It's Not Just Black and White: Classifying Defendant Mugshots Based on the Multidimensionality of Race and Ethnicity

Rahul Kumar Dass*, Nick Petersen†, Ubbo Visser*
University of Miami
*Department of Computer Science, †Department of Sociology
Coral Gables, United States
*{rdass, visser}@cs.miami.edu, †npetersen@miami.edu

Marisa Omori
University of Missouri–St. Louis
Department of Criminology & Criminal Justice
St. Louis, United States
marisa.omori@umsl.edu

***Abstract*—Analyses of existing public face datasets have shown that deep learning models (DLMs) grapple with racial and gender biases, raising concerns about algorithmic fairness in facial processing technologies (FPTs). Because these datasets are often comprised of celebrities, politicians, and mainly white faces, increased reliance on more diverse face databases has been proposed. However, techniques for generating more representative datasets are underdeveloped. To address this gap, we use the case of defendant mugshots from Miami–Dade County's (Florida, U.S.) criminal justice system to develop a novel technique for generating multidimensional race–ethnicity classifications for four groups: *Black Hispanic, White Hispanic, Black non-Hispanic, and White non-Hispanic*. We perform a series of experiments by fine–tuning seven DLMs using a full sample of mugshots (194,393) with race-ethnicity annotations from court records and a random stratified subsample of mugshots (13,927) annotated by a group of research assistants. Our methodology considers race as a multidimensional feature particularly for a more diverse face dataset and uses an averaged (consensus–based) approach to achieve a 74.84% accuracy rate based on annotated data representing only 2% of the full dataset. Our approach can be used to make DLM based FPTs more inclusive of the various subcategories of race and ethnicity as they are being increasingly adopted by various organizations including the criminal justice system.**

***Keywords*-face classification, transfer learning, deep learning, interdisciplinary, race and ethnicity, computational sociology**

I. Introduction

Despite advances in facial processing technologies (FPTs), which broadly encompass various facial classification tasks such as detection, analysis and recognition [1], concerns about algorithmic fairness and perpetuating racist stereotypes persist. For example, Asian faces are often misidentified as blinking [2] and Black faces have been mislabeled as gorillas [3]. Recent impactful auditing studies of commercial FPTs have all consistently showed that FPTs perform worse on women and people with darker skin tone [1], [4], [5]. Consequently, this has led corporations such as Microsoft and Amazon to release updated APIs to be more inclusive for such underrepresented demographics [1].

Before outlining the technical contributions and challenges associated with identifying a person's race and ethnicity based on an image, we must first articulate the societal importance of racial and ethnic labels. Race and ethnicity are both constructed social categories, where race is based on shared perceived physical characteristics, and ethnicity is based on shared cultural qualities [6]. Despite the fact that race and ethnicity are social categories, these categorizations have real consequences for people [7]. The targeting of underrepresented groups by a wide range of social institutions, most notably the U.S. criminal justice system, has led to massive racial inequalities [8].

For example, a recent report by Petersen et al. [9] shows that Black defendants are four times more likely to be incarcerated than their White counterparts in Miami–Dade County's (MDC) criminal justice system. In particular, Black Hispanic defendants are especially disadvantaged. While White Hispanic and White non–Hispanic people make up nearly 75% of the County's population, Table I highlights the overrepresentation of Black people in MDC's justice system. Despite the capabilities and incredible potential of DLM based FPTs, studies show evidence that identification of physical characteristics including race is not a solved problem [1], [4], [5].

II. Related Work

Given the uniqueness of every face, various large–scale face datasets have been curated over the years to help foster research into better understanding physical characteristics including age, gender and skin tone. Those that have investigated race or ethnicity classifications in particular [3], [10]–[21], have proposed a variety of machine learning approaches, such as deep (single or ensemble–based) Convolutional Neural Networks (CNN) with multiple stages of face preprocessing (detection, alignment, cropping). Moreover, facial attribute techniques for measurement of landmarks [22], [23] and skin tone [4], [22] have been used as proxies for racial classification. Novel learning techniques like deep unsupervised domain adaptation [12], multi-task CNN system [15] and joint learning and unlearning [14] have shown promising results leading to greater accuracies for race classification.

Table I
COMPARING THE U.S. AND MIAMI–DADE COUNTY (MDC) GENERAL DEMOGRAPHIC SPREAD OF FOUR RACE-ETHNICITY SUBGROUPS IN 2010, VS. THE MDC DEFENDANTS POPULATION IN 2010–2015.

Race-Ethnic Subgroup	U.S. General	MDC General	MDC Defendants
Black Hispanic	0.4%	1.9%	9.18%
White Hispanic	8.7%	58.4%	39.70%
Black non–Hispanic	12.2%	17.1%	37.96%
White non–Hispanic	63.7%	15.4%	13.14%
Total	**100.0%**	**100.0%**	**99.98%***

* Other racial–ethnic groups represented a very small (0.02%) proportion and were removed from the dataset.

What has remained consistent in all of these studies is the way in which race is considered as a classification task, i.e. an image of a face belongs to *one* of several subcategories of race including White, Black, Indian, East Asian, Southeast Asian, Middle Eastern or Hispanic/Latinx. As Hanna et al. [24] highlight, the notion of race cannot be adopted simply as a singular defining attribute but rather as a multidimensional construct. For instance, within the U.S. many individuals might identify themselves as Black *and* Hispanic as well [25], [26].

III. APPROACH

Here, we consider race (Black and White) and ethnicity (Hispanic and non–Hispanic) labels, combining them to investigate a multidimensional racial–ethnic identification based on four subgroups: Black Hispanic (BH), White Hispanic (WH), Black non–Hispanic (BnH), and White non–Hispanic (WnH). Furthermore, using transfer learning, our goal is to train DLMs not using standard benchmark datasets such as CelebA [18], UTKFace [16] or LFWA+ [18], instead we use a unique dataset from the criminal justice domain.

We consider mugshots from MDC as a case study based on its diversity for the four racial–ethnic subgroups. This mugshot dataset offers a more holistic representation of the general population, as images are not edited and were taken under unconstrained conditions or what others have termed "in the wild" [12], [18]. Despite the standardized procedure involved in collecting a mugshot, as we will describe in Section IV-A, defendants often appear under duress, do not look straight at the camera, and are subject to varying illumination making an image dimmer or darker, as detailed in Section V-A. At the same time, however, research using mugshot databases raises novel privacy concerns that require additional care, which we discuss in Section V-D. Moreover, Figure 1, only displays mugshots of celebrities in our dataset that are readily available online as well to help mitigate such privacy concerns.

Our dataset is also more racially and ethnically diverse than other benchmark datasets. As Table I indicates, MDC has a larger proportion of Hispanic and Black residents than the U.S. at large. Therefore, our dataset is uniquely poised to address claims to more research on underrepresented demographics [4], [10]. Given that Hispanics represent the largest and fastest growing racial–ethnic minority in the U.S., accounting for 52% of national population growth 2008–2018 [27], developing Hispanic ethnicity annotations will become increasingly important in the coming years. Analyses of existing facial benchmark datasets have shown to contain pre–existing biases across a range of attributes particularly for race and gender, resulting in a greater demand for the creation of more inclusive and balanced data [4], [10], [22]. To foster algorithmic fairness, we seek to develop DLMs to generate multidimensional race–ethnicity labels using existing mugshot data to answer the following **three research questions:**

1) How would a DLM's performance vary if the classification task changed from race to race–ethnicity prediction based on the same dataset?
2) Does the performance of DLM race-ethnicity classifications vary based on the model architecture?
3) Does the performance of these DLM tasks vary when using human annotations based on a single rater versus multiple raters?

IV. DATA AND METHODS

A. Data Source and Collection Practice

To answer our research questions, we analyze a novel dataset of 194,393 original MDC defendants' mugshots from 2010–2015. Based on the spread of the four subgroups stated in Table I, there were: 17,860 (BH), 77,177 (WH), 73,803 (BnH), and 25,553 (WnH). Since all MDC defendants are processed and photographed through the same centralized county jail, mugshots are relatively standardized. Mugshots are all in color and of fairly high quality with standardized positioning, lighting, and background, see Figures 1a and 1b as examples from our dataset. However, there was variability in pixel size ranging from 404 x 506 (low–end) to 800 x 1020 (high–end), given that these photographs are originally taken for administrative purposes and not for training and evaluating DLMs. The mugshots and arrest records are publicly available data, and were obtained with IRB approval from the University of Miami. To protect individuals' confidentiality, we still took steps to de–identify the mugshots, as outlined below.

B. Matching Mugshots With Court Race Annotations

After obtaining a county–wide database of defendants during the period of analysis, we worked with MDC Clerk of Courts offices to obtain unique identifiers that allowed us to link each defendants' mugshot with their court records containing information on race, gender, and other demographics annotated by criminal justice officials. Because criminal justice agencies often store data using differing defendant identification numbers, for many jurisdictions linking mugshots and court records is very difficult [28]. Thus, the presence of unique identifiers needed to match mugshots and court records is a novel feature of the database, allowing us to achieve a 99% match rate. To protect confidentiality, we de–identified the dataset by assigning new ID numbers and removing any identifiable information after linking the mugshot and court record data.

C. Information on Race, Gender, and Ethnicity

Information on each defendant's race and gender comes from the arrest form completed by the arresting officer. Since the arrest form does not capture ethnicity, we use the U.S. Census Hispanic Surname List to determine the defendants' ethnicity [29]. Prior research on criminal justice has used this approach [28], [30], which has also been validated with self–reported data [31], [32]. We coded a defendant as Hispanic if 75% or more of the individuals with their surname self–identified as Hispanic or if the defendant was from a Spanish speaking country (other than Spain) according to the arrest form.

Word and Perkins [33] find that nearly all (93%) of the individuals that self-identify as Hispanic in the U.S. census have a surname that is the "Heavily Hispanic" i.e., has a 75% probability or more of being associated with a Hispanic individual. In other words, the "Heavily Hispanic" ($<$ 75%) category captures the vast majority of self–identifying Hispanics, making it the "determining factor why Spanish surname is such an excellent proxy for identifying Hispanics with the United States" [33]. Moreover, subsequent research [32] finds that the "Heavily Hispanic" ($<$ 75%) cut–off dramatically increases the accuracy of racial–ethnic categories. Thus, there is a strong precedent for defining Hispanic ethnicity based on a 75% threshold in the Hispanic Surname List.

D. Consensus–Driven Annotations from Student Raters

In order to compare how DLMs perform on data with annotations derived from a single rater vs. multiple raters, we pulled a stratified random sample of 13,927 mugshots from the full dataset. Specifically, we pulled a stratified random sample by race, ethnicity, and gender (2 race x 2 ethnicity x 2 gender = 8 groups) based on the official records and surname analysis to ensure sufficient representation of racial–ethnic and gender groups. Had we not oversampled White non–Hispanics or Black Hispanics, they would have been underrepresented in our data since they represent a smaller segment of defendants in MDC. The random sample of mugshots was then annotated by undergraduate research assistants at the University of Miami based on the four race–ethnicity subgroups. Less than 2% of mugshots were found to have issues with lighting or clarity and were removed from our analysis.

While viewing each mugshot, each student completed a Qualtrics survey asking about perceived physical characteristics including race–ethnicity annotations. In particular for race and ethnicity, questions such as, *"What race–ethnicity do you perceive this person to be?"* were asked. Responses were limited to Black Hispanic, White Hispanic, Black non-Hispanic, White non-Hispanic, or Other. To minimize priming effects, research assistants were not given names, racial identification, ethnicity, gender, or any other identifying information about the photos based on the court records.

We focused on the raters' subjective notion in categorizing a person's physical characteristics, which is why the raters were not given definitions for the various survey categories. Additionally, the raters independently completed the surveys on their own computers to avoid social desirability effects. Each photo was rated by three individuals, using the modal approach to create final student annotations of race and ethnicity. The use of multiple annotators in order to remove biases is recommended in the sociological literature based on mugshot photographs [28]. Inter–coder reliability for our annotated measure of race–ethnicity is high ($\alpha = 0.89$) and consistent with other research finding high rates of reliability for student–rated criminal mugshots [28]. Therefore, race–ethnicity annotations were based on a high amount of agreement among raters.

E. Image Classification using Deep Learning

Based on the above methods, we obtained two distinct sets of race, and race–ethnicity labels, to train and evaluate DLMs for image classification using mugshot data. Moreover, having a diverse group of research assistants rate our subsample of mugshots, we generated mugshots' race–ethnicity annotations based on established sociological practices that foster transparency within the DLM based classification pipeline.

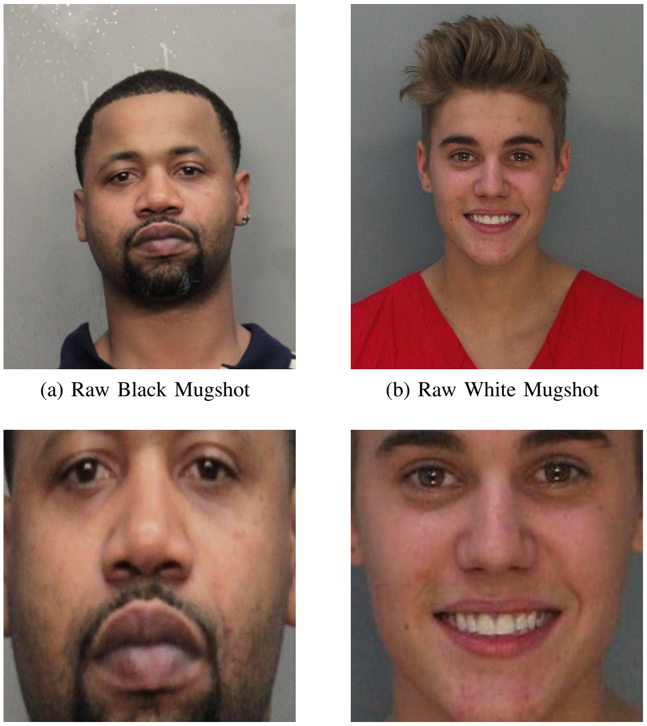
In the following section, we describe our experimental setting, how we prepare our datasets based on varying parameters for classification, discuss challenges faced and our results to answer our research questions from Section III.

V. Experiments and Discussions

A. Experimental Setup

We trained seven PyTorch [34] DLMs: ResNet–50 [35], AlexNet [36], Inception-v4 [37], SE–ResNet–50 [38], SE–ResNext–50_32x4d [38], VGG–16_bn [39], and VGG–19_bn [39] using the fastai library [40]. We conducted two overarching experiments: binary (Black and White) race classification, and four race–ethnicity classification, discussed in Sections V-B and V-C respectively. Moreover, we structure our experiments based on the following varying parameters:

- **Balanced vs. Imbalanced sampling.** For a balanced sampling setting, we first created subsamples consisting of 1,000 mugshots per race and racial–ethnic subgroup from both court and student label sources. To create



(a) Raw Black Mugshot

(b) Raw White Mugshot

(c) OpenFace Black Mugshot

(d) OpenFace White Mugshot

Figure 1. For illustration purposes, actual mugshots of celebrities from our dataset have been used to avoid invasion of privacy of the remaining defendants from 2010 to 2015. These images are readily available online, thereby mitigating privacy concerns associated with their publication. (a) and (b) are raw images with no face segmentation. (c) and (d) are mugshots after applying face segmentation using OpenFace [41].

groupings of 1,000 Black and 1,000 White mugshots for the race subsamples, we randomly pulled 500 samples from each BH and BnH, and WH and WnH racial–ethnic subgroup respectively.

For an imbalanced setting, we used the complete defendant dataset. For the binary race classification task, we combined the BH and BnH to create an imbalanced Black dataset containing 91,663 samples and combined their white counterparts to create an imbalanced White dataset containing 102,730 samples. See Section IV-A for MDC Defendants' population breakdown.

Lastly, we note that for every classification task, we used a constant random seed of 42 to shuffle the respective datasets to ensure no ambiguity in the order by which the images were processed by all DLMs. The dataset was then split into training and validation datasets using an 80:20 ratio. All results shown in Tables II and III are based on the DLMs' accuracies on the validation set.

- **Image Preprocessing.** All mugshots were processed in color, i.e. three channels, and resized to 224x224, resulting in a (3, 224, 224) tensor for each image. To increase DLMs' robustness, we added variations to the datasets by applying default fastai image transformations such as randomly horizontally flipping images with 0.5 probability, random lighting and contrast changes based on 0.2 probability and rotating the image ±10° with 0.75 probability that each affine transform and symmetric warp is applied [40]. For both classification tasks, we investigated how DLMs performed on raw (no pre–processing) and pre–processed mugshot data.

  In our paper, we opted for the OpenFace library [41] which allowed us to crop defendants' faces from the data but also transform the faces based on OpenCV's [42] affine transformation such that the eyes, center of the nose and upper lip appeared in the same location for every pre–processed image. See Fig. 1 for an example of raw vs. OpenFace binary race labeled data.

  However, we encountered certain limitations when using OpenFace, where 217 images could not be pre–processed. Upon further investigation, the majority of the images were due to variations in pose i.e. a person facing sideways, poor illumination making a face either hard to detect or too similar with the clothing worn, or excessive facial hair such as overgrown beards. Moreover, due to mugshots' variability in resolution and zoom, in order to maintain the affine transformation of some images, including the example in Fig. 1c, were stretched. Nevertheless, this only represented approximately 0.1% of the total dataset and so still a viable segmentation tool.
- **Metric, Hyper–parameters, and Fine–Tuning.** The evaluation metric used in all experiments was accuracy. This is calculated by comparing the predicted class categories, represented by indices in a NumPy array found by taking the argmax of the DLM's prediction probability from the set of total possible output classes, with the actual label of the images and then taking an average [40].

  For hyper–parameter settings, rather than guessing a good learning rate, we opted to use fastai's `lr_finder` methods [40] based on Howard's and Ruder's NLP inspired differential learning rates [43] and Smith's [44] highly effective 1-cycle policy that has allowed us to find near optimal learning rates unique to every subsample dataset and DLM combination for faster and fewer training epochs. We use the Adam optimizer and set our batch size to 64 images to update the DLMs' weights. As we are dealing with classification, we use a cross entropy loss function.

  The training process for each classification task was conducted in two–stages: (i) initialized by pre–trained ImageNet [45] weights, we apply transfer learning to train our DLMs for five epochs using the training dataset and outputting the DLM's accuracy for the validation dataset per epoch. (ii) Once tuned, we unfreeze

Table II
COMPARING THE PERFORMANCE OF SEVEN DEEP LEARNING MODELS FOR BINARY (BLACK AND WHITE) RACE CLASSIFICATIONS BASED ON COURT AND STUDENT ANNOTATED MUGSHOTS, 2010–2015

Model	**Raw Images**		**OpenFace**	
	Courts	Students	Courts	Students
ResNet–50	92.00%	93.50%	93.73%	91.72%
AlexNet	92.00%	92.75%	92.73%	89.72%
Inception–v4	94.25%	92.00%	93.98%	88.22%
SE–ResNet–50	93.75%	93.50%	93.98%	91.47%
SE–ResNext–50_32x4d	93.75%	89.25%	94.23%	89.72%
VGG–16_bn	94.00%	92.25%	92.23%	**93.98%**
VGG–19_bn	94.25%	92.50%	**94.48%**	91.47%

(a) Balanced classification: 1,000 samples per race subgroup.

Model	**Raw Images**	**OpenFace**
	Courts	Courts
ResNet–50	97.20%	**97.21%**
AlexNet	97.17%	96.84%
Inception–v4	97.26%	96.79%
SE–ResNet–50	97.37%	97.18%
SE–ResNext–50_32x4d	97.52%	97.12%
VGG–16_bn	97.45%	97.13%
VGG–19_bn	97.50%	97.08%

(b) Imbalanced classification: full Miami–Dade County defendant population.

the entire model and fine–tune with the same training set for five more epochs, validating after every epoch.

Based on our approach for tuning hyper–parameters, after conducting 10–20 experiments, we observed a lack in significantly improved accuracies and occasionally erratic changes in validation losses. We conclude that training our DLMs in two–stages of five epochs each was justified, particularly when training a relatively small dataset within the balanced setting to avoid chances of overfitting.

In total, 80 experiments were conducted and are summarized in Tables II and III. After each experiment, we saved the newly trained DLM with its updated weights, fine–tune parameters etc. in a weight file for testing and future research. As part of the DLMs' evaluation and interpretation process after every stage listed above using fastai's `ClassificationInterpretation` methods [40], we compared our DLMs predictions with ground–truth labels from the validation dataset by plotting confusion matrices with respect to the number of output classes per classification task. Furthermore, we applied Grad–CAM heatmaps [40], [46] on at least nine mugshots to gain deeper insight into which pixels in those images caused DLMs to predict classifications that resulted in the greatest loss.

B. Case 1: Binary Race Classification

Table II shows that within a balanced subsample setting, DLMs achieved comparable accuracies of 92.00% to 94.48% across 28 experiments. However, we should note that DLMs trained using court annotations obtained slightly higher overall classification accuracies than student annotated DLMs. Counter–intuitively, we observed noticeable outperformance by the DLMs when given raw images, such as Fig. 1a and 1b, as opposed to OpenFace segmented images. However, upon interpreting our DLMs' predictions by applying heatmaps to mugshots, we often found no pixels across the face to be highlighted, instead pixels from the background and clothing appeared to be greater contributing factors.

Thus, despite the lower performance in Case 1 and as we will discuss in Case 2, for the remaining analysis we focus on OpenFace mugshots' accuracies. As shown in Fig. 1c and 1d, the face occupies the majority of the image, thereby minimizing external features from the raw image to influence DLMs' predictions. For both Tables II and III, we have highlighted the model that performed with greatest accuracy in bold for both label sources (courts and students). Curiously, our experiments did not result in a singular DLM architecture that performed best under all settings for race classification.

In the case of the imbalanced setting where the entire defendant population was used, we note a definite increase in accuracy by 2.73% when comparing the best models from both settings, VGG–19_bn (balanced courts) and ResNet-50 (imbalanced courts). While an increase in accuracy is a good sign, we wonder about the value of providing almost 100 times more annotated data to yield such a small gain, particularly if these DLMs were to be used in practical applications. Moreover, with the notion of biased demographic datasets being a central concern for DLM based FPTs [4], the gain in accuracy came at the cost of the full dataset being more skewed to the White defendant population 52.85% to 47.14% for the Black population.

C. Case 2: Four Race-Ethnicity Classification

Unlike Case 1, results from Table III indicate that race–ethnicity classification is a lot more complex and subjective than differentiating the binary racial labels (Black/White). When considering court–based DLMs, accuracies on average were only slightly better than chance (56.44%) within a balanced race–ethnicity setting, which is not helpful when we are tasked with predicting four output classes. While it was expected to see greater accuracies when considering almost 50 times more data within the imbalanced setting, we are cautious of the DLMs' generalizability due to nearly 75% of the full dataset belonging to the WH and BnH subgroups from Table I.

Table III
COMPARING THE PERFORMANCE OF SEVEN DEEP LEARNING MODELS (DLMS) FOR FOUR RACE–ETHNICITY CLASSIFICATIONS BASED ON COURT AND STUDENT ANNOTATED MUGSHOTS, 2010–2015

Model	**Raw Images**		**OpenFace**	
	Courts	Students	Courts	Students
ResNet–50	56.20%	73.30%	55.31%	70.71%
AlexNet	58.75%	75.87%	60.95%	73.46%
Inception–v4	59.00%	71.25%	51.43%	67.83%
SE–ResNet–50	61.12%	76.25%	**61.32%**	**74.84%**
SE–ResNext–50_32x4d	61.25%	79.12%	48.31%	70.46%
VGG–16_bn	60.50%	76.37%	58.19%	74.09%
VGG–19_bn	63.87%	77.12%	59.57%	74.09%

(a) Four race–ethnicity classification: balanced (1,000) samples per race subgroup.

Model	**Raw Images**	**OpenFace**
	Courts	Courts
ResNet–50	80.60%	80.93%
AlexNet	79.09%	79.93%
Inception–v4	80.79%	80.18%
SE–ResNet–50	80.61%	**81.05%**
SE–ResNext–50_32x4d	80.40%	80.77%
VGG–16_bn	80.26%	77.92%
VGG–19_bn	80.43%	79.77%

(b) Four race–ethnicity classification: imbalanced full defendant population.

The most important finding from this paper, however, are the accuracies from student rated DLMs for race–ethnicity classification within a balanced subsample setting. They outperformed their court annotated counterparts consistently, ranging from a 12.51% to 22.15% increase in accuracy. Interestingly, unlike Case 1, the pre–trained SE-ResNet-50 model produced the highest accuracies for all OpenFace race–ethnicity annotated data. What makes the student annotated SE-ResNet-50 model even more impactful is that it was trained on only 1,000 mugshots per racial–ethnic category and it underperformed by just 6.21% than the imbalanced court annotated SE-ResNet-50 model that used 50 times more data.

D. Limitations and Future Work

Despite using seven different model architectures, the difference between the lowest and highest accuracies for Case 1 is 2.25% (courts) and 5.76% (students), and for Case 2 is 13.01% (courts) and 7.01% (students), when using a balanced sample size with OpenFace preprocessing. This suggests that the fundamental principle behind the success of deep learning based image classification largely remains consistent and that a model's architecture is less contributory when applying transfer learning techniques and using pre–trained weights. Extending the work of Kornblith et al. [47], future research should investigate the impact of deep learning architectures by training from scratch to see if a difference in architecture yields to different results based on the same mugshot data.

From the remaining 9,927 student annotated mugshots, we will use the student annotated SE–ResNet–50 model and apply inference learning to (i) test its performance in predicting race–ethnicity labels of the remaining mugshots, and (ii) generate new labels based on human perceptions of race–ethnicity for the remaining 180,466 mugshots from the full dataset and compare performance with the 81.05% accurate court annotated SE–ResNet–50 that used the "U.S. Hispanic Surnames" text based approach. In addition to evaluating accuracies on the full MDC defendant population, we will evaluate how our DLMs perform with respect to each race–ethnicity subgroup individually. Thus, building more demographically inclusive DLMs across the multidimensionality of race–ethnicity will promote greater "Fairness in Machine Learning" [23].

While mugshots represent a novel data source for future work, researchers should tread lightly and with care. In contrast to celebrity photo databases, there are additional privacy concerns regarding consent and data sharing. Although mugshots are publicly available in some jurisdictions like MDC, researchers should not necessarily post mugshot databases online because doing so may inadvertently fuel ethnically questionable websites charging fees to remove mugshots from the internet [48], [49]. Since potential employers, lenders, and other institutions can discriminate against individuals with a mugshot online, such unscrupulous website practices are tantamount to extortion [48], [49]. Therefore, researchers using mugshot databases should develop rigorous data sharing agreements that protect against these privacy concerns, while also allowing for replication and transparency pillars of the academic enterprise.

VI. CONCLUSION

We proposed a novel multidimensional approach for understanding and annotating race by looking at race–ethnicity combinations. We conclude that when identifying Black and White race classification, both sets of models achieved greatest accuracies of 94.48% (courts) and 93.98% (students). However, when classifying four–subgroups, we find greater disparity of 61.32% (courts) and 74.84% (students). Despite calls for more research on Hispanic annotations [10], little work exists on the topic. Our multidimensional approach helps to move the literature forward by empirically assessing the validity of DLM race–ethnicity annotations and providing a framework for further research. As the U.S. Hispanic population continues to grow [27], understanding the role of ethnicity in DLMs will become increasingly important.

Our results also yield important insights about the amount of data needed to generate multidimensional race–ethnicity labels. In particular, we find that "bigger is not necessarily better", when comparing the best performing DLMs trained using balanced student annotations (2,000 samples for Case 1 and 4,000 samples for Case 2) against imbalanced court annotations (194,393 samples for both cases), where the "court-DLMs" yield higher accuracies by a few percent, 3.23% (Case 1) and 6.21% (Case 2). Furthermore, given the time–consuming nature of hand–annotation, our results from using student annotated DLMs achieving 93.98% (race) and 74.84% (race–ethnicity) suggests that researchers can produce high–quality DLM generated labels for large–scale databases with only a few thousand hand–annotations.

Finally, our research will also generate new methodological approaches. Using a multidimensional approach for generating race–ethnicity annotations can help understand racial and ethnic disparities in MDC's criminal justice system or be scaled–up to a national level to study criminal justice disparities with large–scale mugshot databases comprising of millions of images. This method could be also be used to fill missing race–ethnicity identification in court databases or other databases using photos to generate race–ethnicity classifications. Given that hand–coding photo data is a time consuming and cost–prohibitive process, our research will help others responsibly unlock the potentials of identifying such patterns present in large–scale face datasets and help advance methods for studying a wider range of social issues.

Acknowledgment

This project was supported by the interdisciplinary U–LINK grant (PG012249) and U–LINK Predoctoral Fellowship. We would like to thank Dr. Tamara Rice Lave and Dr. Cameron Riopelle for their insightful feedback on the project.

References

[1] I. D. Raji *et al.*, "Saving face: Investigating the ethical concerns of facial recognition auditing," in *AIES '20: Conference on AI, Ethics, and Society*. ACM, 2020, pp. 145–151.

[2] A. Rose, "Are face-detection cameras racist?" *Time Business*, 2010. [Online]. Available: https://bit.ly/2Rt5FQR

[3] A. Khan and M. Mahmoud, "Considering race a problem of transfer learning," in *IEEE Winter Applications of Computer Vision Workshops*. IEEE, 2019, pp. 100–106.

[4] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on Fairness, Accountability and Transparency*, vol. 81. PMLR, 2018, pp. 77–91.

[5] I. D. Raji and J. Buolamwini, "Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 429–435.

[6] S. Cornell and D. Hartmann, *Ethnicity and race: Making identities in a changing world*. Sage Publications, 2006.

[7] E. Bonilla-Silva, *Racism without racists: Color-blind racism and the persistence of racial inequality in the United States*. Rowman & Littlefield Publishers, 2006.

[8] B. Western, *Punishment and inequality in America*. Russell Sage Foundation, 2006.

[9] N. Petersen *et al.*, "Unequal treatment: Racial and ethnic disparities in Miami–Dade criminal justice," 2018.

[10] K. Kärkkäinen and J. Joo, "Fairface: Face attribute dataset for balanced race, gender, and age," *CoRR*, vol. abs/1908.04913, 2019. [Online]. Available: http://arxiv.org/abs/1908.04913

[11] S. Nagpal, M. Singh, R. Singh, M. Vatsa, and N. K. Ratha, "Deep learning for face recognition: Pride or prejudiced?" *CoRR*, vol. abs/1904.01219, 2019. [Online]. Available: http://arxiv.org/abs/1904.01219

[12] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang, "Racial faces in the wild: Reducing racial bias by information maximization adaptation network," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 692–702.

[13] H. Han, A. K. Jain, F. Wang, S. Shan, and X. Chen, "Heterogeneous face attribute estimation: A deep multi-task learning approach," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 11, pp. 2597–2609, 2017.

[14] M. Alvi, A. Zisserman, and C. Nellåker, "Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 0–0.

[15] A. Das, A. Dantcheva, and F. Bremond, "Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[16] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5810–5818.

[17] S. Escalera *et al.*, "Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 1–8.

[18] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.

[19] G. Farinella and J.-L. Dugelay, "Demographic classification: Do gender and ethnicity affect each other?" in *International Conference on Informatics, Electronics & Vision (ICIEV)*. IEEE, 2012, pp. 383–390.

[20] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.

[21] K. Ricanek and T. Tesafaye, "Morph: A longitudinal image database of normal aduschmidt lt age-progression," in *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*. IEEE, 2006, pp. 341–345.

[22] M. Merler, N. K. Ratha, R. S. Feris, and J. R. Smith, "Diversity in faces," *CoRR*, vol. abs/1901.10436, 2019.

[23] H. J. Ryu, H. Adam, and M. Mitchell, "Inclusivefacenet: Improving face attribute detection with race and gender diversity," *arXiv preprint arXiv:1712.00193*, 2017.

[24] A. Hanna, E. Denton, A. Smart, and J. Smith-Loud, "Towards a critical race methodology in algorithmic fairness," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 501–512.

[25] J. R. Logan, "How race counts for hispanic americans." 2003.

[26] P. H. Wood, J. D. Forbes, V. M. Gould, and S. D. Greenbaum, *The Afro-Latin@ Reader: History and Culture in the United States*. Duke University Press, 2010.

[27] A. Flores, M. Lopez, and J. Krogstad, "Us hispanic population reached new high in 2018, but growth has slowed," *Pew Research Center. Retrieved July*, vol. 22, p. 2019, 2019.

[28] R. D. King and B. D. Johnson, "A punishing look: Skin tone and afrocentric features in the halls of justice," *American Journal of Sociology*, vol. 122, no. 1, pp. 90–124, 2016.

[29] D. L. Word, C. D. Coleman, R. Nunziata, and R. Kominski, "Demographic aspects of surnames from census 2000," *Unpublished manuscript, Retrieved from http://citeseerx. ist. psu. edu/viewdoc/download*, 2008.

[30] K. Beckett, K. Nyrop, and L. Pfingst, "Race, drugs, and policing: Understanding disparities in drug delivery arrests," *Criminology*, vol. 44, no. 1, pp. 105–137, 2006.

[31] M. N. Elliott *et al.*, "Using the census bureau's surname list to improve estimates of race/ethnicity and associated disparities," *Health Services and Outcomes Research Methodology*, vol. 9, no. 2, p. 69, 2009.

[32] I. I. Wei, B. A. Virnig, D. A. John, and R. O. Morgan, "Using a spanish surname match to improve identification of hispanic women in medicare administrative data," *Health Services Research*, vol. 41, pp. 1469–1481, 2006.

[33] D. L. Word and R. C. Perkins, *Building a Spanish Surname List for the 1990's–: A New Approach to an Old Problem*. Population Division, US Bureau of the Census Washington, DC, 1996.

[34] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012, pp. 1097–1105.

[37] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[38] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR*, 2015.

[40] J. Howard and S. Gugger, "Fastai: A layered API for deep learning," *Information*, vol. 11, no. 2, p. 108, 2020. [Online]. Available: https://doi.org/10.3390/info11020108

[41] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," CMU-CS-16-118, Tech. Rep., 2016.

[42] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[43] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL*, vol. 1, 2018, pp. 328–339.

[44] L. N. Smith, "Cyclical learning rates for training neural networks," in *2017 IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2017, pp. 464–472.

[45] J. Deng *et al.*, "Imagenet: A large-scale hierarchical image database," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[46] R. R. Selvaraju *et al.*, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[47] S. Kornblith, J. Shlens, and Q. V. Le, "Do better imagenet models transfer better?" in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 2661–2671.

[48] S. Schmidt, "This site will remove your mug shot – for a price, authorities say. its owners are charged with extortion." *The Washington Post*, 2018. [Online]. Available: https://wapo.st/2yJZPDX

[49] K. Duffin, "The business of posting mugshots online and charging people to take them down," *National Public Radio*, 2018. [Online]. Available: https://n.pr/2wrUAYQ