images available in the dataset

Table 1: All markers available in the dataset.

Medical history	Densitometry (DEXA)	Advanced glycation end-products (AGEs)
Blood biochemical and immunochemical tests	Three-day nutritional interview	Peripheral blood pressure
Proteomic analysis	Nutritional averages per day	Electrocardiography (ECG)
Urine analysis	Bioelectrical impedance analysis (BIA)	Echocardiography
ATC drugs	PWV, Central pressure, ABI	Carotid arteries ultrasound
Respiratory tests & fractional exhaled nitric oxide (FENO) test	Ergospirometry Cosmed and Vyntus	Thyroid ultrasound
Dental examination	Anthropometry	Liver ultrasound
Hand grip strength	Orthostatic test	Liver elastography
Swelling and varicose veins of the calves	Geriatric assessment 50-64 and +65	Autorefractometry and tonometry
Fundus imaging	Voice recording	MRI

3.1 Dataset Preprocessing

Each image in the dataset underwent a standardized preprocessing pipeline to ensure consistency and compatibility with deep learning models. At the very

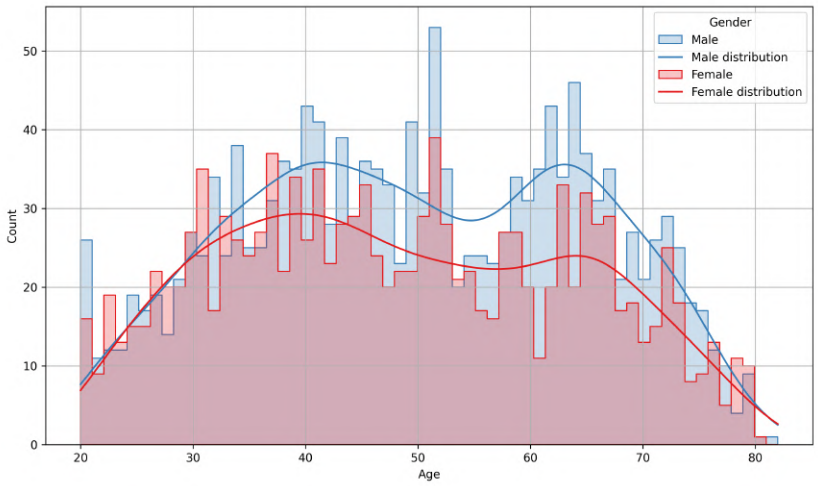


Fig. 2: Age distribution by sex.

beginning, the DICOM format images were converted into pixel-intensity arrays. A normalization process was employed to reduce variation in brightness and contrast, improving model robustness. In cases where the images contained on a single color channel (grayscale), they were expanded to three channels by duplication. That ensures a unified data representation for models that were pre-trained on RGB data. Finally, all images were resized to a consistent resolution - 224×224 pixels, which makes a consistent input format across the dataset.

4 Experimental Evaluation

The experiments were designed and conducted in order to answer the following research questions:

- **RQ1** Can the Polish *Bialystok PLUS* fundus images dataset be used to replicate sex prediction research?
- **RQ2** How efficient is the adaptation of the foundation model for diabetic retinopathy detection (RETFound) to the sex classification task?
- **RQ3** Do the transferability measure values for individual deep models coincide with their generalization ability?

4.1 Set-up

Two distinct models were employed to address the binary sex classification task and verify the exclusive, non-public dataset: ResNet-50 and the RETFound.

ResNet50 model was initialized with *ImageNet* [9] pre-trained weights. The original classification head of this network was replaced with a new fully connected layer with an output of one neuron. Later, this was passed through a sigmoid activation function to predict the probability of the image belonging to the female or male category.

The RETFound model weights are trained on the color fundus photos (CFP) dataset, which is structurally and semantically closer to the medical imaging domain, and these were used. The original head layer was replaced similarly to the previous model; one change was to retain the normalization layer from the backbone to ensure a stable representation before the classification process.

To fine-tune the sex classification models, training was conducted using the Adam optimizer with a learning rate of $1 \cdot 10^{-4}$. The binary cross-entropy loss with logits was set as a criterion that is appropriate for this type of classification. Models described before were trained for 10 epochs with a batch size of 32. That provides a balance between efficiency and gradient stability. The dataset was split adequately with rates of 80% and 20% accordingly for the training and testing sets.

For model performance evaluation, the following metrics were used:

- Accuracy provides a general measure of how many predictions the model got right. However, it may be misleading in imbalanced datasets - in the medical images, healthy cases frequently outnumber pathological ones.
- Precision is important to assess how many predicted DR-positive cases are correct. High precision reduces the risk that patients are wrongly classified as diseased.
- Recall measures how well the model detects actual DR cases. High recall helps prevent biased or imbalanced sex predictions.
- F1 score is the harmonic mean of precision and recall, balancing the trade-off between the two.

The metrics were computed at each epoch, and each model was trained independently following the same experimental setup. The code for experimental setup is publicly available in the GitHub[6].

4.2 Experimental scenarios

Experiment 1 - Fine-Tuning of Deep Models
The first experiment evaluated the ability of two selected architectures to learn sex-related features from fundus images within a consistent training pipeline. Both models were tuned using a specific configuration of preprocessing, training, and evaluation criteria. The training process was designed to investigate the learning behavior of each model over epochs. The evolution of performance was observed under identical conditions. By comparing the two models in this controlled environment, this experiment aims to determine whether the selected dataset is sufficiently informative for sex prediction and how well the baseline

[6] https://github.com/w4k2/biobank_gender_prediction

model, initially developed for diabetic retinopathy tasks, adapts to the goal of sex classification. Therefore, this experiment directly addresses **RQ1** and **RQ2**.

Experiment 2 – Transferability Evaluation
The second experiment aims to evaluate the deep learning models in terms of their ability to transfer knowledge in the sex classification task from fundus images. For this purpose, two transferability measures, TransRate and H-Score, are used to compare the classification performance of ResNet50 and RETFound with their respective potential for transfer learning. This evaluation also allows us to determine whether selecting models based on estimated transferability would lead to choosing the one with the highest generalization capability for this specific task. Such an analysis is particularly insightful given the distinct nature of the two models: ResNet50 is a general-purpose architecture pre-trained on ImageNet, while RETFound is dedicated explicitly to fundus image analysis. Due to the relatively low computational cost of the selected transferability metrics, the experiment follows an evaluation protocol: 5 repetitions of 2-fold cross-validation [24] to stabilize the outcomes. The findings from this experiment aim to address **RQ3**.

4.3 Experiment 1 - Fine-Tuning of Deep Models

Training and validation losses were registered at all epochs, as shown in Figure 3. Both ResNet50 and RETFound showed stable learning during the initial epochs. However, from around the fifth epoch, the validation loss began to increase, indicating the beginning of over-fitting. Despite this, training continued until the 10th epoch to assess how both models fared on the full training schedule.

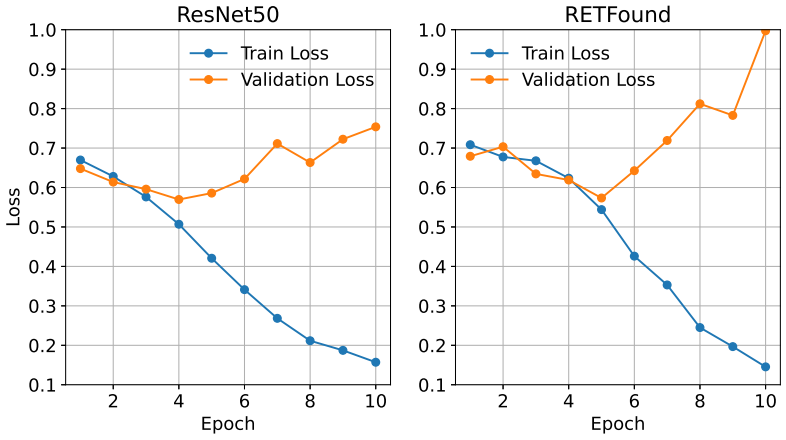


Fig. 3: Training and Validation Loss

Table 2: Models Performance Metrics. Best metric values are marked in **bold**.

Model	**Epoch**	***Accuracy***	***Precision***	***Recall***	***F1***
ResNet50	10^{th} (best)	**0.732**	**0.736**	**0.800**	**0.766**
RETFound	5^{th} (best)	0.716	0.725	0.778	0.751
RETFound	10^{th}	0.677	0.675	0.792	0.729

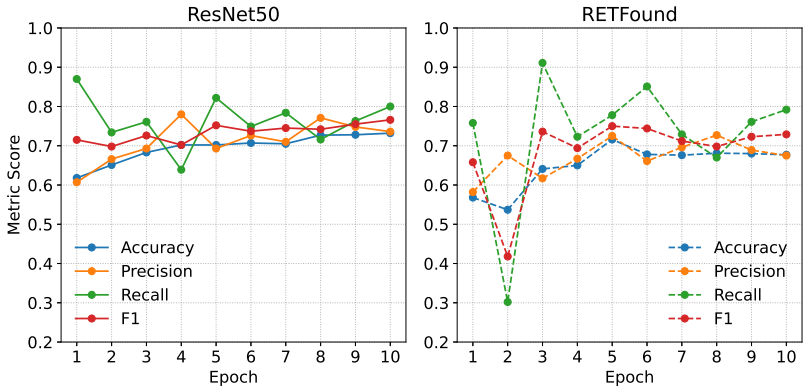


Fig. 4: Fine-Tuning metrics score per epoch

Classification metrics were calculated at each epoch to monitor the quality of the learned representations. The full representation of the metric values for each epoch can be seen in Figure 4. The best model performance results were summarized in Table 2. ResNet50 achieved the highest performance in the last epoch, slightly outperforming RETFound in all metrics. On the other hand, RETFound achieved the best performance earlier, in the fifth epoch, but its metrics dropped later, further confirming early over-fitting. These results can be surprising, given that RETFound is a model dedicated to analyzing fundus images and would probably, in most cases, be the first choice of researchers and practitioners facing pattern recognition tasks based on this type of data. Contrary to intuition, it is ResNet50 – pre-trained on the ImageNet dataset, which does not contain data with characteristics reminiscent of fundus images – that not only shows a comparable early generalization ability to RETFound but also demonstrates an upward trend in subsequent training epochs in terms of all measured metrics.

The results prove that both models can learn sex-related patterns from fundus images, validating the dataset informativeness. Meanwhile, RETFound benefited from pretraining and showed earlier convergence. Ultimately, ResNet50 achieved higher classification performance through continued fine-tuning. The feasibility of using these architectures for the sex classification problem was confirmed, and addressing **RQ1** and **RQ2**. While the results do not diminish the value of

foundation models like RETFound, they raise important considerations about adopting such models automatically for all fundus image-based tasks.

4.4 Experiment 2 - Transferability Evaluation

Table 3 presents both models, TransRate and H-score transferability metrics values, evaluated in their pre-trained states and after additional fine-tuning on the explored *Bialystok PLUS* dataset. It is worth noting that although transferability metrics often show a linear correlation with a model classification performance [13], they should not be considered fully reliable for model selection and must be applied cautiously.

The H-score appears to be relatively uninformative, showing analogous values across each model and matching results after fine-tuning. Interestingly, it still indicates a slight advantage for ResNet50 using only the pre-trained representation.

In the case of TransRate, the discrepancy between the values for the different models is more extensive, and again, one can observe – this time a substantial – advantage in potential knowledge transferability in favor of ResNet50. At the same time, performing fine-tuning for both ResNet50 and RETFound leads to a reduction in estimated transferability. The decrease is particularly noticeable for RETFound and may be related to overfitting and the drop in classification accuracy seen in Experiment 1.Note, however, that ResNet50, which can improve classification quality with successive learning epochs, also exhibits a decrease in transferability in terms of TransRate.

Despite some lack of intuitiveness in the interpretation of the obtained values of transferability measures, TransRate, in this case, could be used to select deep models, pointing clearly to ResNet50, initially offering a slightly higher generalization ability than RETFound, and in subsequent epochs showing the ability to improve the learned representation.

In summary, the results obtained in this experiment cannot be considered definitive, and the need for in-depth research on transferability for foundation and general-purpose models in the sex classification task based on fundus images is evident. At the same time, in the case analyzed of the *Bialystok PLUS* dataset, the TransRate measure can indicate a deep model with potentially better generalizability, answering the **RQ3**. These results may be surprising given the small potential relationship between ImageNet and fundus images.

Table 3: Comparison of average TransRate and H-Score across models.

Model	**Transrate**	**H-Score**
ResNet50 (Image-Net)	7.920 ± 0.032	0.998 ± 0.004
ResNet50 (Fine-tuned)	7.141 ± 0.043	1.000 ± 0.000
RETFound (CFP)	5.109 ± 0.076	0.986 ± 0.033
RETFound (Fine-tuned)	3.377 ± 0.044	1.000 ± 0.001

5 Generalizable Insights about Responsible Application of Machine Learning in Healthcare

While ophthalmologists often find it challenging to recognize sex from fundus camera images, deep learning models have achieved high accuracy rates in this task. This suggests that such models can identify subtle differences that are not readily apparent to the human eye, e.g., they can utilize features related to retinal vascularization, the optic disc region, and the macula to differentiate between male and female eyes [7]. Some studies suggest [15] male retinas may have greater vascularization and a darker ring around the optic disc, while the macula might contain more distinctive female features.

Explaining exactly how it is possible to identify a patient's sex or age is still a significant challenge. However, the use of advanced explainable AI mechanisms may bring us closer to understanding, among other things, sex differences in fundus camera images. To better understand the prediction process of the models in this study, the Grad-CAM explainability method was employed. Images in relation to which the models have demonstrated the greatest confidence (maximum classifier support value) were selected. These confident predictions served as examples to visualize the areas of activation most relevant to classification decisions, visible in Figure 5. Activation maps reveal a noticeable difference between the two models: ResNet50 shows more concentrated and localized focus areas, while RETFound produces more diffuse activations in multiple areas of the retina.

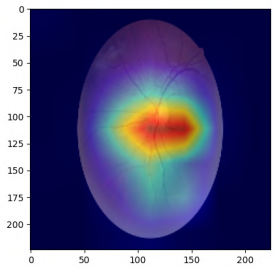
It may also bring us closer to proposing a set of biomarkers useful in this problem. Further challenges will be to propose other biomarkers extracted from fundus camera images that may help diagnose different diseases.

The indicated observation, although already confirmed by other researchers, shows that when sharing imaging data, one should pay close attention to the possibility of leakage of sensitive data. It could be dangerous because of possible privacy violations and can be used, for example, for patient re-identification. Thus, one should consider such mechanisms that would prevent such activity. Perhaps tools inspired by solutions that protect artists' rights to images by their use by generative AI systems to learn to mimic the artistic style. Such methods use poisoning attack techniques, such as *Nightshade* [22], that are designed as an offensive tool to distort feature representations inside generative AI image models or *Glaze* [21].

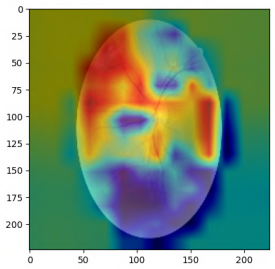
6 Conclusion

This article aimed to examine how the established deep learning models – both general-purpose and dedicated to analyzing fundus images – handle sex classification task based on a Polish *Bialystok PLUS* dataset.

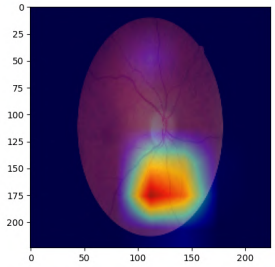
[7] https://www.vchri.ca/stories/2024/03/20/novel-ai-model-explains-retinal-sex-difference



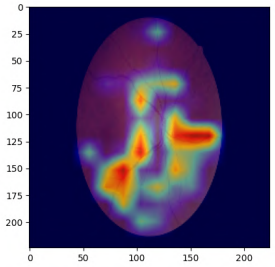
(a) ResNet50 - participant 1 (male)



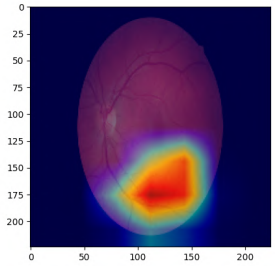
(b) RETFound - participant 1 (male)



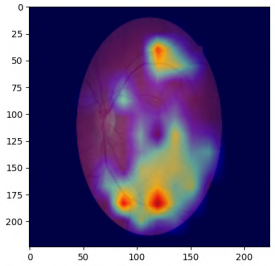
(c) ResNet50 - participant 2 (female)



(d) RETFound - participant 2 (female)



(e) ResNet50 - participant 3 (female)



(f) RETFound - participant 3 (female)

Fig. 5: Grad-CAM visualizations for models

Both ResNet-50 and RETFound demonstrated the ability to learn meaningful representations for sex prediction from fundus retinal images. Notably, the RETFound model that was developed for diabetic retinopathy detection achieved slightly lower performance after fine-tuning to the sex classification task. This proves that using foundation models for a particular type of data is not always obvious, and fine-tuned models with a more general purpose may perform better (in regards to both generalization ability and knowdledge transfer) even in med-

ical tasks. Overall, the results demonstrate the viability of using deep learning on fundus images for supplementary classification tasks.

Further research will focus on identifying domain-specific biomarkers that will improve model generalization across different populations and discover correlations between patient clinical data and retinal images.

Acknowledgement

The study is partially supported by the statutory fund of the Faculty of Computer Systems and Networks at Wrocław University of Science and Technology. The study is part of the *Bialystok PLUS* project. The study was supported by the National Centre for Research and Development project no POIR.04.01.04-00-0052/18, supported by the European Regional Development Fund (to KAK – funduscopy imaging), and statutory funds of the Medical University of Bialystok - data acquisition.

References

1. Achille, A., Lam, M., Tewari, R., Ravichandran, A., Maji, S., Fowlkes, C.C., Soatto, S., Perona, P.: Task2vec: Task embedding for meta-learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6430–6439 (2019)
2. Alyoubi, W., et al.: Diabetic retinopathy detection through deep learning techniques: A review. Informatics in Medicine Unlocked **20**, 100377 (2020)
3. Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., Hoebel, K., Gupta, S., Patel, J., Gidwani, M., et al.: Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. Radiology: Artificial Intelligence **3**(6), e200267 (2021)
4. Bao, Y., Li, Y., Huang, S.L., Zhang, L., Zheng, L., Zamir, A., Guibas, L.: An information-theoretic approach to transferability in task transfer learning. In: 2019 IEEE international conference on image processing (ICIP). pp. 2309–2313. IEEE (2019)
5. Butt, M., Awang Iskandar, D., Khan, M.A., Latif, G., Bashar, A.: Medcnet: A memory efficient approach for processing high-resolution fundus images for diabetic retinopathy classification using cnn. International Journal of Imaging Systems and Technology **35**(2), e70063 (2025)
6. Chen, R.J., Wang, J.J., Williamson, D.F., Chen, T.Y., Lipkova, J., Lu, M.Y., Sahai, S., Mahmood, F.: Algorithmic fairness in artificial intelligence for medicine and healthcare. Nature biomedical engineering **7**(6), 719–742 (2023)
7. Chinta, S.V., Wang, Z., Zhang, X., Viet, T.D., Kashif, A., Smith, M.A., Zhang, W.: Ai-driven healthcare: A survey on ensuring fairness and mitigating bias. arXiv preprint arXiv:2407.19655 (2024)
8. Cui, Y., Song, Y., Sun, C., Howard, A., Belongie, S.: Large scale fine-grained categorization and domain-specific transfer learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4109–4118 (2018)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)

10. Dwivedi, K., Roig, G.: Representation similarity analysis for efficient task taxonomy & transfer learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12387–12396 (2019)
11. Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. jama **316**(22), 2402–2410 (2016)
12. Hall, W.J., Chapman, M.V., Lee, K.M., Merino, Y.M., Thomas, T.W., Payne, B.K., Eng, E., Day, S.H., Coyne-Beasley, T.: Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: a systematic review. American journal of public health **105**(12), e60–e76 (2015)
13. Huang, L.K., Huang, J., Rong, Y., Yang, Q., Wei, Y.: Frustratingly easy transferability estimation. In: International conference on machine learning. pp. 9201–9225. PMLR (2022)
14. Khalifa, M., Albadawy, M.: Ai in diagnostic imaging: Revolutionising accuracy and efficiency. Computer Methods and Programs in Biomedicine Update p. 100146 (2024)
15. Liu, S., Zhao, H., Huang, L., Ma, C., Wang, Q., Liu, L.: Vascular features around the optic disc in familial exudative vitreoretinopathy: findings and their relationship to disease severity. BMC Ophthalmology **23** (04 2023). https://doi.org/10.1186/s12886-023-02884-7
16. Maxim, L.D., Niebo, R., Utell, M.J.: Screening tests: a review with examples. Inhalation toxicology **26**(13), 811–828 (2014)
17. Nguyen, C., Hassner, T., Seeger, M., Archambeau, C.: Leep: A new measure to evaluate transferability of learned representations. In: International Conference on Machine Learning. pp. 7294–7305. PMLR (2020)
18. Niu, S., Liu, Y., Wang, J., Song, H.: A decade survey of transfer learning (2010–2020). IEEE Transactions on Artificial Intelligence **1**(2), 151–166 (2021)
19. Scanlon, P.H.: The english national screening programme for diabetic retinopathy 2003–2016. Acta diabetologica **54**(6), 515–525 (2017)
20. Settles, B.: Active learning literature survey (2009)
21. Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R., Zhao, B.Y.: Glaze: protecting artists from style mimicry by text-to-image models. In: Proceedings of the 32nd USENIX Conference on Security Symposium. SEC '23, USENIX Association, USA (2023)
22. Shan, S., Ding, W., Passananti, J., Wu, S., Zheng, H., Zhao, B.Y.: Nightshade: Prompt-specific poisoning attacks on text-to-image generative models (2024), https://arxiv.org/abs/2310.13828
23. Song, J., Chen, Y., Ye, J., Wang, X., Shen, C., Mao, F., Song, M.: Depara: Deep attribution graph for deep knowledge transferability. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3922–3930 (2020)
24. Stapor, K., Ksieniewicz, P., García, S., Woźniak, M.: How to design the fair experimental classifier evaluation. Applied Soft Computing **104**, 107219 (2021)
25. Ting, D.S.W., Cheung, C.Y.L., Lim, G., Tan, G.S.W., Quang, N.D., Gan, A., Hamzah, H., Garcia-Franco, R., San Yeo, I.Y., Lee, S.Y., 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**(22), 2211–2223 (2017)

26. Tran, A.T., Nguyen, C.V., Hassner, T.: Transferability and hardness of supervised classification tasks. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1395–1405 (2019)
27. Wu, D., Xu, Y., Lu, B.L.: Transfer learning for eeg-based brain–computer interfaces: A review of progress made since 2016. IEEE Transactions on Cognitive and Developmental Systems **14**(1), 4–19 (2020)
28. Wu, J., Hu, R., Xiao, Z., Chen, J., Liu, J.: Vision transformer-based recognition of diabetic retinopathy grade. Medical Physics **48**(12), 7850–7863 (2021)
29. You, K., Liu, Y., Wang, J., Long, M.: Logme: Practical assessment of pre-trained models for transfer learning. In: International Conference on Machine Learning. pp. 12133–12143. PMLR (2021)
30. Zamir, A.R., Sax, A., Shen, W., Guibas, L.J., Malik, J., Savarese, S.: Taskonomy: Disentangling task transfer learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3712–3722 (2018)
31. Zhang, W., Deng, L., Zhang, L., Wu, D.: A survey on negative transfer. IEEE/CAA Journal of Automatica Sinica **10**(2), 305–329 (2022)
32. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)
33. Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., Liu, T., Xu, M., Lozano, M.G., Woodward-Court, P., et al.: A foundation model for generalizable disease detection from retinal images. Nature **622**(7981), 156–163 (2023)