University of Nevada, Reno

Taking Politics at Face Value: How Features Expose Ideology

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Social Psychology

by Christopher Copp

Dr. Markus Kemmelmeier, Dissertation Advisor

December, 2022

Abstract

Previous studies using computer vision neural networks to analyze facial images have uncovered patterns in the feature extracted output that are indicative of individual dispositions. For example, Wang and Kosinski (2018) were able to predict the sexual orientation of a target from his or her facial image with surprising accuracy, while Kosinski (2021) was able to do the same in regards to political orientation. These studies suggest that computer vision neural networks can be used to classify people into categories using only their facial images.

However, there is some ambiguity in regards to the degree to which these features extracted from facial images incorporate facial morphology when used to make predictions. Critics have suggested that a subject's transient facial features, such as using makeup, having a tan, donning a beard, or wearing glasses, might be subtly indicative of group belonging (Agüera y Arcas et al., 2018). Further, previous research in this domain has found that accurate image categorization can occur without utilizing facial morphology at all, instead relying upon image brightness, color dominance, or the background of the image to make successful classifications (Leuner, 2019; Wang, 2022).

This dissertation seeks to bring some clarity to this domain. Using an application programming interface (API) for the popular social networking site Twitter, a sample of nearly a quarter million images of ideological organization followers was created. These images were followers of organizations supportive of, or oppositional to, the polarizing political issues of gun control and immigration. Through a series of strong comparisons, this research tests for the influence of facial morphology in image categorization. Facial images were converted into point and mesh coordinate representations of the subjects'

faces, thus eliminating the influence of transient facial features. Images were able to be classified using facial morphology alone at rates well above chance (64% accuracy across all models utilizing only facial points, 62% using facial mesh). These results provide the strongest evidence to date that images can be categorized into social categories by their facial morphology alone.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

In 2018, Stanford researchers made a remarkable claim. Using images from a dating website, Wang and Kosinski (2018) found that they were able to assess a person's sexual orientation, solely from a picture of their face. The researchers used a deep neural network (DNN) to extract "features," an array of numbers, from over 35,000 images retrieved from a dating website (Wang & Kosinski, 2018). A logistic regression classifier was trained separately for men and women, and accuracy of predictions was recorded. When tested on a novel set of images, the classifier was able to predict the sexual orientation of the individual in the image at an accuracy of 81% for males and 71% for females (Wang & Kosinski, 2018). For each model, the classification algorithm was superior at identifying the sexual orientation of the subject to a sample of human counterparts, who demonstrated considerably lower accuracy in their study, 61% and 54% for male and female images respectively (Wang & Kosinski, 2018). Further, additional pictures of a subject increased the accuracy of the DNN, reaching 91% for men with five images (Wang & Kosinski, 2018).

The results of the study, perhaps unsurprisingly, created a great deal of controversy. News media claimed that we were entering the era of "Minority Report", a 2002 movie in which legal judgments are delivered for future crimes before they even take place (Levin, 2017). While some critics claimed the authors were delivering a tool to oppressive regimes around the world to identify and persecute homosexuals, others suggested that the work was illegitimate, believing that the study's methodology was suspect and that the authors' claims were overstated.

The research also hearkens back to a time when physiognomy, the practice of

discerning characteristics about a person from their facial features, was considered mainstream science. This is not a comparison to be taken lightly. Physiognomy has a troubled history, as it has often been used to justify the oppression of people whom those in power deemed to be undesirable. The mere mention of physiognomy is seen as being proximal to the belief that the internal character of an individual is secondary in comparison to their external, superficial characteristics. These are unpopular notions in societies where the "content of one's character" is generally believed to be of paramount importance.

An additional concern is whether the findings from the 2018 study were even valid. Opposing scholars pointed out that the features extracted from the images did not differentiate between the structure of the face itself and the other contextual clues in the image. Wang and Kosinski (2018) themselves acknowledge that their research suggests heterosexual men are more likely to wear beards than homosexual men, for example. At the same time, heterosexual women are more likely to wear makeup than homosexual women (Wang & Kosinski, 2018). It is possible that there are systematic differences in the presentation of individuals based on their sexual orientation. It then follows that these environmental differences, rather than differences in facial structure, are what allows for the illusion of predictive power in categorization by facial morphology.

Doubling down on the controversy surrounding the initial paper, Kosinski (2021) published a follow-up study differentiating people by their political orientation. Using over one million images gleaned from dating websites as well as Facebook, Kosinski (2021) attempted to address some of the criticisms from the previous study. For example, the author contrasts between facial morphology, stable facial features due to the size, shape, and proportions of the face, and transient facial features, which are dynamic features such

as facial expression, adopting makeup or facial hair, or head tilt and positioning (Kosinski, 2021). The study was also more inclusive, with over 300,000 images of non-white subjects.

This second study achieved impressive success in predicting people's political orientation solely from an image of the person's face, achieving an accuracy rate between 65% and 73%, depending upon the sample and whether demographics were controlled for or not (Kosinski, 2021). The accuracy of the model decreased slightly when controlling for demographic variables, although transient features of the face (such as adopting facial hair or wearing sunglasses) seemed to have little predictive impact.

Currently, there is a great deal of uncertainty around this subject matter. Because of the novelty of machine learning techniques and computer vision, many are confused by the meaning and limitations of this type of technology. Further, activist organizations seem to be torn between condemning the research for its success in being able to identify minority populations without their consent while at the same time criticizing the research as pseudoscience. While the results of the two studies are intriguing, some fear the potential implications of the findings.

For one, the resurgence of something resembling physiognomy seems to indicate the belief in a deterministic world where people are confined to their biological limitations, independent of individual effort, character, or merit. The implications of that are startling to say the least. Human beings thrive on the belief that they have agency, and that they can affect their environment in a constructive way. This type of research threatens that belief system.

Further, we all have firsthand knowledge of technological growth on an exponential scale. While it took tens of thousands of years for humans to gain the ability to fly, it was

only 66 years after that flight that human beings walked on the moon (Sheth, 2017). If technology can currently predict sexual or political orientation with the accuracies presented in the two studies, surely this accuracy will only increase over time. Further, machine learning techniques are becoming more ubiquitous by the day, as well as easier to implement. This portends a novel and easy way in which to oppress undesirable populations, as has been the history of physiognomy previously, which is quite vexing in and of itself.

Perhaps even more problematic is the potential for models such as these to be misused or misinterpreted. For example, governments oppressive to sexual minorities, religious minorities, or political dissidents could claim model accuracy rates approaching 100%, tacitly providing a sense of legitimacy for the government in question to oppress. In more progressive societies, such models might be interpreted incorrectly, misclassifying people and potentially causing distress in a variety of ways. For example, we know that machine learning algorithms have been employed to discover pregnancies through shopping habits in order to better serve customers, resulting in at least one instance of a father being notified of his teenage daughter's pregnancy via an abundance of coupons arriving through the mail (Hill, 2012). In this same vein, we might one day expect to see the sexual orientation of a teenager exposed to his or her socially conservative parents.

Additionally, there are questions related to the underlying methodology utilized in Wang and Kosinski (2018) and Kosinski (2021). Perhaps most controversial is the question of whether the classifiers are categorizing these images based on the morphology of the face, that is, the specific facial structure itself, or whether the classifiers are being influenced by additional information present in the images, such as demographic

information (e.g., age, gender, ethnicity) or transient facial features (e.g., wearing makeup, facial hair, facial expression). Wang and Kosinski (2018) claim that their results are due in large part to facial structure while at the same time acknowledging group differences in transient facial features. Specifically, they found that heterosexual men were more likely to wear facial hair and heterosexual women were more likely to wear makeup than their homosexual counterparts. At the same time, the researchers discovered there were group differences in the way the subjects photographed themselves, with heterosexual men and homosexual women shooting their photographs from below while homosexual men and heterosexual women tended to shoot their selfies from above.

Kosinski (2021) attempted to remedy some of this criticism by taking into account head pose as well as facial expression, two transient facial feature sets left unexamined in the previous research. He also examined whether subjects were wearing glasses or sunglasses or displaying facial hair. Kosinski found that these factors did explain a considerable amount of the variance in the success of the model, although these features did not account for the entirety of it. This strongly suggests that facial morphology was contributing, at least partially, to the success of the model.

Because Kosinski did not explicitly isolate facial morphology, however, the influence of these transient features remain fairly ambiguous. Although Kosinski (2021) controlled for facial expression, head pose, and whether the participant was wearing glasses or not, he did not control for other transient features that might also have had influence, such as how tan the person was, whether or not they were wearing makeup, how their hair was maintained, or countless other phenomena that might have contributed to model success. Because of this, it is unclear as to whether the success demonstrated in Wang and

Kosinski (2018) and Kosinski (2021) is due to contextual clues within the image, or rather the structure of the participant's face itself.

This dissertation serves to bring some badly needed clarity to the type of research Wang and Kosinski (2018) and Kosinski (2021) performed. It seeks to replicate and extend the findings of Wang and Kosinski (2018) using a new sample gained from Twitter profile pictures, specifically followers of political organizations. A study of the type proposed would serve many benefits. First, it contributes to an existing line of research that has investigated the relationship between facial appearance and political ideology. Second, it utilizes an uncommon methodology in a novel way that could potentially reveal more about the link between faces and politics. Third, it could foment an increase in the breadth and depth of analysis for this field, as it has mostly been overlooked, misunderstood, or underappreciated. Fourth, it seeks to clear up some of the ambiguity in previous research, specifically the failure of Wang and Kosinski (2018) and Kosinski (2021) to isolate for facial morphology exclusively, making it unclear as to whether the models are utilizing facial structure (e.g., jaw size) to make their predictions, or rather utilizing transient features that are unrelated to facial structure (such as head position, emotional expression, or adopting facial hair or makeup). While Kosinski (2021) attempted to control for some of these variables, the proposed methodology would go much further in isolating facial morphology.

Chapter 2 briefly discusses the historical practice of physiognomy and the shortcomings of research related to physiognomy in general.

Chapter 3 reviews the extant literature on categorizing faces in a broad sense. The chapter begins with research describing how people can process faces for information with

high reliability, categorizing people by emotional expression, age, sex, and kinship (Biehl et al., 1997; Brewer, 1998; Kazem & Widdig, 2013). The chapter then narrows its focus, highlighting research that suggests people are also adept at categorizing people by their sexual orientation, socio-economic status, and religion (Skorska et al., 2015; Kraus and Keltner, 2009; Rule et al., 2010). Finally, the chapter goes into research highlighting people's ability to categorize others in regards to their personality characteristics, including "Big Five" traits such as extraversion and agreeableness, as well as other personality characteristics, such as baby-facedness or trustworthiness (Carney et al., 2007; Ambady & Rule, 2010; Todorov et al., 2008).

Chapter 4 delves into political orientation. This chapter has three primary focuses. It begins by highlighting some of the differences in personality disposition observed between conservatives and liberals. Next, it illuminates research describing the heritability of political ideology. Finally, it describes research in which experimental participants accurately categorize people into political groups merely by their appearance, similar to the research in Chapter 3 but for political affiliation.

Chapter 5 gives a basic background on neural networks, computer vision, and the training/testing methodology of machine learning algorithms.

Chapter 6 highlights some of the current social science research that utilizes neural networks in its methodology, with the most relevant works being Wang and Kosinski (2018), Leuner (2019), and Kosinski (2021). It also highlights some of the shortcomings of both Wang and Kosinski (2018) and Kosinski (2021), as well as how the current research remedies some of those shortcomings.

Chapter 7 reviews some of the broader ethical considerations for research involving

facial recognition as well as its application in real-life. Chapter 8 details the method of the research, as well as the theory behind the design. It also includes several hypotheses.

Chapter 9 goes over the results of the analyses. This chapter is divided into two parts. The first section of chapter 9 runs a series of inferential tests to examine the variables gleaned from the photographs, including sex, race, age, head positioning, and emotional expression. Section 2 uses a logistic regression classifier to predict and categorize images per their group membership across a series of models. By comparing model success, this research hopes to illuminate the process by which the classifier is making its predictions, as well as revealing whether facial morphology is implicated in how images are categorized.

Chapter 10 discusses the implications of the results. It also highlights the benefits of the current line of research, the limitations of this research, the role of legislative intervention, and the future of research incorporating this methodology.

Chapter 2: Physiognomy

Physiognomy, the practice of discerning characteristics about a person from their physical features, has had a long and contentious history across many cultures. Literature dating back to antiquity is replete with examples of physiognomy. For example, treatises on physiognomy were written by both Plato and Aristotle, who believed that physical beauty was associated with moral virtue (Twine, 2002). The Greeks believed that a person's character could be divined by observing their features and identifying the animal that they most resemble (Jenkinson, 1997). Similarly, a person's temperament was seen to be determined by their relationship to the four bodily humors, which were in turn related to the four elements of the earth (NIH, n.d.).

Lavater, a Swiss Protestant pastor, carried this idea forward to the 18th century when he wrote a four-volume treatise on physiognomy entitled *Essays of Physiognomy* (Hassin & Trope, 2000). Lavater claimed scientific rigor in his analyses of people's faces, and believed moral virtue was represented by physical beauty in individuals. He extolled and reiterated the ancient Greek idea that a surplus of one of the four humors would create a physical impression in the face in the form of lines or wrinkles that could be read and interpreted (Jenkinson, 1997).

Lavater's work was extremely popular in his time. His work *Essays of Physiognomy* was printed in at least 55 different editions, and included artwork from famous intellectuals and artists of the time including William Blake and Henry Fuseli (Twine, 2002). The impact of physiognomy on culture can also be seen through its inclusion in popular works of 19th century literature. Physiognomy was included in the works of Dickens, Balzac, Brontë, Austen, Eliot, and Wilde (Twine, 2002). For example,

in Charlotte Brontë's work, *The Professor*, there are no less than 30 mentions of physiognomy or phrenology (the notion that characteristics of one's skull reveal psychological characteristics), and the author explicitly mentions the link between her characters' physical appearances and their morality (Twine, 2002).

Despite this popularity, physiognomy and its cousin phrenology eventually fell out of favor. This occurred for several reasons. First, although some socialists and feminists at the time utilized these ideas as tools for social justice, their more common use was to reinforce existing social hierarchies and to legitimize racial and ethnic prejudice under the guise of scientific inquiry (DeLisi, 2013; Twine, 2002). Despite being pseudoscience, this type of research could have been interpreted in such a way as to offer struggling populations more assistance to reach their full potential. For example, the belief that some populations had a proclivity for criminal behavior could have been used as a justification for additional social spending to curb such behaviors. Instead, it was common to argue that these populations were simply inferior from birth and deserved nothing but reproach. Second, the movement of eugenics was irrevocably linked to the study of physiognomy, and as eugenics fell out of favor, physiognomy did as well (Twine, 2002). Third, the study of physiognomy primarily focused on the link between physical attractiveness and moral virtue, two qualities which critics correctly identify as being, at least in part, socially constructed (Twine, 2002). In other words, both attractiveness and virtuousness are both influenced by the norms present in the culture under examination. Further, moral virtue is inherently more subjective than other, more objective measures of personality, such as extraversion or conscientiousness.

However, the primary problem with physiognomy research is not often mentioned

in works that are critical of the practice. In a very real sense, physiognomy bears a striking resemblance to work that researchers traditionally carry out with little or no controversy. For example, on average men and women differ by temperament, with women being more agreeable, more subject to neuroticism or negative emotion, and less aggressive than their male counterparts (see Hyde, 2014, for a review). At the same time, men and women demonstrate biological differences in facial structure, with women typically having larger eyes, proportionately, as well as smaller foreheads, chins, and noses (Burke & Sulikowski, 2010). Thus, we might expect that a classifier trained to interpret psychological temperament via facial structure would find that people with larger eyes and smaller facial features are, in general, more agreeable, less aggressive, and demonstrate more neuroticism. When researchers control for biological sex in their research, they are controlling for both the personality differences between the sexes as well as their differences in facial attributes, regardless of whether that is understood or acknowledged at the time. In a very real way, this type of research is not dissimilar to controlling for sex, for example, with the sex-characteristic facial features being a proxy for self-reports of the participant's biological sex.

Despite its potential accuracy, the aforementioned hypothetical model does not provide information about specific individuals. It is unclear, for example, whether a woman picked at random would demonstrate more or less agreeableness than a man picked at random. The primary shortcoming of traditional physiognomy research occurs when characteristics that have been observed at a group level of analysis are mistakenly applied to people at the individual level. The fact that one can measure mean scores on any metric and organize those averages into demographic categories tells us nothing about the specific

scores of any individual in any demographic group, because demographic groups typically demonstrate wide within-group variation and considerable overlap with one another (Rosenberg et al., 2002).

Additionally, while these analyses might correctly observe that there are naturally occurring differences between groups that can be measured and replicated, they must also acknowledge that the way these groups are treated could result in differential outcomes. Women might demonstrate more neuroticism because they are treated more poorly by patriarchal societies, for example.

Whether group differences can be attributed to differential treatment is a difficult question to unpack. To take another example, Lavater believed moral virtue was related to physical attractiveness (Twine, 2002). Even if this were to be true, it would still be unclear whether attractive people are treated in a more positive way by others, and that positive treatment is a meaningful covariate.

Chapter 3: Categorizing Faces

The most immediate and omnipresent social inferences that people make from faces are often overlooked. For example, it may seem obvious, but human beings are excellent at categorizing the emotional expressions on the faces of others. In his work *The Expression of the Emotions in Man and Animals*, Darwin (1872) found evidence for the universality of facial expressions in human beings and posited that emotions are also visible in animal facial physiology (Darwin & Prodger, 1998; Ekman, 2003). This tradition of research has continued into more contemporary times. For example, Biehl et al. (1997) used a dataset of Japanese and Caucasian faces demonstrating the emotions of anger, contempt, disgust, fear, sadness, and surprise. The authors had participants from six different countries categorize the facial images by their displayed emotions. Accuracy across the images was very high, and the displayed emotion was the most common response for 321 out of 336 images (Biehl et al., 1997). While not often interpreted in such a way, this type of research demonstrates that people can accurately categorize people by their emotional state simply by looking at their face.

Although these results are perhaps not particularly astonishing, they demonstrate a remarkable achievement for human cognition. We can observe a complete stranger, and discern with a high level of accuracy what they are feeling, and potentially, what their intentions are. Not only that, but we do so unconsciously, quickly, and with little cognitive load (Willis & Todorov, 2006). Obviously, the context in which these emotions are displayed is likely more important than the expressions themselves. However, to our more primitive ancestors, a quick determination about emotion could mean the difference between life and death.

Additionally, we instantaneously process the faces of others for indicators of age and sex (Brewer, 1988). For example, Fink et al. (2006) were curious to discover if skin color homogeneity was related to perceived attractiveness. Even without the facial indicators traditionally associated with age, such as wrinkles or furrows, people were able to differentiate between older and younger skin colorings on neutral facial templates (Fink et al., 2006). Participants rated the images with younger skin tones as significantly more attractive, more youthful, and healthier (Fink et al., 2006). These traits were likely very adaptive in our evolutionary history for determining potential sexual partners.

Similarly, research has shown that individuals are rather adept at determining familial relationships simply by observation, including the links between grandparents and grandchildren (Kaminski et al., 2009). This adaptation would have many benefits, including differential treatment towards those that carry one's own genes, ascertaining tribal allies or adversaries, preventing inbreeding, and so on. The ability is also present in the assessment of different species, going so far as to allow human beings to ascertain familial relationships in other primates (Kazem & Widdig, 2013). Further, our ability to automatically process the race of others might have assisted our in differentiating members of their own group and that of outsiders (Brewer, 1988). This would give us a natural advantage in making new allies or preventing potential threats.

Additional avenues of facial processing include gaze and attraction (Leopold & Rhodes, 2010). In human beings, the white sclera provides ample contrast against the darker irises, allowing others to easily orient their gaze to match our own (Kobayashi & Kohshima, 1997). Human infants have been shown to rely upon the gaze of another when searching for information, in contrast to other primates that typically follow the positioning

of the head (Tomasello et al., 2006). In terms of attraction, people unconsciously process features like face averageness and symmetry, a rudimentary but important proxy for biological fitness in sexual selection (Rhodes & Simmons, 2007).

While the importance of first impressions is well-known, research suggests that people make inferences from people's faces in as little as 100 ms, with increased exposure to the faces resulting in no overall change in judgment (Willis & Todorov, 2006). It appears as though interpreting faces is hard-wired into human beings' cognition, as indicated by experiments using inverted faces, or even the Thatcher effect, where a face with jumbled facial features is difficult to observe when inverted (Willis & Todorov, 2006; Farah et al., 1995; Carbon et al., 2005; see Appendix A).

These findings point to abilities that may go unappreciated. The ease with which we process this type of information has deceived us into thinking that something that should be very difficult is, for us, very easy. The fact that human beings can perform these acts universally, automatically, with little cognitive load, and with greater than chance accuracy, is, in itself, suggestive of the utility of physiognomy. However, there is additional evidence that people can even infer more subtle social traits from facial indicators.

Social Traits

People make inferences about people's social traits and affiliation from their faces. For example, Rule et al. (2008) found that people were able to predict sexual orientation of subjects across five studies at a better than chance level. They also attempted to uncover which facial indicators people relied upon when making their determinations. They purposefully occluded certain parts of the sample images such as the eyes, hair, and mouth,

and found that these occlusions did not prevent better than chance accuracy at predicting the sexual orientation of the target (Rule et al., 2008). This finding was reinforced by Skorska et al. (2015), who reported systematic differences in the faces of lesbian women in comparison to heterosexual women and gay men in comparison to heterosexual men.

As previously mentioned, Wang and Kosinski (2018) utilized a neural network to categorize facial images by their extracted features. Using images from a dating website, they were able to analyze the images and accurately predict the sexual orientation of the target at levels better than chance. Furthermore, additional images of the target increased the accuracy of their model.

Kraus and Keltner (2009) examined nonverbal cues of social class. They took 60-second video clips of low-SES and high-SES people interacting with a stranger. Raters analyzed the clips examining engagement and disengagement behaviors, and estimating SES based on the interactions. Engagement behaviors are non-verbal behaviors that implicitly imply engaging in conversation, while disengagement behaviors do the opposite. They found that high-SES was associated with disengagement cues like self-grooming, fidgeting with objects, and doodling, while low-SES was associated with engagement cues in the face such as laughing, head nods, and raised eyebrows (Kraus & Keltner, 2009). Bjornsdottir and Rule (2017) also found an effect in regards to people's faces and their social class. Across multiple studies, they examined participants' abilities to determine social class from photographs taken in a laboratory setting as well as facial images retrieved from a dating website. They found that people were able to accurately judge the social class of the person at a rate above chance. The ability to do this was unrelated to the evaluator's social class or their attitudes towards social classes in general (Bjornsdottir &

Rule, 2017). Interestingly, this suggests that this effect is not merely confined to experiments, nor can it be due to environmental indicators inherent in real-world images or selective differences in how the images were taken or presented.

Similar research has been applied to religious affiliation. For example, Rule et al. (2010) studied members of the Church of Latter Day Saints. They discovered that Mormons could be differentiated from non-Mormons simply by their physical appearance. This was true amongst both Mormon and non-Mormon judges.

There has also been a large literature examining more specific social traits, unrelated to group membership or social class. For example, Rezlescu et al. (2012) reported that people participating in a trust game were likely to value faces that appeared more trustworthy over faces that appeared less trustworthy. This trend held across negative reciprocating conditions, suggesting that, even when behavioral evidence indicated a lack of trust, people still perceived trustworthiness in faces as having indicative value (Rezlescu et al., 2012). Mueller and Mazur (1996) found that perceived dominance in facial features predicted people's success in military organizations. More dominant facial features were related to higher ranks and more promotions (Mueller & Mazur, 1996).

Evidence suggests that natural and spontaneous photographs might provide higher accuracy ratings in predicting personality characteristics. Naumann et al. (2009) photographed participants with constrained posture and expression and spontaneous facial expressions. Participants rated these photographs on 10 dimensions related to personality traits. Evaluators demonstrated lower accuracy at assessing personality traits when participants were photographed in a static position, only achieving significant accuracy in extraversion, self-esteem, and religiosity (Naumann et al., 2009). Contrastingly, the judges

were accurate for nearly all traits when assessing images from the spontaneous condition, meaning that when participants were allowed to 'be themselves', the characteristics of their personalities became more apparent (Naumann et al., 2009).

In their meta-analysis involving 47 journal articles, 131 independent effects, and over 6,000 judges, Tskhay and Rule (2013) demonstrated that observers could accurately perceive group membership in ambiguous photographs. For example, people were able to accurately classify people into categories of sexual orientation, religious affiliation, and political orientation at rates that were significantly better than chance. They estimated the total accuracy rate across all dimensions at 64.5% (in comparison to a chance rate of 50%).

From their meta-analysis, Tskhay and Rule (2013) documented positive effects in 92% of studies. The aggregate effect was positive, moderate-to-small, and statistically significant. They also investigated for null-result studies that were left unpublished, the so-called file-drawer problem. More than 20,000 null effect studies would be required to bring the aggregate results to a level approaching non-significance (Tskhay & Rule, 2013). For political orientation specifically, the researchers found a small but significant correlation between group categorization and group belonging (r = .18) (Tskhay & Rule, 2013).

These findings suggest that people are modestly adept at categorizing people into social groups with relatively little information about a target. It also suggests that people are able to discern specific traits about the individual at levels above chance.

Assessment of Personality Traits

People also have the ability to discern personality characteristics from facial features. Little and Perrett (2007) photographed 191 participants' faces after assessing their personalities on the Big Five. Separately for men and women, they took the facial

images from the 15 highest scorers and 15 lowest scorers in each of the five factors to isolate the common facial features among those high or low these personality dimensions. They then created a composite image of each group of 15, creating a set of ten composites. When images of individuals are merged in this way, facial characteristics held in common by individuals that are high or low in a certain trait will be maintained in the composite, while differences are averaged out (Boothroyd et al., 2008). Little and Perrett (2007) then asked 40 participants to rate the 10 composite images for both sexes, evaluating each of the 20 images for openness, conscientiousness, extraversion, agreeableness, and neuroticism. Female faces were rated accurately at better than chance levels in all categories with the exception of openness, while male faces were predicted at greater than chance levels in their extraversion (Little & Perrett, 2007). Biel et al. (2012) also found support for Big-Five trait accuracy in participant assessment of YouTube videos, in particular the extraversion dimension.

Carney et al. (2007) observed that people could accurately judge Big Five personality traits after relatively brief exposures to video clips. Interestingly, the researchers differentiated personality characteristics by emotional affect. They observed that accuracy in rating positive affect personality traits such as extraversion and agreeableness increased as the duration of the video increased (Carney et al., 2007). In contrast, additional evaluation time did not increase the accuracy for personality characteristics associated with negative effect, such as neuroticism, openness, and intelligence (Carney et al., 2007). In other words, people seem predisposed to quickly ascertain whether someone is exhibiting negative emotional affect, whereas correctly establishing positive affect quickly seems to be less important, evolutionarily.

Ambady and Rosenthal (1992) performed a meta-analysis on 44 research articles that examined evaluator accuracy on thin slices of expressive behavior in videos. They found that participants across studies achieved correct classifications in their assessments nearly 70% of the time. Further, they found that length of recording was unrelated to performance, with a 30-second clip being as informative as a 5-minute long clip (Ambady & Rosenthal, 1992). However, this effect might be variable, with some people having the ability to perceive personality traits better than others. Kenny et al. (1994) also performed a meta-analysis on evaluator consensus regarding perceptions of Big Five traits in others. Across their 32 studies, they found considerable variability, with evaluators' interpretations correlating with one another at a range of 0 to .3 (Kenny et al., 1994). These correlations did not change if the target was a stranger, short-term acquaintance, or long-term acquaintance (Kenny et al., 1994).

Cheung et al. (2010) used EEG to study whether extroverted people processed faces differently than introverted people. They utilized an inverted face paradigm, a commonly used technique to assess a person's ability to perceive faces. Human beings are typically better at recognizing faces than other objects when they are upright, but worse at recognizing them when they are inverted (Civile et al., 2014). The researchers found that extroverts and introverts showed no neurophysiological differences in their assessment of faces in the upright condition, but extroverts were better at recognizing faces in the inverted face condition (Cheung et al., 2010).

While it is unclear precisely how people are assessing personality traits from facial features, several studies have attempted to examine which facial features are indicative of personality traits. Wolffhechel et al. (2014) manipulated photographs of peoples' faces to

accentuate specific facial landmarks. For example, they modified facial images to have +/- 2 standard deviations of the average face width. They discovered that ratings of these faces across many personality dimensions correlated with their predictions based on a linear model. Facial width to height ratio was important in assessing dominance, while the shape of the mouth being neutral or curved slightly upward was associated with positive traits such as trustworthiness and intelligence (Wolffhechel et al., 2014).

Research on baby-facedness has demonstrated that people tend to associate certain facial features with specific personality characteristics. Baby-faced people tend to have faces with neonatal features, such as large eyes, raised and thin eyebrows, round faces, small chins, and small nose bridges (Berry & Brownlow, 1989; Zebrowitz & Montepare, 1992). People tend to associate those with so-called "baby faces" with warmth, approach-ability, and honesty, but also physical and social weakness as well as naiveté (Berry & Brownlow, 1989). Berry and McArthur (1985) discovered that 57% of the variance in ratings of baby-facedness were derived from eye size and chin width exclusively, suggesting that these facial characteristics are positively related to perceptions of honesty, kindness, and warmth. These same ratings could not be explained by either attractiveness or perceived age (Berry & McArthur, 1985).

There is also research investigating the relationship between facial masculinity and perceived personality traits. Kruger (2006) examined whether facial masculinity has an influence on reproductive strategy. The researcher discovered a two-factor model of mating based on the associations between perceived facial masculinity or femininity in male faces and the individuals' perceived personality traits. Masculine male faces were perceived as more likely to be aggressive, promiscuous, focused on short-term

relationships, and fun at parties, while feminine male faces were more likely to be perceived as a good husband, emotionally supportive, responsible at work, and good with children. Kruger (2006) surmises that masculine faces might be evolutionarily associated with riskier mating choices, which might explain why women prefer highly masculine faces for extra-relationship affairs and less masculine faces as marriage partners.

Todorov et al. (2008) describe the difficulty in identifying specific facial features that are associated with personality characteristics. The authors caution against studying trait inferences such as trustworthiness through perceptions of faces. They describe that trustworthiness can be explained by two other trait judgments, attractiveness and how caring a person is perceived as being. Indeed, according to the authors, these two characteristics account for 84% of the variance in trustworthiness activation in the amygdala as determined by fMRI studies (Todorov et al., 2008). As such, it is unclear whether perceptions of trustworthiness are influenced by perceptions of attractiveness and caring, or whether perceptions of attractiveness and caring are influenced by perceptions of trustworthiness. Regardless, it appears that perceptions of certain traits within a face might influence the perception of other traits, making it difficult to isolate specific facial features that are correlated with solely one trait.

These studies strongly suggest a general ability for humans to discern a variety of characteristics about others, including their sex, gender, age, attractiveness, personality traits, religiosity, sexual orientation, and socioeconomic status. But this is not the only domain where people have this ability. The next chapter will discuss differences in political temperament, their origins, and humans' ability to discern political ideology from faces.

Chapter 4: Political Orientation

To date, it is unclear as to whether the research Wang and Kosinski (2018) and Kosinski (2021) performed is successful at categorizing images by their facial morphology, i.e., the structure of the face, or due to some other information in the image such as demographic information or transient facial features. The proposed research project seeks to isolate facial morphology in categorizing individuals by their political position taking. As such, there should be evidence of several factors. First, we should establish that there are personality differences between people with different ideologies. This would strongly suggest that political ideology is influenced, at least in part, by interpretation and perspective, rather than derived solely from specific facts, media, or group affiliation. Second, because facial features are heritable, one should believe that some of the variability present in political ideology is also heritable. Third, there should be evidence that people can identify another's political orientation simply from their appearance.

Personality Differences across the Political Spectrum

Researchers examining the Big Five traits in relation to political ideology have found that liberals tend to demonstrate more openness in regard to new experiences (Gerber et al., 2011). For example, liberals are more willing than conservatives to try ethnic foods, attend concerts, and create art (Hibbing et al., 2013). Conservatives, on the other hand, tend to be more conscientious (Gerber et al., 2011). This might represent itself in a person being orderly, faithful, loyal, and patriotic (Hibbing et al., 2013).

Furnham et al. (2018) analyzed the survey data of participants who filled out two surveys: one related to financial affairs which gave data on political orientation and demographic variables, the other with personality measures. They discovered that

demographic information, such as age, gender, education, and social class, provided less predictive power in regards to political affiliation than personality characteristics (Furnham et al., 2018). They found liberal self-description to be associated with increases in extraversion, agreeableness, and openness to new experiences, while conservative ideology was associated with increases in conscientiousness, disagreeableness, introversion, stability, and being closed to new experiences (Furnham et al., 2018). Neuroticism was also associated with political affiliation but moderated by social class. As neuroticism increases, low social class people tend to become more right-leaning, while high social class people tend to become more left-leaning (Furnham et al., 2018).

These types of personality characteristics have been identified in children as young as age three and appear to be persistent over time. For example, Block and Block (2006) had nursery school instructors assess their students. They then evaluated those same children twenty years later when they were young adults. They discovered that a child's personality characteristics were highly associated with identifying as a liberal or conservative in the future. Future liberals were characterized as having close friendships, being self-reliant and energetic, resilient, and being somewhat dominating and relatively under-controlled (Block & Block, 2006). Future conservatives, in contrast, were more likely to feel victimized, more easily offended, more indecisive, fearful, rigid, and inhibited (Block & Block, 2006). IQ was unrelated to these factors, suggesting that political orientation is unrelated to intelligence (Block & Block, 2006).

Additionally, evidence suggests that liberals and conservatives come to their political conclusions in different ways. For example, Haidt and Joseph (2004) found common themes for the roots of morality present in all human cultures across the world.

The virtues of Harm, Fairness, Ingroup, Authority, and Purity were discovered to be these bases, derived from anthropological and evolutionary accounts of morality (Graham et al., 2009).

Moral Foundations Theory suggests that those who endorse themselves as strongly liberal or strongly conservative embrace their moral foundations in different ways (Graham et al., 2009). People with a strongly liberal political identity are more likely to generate their morality from harm and fairness than the other three foundations, while people with strongly conservative political identities endorse all five relatively equally, but are more influenced by ingroup, authority, and purity than liberals (Graham et al., 2009).

Thus, liberals and conservatives tend to have measurable personality differences. These differences are observable at a very young age and seem to be relatively stable across adulthood (Costa & McCrae, 1992; Costa & McCrae, 2002). Further, there is evidence to suggest that right- and left-leaning people tend to process their moral decisions in different ways. These findings suggest that political ideology is not merely derived from what books one reads or what news media one consumes, but rather is influenced by and associated with other facets of personality. While some might lament over the perception of an inescapable partisan divide, these differences could also be seen as the two necessary components to the formation of a civil society, with the moderate left representing innovation, creativity, and modernity, and the moderate right representing law and order, fortitude, and ties to tradition.

Genetic Origins of Political Orientations

If there are evolutionary differences across politically ideological lines, we must infer that these would be represented in genetic attributions of political ideologies. Multiple

studies analyzing twins have found that a substantial part of the variance explained in differences in political ideology can be attributed to hereditability. For example, Hatemi et al. (2014) analyzed data from over 12,000 pairs of twins that were collected from five different countries, sampled over four decades. Their results suggest that around 40% of the variance in political attitudes can be explained by genetics. A study from Eaves et al. (1999) examined the effects of heritability by gender, and found that genetic influence on conservatism explained 64.5% of the variance in males and 44.7% of the variance in females.

Funk et al. (2013) performed a similar analysis, but in addition to political ideology measures, they accounted for personality traits from the Big Five and Right-Wing Authoritarianism. They uncovered that all of these features were highly heritable (Funk et al., 2013). Alford et al. (2005) found that individuals' positions on specific social and political issues were three times more heritable than party affiliation. This is perhaps unsurprising, as party affiliation is often a loose proxy for political beliefs. Many individuals belonging to the same party might have different reasons for group membership, some of which are more related to genetic dispositions, while others are influenced more by environmental factors. Put another way, people adopting the same party affiliation often demonstrate more within-group variation than people espousing the same political position (Alford et al., 2005).

Similarly, Hatemi and McDermott (2012) aggregated results of twin and kinship studies and aggregated the results into 26 domains. "Political knowledge and sophistication" and "overall ideology" were the most related to genetics, explaining almost 60% of the variance (Hatemi & McDermott, 2012). Political party identification was the least

related to genetics, with over 95% of the variance explained by environmental factors (Hatemi & McDermott, 2012).

Kandler et al. (2012) examined the influence of genetics, culture, mating, and personality through the lens of political orientation. Using data from 1,992 twins, they found that political attitudes were transmitted genetically, rather than environmentally, from parents to their children. They discovered two dimensions to political orientation that are typically related to one another and that have a genetic basis: attitudes toward inequality and endorsement of system change.

Differences in Appearance by Political Affiliation

Social psychologists who have examined the relationship between appearance and political affiliation have demonstrated some surprising accuracy in initial categorizations. Samochowiec et al. (2010) showed across four experimental studies that people were accurate above chance in correctly identifying the political affiliation of a member of parliament whom they were otherwise unfamiliar with. These authors further demonstrated that they were more accurate in identifying those whose political affiliation was different from their own (Samochowiec et al., 2010).

Rule and Ambady (2010) demonstrated similar results in an American population. Further, they applied the same methodology to yearbook pictures of university students. Participants were able to accurately discern whether the college students affiliated themselves with the Democratic or Republican Party. Participants were also accurate across yearbook images, strongly suggesting that there is a general ability to infer political affiliation from faces, rather than certain faces being particularly revealing while others remained stubbornly mysterious (Rule & Ambady, 2010).

While findings such as these are well-established, it is less clear what types of features or behaviors are driving these accurate appearance-based identifications. A survey of the available research suggests it is likely a combination of factors. For example, Peterson et al. (2017) discovered a link between emotional expressivity in images and the political identification of the target, with liberals being more emotionally expressive. Carpinella and Johnson (2013) found that there were sex-typical differences in elected officials, with female Republicans being more sex-typical than female Democrats, and male Republicans being less sex-typical than male Democrats. Rule and Ambady (2010) discovered that ratings of Republicans and Democrats images differed on personality traits, with Democrats appearing more likeable and Republicans appearing more dominant. These differences might have real-world implications in regard to political decision-making, such as electing war-time or peace-time leader, with dominance potentially being a more desirable trait during times of war (Little, 2012).

Antonakis and Eubanks (2017) posit a self-fulfilling prophecy explanation regarding people's faces. If a person has facial features that are more associated with leadership, he or she might be reinforced for dominant, empowered, or extraverted behaviors. This might encourage the person to continue these behaviors, especially in environments where leadership is in high demand. Further, these facial features might encourage others to bestow trust upon the person, or even be deferent towards them. Similar reinforcement mechanisms might take place with people who have facial features associated with other characteristics, such as warmth, aggression, or introversion. We might thus assume that conservatives would favor traits traditionally considered to be indicative of biological fitness, namely facial attractiveness, health, and masculinity.

Evolutionarily, dominant and protective leadership would likely have demonstrated good facial symmetry as a proxy for fitness. Healthy leaders would be more likely to survive and assist in conflict situations, as well as in resource acquisition. Patriarchal societies would have preferred their leaders to demonstrate masculine features, suggestive of higher testosterone, aggression, and strength (Penton-Voak & Chen, 2004; Montoya et al., 2012).

Analyses of politicians across countries support this argument. Berggren et al. (2017) found that conservative politicians in Europe, Australia, and the United States are, on the whole, more attractive than liberal politicians. Bull and Hawkes (1982) also found that conservative politicians were more attractive than their more liberal opposition. Assessments of political attitudes across three decades show a clear trend of people associating liberal policies with femininity and conservative policies with masculinity (Winter, 2010). Subramanian and Perkins (2009) found that Republicans, in general, demonstrated greater health than both Democrats and Independents. Democrats, for example, were more likely to be in poor health and more likely to smoke than Republicans (Subramanian & Perkins, 2009).

Of particular evolutionary benefit might be ascertaining qualities about leadership from faces. Ballew and Todorov (2007) had participants assess photographs of political candidates. Participants were shown the faces of two candidates who had run against each other in past gubernatorial elections and asked to make a determination about which candidate appeared more competent for the position. The images were shown for 100 ms, 250 ms, or an unlimited amount of time. The researchers discovered that participants were able to successfully determine which candidate won the race simply by viewing the photographs of the two candidates. Participants were more successful at guessing the

winner when making snap judgments than they did upon lengthier deliberation. They also found that higher competence ratings for the winning candidate correlated with greater vote discrepancies for that candidate. In other words, higher ratings of perceived competence from photos translated into more votes at the polls. This finding explained over 7% of the variance in vote share for the candidate, a small but meaningful effect given that they had no information about the candidate or their policies (Ballew & Todorov, 2007). The authors replicated this effect in a prospective study, which included potential senators in addition to potential governors (Ballew & Todorov, 2007).

One potential shortcoming of the Ballew and Todorov (2007) study is that the authors did not measure the actual competence of the candidates; instead, they asked participants to choose the more competent candidate and used being elected as a proxy for the competence of the candidate. In other words, we cannot be sure that the candidates chosen as competent in the Ballew and Todorov (2007) study actually are competent, rather than just being perceived that way. In order to determine if competence itself could be predicted from facial appearance, Rule and Ambady (2008) took the top and bottom 25 companies from the Fortune 500 list of companies and obtained images of their respective CEOs from their company websites. Participants were asked to rate the faces on a variety of aspects, including leadership, competence, dominance, facial maturity, likeability, and trustworthiness. They found that the CEO's 'power' rating, a composite score of the competence, dominance, and facial maturity dimensions, accurately predicted company profits. In other words, participants' perceptions of competence were correlated with competent CEO performance. While not explicitly about political qualities, the Rule and Ambady (2008) study demonstrates ecological validity in assessing leadership from faces.

Little et al. (2012) also added an ecological element. The researchers had participants evaluate the faces of potential leaders in the context of peacetime and wartime. They found that people, in general, preferred attractive leaders during wartime but preferred trustworthy looking leaders in peacetime (Little et al., 2012). The authors surmised that these characteristics might be "independently valued traits in leader choice" and that attractiveness is likely to be valued because it is indicative of fitness and health (Little et al., 2012, p. 2031; see also Little et al., 2007).

Finally, because liberals have a preference for more complex and abstract thinking (Eidelman et al., 2012), we might imagine that they would select their leaders less by how they look and more by the way they think. Antonakis and Eubanks (2017) found that people use whatever information they have about a leader to evaluate them. People who have little information about a leader will judge their competence and character from their facial features, while people who have more information about the individual will be less swayed by facial cues. Because liberals have been shown to demonstrate more abstract thinking, it follows that liberals will be less swayed by physically attractive candidates than their conservative counterparts (Antonakis & Eubanks, 2017). An alternative explanation is posited by Brown et al. (2021) in which they had participants rate photographs of physically strong and physically weak men for political conservatism. They discovered that people tend to use physical strength as a proxy for conservative ideology, labeling the stronger men as more conservative, on average, while weak men were perceived as neither liberal nor conservative. These differences in assessing leadership qualities from physical appearance might help explain the attractiveness difference in liberal and conservative leadership.

There is some evidence to suggest that this difference in attractiveness might be due, in some part, to grooming practices. Lönnqvist (2017) studied right-leaning and left-leaning scholars to see if attractiveness was related to success in academia. Further, the scholars were rated for how well-dressed or well-groomed they were. He found that attractiveness was related to better grooming in right-leaning scholars but unrelated to left-leaning scholars (Lönnqvist, 2017). Other studies have shown that SES might be related to an increase in grooming behaviors in the form of disengagement cues (Kraus & Keltner, 2009).

This is a very brief review of the extensive literature on categorization accuracy from facial analysis in regards to social traits, personality features, and political ideology. While far from comprehensive, it serves to demonstrate the size and scope of this type of research.

However, the research described previously has several aspects which could be improved upon. Aside from utilizing more modern techniques and applying them to a field where they have not been used before, there are several additional benefits of this proposal that could contribute to the field of facial analysis in the political arena.

First, there are benefits associated with employing computer vision within this paradigm. For example, a neural network will process faces into numerical features, whereas human beings tend to process faces holistically (Tanaka & Farah, 1993). As a result, assessments from human beings in facial analysis studies are subjective and error-prone. Utilizing a neural network makes analyzing images objective. Because this methodology is not typically used, it has the potential to inform our understanding of facial feature analysis more deeply.

Second, this work can be replicated exactly. A researcher working in this domain might attempt design replications, but the images and techniques used might be different from the original studies. Even given the original materials to work with, they would still require novel human judges to assess the targets. A benefit of employing this methodology is that, given the same data and the same method, the output should be precisely the same, allowing for greater validation of the results achieved.

A closely-related third benefit is that the rater of the images in this proposed study should not demonstrate any unaccounted for bias. For example, the neural network does not get tired or lose focus. It does not get hungry or grumpy. The neural network does not inherently prefer blue eyes or long hair, a cleft chin or a broad nose. While these features might potentially be indicative of some type of trend, any bias within the machine is due to bias in the data rather than externally from the individual preferences or disposition of the rater.

Finally, despite the evidence presented here and elsewhere, some researchers believe that humans might not be particularly great at assessing trait characteristics from faces. For example, while accuracy rates in meta-analyses are often significantly above chance levels, the overall effect might be rather weak (Carpinella & Johnson, 2013; Olivola et al., 2014), making it difficult to investigate. Some researchers have found that overreliance on facial characteristics might actually undermine accurate assessments, because additional, more valid information might be ignored (Todorov et al., 2015). The current line of research could help illuminate how much an observer should be relying on facial features.

Chapter 5: Neural Networks and Computer Vision

Background

Neural networks were first proposed in the mid-1940s by Warren McCullough and Walter Pitts, two researchers from the University of Chicago (Hardesty, 2017). The first rudimentary neural networks were used experimentally to differentiate simple shapes such as circles and squares, but by the 1970s they were being used commercially in optical character recognition, reading written or typed text for the blind (SAS, n.d.). These neural networks were time-consuming to train and expensive to maintain (Babich, 2020).

Although neural networks would fluctuate in their usefulness over the next several decades, they found their biggest resurgence in the first and second decades of the 21st century. This is mainly due to the availability of powerful and dedicated graphics cards, the access to large repositories of images made available by widespread adoption of the internet, and the advent of cell phone camera technology that increased the facility of individuals to take and share photographs (Hardesty, 2017; SAS, n.d.; Babich, 2020).

Inspired by the biological development of the human brain, neural networks consist of thousands or millions of nodes that are connected to one another across layers (Read et al., 2017; Zou et al., 2008). Early neural networks had one layer of nodes, a "shallow" network of artificial neurons used to analyze complex data through simple formulas that were iteratively updated to perform tasks better over time (Schmidhuber, 2015). Today's neural networks have multiple layers of nodes, meaning they are "deeper" than their ancestors from decades earlier (Schmidhuber, 2015). To put this in context, VGGFace, a neural network employed in this project, has 16 layers and approximately 140 million parameters (Shaikh, 2017). Each node receives data from prior layers of the network and

sends data to the next layer.

Training/Testing

To create a neural network, you must train it with data; lots of it. Most neural networks are 'supervised' learners, meaning that the data being analyzed has been identified and labeled by a human being. In the training phase of a neural network specialized in computer vision, a majority of the labeled photographs would be translated individually by the neural network into an array of values. The first layer of nodes will represent edges at certain locations in the image, the second layer will recognize patterns in those edges, the third would recognize larger motifs, and so on, with each successive layer manipulating the data from the previous (LeCun et al., 2015). By the end of the training process, the numerical output from the photographs will be analyzed for differences in relation to the categories of images.

The neural network will then test their predictions on the unidentified images, the remainder of the labeled dataset, with the identifications hidden to the neural network. The output for these "test" images will resemble a propensity score between one and zero. This is the neural network's best guess about whether the image belongs to one category or the other. Based on how well it performs, some connections between the nodes are weighted to give them more importance, while others are pruned because they are less relevant to the task at hand (Read et al., 2017). In this way, the neural network can improve iteratively over time.

However, researchers and scholars have been relatively slow to adopt neural network and computer vision methodology into social science research. This is likely for several reasons. First, neural networks are typically evaluated by the accuracy of their

predictions, rather than the computations by which they came to their conclusions. This is because neural networks are so mathematically complex, deciphering the process of *how* they make their predictions is incredibly difficult. For this reason, neural networks are typically considered "black box" systems, where the internal mathematics behind the outputs are largely unknown or beyond interpretation. This could present a challenge in describing the results of social experiments. Second, utilizing neural networks is more technically challenging to implement in comparison to simpler, more commonly used methods like ANOVA or linear regression. Because the barrier to entry is relatively high, researchers need to have a fair amount of technological savvy to employ these techniques. Third, morphological effects that differentiate participants might be small, requiring a massive amount of data, much larger than sample sizes routinely utilized by social scientists. Fourth, utilizing computer vision in research requires a specific type of data (images), rather than the more typical quantitative or qualitative data.

Chapter 6: Neural Network and Computer Vision Research in Psychology

Despite these challenges, there is a small but burgeoning line of research in the social sciences that employ computer vision and neural networks in ways similar to those proposed here. Using a neural network/machine learning approach, Segalin, Celli and colleagues (2017) examined the profile pictures of 11,736 Facebook users in order to see if inferences could be made about their personalities. Their technique was especially interesting as the researchers used both a convolutional neural network to extract features from the image, but also used additional neural networks to glean properties of the images themselves. For example, they used "computational aesthetics" based features, features that describe low level information such as color range, relative lightness of the pixels, and amount of color present in the image. They discovered that the Big Five personality traits of the person often demonstrated slight correlations with properties present in the images. For example, extraverts tended to be in a group of people in their profile images, as well as having brighter photos. Additionally, there were color effects present in the images, with extraverts being statistically more likely to have the colors pink, purple, red, and/or yellow in their images. The researchers then tested the prediction ability of their model on images that were mean split (divided at the mean) and split at the quartile (lowest 25% in score versus highest 25%). Their model achieved modest success with an accuracy rate of about 55% across all five traits in the mean split condition and 60% accuracy in the quartile split condition (Segalin, Celli et al., 2017). This is perhaps unsurprising, as personality traits had previously been shown to affect profile picture choice (Wu et al., 2015). Further, observers of Facebook profiles have demonstrated accuracy in their estimation of the owner's personality characteristics (Hall et al., 2014).

Similar work has been performed, not on images of the participant themselves, but by images favored by the participant. By analyzing photograph preference as indicated by 'likes', researchers have uncovered patterns on sites like Flickr and Instagram that illuminate personality traits among the users (Segalin, Cheng, et al., 2017; Ferwerda et al., 2016). For example, users who are high in openness tend to prefer pictures with high saturation and vivid colors, and tend to share pictures that are low in brightness (Ferwerda et al., 2016).

In a small sample of facial photographs, Zhang et al. (2017) were able to create a neural network that was accurate at predicting some personality features. Using 186 photos, they were able to predict the personality characteristics of "Rule-consciousness", "Vigilance", and "Tension" from facial images. While their overall results were somewhat mixed (potentially due to sample size), their results suggest that neural networks can be trained to reliably assess personality characteristics purely from facial images.

Leuner (2019), in a Master's thesis, attempted a replication of the Wang and Kosinski (2018) study. Using a novel dataset of nearly 21,000 images, he was able to replicate with some success the prior study's findings. Using VGGFace, he extracted features from these photographs and predicted sexual orientation with 68% accuracy for males and 77% accuracy for females. Curiously, Leuner's model was more successful with females, whereas the opposite was true for Wang and Kosinski (2018).

Leuner (2019) used a facial morphology classifier, which only accounts for facial structure. Using the Face++ algorithm, he identified 83 landmark points on each face in his dataset. He then generated Euclidian distance measurements between the points and scaled them to be proportionate to the face by dividing these numbers by the distance

between the eyes in the image. It should be noted that this method eliminates all extraneous information from the image except for facial structure. This model, too was successful, achieving accuracy rates of 68% for males and 81% for females, with three images per subject (Leuner, 2019). The author was also able to estimate model success using just the landmarks of specific feature sets, namely the eyes, eyebrows, contour of the face, the mouth, and the nose. He discovered that eyes and eyebrows are the features that are most predictive for males, while the nose has no predictive value at all (Leuner, 2019). For females, the eyes were most predictive and the facial contour was least predictive (Leuner, 2019).

Finally, he utilized a blurring procedure to eliminate virtually all data in each of the photographs save for a dominant color. The model was still able to classify images correctly at a rate of 63% for males and 72% for females (Leuner, 2019).

In an attempt to apply his previous success to a different subject, Kosinski (2021) examined over one million images to determine if he could accurately categorize people by their political orientation, rather than sexual orientation, using only images of their faces. To address critics, he also attempted to discern between stable facial features (the morphology of the face), transient facial features (such as sunglasses, head orientation, or facial expression), and demographic traits that can be inferred from the image (age, gender, ethnicity) in the model. He found that some transient features offered little predictive value. This included things such as wearing glasses or sunglasses and adopting facial hair (Kosinski, 2021). In contrast, head pose and facial expression were related to political affiliation; however, the facial recognition algorithm had a much larger overall effect than what could be attributed to just the transient features of the face, suggesting that facial

structure itself had a strong impact on the success of the model (Kosinski, 2021). Accuracy of the models was high, reaching a high of 73% accuracy when utilizing the entire sample and 71% when controlling for demographics (Kosinski, 2021).

Wang (2022) attempted to replicate the findings of Wang and Kosinski (2018) while examining the differences in how heterosexuals and homosexuals present themselves in images. Utilizing images from a U.S. dating website, Wang (2022) was able to replicate the original effect, differentiating between heterosexuals and homosexuals with an accuracy of 65% for women and 61% for men. He discovered that, in his sample, images did differ systematically by group, with homosexuals being more likely to wear glasses in their images. He also found that gay men specifically had brighter images in comparison to heterosexual men.

The author was also curious about the effect that the background of the image had on the classifier. He utilized a masking technique to cover the face in the image, gradually increasing the size of the mask to encompass more and more of the image. Even with no facial information at all, the classifier was still able to determine the sexual orientation of the person in the image at rates better than chance.

Critiques of Wang and Kosinski – Background and Methodology

There has been a fair amount of controversy around Wang and Kosinski (2018), as well as the employment of neural networks and computer vision in general. Kosinski (2021) sought to address some of this criticism; however, because the author failed to isolate for facial morphology, many of the criticisms leveled against Wang and Kosinski (2018) still apply to the follow-up research. This section serves to acknowledge these sources of criticism and to address some of the more common critiques with these types of

studies and with this type of technology.

While the results of the Wang and Kosinski (2018) study are intriguing, the explanation for the success of their model is controversial. The authors rely on prenatal hormone theory (PHT) in describing the differences in facial features between heterosexuals and homosexuals. PHT suggests that an important factor in the differentiation of biological sex is due to the introduction of androgenic hormones during fetal gestation, and that the availability or absence of these hormones might be influential in sexual orientation (Ellis & Ames, 1987). The availability of androgenic hormones might, in turn, cause differentiation in the facial morphology of individuals, with homosexual men demonstrating more feminine facial features and homosexual women demonstrating more masculine facial features.

Although a deep dive into this literature is beyond the scope of this research, PHT is not without its criticisms. For example, much of the research performed on this subject matter is restricted to animals, which might not adequately translate to human beings which are decidedly more complex (Breedlove, 2017). Perhaps an objective estimation of the current state of literature on this topic would assert that there is an abundance of research showing biological differences between homosexuals and heterosexuals, but that much of it lacks generalizability or is contradictory with other evidence. In order to not become embroiled in these debates, it is possible that Wang and Kosinski (2018) suggested a lone biological theory that might lend credibility to their hypotheses, although that theory is, in itself, insufficient in explanatory power.

The authors also acknowledge that transient feature differences, such as the presence or absence of facial hair in the subjects' self-portrayal, may have influenced the

model. They attribute these differences to "androgenic hair growth, grooming style, or both" (Wang & Kosinski, 2018, p. 251). Other differences observed included the presence or absence of eye makeup, the presence or absence of caps or hats, the darkness of skin tone, and the facial expression of the individual. Whereas some of these shortcomings were addressed by future works (Leuner, 2019; Kosinski, 2021), there remains a great deal of ambiguity in regards to what information these models are utilizing in making their predictions.

Despite this ambiguity, the authors often relate these differences to genetic points of origin, with little confirmatory evidence. While potentially a plausible explanation for the results of the experiment, the authors overstate their case in regards to genetic determinism. It could be, for example, that a person's facial features influence how they are treated by others. Men with softer facial features might be tacitly and unconsciously encouraged by observers to act in ways that highlight their femininity, for example.

Agüera y Arcas et al. (2018) criticize the Wang and Kosinski (2018) study. They surveyed 8,000 Americans with simple "yes" or "no" questions, such as "Do you wear glasses?", "Do you have a beard?", and "Do you wear eye makeup?". In other words, they attempted to capture information via survey that might have been gleaned by a neural network analyzing photographs. They discovered that they could achieve rates of accuracy similar to Wang and Kosinski by including just a few of these questions in a simple linear classifier (Agüera y Arcas et al., 2018).

The comparison by Agüera y Arcas et al. (2018) is not completely analogous. One might point out that there is a large difference between the item "Do you ever use makeup?" and a photograph that has a person either wearing makeup or not. Specifically, the former

includes a temporal certitude that the latter does not. Regardless, the authors' underlying point that the neural network is likely gleaning some of its predictive ability from phenotypic differences present in photographs is important, likely, and worthy of consideration.

The authors propose that the facial pose of the person might also be indicative of sexual preference. They argue that the comparison images in the Wang and Kosinski (2018) study demonstrate wider nostrils and flatter eyebrows for heterosexual males and homosexual females than the others (see Appendix B). They posit that the features extracted might indicate a difference in facial pose within the photograph, rather than structural cues from the faces themselves.

Indeed, research on 'selfies' suggests that people tend to portray themselves in ways that are flattering to their potential mates, specifically in regards to the vertical orientation of the photograph (Sedgewick et al., 2017). Heterosexual women tend to shoot their 'selfies' from above, while heterosexual men tend to shoot more from below. It is likely that these behaviors are intended to illustrate, consciously or unconsciously, the difference in height between men and women, on average. However, there is some ambiguity, as males and females have evolved sexually dimorphic facial features that are amplified by these facial poses (Burke & Sulikowski, 2010). For example, males typically have larger jaws than females, while females typically have larger eyes than males (Burke & Sulikowski, 2010).

As readers, we are left with the somewhat opaque conclusion that the successful analysis presented in Wang and Kosinski (2018) is due to the difference in facial features present in the photographs, whether those features are transient or related to facial

morphology. It seems plausible that the authors overstated the influence of genetics in the success of their model, while underestimating the influence of environmental cues present in the photographs.

However, since the publication of Wang and Kosinski (2018), both Leuner (2019) and Kosinski (2021) worked to reduce the potential for transient factors to have influenced their models. Leuner (2019) used a facial morphology classifier that eliminated all extraneous information from the image, including only facial points and the distances between them. This model, using only facial morphology features, still accurately predicted sexual orientation at levels above chance. Leuner (2019) also examined head pose, and found no evidence that head pose was correlated with sexual orientation. Kosinski (2021) took into account demographic information, transient facial features, facial expression, and head pose. After controlling for these variables, the model still predicted political orientation at levels well above chance (Kosinski, 2021).

Despite these advances, the relationship between facial morphology and political ideology has not yet been fully explored. Leuner (2019) utilized facial morphology classifiers for sexual orientation, while Kosinski (2021) controlled for some transient features but did not assess facial morphology exclusively. This leaves open an exciting opportunity to provide some much needed clarity around this type of research broadly as well as how it relates to this population specifically.

Popular media sources were quick to criticize the Wang and Kosinski (2018) article. The New York Times described the study as raising "knotty questions about perceptions of sexual orientation" (Murphy, 2017). The Guardian compared the research to the "science-fiction movie *Minority Report*" and lamented about the day when "people can be

arrested based solely on the prediction that they will commit a crime" (Levin, 2017). A writer for Vox believed that the results were "the first stone on a path to a *Black Mirror* future" (Resnick, 2018).

Activist groups, too, were unhappy. To take just one example, the Gay and Lesbian Alliance against Defamation (GLAAD) had the somewhat confusing reaction of describing the research as a "weapon to harm… gay and lesbian people" while simultaneously admonishing the study's authors for not being more inclusive in their participant selection (Anderson, 2017). This appears to be paradoxical in several respects. First, greater inclusivity would seem to mean that more participant images should be included in the study, meaning that more people would be participating in a study that the critics believe is harmful. Second, more inclusivity in the model would likely serve to make the model better at predicting sexual orientation, rather than worse. For these reasons, it is curious to condemn the authors for not being more inclusive, as increased inclusivity would hypothetically improve model quality and make the model better suited to perform the task that GLAAD is outraged over.

To these critics, governments utilizing this technology as a means to identify whom to persecute represents a real threat to liberal democracies worldwide. However, to put these criticisms in the proper context, it is worthwhile to discuss the severe limitations of this type of technology. Perhaps most importantly, the methodology present in this type of study does not necessarily generalize well to the real world, which the New York Times article correctly observed.

"Let's say 5 percent of the population is gay, or 50 of every 1,000 people.

A facial scan that is 91 percent accurate [in a 50/50 classification paradigm]

> would misidentify 9 percent of straight people as gay; in the example above, that's 85 people. The software would also mistake 9 percent of gay people as straight people. The result: of 130 people the facial scan identified as gay, 85 actually would be straight" (Murphy, 2017).

In other words, because the underlying base rate of homosexuality is rather low, an applied effort to detect it under these precise conditions would result in an error rate approaching two-thirds. Although the Wang and Kosinski (2018) article touted a 91% accuracy rating for a specific subset of their sample, it is important to realize that in real-world conditions, one could achieve a 95.5% accuracy rate by merely identifying every target as heterosexual (Newport, 2018).

Further, translating this type of research to "the wild" (as it is referred to in machine learning parlance) would almost certainly result in greatly reduced accuracy. In spite of their erroneous conclusion, GLAAD observed that the sample for Wang and Kosinski (2018) was quite limited in scope (Anderson, 2017). Putting aside the issues of inclusivity pertaining to alternative identity groups, the samples for Wang and Kosinski (2018) and Kosinski (2021) were derived from dating site images and Facebook images. This presents two potential problems. First, because these images were not collected in an effort to measure political orientation specifically, the measure of political ideology associated with these images undoubtedly contains a lot of error. People on dating sites could be motivated to alter their true political ideology in order to cater to prospective mates. In other words, people might exaggerate their political orientation as being more extreme in order to appeal to certain types of people. Conversely, polarized believers might downplay their political ideology to appeal to a broader group of potential partners.

Further, we know that political orientation is somewhat complex to measure. For example, Stenner (2009) describes three different types of "conservatism": one being an inclination to favor tradition, "status quo conservatism", one favoring free market economies, "laissez-faire conservatism", and one favoring obedience and conformity, or "social conservatism". These subtle distinctions are completely absent in a unidimensional measure on the liberal-conservative spectrum.

Additionally, unidimensional, self-report measurements of political orientation have been shown to have severe limitations. Bauer et al. (2017) examined the left-right scale in regards to its validity. They found that people anchor the endpoints of such scales differently, with some people viewing the term "left" as representing the left-leaning party in their country, while others view it as something much more extreme, like a belief in communism, for example. Although the images in Kosinski (2021) are associated with a measure indicating left-right political orientation, this tells us very little about the specific political position taking of any individual member. It is entirely possible that two people at the opposite ends of a self-report measure on the left-right continuum could nevertheless align on views for any specific political position. Due to these facts, we might suppose the political ideology measure utilized by Kosinski (2021) carries a fair amount of error in its measurement. Retrieving images from followers of politically motivated twitter groups will better isolate on a specific political belief, thus ensuring that the two comparison groups are actually oppositional in regards to the political belief in question.

As previously mentioned, Wang and Kosinski (2018) and Kosinski (2021) failed to isolate for facial morphology. We cannot be sure that that their models would generalize well to an equivalently restricted sample of individuals across other contexts. Because the

information that the model might be relying upon may be based on features unrelated to facial morphology, we cannot be sure that these unrelated factors would remain consistent when applied to a novel sample of images. For example, because Kosinski did not control for color propensity in the image, one cannot be sure that the classifier is not relying upon color dominance in the image to indicate group belonging. We also cannot be sure that the indicator, color dominance, would yield the same effectiveness on a different sample of images.

Additionally, it is reasonable to suspect that any sample of images used to produce any sufficiently complex classification algorithm would have similar shortcomings in their attempts at broader generalizability, including the ones proposed in this document. Because the transient facial features in Kosinski (2021) explained a great deal of the variance, one might presume that there is simply not a great deal of information provided by facial morphology itself. Although it might be enough to make somewhat accurate predictions in the aggregate, one would expect any models entirely reliant upon facial structure when making group membership predictions to contain sufficient error that when selecting for any lone individual, the accuracy rate would be low enough so as to be unreliable for pragmatic purposes.

There are also easier ways for oppressive regimes to obtain this type of information. For example, many countries, including the world's most populous ones, heavily monitor the internet traffic of their citizens (Mitchell & Diamond, 2018). Kosinski et al. (2013) showed that a logistic regression model could correctly classify individuals merely from their Facebook "likes". From these self-reported likes, the authors could successfully predict race (white or black) with 95% accuracy, gender with 93% accuracy,

homosexuality with 88% and 75% accuracy for men and women respectively, and political orientation with 85% accuracy. Less accurate, but still well above chance, were predictions for relationship status, drug use, and religion. The same could be said of other social networks, internet search engines, and online purchases, all of which were available before the widespread employment of neural networks (e.g., Hill, 2012). The amount of data that is tacitly given can, predictably, reveal a great deal about the personal characteristics of any individual (see Vinciarelli & Mohammadi, 2014 for a review). At the same time, the computational and technological hurdle is much lower for analyses of this type, with equivalent or superior prediction rates. Organizations that might one day hope to utilize facial feature analysis to derive information regarding a person's sexual orientation, political ideology, or personality characteristics, likely already have superior means for identifying members of groups they wish to oppress, that likely would be more accurate as well as less computationally taxing.

Furthermore, we should expect that relying purely on facial morphology features will reduce model accuracy substantially. For example, accuracy dropped between two and five percentage points when Kosinski (2021) controlled for demographic information. Further, facial expression accounted for 59% of the variance in model accuracy, suggesting that controlling for these factors might also substantially reduce model accuracy as well as suggesting that these factors, rather than facial morphology, were responsible for the success in model categorization. This implies that ecological clues within the images are providing at least some of the information relied upon to make the predictions. Models categorizing images for these transient and demographic cues will be easier to develop as well as more accurate in categorization in comparison to those detecting minor differences

in facial structure. At the same time, features that are unrelated to facial morphology might demonstrate greater variability in regards to external validity. In other words, features that aided in categorization in Wang and Kosinski (2018) or Kosinski (2021) such as head pose or facial expression might be specific to the sample used and unrelated or even inversely related in a novel sample.

Finally, it should be noted that any government that is attempting to oppress its own people needs no excuse to carry out its nefarious acts. Government oppression existed long before neural networks, computer vision, the internet, or computers. The idea that facial feature analysis provides some air of legitimacy to that oppression is irrelevant. As Agüera y Arcas et al. (2018) correctly point out, the Stasi in East Germany oppressed their people with "nothing but paper files and audiotapes". The problem, in other words, is the abuse of power, rather than the means by which that power is abused.

Chapter 7: Ethical Considerations of Widespread Adoption of Facial Analysis

Despite these reassurances, facial analysis through computer vision more broadly has sometimes resulted in breaches of ethics. The ethical issues concerning facial recognition technologies largely fall into two camps; that of ignorance and that of overreach. A model that is poor, poorly applied, or poorly interpreted, is an example of ignorance. A model that is accurate but that is used in an inappropriate way is an example of overreach.

Amazon's "Rekognition" has been leased by county and city governments, who have attempted to use the program as a tool for law enforcement. Initially, Amazon spokespeople suggested that a confidence rating of 85% was sufficient to indicate a match for the software (ACLU, 2018). Using Rekognition under those specifications, the ACLU was able to match 28 members of Congress to a mugshot database, illustrating the dangerous precedent of using facial recognition software when assessing guilt and innocence as well as the potential danger and prevalence of mistaken identifications (Snow, 2018). Since the ACLU's investigation, Amazon spokespeople have changed their recommendation to 95%, before finally asserting a 99% confidence is necessary to ensure an accurate match (ACLU, 2018).

Adding fuel to the fire, a study published by MIT and University of Toronto researchers demonstrated a racial and gender bias within the program, determining that darker-skinned females were more prone to being misclassified than white males (Kelion, 2019). Amazon has demonstrated previous problems with bias in their machine learning algorithms, specifically in regards to their automated hiring process, which learned to filter out resumes with female indicators (Goodman, 2018).

These are examples of ignorance regarding the application of this technology. By setting an artificially low confidence level, the models were over-inclusive, and produced Type I errors. The racial and gender bias demonstrated by their program strongly suggests that the quality of their training data differed across representation of gender and race. Supporters of Rekognition seemed to be so eager to implement a program that their model was poorly conceived, their data underdeveloped, and their outcome unreliable.

At least one potential city government has since discontinued their use of Amazon Rekognition (Roulette, 2019). In June, 2019, Amazon Web Services's (AWS) then CEO Andy Jassy expressed some concern regarding the misuse of facial recognition software, and even called for federal regulations on facial analysis technology (Hellman, 2019). Both Microsoft and Google representatives have previously called for similar regulations (Smith, 2018; Hellman, 2019). As of 2022, at least 17 municipalities have administered local bans on the government use of facial recognition technologies, including San Francisco, New Orleans, Minneapolis, and Boston (Sheard & Schwartz, 2022).

There are additional examples. Other private companies such as Faception or Terrogence are also developing facial recognition software for biometric security applications (Bendel, 2018; Stanley, 2018). While Amazon might be constrained by shareholder or consumer pressure, these companies are largely unaccountable, their company processes are opaque, and they are not beholden to the public (who are not their primary customers). As such, it is unclear how their analysis is performed, where their data comes from, or what quality control has been imposed in order to prevent inaccurate assessments from being made.

The Immigration and Customs Enforcement (ICE) agency has been caught mining

state driver's license databases in order to obtain information about undocumented immigrants, potentially violating existing privacy laws and undermining state sovereignty (Edmondson, 2019). Critics have denounced the use of facial recognition technology by governments and private industries, arguing that peoples' civil liberties could be violated and condemning the lack of transparency, two dramatic examples of overreach (Stanley, 2018).

Perhaps unsurprisingly, many are calling for the development of ethical guidelines for the use of facial analysis and recognition. For example, the London Policing Ethics Panel (LPEP) has recommended conditions the Metropolitan Police must adhere to in order to adopt facial recognition technology (Government Europa, 2019). Among other requirements, these conditions insist that no racial or gender bias be present in the software, and that "the benefits afforded to public safety by deploying the software must objectively outweigh levels of potential public distrust of the technology" (Government Europa, 2019).

Microsoft, in their "six pillars of ethical uses" for facial recognition technology, reiterates this sentiment (Spirina, 2019). They focus on fairness, reliability and safety, privacy and security, inclusiveness, transparency, and accountability (Spirina, 2019). Similarly, the ACLU delivered an Ethical Framework for Facial Recognition to the U.S. Department of Commerce (ACLU, 2014; Martin, 2014). In their memorandum, they specified that consent must be obtained by an individual before including their image in a facial recognition database and that the individual has the right to delete their information from that database at any time (ACLU, 2014). They specify that social networks should take all available action to prevent others from creating a "faceprint" database, meaning a database that will be used for recognition and identification purposes (ACLU, 2014).

Perhaps the moral foundation that is most apropos to the proposed line of research are the guidelines put forth in the Belmont Report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979). The Belmont Report was a response to the injustices perpetrated by the medical and psychological communities on the public, specifically the medical atrocities committed by the Nazis, the Tuskegee Study, and the Milgram study (Paxton, 2020). The Belmont Report described three principles to follow during the research process: respect for persons, beneficence, and justice (Belmont, 1979).

Wang and Kosinski (2018) describe reticence at publishing their findings, because governments hostile to homosexuals might attempt to use such facial analysis as an avenue for persecuting their own citizens. They decided to proceed in the hopes that they would spread awareness about the issue to the public and policymakers, bringing the issue to light so that it could not be used in the dark (Wang & Kosinski, 2018).

This demonstrates an apparent contradiction in this line of research. The researchers studying facial analysis and lauding its efficacy are quick to assert the ethical dangers surrounding the topic. The companies providing the software to law enforcement organizations are concurrently calling for its regulation at the federal level. While some county and city governments seem eager to utilize the software, other cities and even states are excoriating its use and either banning, or considering banning, the practice.

It is not hard for an imaginative person to contemplate a dystopian future where privacy is dead and greedy corporations and oppressive governments collaborate to maintain tight control of a helpless and docile populous. However, evidence suggests there is reason to be both cautious and optimistic.

First, in spite of Rekognition and the other opportunities for malfeasance, facial analysis could actually have a positive, practical impact in a law enforcement capacity if used correctly and conservatively. For example, this type of software could excel at finding missing children (Smith, 2018). In such a scenario, database inclusion would likely not be related to privacy concerns and false positives would have minimal detrimental effects.

Second, Amazon has tempered their stance on facial recognition software substantially, and much of that change has been in reaction to public sentiment. Additionally, Google and Microsoft are both encouraging federal regulation. This suggests that consumers, activist groups, and shareholders wield some power over these larger industries (ACLU, 2018).

Third, cities and states are considering exercising their ability to ban the use of facial recognition software for law enforcement and other government agencies. It appears that some governments are listening to public concern and responding with decisive action. We might anticipate a more healthy and vibrant debate on this subject from policy makers in the near future.

Fourth, multiple frameworks already exist for ethical regulations provided by the ACLU, the Belmont Report, and other sources. Applying these guidelines to emerging technologies might not prove as difficult as one might initially imagine upon cursory inspection.

Finally, researchers seem to recognize and acknowledge the gravity of the subject matter. Although facial analysis is on the razor's edge of technology, research on the subject has often been tentative, thoughtful, and a great deal more circumspect than one might predict, although with occasional exceptions.

Chapter 8: Method

Four large samples of images were drawn from Twitter users using the Twitter application programming interface (API). An API is a software intermediary which allows users and developers to make requests for data from organizations. For example, online travel services use API's to pull airline company data, allowing customers to comparison shop for flights based on price, destination, times, dates, and number of connections. Many organizations, such as Reddit, Twitter, and the U.S. Government, utilize APIs to allow end-users access to large sources of data (Chen & Wojcik, 2016).

Using the Twitter API, nearly two million usernames were harvested from followers of partisan organizations related to gun control and immigration. To see the specific organizations represented in tabular form, see Appendix C. These topics were selected for two reasons. First, they are issues on which the American public demonstrates sharp divisions (Oliphant, 2017; Reinhart, 2018; Daniller, 2019). In a recent article highlighting the widest partisan gaps in the United States by polling, both immigration and gun control ranked among the issues most polarizing for American citizens (Johnson, 2020). For example, a recent Gallup poll found that 52% of respondents wanted stricter gun laws, whereas 46% of the sample wanted gun laws that were less strict or kept the same as they are now (Gallup, n.d.). Similarly, the polling organization found that 33% of respondents believe immigration should increase, 31% believe it should decrease, and 35% believe it should remain at its current level (Jones, 2021).

Second, because political organizations such as the ones listed in Appendix C typically support a narrowly focused agenda, we can expect that followers of such an organization will overwhelmingly endorse the same political position taken by the

organization. In other words, followers of gun control organizations are more likely to be in favor of gun control than followers of ideologically liberal organizations in general or supporters of the Democratic party specifically. Similarly, we would expect followers of anti-immigration organizations to demonstrate greater animosity towards immigration than the average follower of Conservative organizations not dedicated to restricting immigration. Analyzing the followers of these organizations and comparing them across narrow policy positions should allow us to isolate on the factor participants disagree over, rather than more general and ambiguous positions of liberalism or conservatism.

To appropriately compare the sets of images, they needed to be as uniform as possible without any extraneous information that might differ across groups. Because this study is primarily focused on facial morphology rather than environmental indicators, each facial image was cropped so that only the face is represented in the image and most of the ancillary background information is removed. Further, each image was rotated and scaled such that each face is vertically aligned and the eyes are similarly spaced. To perform this task, the "dlib" package in Python was used (King, 2009). Faces were identified and aligned by determining 68 "landmark points" that identify facial features such as the curve of the jaw, eyebrows, eyes, nose, and mouth (see Appendix D). Images with no detectable faces or more than one detectable face were eliminated from the sample.

Control Variables – Age, Sex, Race, Emotional Expression, Head Position

It is probable that each comparison group differed from its counterpart in a multitude of ways. For example, we might expect conservative groups to have more males, less minority representation, and be older on average than the liberal organizations. If we did not control for these variables, the classifier employed for categorization could have

used this ancillary information to determine which group a participant belongs to, confounding the experiment. For example, if 70% of men in a sample belong to conservative groups, the classifier could "cheat" and achieve a model accuracy rate of 70% just by categorizing all men in conservative groups and all women in liberal groups.

Because of this, the images needed to be sorted into basic demographic categories and compared only to members of the same gender and race. Additionally, because groups might differ by facial expression, facial expression needed to be assessed for each image.

To perform these actions, we used the lightweight facial recognition framework "DeepFace" for Python (Serengil & Ozpinar, 2020). DeepFace is a facial recognition system employed and maintained by FaceBook and described by its authors as a "hybrid face recognition framework". Trained on over 9 million images, DeepFace has been shown to have accuracies in facial categorization which approach that of human beings (GeeksforGeeks, 2021). Specifically, DeepFace was used to identify subjects' ages, genders, races, and facial expressions. With this information identified, we could then compare across ideological groups while holding these confounding variables constant.

The python packages dlib and opencv were used to estimate head pose and position. Critics of Wang and Kosinski (2018) have suggested that head pose and position might be indicative of group membership (Agüera y Arcas et al., 2018). Although the groups of analyses are different across the two studies (sexual preference vs. political affiliation), it is possible that political groups systematically differ in head positioning. As such, the roll, pitch, and yaw of the head was estimated and controlled. Images had already been rotated to align the faces across photographs, functionally negating the need for the roll variable. Pitch and yaw are measured in degrees, with a head facing straight towards the camera

demonstrating a value of zero for both pitch and yaw. Positive numbers indicate a face tilted more upward for pitch and a face turned towards the viewer's left for yaw, while negative numbers indicate the opposite.

Feature Extraction and Singular Value Decomposition

After the images were uniform and the image set had been reduced, VGGFace was used to extract features from the facial images. VGGFace is a neural network trained on over 3.3 million images of over 9,000 individuals by researchers from the Visual Geometry Group (VGG) at the University of Oxford. Facial features of the edited images were extracted, creating a vector of length 4,096 for each image.

Singular Value Decomposition (SVD) is a process similar to factor analysis whereby a dataset is reduced to its most influential components. Baker (2005) describes SVD as "a method for transforming correlated variables into a set of uncorrelated ones that better expose the various relationships among the original data items" (p. 14). SVD utilizes matrix algebra to decompose a matrix into three component matrices, which allows for the extraction of the most important components of the data (Bagheri, 2020). Because these important components might differ across groups, SVD must be performed on each group's data individually, reducing the 4,096 long feature vector for each image to just 500.

Point and Mesh Coordinates, Image Masking, Classifier Training

Finally, facial point coordinates and facial mesh coordinates were collected using the python libraries dlib and mediapipe, respectively. The libraries provided facial coordinates for each image, a total of 68 points for dlib and 468 points for mediapipe. In the same manner as utilized for features, one can use the facial point coordinates and facial mesh coordinates as information in the logistic regression classifier models. That is, the

models can be trained on the point coordinates themselves, reducing or eliminating entirely the influence of anything unrelated to facial morphology.

By utilizing point and mesh coordinates, we control for the influence of a great number of things concurrently. For example, by utilizing facial point or mesh coordinates, the influence of transient facial feature properties, such as makeup, glasses, sunglasses, eye positioning, tans, acne, beards, and the like, are eliminated. At the same time, the influence of image properties, such as dominant color or colors, brightness, pixilation, sharpness or blurriness, contrast, irrelevant edge detection, and so on, is also eliminated.

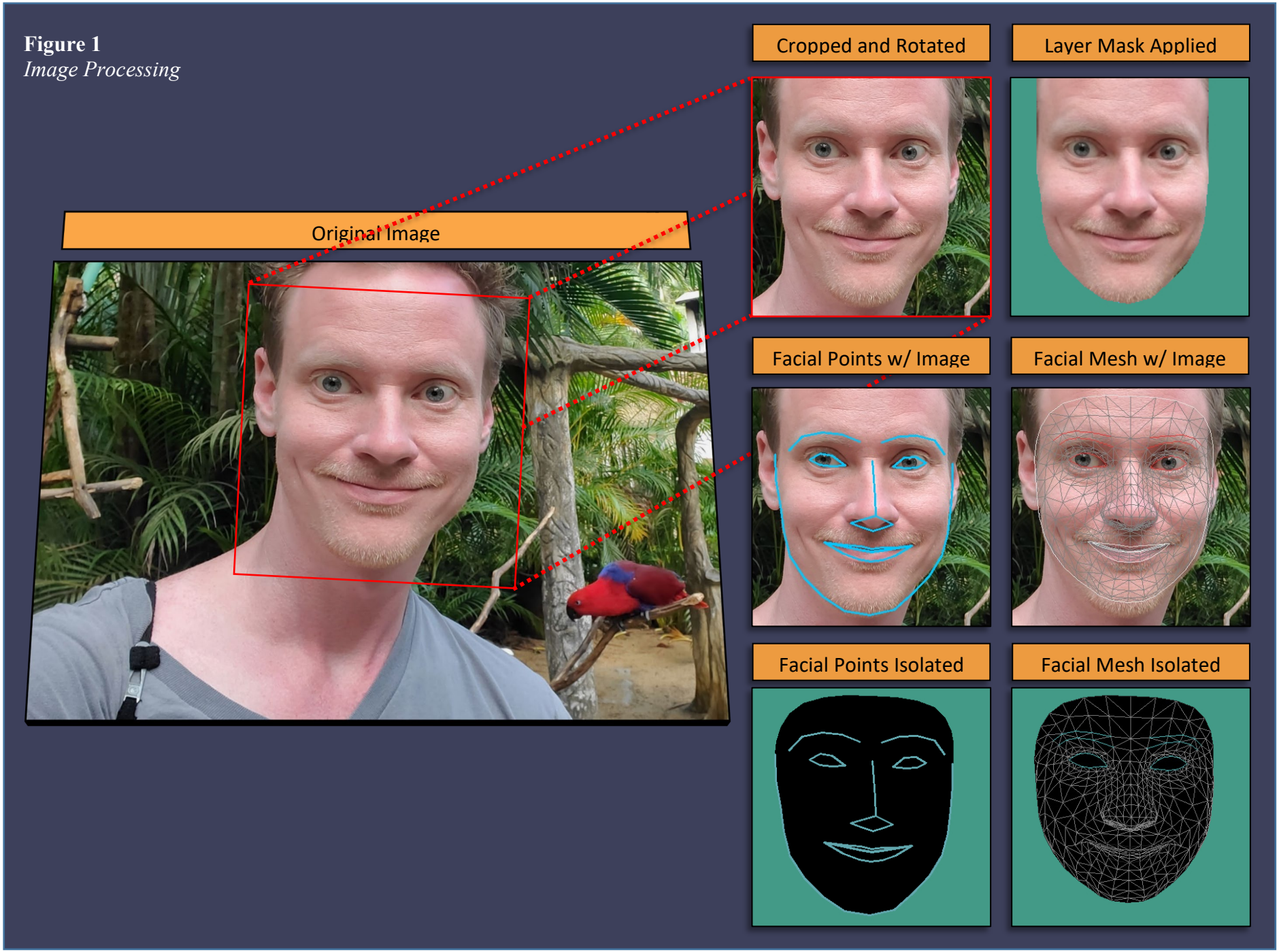
Additionally, to test the influence of the background in image classification, each image was subjected to a masking technique. This technique utilized the 68 dlib facial points to draw a border around the face in the image and to mask every part of the image outside of the contours of the face. By comparing the accuracy of the whole image model to that of the model with the background removed, we can determine how much the background is influencing the classifier.

Signal Detection Theory and Analysis

To analyze the success of the logistic regression models, signal detection theory (SDT) was employed. SDT was developed to study human sensory discrimination in the 1950s (Swets, 2014). Participants in an experiment might have been tasked between distinguishing between hearing a tone, and not hearing a tone. Categorizing the output from this simple task results in a 2 x 2 matrix called a classification table (Peng et al., 2002; see Appendix E). Thus, outcomes are classified in four ways: true positives (reporting a tone when one is presented), true negatives (not reporting a tone when no tone is presented), false positives (reporting a tone when none is presented), and false negatives (not reporting

a tone when one is presented) (Flach, 2016).

Figure 1
Image Processing



The success of a model can be visualized by using a receiver operating characteristic (ROC) curve, which plots the true positive rate and false positive rate for each of a variety of confidence levels (Stanislaw & Todorov, 1999). For the current study, the confidence levels will be the logistic regression output for any specific image, which will be the probability that the image belongs to the category in question (between 0 and 1). The category to which the images will be tested for, i.e., either "left" or "right" position taking, is arbitrary, as selection for inclusion in, or exclusion from, one category determines the outcome for the remaining category.

Performance for models using SDT is measured using accuracy, defined as the fraction of successful categorizations over the sample size, as well as area under the curve (AUC), defined as the true positive rate divided by the false positive rate. AUC is a useful metric for being able to tell how well a model is at distinguishing between categories, with an AUC of 1 (100%) representing a perfect classifier and an AUC of .5 (50%) representing a classifier no better than chance (Narkhede, 2018).

Strong Inference Testing

In one of the most widely cited scientific papers of the 20th century, Platt (1964) introduced the concept of 'strong inference'. Strong inference was proposed as an optimal way of performing scientific inquiries. The process occurs in the experimental design phase, where researchers specifically design their experiments with competing or alternative hypotheses. According to Platt, experiments should be designed "with alternative possible outcomes, each of which will, as nearly as possible, exclude one or more of the hypotheses" (Platt, 1964, pg. 347). In other words, experiments can sometimes be organized in such a way so that supporting evidence for one hypothesis necessarily

precludes support for the alternative hypotheses. Despite not being experimental, this research utilizes a series of strong inferences in order to determine if facial morphology is being utilized during image classification.

Recall that there are several ambiguities in previous works of this nature. Wang and Kosinski (2018) and Kosinski (2021) both emphasized the fact that the success of their model is due primarily to facial features, that is, the specific facial morphology of the subjects in the photographs. However, despite the common nomenclature, the 'features' extracted from a facial image contain more information than the facial features themselves. As such, it is possible that the accuracy of feature only models is not due to facial morphology, but instead other information. For example, Leuner (2019) found he could accurately predict group membership based on color dominance in the image, while Wang (2022) discovered he could classify images based on brightness of the image as well as from the background of the image but with no facial data at all.

Restated, there is a great deal of controversy over whether the accuracy of these feature models is due to facial morphology itself, or contrastingly, if it is due to other sources of information within the image. These sources might include the background of the image, the transient characteristics of the person within the image, or other features that might be 'observed' by the computer vision algorithm but are not necessarily dependent upon facial morphology.

In this vein, one might establish some competing hypotheses in relation to the research presented. For example, there is concern that the background of the image might in some way be indicative of group belonging (Wang, 2022). By comparing the accuracy of the model created from the masked images to the accuracy of the model created from

whole images, we can infer that any difference in predictive ability between models is due solely to the influence of the background across the image set. Any reduction in the masked model's accuracy in comparison to the whole image feature model must be due to the lack of contextual clues in the background of the images. If model accuracy is consistent across these two models, we are compelled to conclude that the whole image feature model does not rely upon the background information to make its classifications.

Similarly, there is concern that the accuracy of the feature classifiers is due to components unrelated to facial morphology. For example, Agüera y Arcas et al. (2018) suggest that much of the variance explained in the model might be due to variables that are environmentally influenced, including differences in presentation style like donning a beard, makeup, glasses, or a tan. By utilizing the point and mesh coordinates and isolating on facial morphology, we also create a strong comparison. If the feature models are successful classifiers but the point and mesh models are not, we know that the feature models are not using information that is related to facial morphology, because facial morphology is not predictive. If, contrastingly, the point and mesh models *are* predictive, it is certain that facial morphology can be utilized to classify images.

Finally, groups might still differ by their elective facial morphology, namely, their facial expression. That is, even when eliminating all information other than facial morphology, smiles, frowns, and other facial expressions might still be indicative across groups (Agüera y Arcas et al, 2018). This research has attempted to correct for this in two different ways.

First, by removing the points related to the mouth, one might determine if point models can be successful with only the remaining facial morphology. By comparing the

whole facial point model to the facial point model with the mouth points removed, we set up another strong comparison. If the whole facial point model achieves accuracy in predictions while the no-mouth models do not, we conclude that the success of the whole facial point model is due solely to the morphology of the mouth across images. If the no-mouth model remains predictive, we know that the structure of the face, rather than the positioning of the mouth, can be used to predict group belonging.

Second, in order to determine whether model success is due to facial morphology or, instead, facial expression, one might utilize the point or mesh models (using only facial morphology) while constraining the sample to only subjects demonstrating the same facial expression. Doing so should further elucidate whether model success is due to facial morphology or, instead, facial expression. If a model limited to facial morphology and constrained by facial expression is still predictive, we must consider that the success of that model is due to the facial morphology contained within the image set. If, on the other hand, the model's accuracy reduces substantially during this analysis, it is strongly supportive of the idea that the algorithm is predicting on facial expression instead of facial morphology.

Hypotheses

H1: Images of subjects will be able to be categorized by utilizing only their features, replicating the primary finding by both Wang and Kosinksi (2018) and Kosinski (2021).

H2: The accuracy of the whole image feature models in distinguishing between left and right groups will not differ from the accuracy of the masked image feature models. If confirmed, we can be certain that the background of the image is not influential in regards to image classification. If not confirmed, we must conclude that the feature models rely

upon the background information in the image to make their classifications.

H3: The accuracy of the point models in distinguishing between left and right groups will be better than chance. If confirmed, we know that images can be classified into group membership utilizing only facial point morphology. If not confirmed, we know that the feature models are not utilizing morphology in their classification strategy.

H4: The accuracy of the no-mouth point models will be better than chance. If confirmed, we know that images can be categorized solely by facial morphology, even when ignoring the mouth, a potential source of morphological divergence between groups. If not confirmed, we know that any success in classification of the point models was due solely to the variation in the mouth of the person in the image.

H5: The accuracy of the point models constrained by facial expression (happy, neutral) will be better than chance. If confirmed, we know that images can be categorized by facial morphology and that classification accuracy is unrelated to facial expression. If not confirmed, we know that the success of the whole image point model is due to the facial expression of the person in the image.

H6: The accuracy of the mesh models in distinguishing between left and right groups will be better than chance. If confirmed, we know that images can be classified into group membership utilizing only facial mesh morphology. If not confirmed, we know that the feature models are not utilizing facial morphology in their classification strategy.

H7: The accuracy of the mesh models constrained by facial expression (happy, neutral) will be better than chance. If confirmed, we know that images can be categorized by facial morphology and that classification accuracy is unrelated to facial expression. If not confirmed, we know that the success of the whole image mesh model is due solely to

the facial expression of the person in the image.

Creating the Sample

To attempt to discern whether political orientation could be derived from features extracted from profile features, four groups were examined. The National Rifle Association (NRA) and Everytown (ET) are two activist organizations related to gun control, the NRA being against gun regulations (right-leaning) and Everytown being for restrictions on firearms (left-leaning). Additionally, two activist groups related to immigration policy were selected. The Federation for American Immigration Reform (FAIR) is an organization dedicated to reducing immigration 'to a more normal level' (right-leaning), while United We Dream (UWD) is an organization committed to ensuring immigrants have a say in immigration policies (left-leaning).

Over 1.6 million Twitter followers of these organizations were pulled using Twitter's application programming interface (API). Despite the large number of followers, many subjects were removed in the process of standardizing and categorizing the images. These eliminations are illustrated in the Sankey plot in Figure 2. First, users with the default profile image were removed from the sample. Then, each image was analyzed to obtain the categorical variables including race, sex, and facial expression. At this stage, images with zero faces detected or with more than one face detected were eliminated. Next the images were cropped tightly to the face and rotated so that the facial features across images would be aligned. At this point, the image set was reduced to around 425,000 images.

To prevent the undue influence of profile images that were not of the account holder, every image was viewed individually by at least one rater. Two raters viewed

50,000 of the images one at a time, and a third rater viewed all images in this way. Raters were tasked with eliminating 'imposter' images that they suspected of not being the account holder. For example, many right-leaning followers had profile images of Donald Trump. If the profile picture contained an image of a famous person, it was assumed that the image was an 'imposter'. Additionally, images were removed if the subject appeared to be underage, if there were multiple people in the frame (despite the algorithm having identified only one face), or if the image had a significant portion of the face covered (more than half), amongst other criteria. While this process was highly subjective, interrater reliability was high across raters (ICC = .854, $F(49999, 100000) = 18.49$, $p < .001$, $CI_{95\%} = [.85, .86]$).

Nearly 72,000 images were removed in this fashion. To give an impression of the removal process, a random sample of 200 images was taken from the removed images. Of the 200 removed images, 25.5% were removed because they were artistic renderings of faces, whether line drawings, 3-dimensional models, face-altering filters, or sculptures. A total of 18% of the drawn sample were images of children, while 17.5% of the sample were of politicians (77% Trump, 6% Biden). Another 15% of the sample were celebrities, and 9% of the sample were removed because the images had multiple people in the foreground of the frame. The remaining 15% of images were removed for having their face covered (3%), being a still from a movie or television program (3%), the image being heavily manipulated (2.5%), the image being a meme (2.5%), the image appearing to be excessively dated (2%), the image being historic in nature (1%), the image being of an animal (0.5%), or the image having no face at all (0.5%).

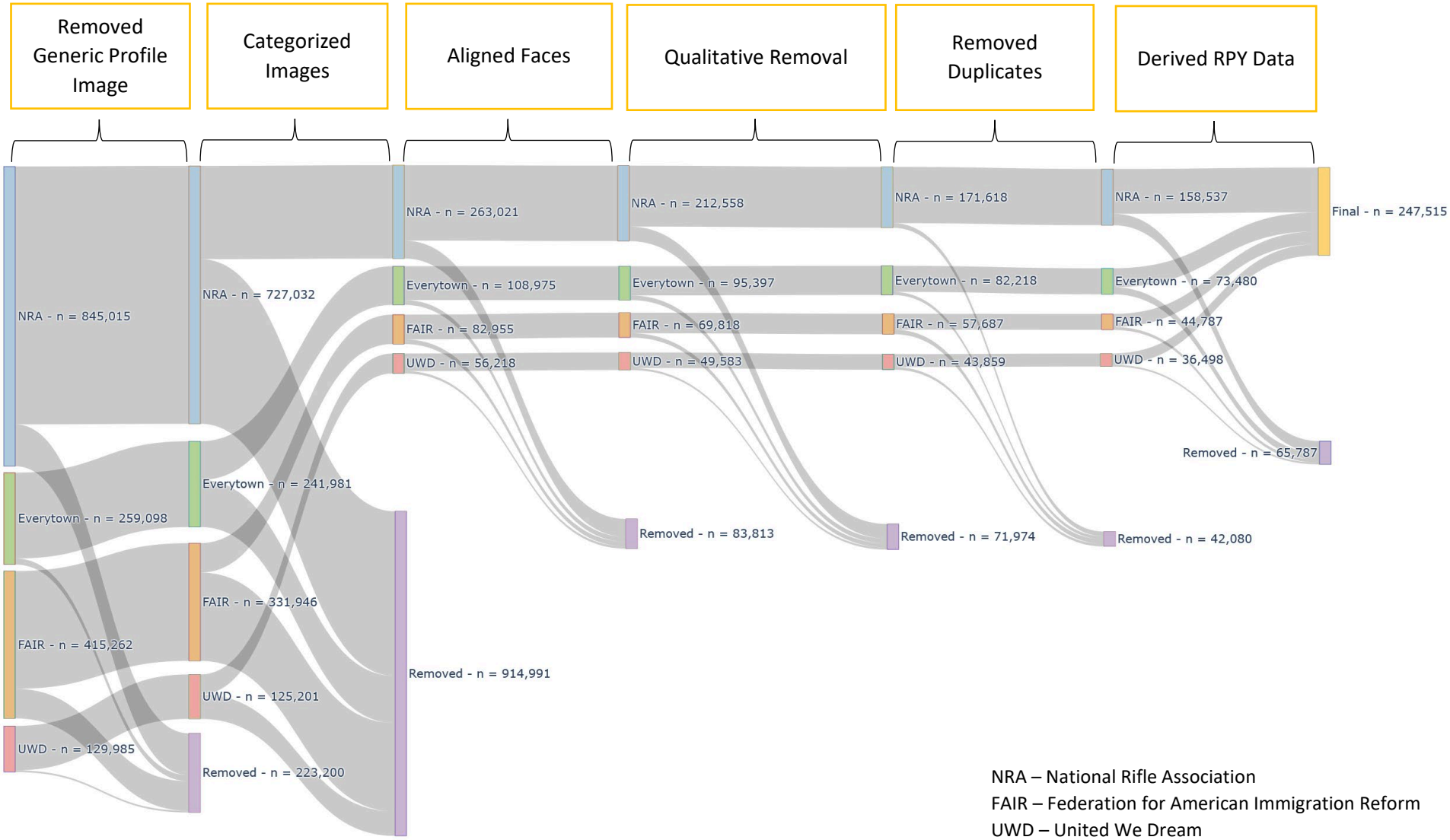
After the qualitative removal process, duplicate images across the sample were

eliminated. This was performed by matching exactly on the 4,096 features extracted. Any images sharing an identical feature set were removed. Finally, roll, pitch, and yaw data was extracted from the images if possible. This left us with a total sample of 247,515 images.

Wang and Kosinski (2018) removed faces with a yaw of greater than 15 degrees and a pitch greater than 10 degrees. Appling these standards to the present image set reduced its size by nearly 39% to 151,377. Thus, it was decided to perform two sets of analyses, one without reducing the sample by pitch and yaw, and one where the sample was reduced in the same way as in Wang and Kosinski (2018).

Figure 2

Sankey Process Plot for Sample Creation



NRA – National Rifle Association
FAIR – Federation for American Immigration Reform
UWD – United We Dream

Chapter 9: Results

This chapter is divided into two sections, broadly. The first section examines the control variables that were gleaned from the images and attempts to elucidate differences in these variables across groups using traditional inferential statistics. Although these findings are not the primary focus of this research, identifying differences across subgroups should be illuminating from a theoretical perspective while giving the reader an overall impression of the data. These analyses were performed on the entire corpus of images.

Section two will utilize a machine learning approach to determine the role that facial morphology plays in group categorization. To eliminate the possibility of a classifier categorizing subjects based on their race or gender, all groups of analysis were constrained by these two variables for this section. In other words, the overall sample was broken into subsamples, each specific to a particular group membership, as well as the sex and race of the subject. With four different issue groups, two sex categories, and six ethnic-racial categories, this resulted in a total of 48 subgroups (e.g., white female followers of the NRA; Hispanic male followers of Everytown). Given that this approach was predicated on comparisons of different group members with different political leanings but identical demographic characteristics, this implies a maximum of 24 pairwise comparison.

Note that this approach relies upon an ample sample size in order to detect potentially small effects. Using a similar method, Wang and Kosinski (2018) previously relied on a minimum sample size of 3,441. In the present analyses, the intent was to include any subsamples that included at least 3,000 subjects. Because this resulted in only five viable pairwise compairsons, the lower bound was reduced to 2,900 to allow for the inclusion of two additional analyses. These are the sizes for the sample that is unrestricted

in terms of pitch and yaw; however, companion analyses were included in the Appendices for the reduced sample set regardless of sample size. As illustrated in Table 1, this expanded the number of viable comparisons to seven.

Table 1
Groups of Analysis

Group	Everytown		United We Dream		National Rifle Association		Federation for American Immigration Reform	
DomSex	Males	Females	Males	Females	Males	Females	Males	Females
Asian	3,010	1,547	2,654	1,941	7,110	1,790	2,372	934
Black	2,346	527	2,024	483	5,815	476	3,221	451
Indian	638	170	610	196	1,803	154	629	95
Hispanic	2,921	1,980	3,383	2,842	10,074	2,261	2,991	1,105
Middle Eastern	1,219	70	783	109	5,623	99	1,231	43
White	20,531	18,416	8,742	7,782	70,715	20,223	15,199	8,177

Note: Fields with identical colors identify the viable comparison between a left-learning and a right-leaning group pertaining to the same issue. Everytown and United We Dream are left-leaning groups; National Rifle Association and the Federation for American Immigration Reform are right-leaning groups.

Because the sample skewed male (as does Twitter in general), the subgroup selection did as well, with five pairwise subgroup comparisons for men and only two such comparisons for women. Only women whom the classifier deemed white could be included in the present analyses.

For the sake of brevity, only one of the subgroups is examined in detail. For this portion of the analysis, the subgroup with the largest sample size was utilized: white males in the gun topic. It is important to understand that this subgroup is not of any particular interest in regards to these analyses. Rather, this subgroup is utilized merely as an illustration for the reader of how analyses were performed on all seven subgroups. However, more important than any specific subgroup comparison, we are looking for overall trends in regards to this methodology. Thus, after the detailed analysis of the white

male gun subgroup, broader trends and realizations across all subgroup analyses will be presented. Metrics for all models are presented in the Appendices.

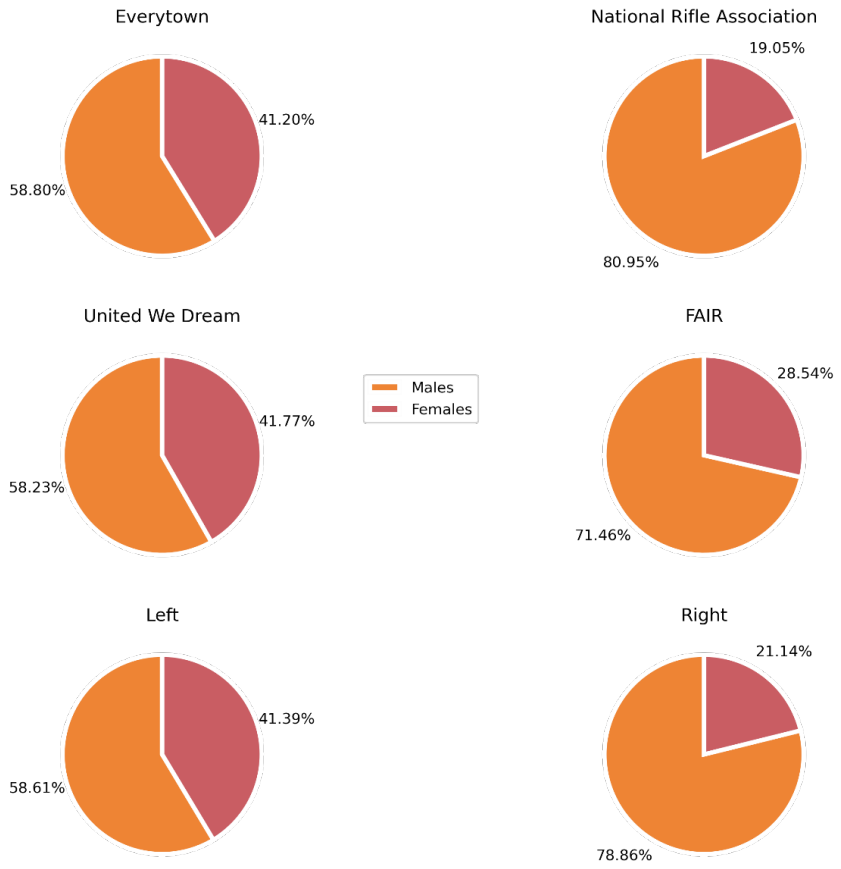
Section 1 – Control Variables

Each of the images in the sample was evaluated by DeepFace, a deep learning facial recognition network that was developed by a team at Facebook. DeepFace utilizes a nine-layer neural network that was trained on four million images from Facebook users (Serengil, n.d.). The network involves over 120 million parameters and describes itself as 'the most lightweight face recognition and facial attribute analysis library for python'(Serengil, n.d.). Using this architechture, attributes for sex, race, age, and emotional expression were extracted from each image. These factors were examined for Section 1, with all images being included in these analyses.

Sex

Research on Twitter demographics has previously shown that Twitter users are majority male (Dixon, 2022, 56.4%; Yildiz et al., 2017, 73%). This sample did not differ in that respect, with the sex identifier classifying nearly 71% of subjects as male. See Figure 3.

Figure 3
Sex by Organization Barplot



Note: FAIR = Federation of American Immigration Reform. Everytown and United We Dream are left-leaning groups; National Rifle Association and FAIR are right-leaning groups.

Several chi-square tests of independence were used to determine whether followers of organizations differed by sex. In comparing all images, right-leaning subjects were significantly more likely to be male than left-leaning subjects ($\chi^2(4, N = 247{,}515) = 11{,}311.91$, $p < .001$, $\Phi = .21$). The same held true for both the gun subgroup ($\chi^2(4, N = 179{,}518) = 9{,}925.86$, $p < .001$, $\Phi = .24$) and for the immigration subgroup ($\chi^2(4, N = 67{,}997) = 1{,}186.43$, $p < .001$, $\Phi = .13$), although the effect was weaker in the immigration

subgroup. The phi value of .21 translates to a weak overall effect according to statistical conventions (Bhandari, 2021; Zaiontz, n.d.).

The sex classifier provided a propensity score to classify subjects in images as being male or female with some overall likelihood. Each image was provided a 'score' as to whether the individual was male or female, on a scale from zero to one. If the score is closer to zero, the algorithm predicts that the individual in the photo is a woman; if it is closer to one, the algorithm predicts that the individual is a man (or vice versa). In the present context, these categories are mutually exclusive: if the algorithm is predicting the image is a woman at a likelihood of 10%, this necessarily means that it is also predicting the image is a man with a likelihood of 90%. Thus, one can simply subtract the propensity score for women (in this example) from one, providing a proxy score for 'sex typicality' for both sexes with a scale from .50 to 1.

Analyzing the data in such a way, we identified if followers of organizations differed in their sex typicality, at least in terms of what the sex classifier deemed as sex typical qualities. Previous research has suggested that conservative leaning individuals might endorse more traditional gender roles, suggesting that males and females that are right-leaning might adopt appearances that are more 'sex typical' in nature (Duncan et al., 1997). At the same time, left-leaning individuals tend to score higher on the personality trait openness, to demonstrate a greater propensity to break with tradition, and to espouse for equity across social constructs, suggesting that the 'borders' between the sexes might be more traversable for left-leaning individuals (Carney et al., 2008).

A linear model was created using all images with the 'sex typicality' variable being independent and sex, orientation, topic, and the interactions of the three being dependent

variables. All linear models were examined with a Type-3 analysis of variance test, and all categorical variables utilized 0/1 coding.. The results for this sex typicality ANOVA model were significant ($F(7, 247{,}507) = 1{,}335.00$, $p < .001$). Individuals on the right were significantly more likely to demonstrate 'sex typicality' than individuals on the left ($b = 3.45$, $CI_{95\%} = [3.33, 3.57]$, $t(247{,}507) = 57.11$, $p < .001$). There was also an effect by sex, with males on average across all images demonstrating greater sex typicality than females ($b = -0.72$, $CI_{95\%} = [-.88, -.56]$, $t(247{,}507) = -8.86$, $p < .001$). There was an interaction effect between orientation and sex, with males demonstrating a significant increase in sex typicality while moving from left to right in orientation, but with females being relatively consistent across the ideological gap ($b = 2.86$, $CI_{95\%} = [-3.06, -2.65]$, $t(247{,}507) = -27.39$, $p < .001$). Topic (gun vs. immigration) was non-significant ($b = .029$, $CI_{95\%} = [-.14, .20]$, $t(247{,}507) = .34$, $p = .74$), as was the interaction with topic and sex ($b = 0.050$, $CI_{95\%} = [-.31, .21]$, $t(247{,}507) = -.37$, $p = .71$). The interaction of topic and orientation was significant, however ($\beta = -1.72$, $CI_{95\%} = [-1.93, -1.51]$, $t(247{,}507) = -15.86$, $p < .001$). Left-leaning subgroups did not differ substantially in their sex typicality across the topics of gun control and immigration. Right-leaning subgroups, contrastingly, demonstrated greater sex typicality in the gun control conditions than the immigration conditions. Finally, the three-way interaction between sex, orientation, and topic was also significant ($b = 0.92$, $CI_{95\%} = [0.56, 1.28]$, $t(247{,}507) = 5.03$, $p < .001$), although the effect was rather modest. The R^2 for the model was .036, explaining just over 3.5 percent of the variance in the model. Figure 4a summarizes the pattern of findings for the collapsed data. Figures 4b and 4c display the data separately for gun and immigration subgroups, while 4d demonstrates the interaction between orientation and topic. Results for this analysis are presented in Appendix F.

Figure 4
Sex Typicality Line Plots

Figure 4a

Orientation x Sex Typicality – All

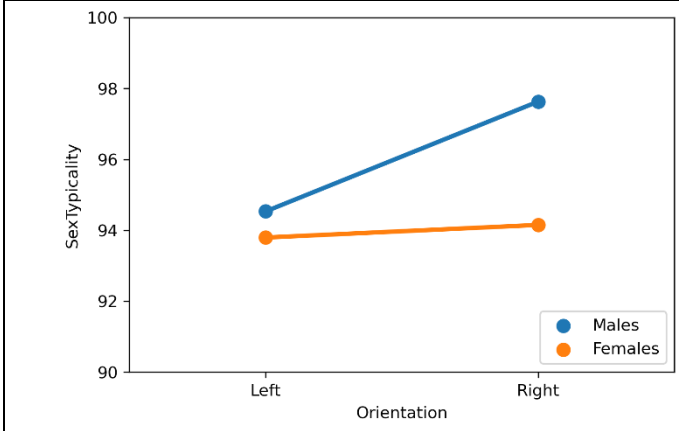


Figure 4b

Orientation x Sex Typicality – Gun

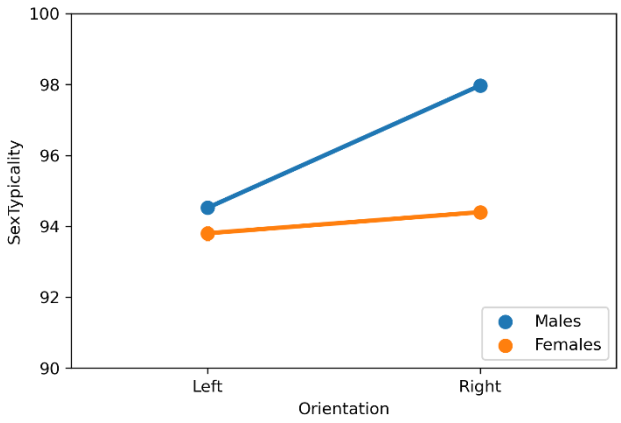


Figure 4c

Orientation x Sex Typicality – Immigration

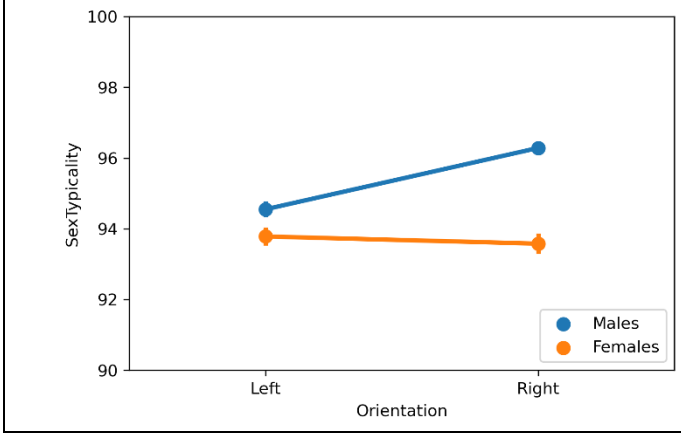
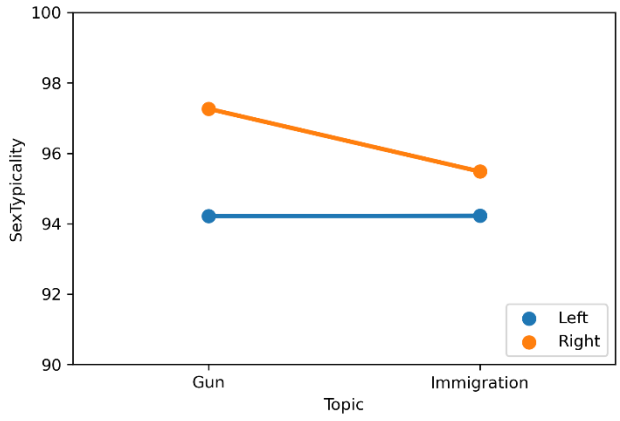


Figure 4d

Orientation x Topic - All



Note: All plots contain confidence intervals for each point estimate. Due to the large sample sizes, they are very small and difficult to see.

Race

The python library DeepFace was used to retrieve the racial information for subjects in the images. The sample was majority white, with approximately 69% of the sample being Caucasian. The next largest ethnicity represented was Hispanics, with nearly 11% of the sample. Asians, African-Americans, Middle-Easterners, and Indians comprised

the rest of the sample, each comprising 9%, 6%, 4%, and 2% of the sample, respectively (rounded to the nearest integer) (see Table 2).

Table 2

Percent Race by Group

	Everytown	National Rifle Association	United We Dream	FAIR	Left Groups	Right Groups	All
Asian	8.54%	7.06%	14.56%	9.07%	10.78%	7.51%	8.63%
Black	5.38%	4.99%	7.95%	10.07%	6.34%	6.13%	6.20%
Indian	1.51%	1.55%	2.55%	1.99%	1.90%	1.65%	1.74%
Hispanic	9.18%	9.78%	19.73%	11.24%	13.10%	10.11%	11.13%
Middle Eastern	2.41%	4.54%	2.83%	3.50%	2.57%	4.30%	3.71%
White	72.97%	72.09%	52.38%	64.14%	65.32%	70.31%	68.60%

To determine if left- and right-followers differed significantly in their racial composition, two sets of analyses were run. First, a chi-square test was performed on all of the images, with all of the ethnicities represented. Left-leaning subgroups were significantly more diverse than right-leaning subgroups, ($\chi^2(12, N = 247{,}515) = 1{,}820.07$, $p < .001$, $\Phi_c = 0.09$), although the effect was rather weak. The effect was the same albeit greatly reduced when looking at just the gun subgroups, with Everytown being more ethnically diverse than the NRA ($\chi^2(12, N = 179{,}518) = 571.71$, $p < .001$, $\Phi_c = 0.06$). The effect was much stronger for the Immigration subgroups, with the images from United We Dream being significantly more diverse than FAIR and with the strongest effect overall in regards to ethnicity, although overall still relatively weak ($\chi^2(12, N = 67{,}997) = 1{,}773.81$, $p < .001$, $\Phi_c = 0.16$).

Second, a chi-square test was carried out on a dichotomized ethnicity variable with Caucasians representing one subgroup and members of all other ethnicities representing a non-Caucasian/'minority' subgroup. Results were similar although less pronounced.

When accounting for all images, left-leaning subgroups demonstrated greater racial diversity than right-leaning subgroups ($\chi^2(4, N = 247{,}515) = 644.41$, $p < .001$, $\Phi = 0.05$). For gun subgroups, the test result was significant but the effect was negligible ($\chi^2(4, N = 179{,}518) = 14.39$, $p < .001$, $\Phi < 0.01$). Similar to the previous analysis, the immigration subgroups demonstrated the largest effect ($\chi^2(4, N = 67{,}997) = 964.00$, $p < .001$, $\Phi < 0.12$).

Age

Each of the images was evaluated by DeepFace to provide an age estimation for each participant. Mean predicted age across all images was 31.54 years old ($SD = 5.48$). Research on Twitter users has previously shown that, on average, they are relatively young (Dixon, 2021, 64% ≤ 34). This sample was no different, with 75% of subjects being deemed 34 years of age or lower. The maximum age predicted by the classifier was 62, and the minimum age was 19.

A linear model was used to assess if there were significant differences in age across the images, with sex, orientation, and topic as the independent variables. The model proved significant overall ($F(3, 247{,}507) = 841.10$, $p < .001$). Females appeared significantly older than males in the sample ($b = 1.20$, $CI_{95\%} = [1.11, 1.29]$, $t(247{,}507) = 25.28$, $p < .001$). At the same time, right-leaning individuals appeared significantly older than left-leaning individuals ($b = 0.20$, $CI_{95\%} = [0.13, 0.27]$, $t(247{,}507) = 5.70$, $p < .001$). This is perhaps not surprising, as previous research has suggested people become more conservative with age (Truett, 1993). Topic was also significant in the model, with subjects in the immigration subgroups appearing older on average than those in the gun subgroups ($b = -1.04$, $CI_{95\%} = [-1.14, -0.95]$, $t(247{,}507) = -20.63$, $p < .001$). There was a significant interaction effect between topic and orientation ($b = 2.62$, $CI_{95\%} = [2.50, 2.74]$, $t(247{,}507)$

= 41.43, $p < .001$), with left-leaning and right-leaning followers being very similar in age for the gun topic but right-leaning subjects being around 2.5 years older on average than left-leaning subjects for the immigration topic. The interaction between sex and orientation approached significance (b = -0.12, $CI_{95\%}$ = [-0.24, 0.00], t(247,507) = --1.93, p = .054), illustrating a slight decrease in age difference between the sexes when moving from left to right in orientation. The interaction between sex and topic was non-significant, (b = .02, $CI_{95\%}$ = [-.13, .18], t(247,507) = .29, p = .77), although the interaction between the three independent variables was significant, (β = -.37, $CI_{95\%}$ = [-.58, -.17], t(247,507) = -3.51, $p < .001$). The R^2 for the model was .023. See Figures 5a-5d. Results for this model are in Appendix G.

Figure 5

Age Line Plots

Figure 5a	Figure 5b
Age x Orientation x Sex – All	Age x Orientation x Sex – Gun

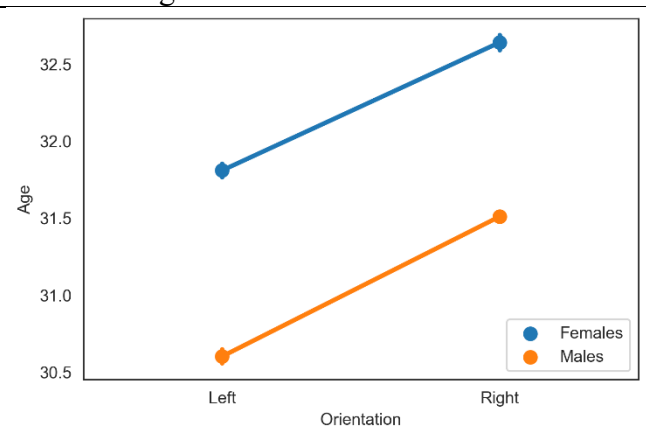
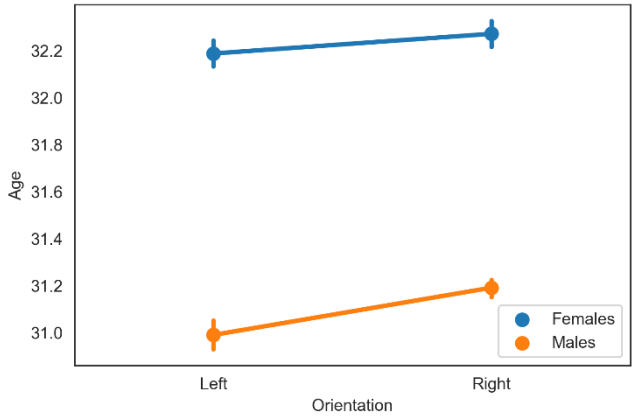
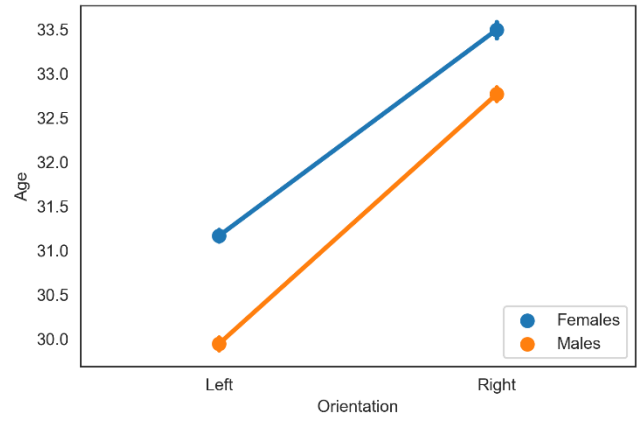
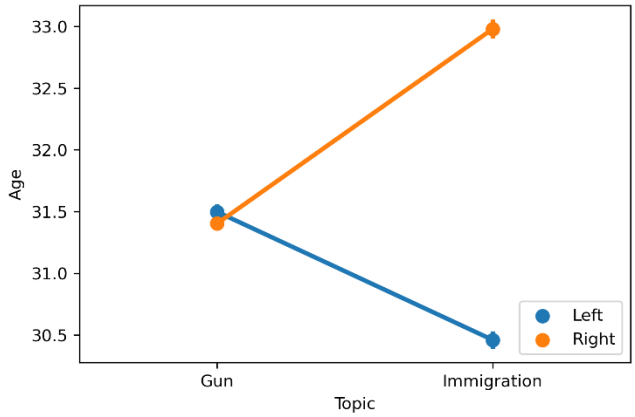

	

Figure 5c	**Figure 5d**
Age x Orientation x Sex – Immigration	Age x Orientation x Topic - All
	

Emotional Expression

Images in the sample were evaluated by DeepFace to provide a measure for subjects' facial expression. Every image was given a propensity score on each of seven emotions: happy, neutral, sad, fear, angry, surprise, and disgust. Over 90% of the sample was covered by just three of those emotions: happy, neutral, and sad. A chi-square goodness of fit test was performed on the sample, testing to see if emotional expression differed by political orientation. It did ($\chi^2(14, N = 247{,}515) = 2{,}384.25$, $p < .001$, $\Phi = 0.10$). See Table 3.

Table 3
Percent Emotional Expression by Organization, Orientation

	Everytown	National Rifle Association	United We Dream	FAIR	Left Groups	Right Groups	All
Happy	73.07%	60.32%	65.96%	61.95%	70.43%	60.68%	64.03%
Neutral	14.07%	20.97%	19.21%	19.28%	15.98%	20.59%	19.01%
Sad	5.56%	8.53%	6.64%	8.91%	5.96%	8.62%	7.71%
Fear	3.62%	4.54%	4.11%	4.60%	3.80%	4.55%	4.29%
Angry	3.11%	4.93%	3.47%	4.57%	3.24%	4.85%	4.30%
Surprise	0.40%	0.49%	0.45%	0.46%	0.42%	0.48%	0.46%
Disgust	0.17%	0.22%	0.16%	0.24%	0.16%	0.22%	0.20%

However, this did not reveal much about the individual emotions at play. Thus, binary variables were created for happy and sad, the two most common facial expressions aside from neutral, to see if followers differed by these emotions. Results were mixed.

Happy. A chi-square test for independence was run on the sample to see if followers differed significantly by proportion of happy subjects. Right-leaning followers had significantly fewer people per capita in their groups with happy expressions than left-leaning followers ($\chi^2(4, N = 247{,}515) = 2{,}300.13$, $p < .001$, $\Phi = 0.10$). Both of the subgroups were tested and both results were significant, although the results for the immigration subgroups were much weaker overall (gun: $\chi^2(4, N = 179{,}518) = 2{,}650.13$, $p < .001$, $\Phi = 0.12$, immigration: $\chi^2(4, N = 67{,}997) = 117.88$, $p < .001$, $\Phi = 0.04$).

Mean happiness propensity scores were also examined. The sample was reduced to just those subjects whom the classifier believed demonstrated a happy expression on their face. Among only these subjects, a linear model was fit to the data with happy propensity scores for the dependent variable and orientation, sex, topic, and the interactions as the independent variables. The model was significant overall ($F(7, 158{,}471) = 307.40$, $p < .001$, $R^2 = .01$). Among those who were happy, orientation was significant, with happiness scores being significantly higher for those on the left than for those on the right ($b = -1.63$, $CI_{95\%} = [-1.81, -1.45]$, $t(158{,}471) = -17.48$, $p < .001$) . At the same time, female subjects were scored as expressing more happiness in their photos than male targets on average ($b = 1.75$, $CI_{95\%} = [1.53, 1.98]$, $t(158{,}471) = 15.09$, $p < .001$). Happy subjects in the immigration topic demonstrated significantly lower happiness scores on average than those in the gun topic, ($b = -.83$, $CI_{95\%} = [-1.09, -0.57]$, $t(158{,}471) = -6.16$, $p < .001$). Both the interactions between orientation and sex as well as orientation and topic were significant, ($b = 1.04$, $CI_{95\%} = [0.75, 1.34]$, $t(158{,}471) = 6.89$, $p < .001$) and ($b = 0.60$, $CI_{95\%} = [.26, .93]$, $t(158{,}471) = 3.47$, $p = .001$), respectively. Females were happier on average than males, and the magnitude of this difference increased when moving from left to right.

Similarly, happy left subjects were happier than their right counterparts, and this difference was much larger across the gun topic than it was across the immigration topic. Neither the interaction between topic and sex (b = -0.20, $CI_{95\%}$ = [-0.59, 0.18], t(158,471) = -1.03, p = .30) nor the three way interaction (b = 0.30, $CI_{95\%}$ = [-0.23, 0.83], t(158,471) = 1.11, p = .27) were significant. See Figures 6a-6d. Model results are presented in Appendix H.

Figure 6

Happy Line Plots

Figure 6a

Orientation x Sex – All

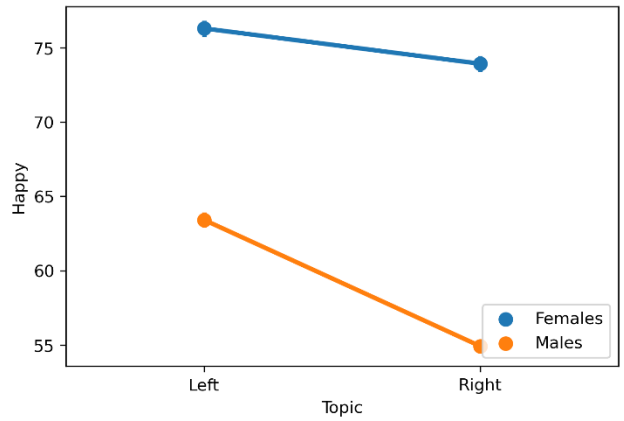


Figure 6b

Orientation x Sex – Gun

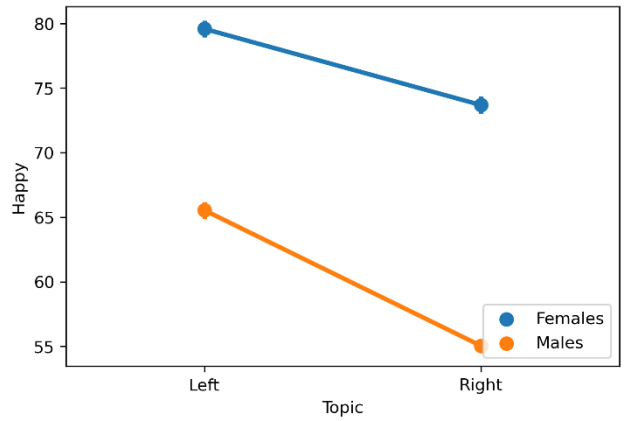


Figure 6c

Orientation x Sex – Immigration

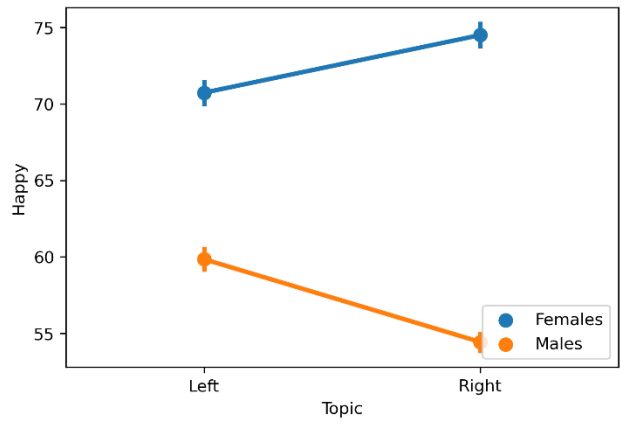
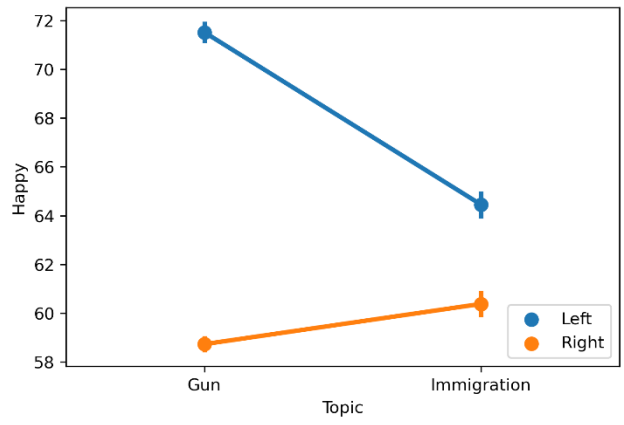


Figure 6d

Orientation x Topic - All



Sad. Results were more opaque in regards to the 'sad' emotional expression. A chi-square test for independence found that right leaning followers had proportionally more photos with sad expressions than left leaning followers, $\chi^2(4, N = 313{,}302) = 745.44$, $p < .001$, $\Phi = 0.05$. This effect remained when testing only the gun subgroups ($\chi^2(4, N = 232{,}017) = 648.99$, $p < .001$, $\Phi = 0.05$) and the immigration subgroups ($\chi^2(4, N = 81{,}285) = 133.43$, $p < .001$, $\Phi = 0.04$).

Similar to the 'happy' expression, sad subjects were isolated, and a linear model for sadness was fit with orientation, sex, topic, and their interactions as predictors. The model was significant, although it explained little of the variance ($F(7, 19{,}065) = 2.29$, $p = .03$, $R^2 < .01$). Only orientation was significant in this model, with subjects on the right demonstrating significantly more sadness in their sad pictures than subjects on the left ($b = -1.29$, $CI_{95\%} = [-2.21, -0.36]$, $t(19{,}065) = -2.72$, $p < .01$). Results for the linear model are presented in Appendix I.

Pitch and Yaw

One criticism of the original work by Wang and Kosinski (2018) was that the researchers did not account for head positioning. Groups of analysis might differ by head positioning, critics argued, in which case the classifier might be predicting based on head position rather than facial morphology itself (Agüera y Arcas et al., 2018). To test for this, all images were subjected to an algorithm that estimated the roll, pitch, and yaw of the head. Because all facial images were cropped and rotated, the roll variable demonstrated little variation across images. Thus, the factors of interest are primarily pitch and yaw.

Pitch. Critics of Wang and Kosinski (2018) describe that heterosexual males and females might orient their heads in different positions when taking their photographs,

specifically in the pitch orientation, the axis used when nodding your head up and down in agreement. It is posited that this is due to sexual dimorphism regarding height in human beings and subconscious sexual signaling to potential mates (Sedgewick et al., 2017). If this were true, we would expect female faces to demonstrate a larger pitch number, indicating that women had taken pictures from above with their faces tilted up, while men would have a lower pitch number indicating that their pictures on average were taken from a lower angle.

To test for such a hypothesis, a linear model was created with pitch as the dependent variable and orientation, sex, topic, and their interactions as predictors. The model was significant overall ($F(7, 247{,}507) = 1{,}820.00$, $p < .01$, $R^2 = .05$). Women demonstrated a significantly higher pitch value in comparison to men ($b = 3.90$, $CI_{95\%} = [3.69, 4.10]$, $t(247{,}507) = 37.48$, $p < .001$), as did people on the left in comparison to people on the right ($b = -2.18$, $CI_{95\%} = [-2.33, -2.03]$, $t(247{,}507) = -28.17$, $p < .001$). Topic was also significant ($b = -0.63$, $CI_{95\%} = [-0.85, -0.42]$, $t(247{,}507) = -5.70$, $p < .001$), with people in the gun groups demonstrating a slightly higher pitch than those in the immigration groups. The interaction between orientation and sex was significant ($b = 2.96$, $CI_{95\%} = [2.70, 3.22]$, $t(247{,}507) = 22.18$, $p < .001$). Females demonstrated an increase in pitch angle when moving from left to right in orientation, while males experienced a decrease in pitch angle. Put another way, right-leaning subjects demonstrated greater variation in pitch between the sexes, while left-leaning subjects exhibited less variation. These findings parallel the findings concerning sex typicality, and provide some additional support for increased gender role adoption among right-leaning subjects.

The interaction between orientation and topic was significant ($b = 0.66$, $CI_{95\%} =$

[0.38, 0.93], $t(247{,}507) = 4.73$, $p < .001$). Whereas participants on the left demonstrated greater mean pitch angles than those on the right, these differences were reduced in the immigration subset in comparison to the gun subset. The final two-way interaction between topic and sex was non-significant, ($b = -0.12$, $CI_{95\%} = [-0.46, 0.21]$, $t(247{,}507) = -.71$, $p = .48$). However, the three-way interaction between all of the variables did reach significance, ($\beta = -0.68$, $CI_{95\%} = [-1.13, -0.22]$, $t(247{,}507) = -2.89$, $p < .01$). See Figures 7a-7d. Model results are presented in Appendix J.

Figure 7
Pitch Line Plots

Figure 7a
Orientation x Sex – All

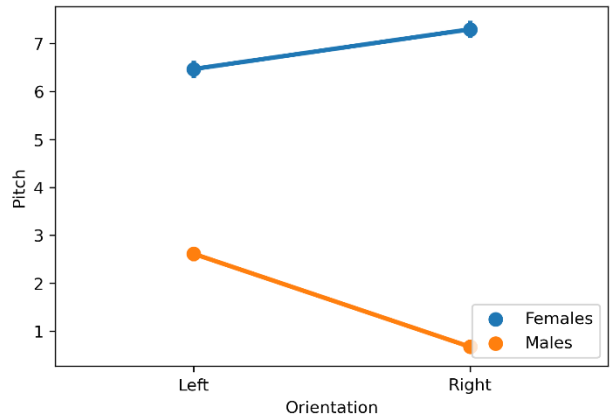


Figure 7b
Orientation x Sex – Gun

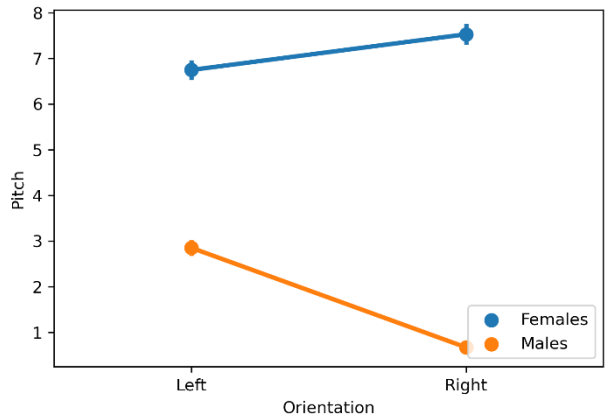


Figure 7c
Orientation x Sex – Immigration

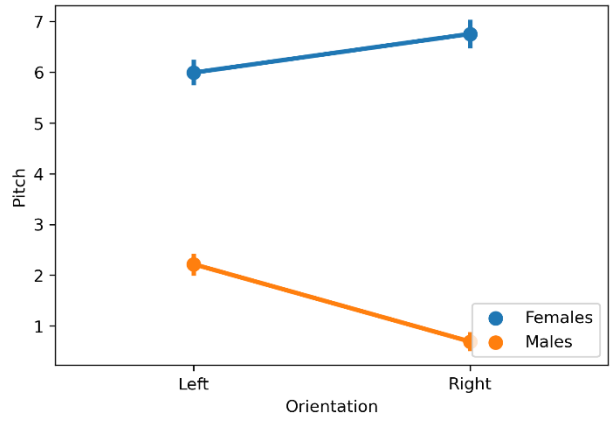
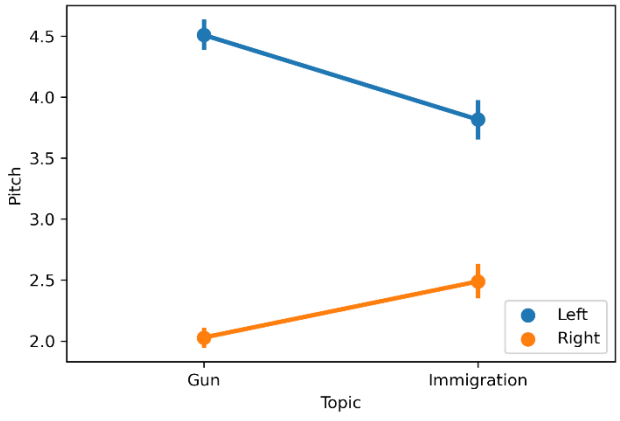


Figure 7d
Orientation x Topic - All



Yaw. Although there was no reason to believe that groups differed by the yaw of their head position (shaking your head no in disagreement), a linear model was created to determine if groups differed in this dimension, with sex, orientation, topic, and their interactions as predictors. Recall that positive numbers in the yaw dimension translate to a head position facing towards the subject's right, the viewer's left. Results for the model were significant ($F(7, 247{,}507) = 45.84$, $p < .001$, $R^2 = .001$), but the small amount of variance explaine suggests that yaw does not aid much in classification of images.

There was a main effect for sex ($b = -0.89$, $CI_{95\%} = [-1.08, -0.71]$, $t(247{,}507) = -9.67$, $p < .001$), with women on average, facing slightly more to their left than men. There was also a main effect for orientation ($b = -0.40$, $CI_{95\%} = [-0.53, -0.26]$, $t(247{,}507) = -5.79$, $p < .001$), with subjects on the political right facing slightly more to their left than those on the political left. The main effect for topic approached significance ($b = 0.19$, $CI_{95\%} = [-0.01, 0.38]$, $t(247{,}507) = 1.89$, $p = .06$), with those subjects in the gun groups demonstrating a slightly lower yaw than those in the immigration groups, translating to subjects in the immigration groups facing slightly more to their left than those in the gun groups.

Two interactions were significant. First, there was a significant interaction between sex and orientation ($b = 0.28$, $CI_{95\%} = [0.04, 0.51]$, $t(247{,}507) = 2.33$, $p = .02$). Females on average had a lower yaw than males, but this difference was reduced among right leaning subjects in comparison to left leaning subjects. Second, there was a significant interaction between sex and topic, ($b = -0.38$, $CI_{95\%} = [-0.68, -0.08]$, $t(247{,}507) = -2.50$, $p = .01$). Males and females in the immigration groups differed more than strongly than males and females in the gun groups.

Neither the two-way interaction of topic and orientation ($b = 0.01$, $CI_{95\%} = [-0.23,$

0.25], $t(247{,}507) = .07$, $p = .94$) nor the three-way interaction of topic, orientation, and gender ($b = 0.27$, $CI_{95\%} = [-0.14, 0.68]$, $t(247{,}507) = 1.30$, $p = .19$) were significant. See Figures 8a-8d. Results for this model are presented in Appendix K.

Figure 8
Yaw Line Plots

Figure 8a
Orientation x Sex – All

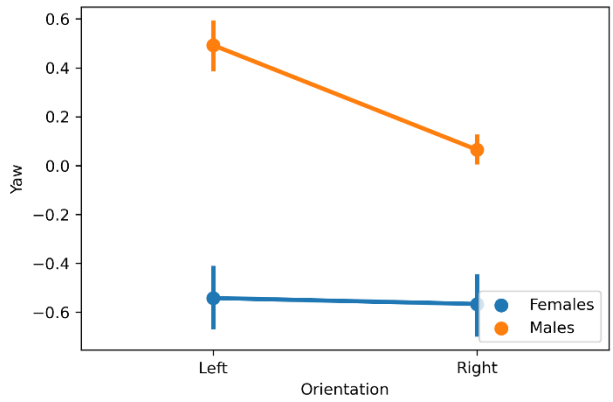


Figure 8b
Orientation x Sex – Gun

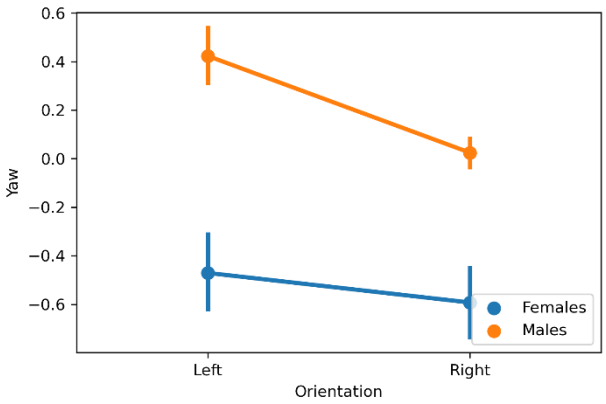


Figure 8c
Orientation x Sex – Immigration

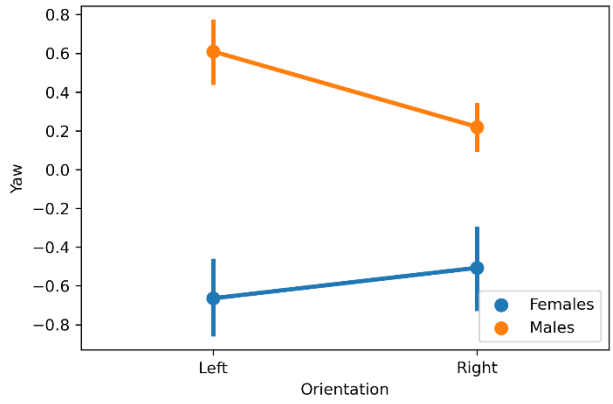
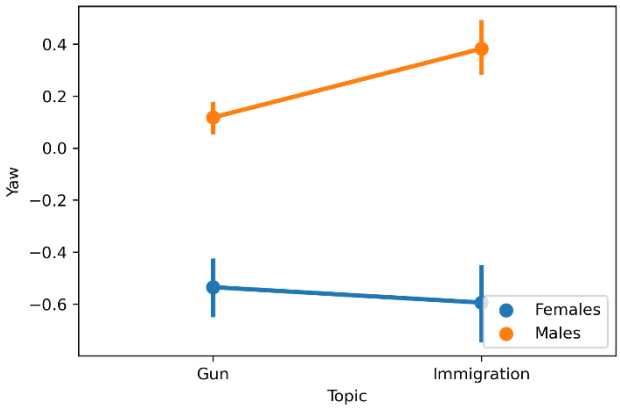


Figure 8d
Topic x Sex – All



Section 2 – Hypothesis Testing

To test for the hypotheses presented in Chapter 8, the sample of images was grouped by race, sex, and topic, so that only members of the same race, sex, and topic were compared to one another in (e.g., white female pro-immigration v. white female anti-immigration). To illustrate the process of image analysis, we begin with the group demonstrating the largest sample size as an example. Our largest group of comparison was white males confined to the gun topic, with a sample size of 20,531 for the smaller sample.

White Males – Gun

White males in the gun domain was the largest subsample. Two sets of parallel analyses were run, one on the entire corpus of images for white males in the gun domain, and a second set of analyses with the pitch and yaw data constrained to the limits proposed by Wang and Kosinski (2018). The results for two sets of analyses do not differ by much, and as such only the larger sample is presented here.

In order to compare white males following Everytown to white males following the NRA, the two subgroups needed to be equivalent in number of observations. To make these two groups of analysis even in sample size, the larger group was randomly sampled to reach the same number of subjects that were available in the smaller group. For every logistic regression analysis, 10-fold cross-validation was performed, a sampling procedure that ensures that every element of the data is part of both the train and test sets. For each model the data were standardized, and each logistic regression used a Least Absolute Shrinkage and Selection Operator (LASSO) penalty for regularization. LASSO regression is best utilized when attempting to reduce overfitting in a model, when features or columns might outnumber sample size, or when many of the components of the model can be

reduced to zero without losing much information (Maina, 2021; McNeish, 2015). Additionally, models employed a 'SAGA' solver, an optimization method related to stochastic average gradient (SAG) but with better convergence (Defazio et al., 2014). SAGA solvers are optimal for sparse regression matrices as well as large data, and are often the best choice for solvers according to sklearn documentation (Defazio et al., 2014).

For each model, metrics related to both accuracy and area under the curve (AUC) are reported as measures of classification power. Accuracy is defined as the ratio of correct predictions to total predictions, while area under the curve is the ratio of true positives to false positives. AUC is typically seen as a superior metric for model fit in comparison to accuracy, because models with high accuracy can sometimes be poor classifiers. Despite this, accuracy is perhaps more intuitive, so both are presented here. Error is presented in standard error of the mean (σ_M) of cross validated scores, and is a measure of how closely the model mean approximates the probable population mean.

Age, Pitch, Yaw, Emotions, Sex Typicality – 'Apyest'. Of the variables for these analyses, age was the least predictive, with the model demonstrating an average AUC = .505 ($CI_{95\%}$ = [.494, .517], σ_M = .0051). The accuracy of the model was 51%, not much better than chance. The model for pitch and yaw was a better classifier, AUC = .569 ($CI_{95\%}$ = [.562, .577], σ_M = .0033), accuracy = 55%. Using the emotional expression variables as predictors was comparable to the success of the pitch and yaw model, with AUC = .586 ($CI_{95\%}$ = [.578, .594], σ_M = .0036), achieving an overall accuracy in predictions of 55%. Sex typicality was the most effective predictor of these variables, AUC = .613 ($CI_{95\%}$ = [.598, .629], σ_M = .0067). Sex typicality proved an accurate classifier for 59% of the images.

These predictors were combined into one logistic regression model (age, pitch, yaw, emotions, and sex typicality, or together 'Apyest') with orientation being the criterion variable. This 'Apyest' model achieved a better AUC score than any of the previous models, AUC = .646 ($CI_{95\%}$ = [.632, .659], σ_M = .0060). Perhaps unsurprisingly, this model also had the best accuracy thus far, correctly categorizing 60% of the data.

ROC curve plots are presented in Figure 9a – 9e. For each of these plots, the dashed red line represents a model that performs no better than chance. Each of the semi-transparent lines represents one fold of the 10-fold cross validated model, while the solid blue line represents the mean across all folds. One can interpret the classification power of the model by assessing the curve of the blue line. Models that are better classifiers will have blue lines that arc towards the upper left corner of the plot, while models that are poor classifiers will have mean lines that hug the random chance line. Models with less error will have the fold lines tightly surrounding the mean line, while models with more error

will demonstrate a greater spread around the mean. ROC curve plots for all analyses are presented in Appendix L.

Figure 9

ROC Plot – Apyest – White Male Gun

Figure 9a - Age

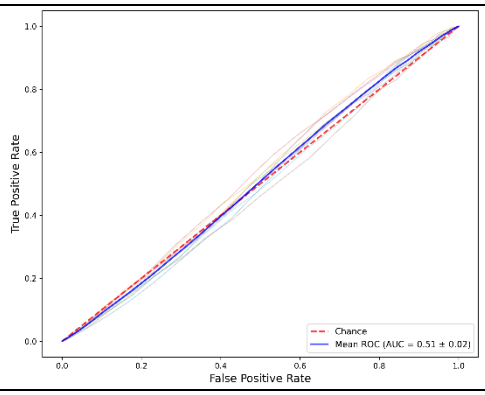


Figure 9b - Pitch, Yaw

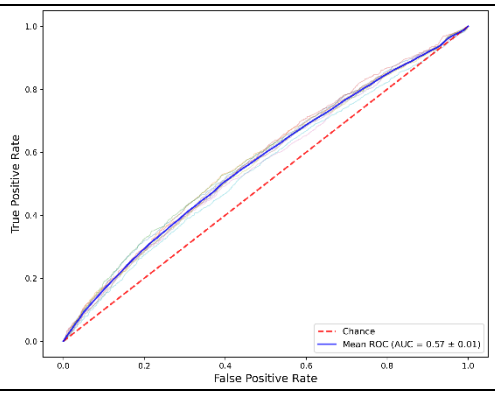


Figure 9c - Emotions

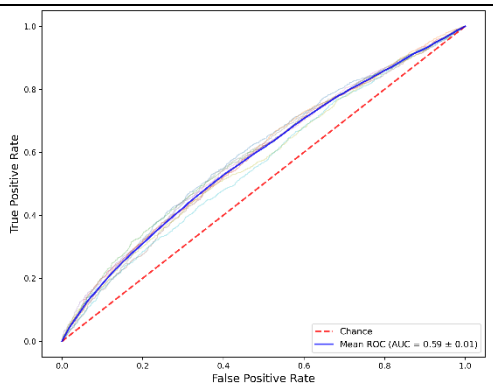


Figure 9d - Sex Typicality

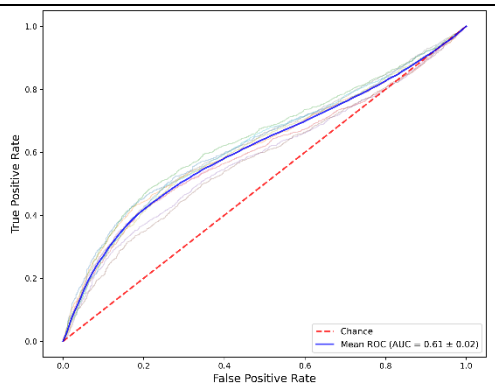
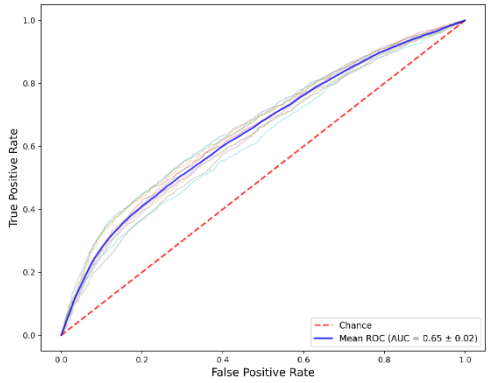


Figure 9e - Combined - Apyest



Feature Analysis. Recall that for each image, 4,096 features were extracted. Singular Value Decomposition (SVD) was performed on these features for just the group of interest (white males in the gun topic, in this case), leaving the 500 feature columns that were most important to classification for the group of interest.

Results from the cross-validated feature model were more impressive than any of the previous models, AUC = .746 ($CI_{95\%}$= [.736, .756], σ_M = .0044). Accuracy on the model using only features averaged 68% across all folds. Adding age, emotions, pitch, yaw, and sex typicality to the feature model did not improve classification substantially, AUC = .750 ($CI_{95\%}$ = [.740, .761], σ_M = .0045), accuracy = 69%. This confirms Hypothesis 1 for the white male gun analysis, replicating the primary finding presented in Wang and Kosinski (2018) and Kosinski (2021). Features alone were a superior classifier than the 'Apyest' model, and the inclusion of the 'Apyest' data to the feature model did not offer much new information in regards to classification.

The same analysis was performed on the masked image set. A total of 4,096 features were extracted from each masked image, and SVD was performed, reducing the feature set to the 500 most influential features. Comparing the classification power in the masked model to the classification power of the whole image model should reveal the importance of the background in image classification.

The removal of the background data did not appear to be very important to the model, causing only a minor reduction in predictive power, AUC = .735 ($CI_{95\%}$ = [.721, .748], σ_M = .0059). The accuracy of this model using only masked features was 67%, a slight reduction from the whole image feature model, but still far superior to the 'Apyest' model constrained on a sample by sex and race. Adding the 'Apyest' data to the masked

feature model had a negligible effect, AUC = .741 ($CI_{95\%}$ = [.727, .755], σ_M = .0063), accuracy = 68%. These findings confirm Hypothesis 2 for this comparison, and suggest that the background of the image is not particularly effective in regards to being a classifier, at least not in comparison to features related to the face in the image. See Figure 10a – 10d.

Figure 10

ROC Plot – Features – White Male Gun

Figure 10a - Whole Image Features	Figure 10b - Whole Image Features + Apyest

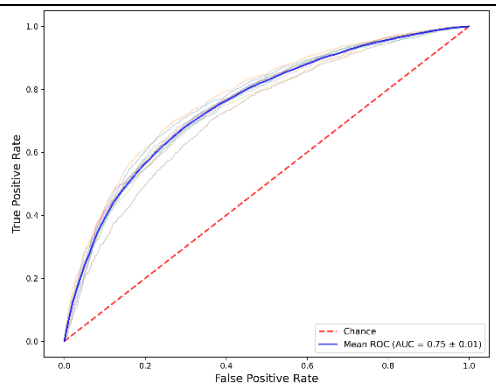
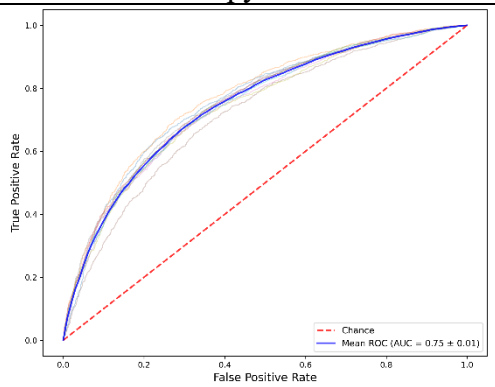
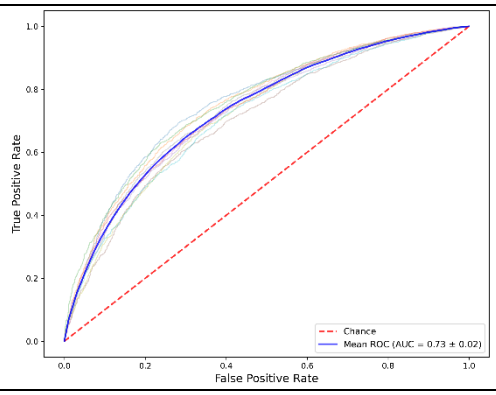
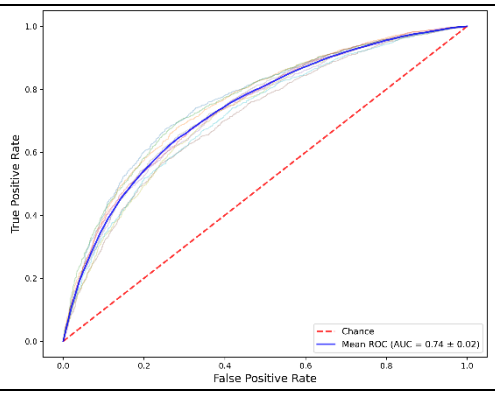

	

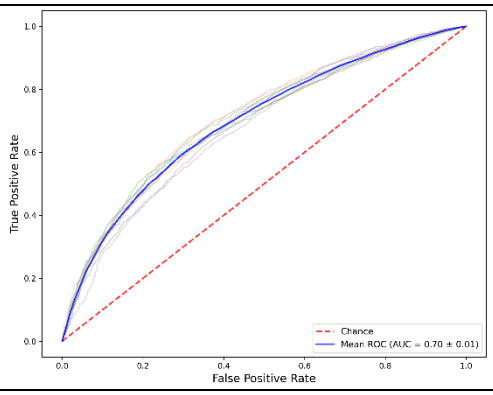
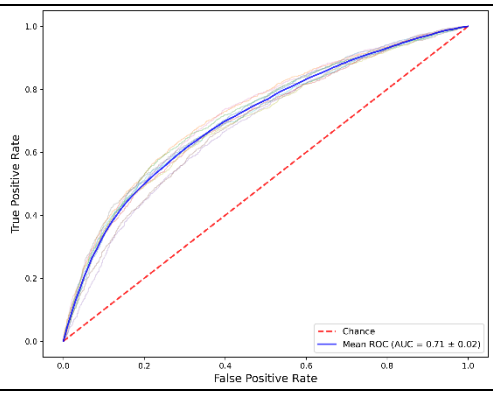
Figure 10c - Masked Features	Figure 10d - Masked Features + Apyest
	

Point Coordinates. Using the dlib library, 68 facial point coordinates were taken for each image. These point coordinates were then fit to a logistic regression model, utilizing only the point coordinates for prediction. By utilizing these point coordinates in such a manner, we can reduce or eliminate the influence of anything unrelated to feature morphology.

This model was successful in classification, AUC = .701 ($CI_{95\%}$ = [.689, .712], σ_M = .0049), although not as successful as the feature models. The accuracy of the model was 65%, slightly reduced from the feature models but far enough from chance to demonstrate that facial morphology is almost certainly influential in terms of model success, at least for white males in the gun topic, confirming Hypothesis 3 for this comparison. Adding 'Apyest' to the model improved metrics slightly but not dramatically, AUC = .710 (CI = [.698, .721], σ_M = .0051, accuracy = 66%). See Figure 11a – 11b.

Figure 11

ROC Plot – Facial Points – White Male Gun

Figure 11a - Facial Points	Figure 11b - Facial Points + Apyest
	

Previously, neural networks have been described as 'black box' systems, because researchers do not necessarily have easy access to their internal workings. In other words,

researchers often know their neural networks are working from the output of the model, rather than understanding each transformation the data is going through at each layer of the network itself. This provides an opaque understanding of how the classifier is coming to make its decisions. However, we might reverse engineer some data in order to see what the classifier is basing its decisions on.

To do this, the probability likelihood ratio of belonging to either the left or right orientations was captured for each image. These probabilities were then divided into quartiles. By comparing those images most likely to be classified into both left and right groups, that is, the first quartile compared to the fourth, one might get an idea of what differentiates left and right subjects, according to the classifier. Mean point coordinates are demonstrated for the left, right, and combined white male gun groups in Figure 12. Blue points and lines represent the mean point coordinates for subjects in the left-most quartile, while red points and lines represent those of the right-most quartile. Facial area in the left groups appears to be quite a bit smaller than in the right groups, as well as left groups demonstrating a more open mouth and right groups demonstrating a slightly shorter nose in length.

Figure 12

Facial Quartile Points Plot – White Male Gun

Figure 12a - Left (w/ right points)

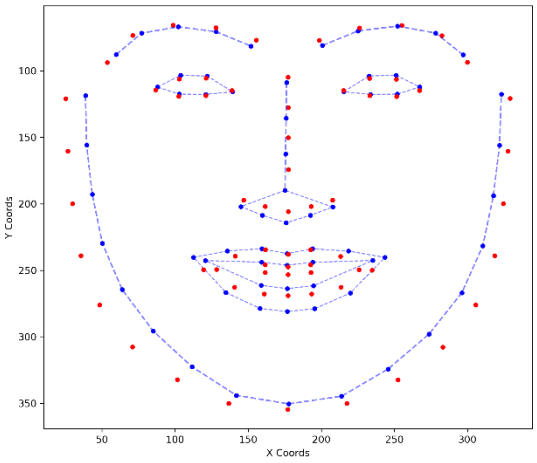


Figure 12b - Right (w/ left points)

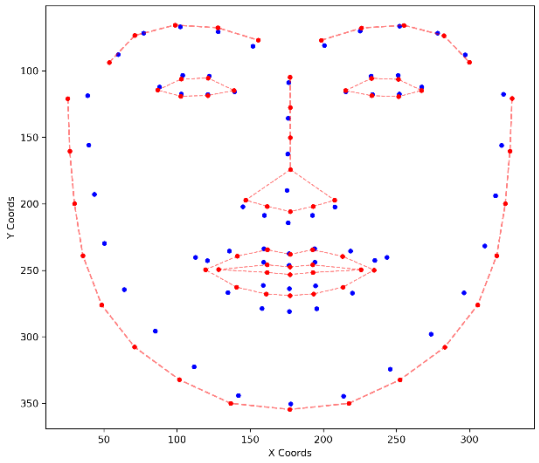
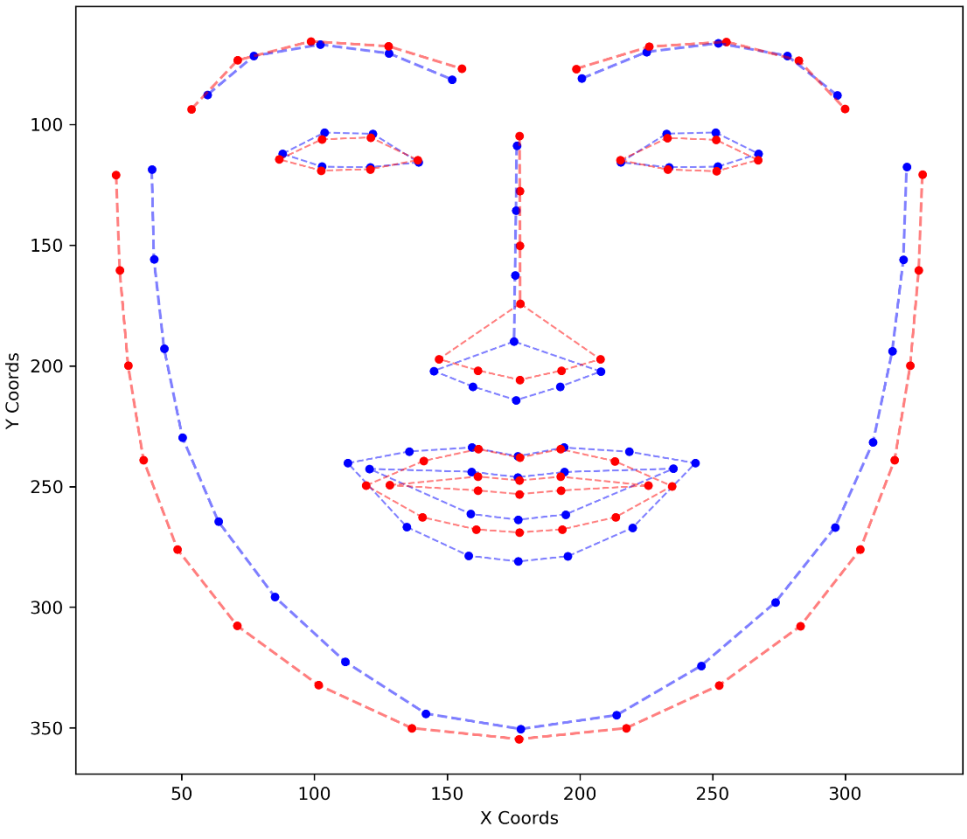


Figure 12c - Combined

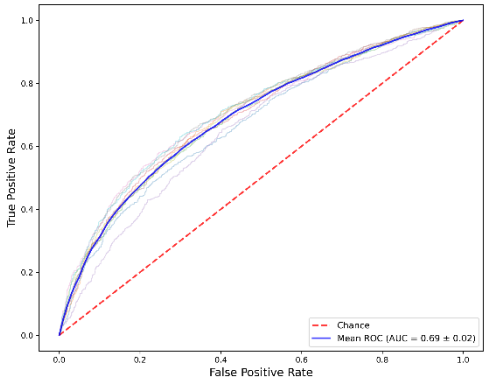
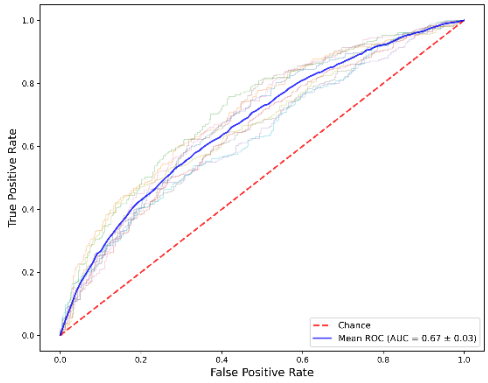


Because the mouth can vary so much as a function of emotional expression, points related to the mouth were eliminated to determine if a model relying upon facial morphology but not facial points would be successful in image categorization. Utilizing only the remaining points resulted in a rather negligible decrease in model success, AUC = .692 ($CI_{95\%}$ = [.681, .701], σ_M = .0045). The accuracy of the facial points without mouth coordinates was 64%. This confirms Hypothesis 4. Even when eliminating the mouth points, this model was still able to accurately categorize images, adding further support to the idea that images can be categorized by facial morphology alone. Adding the 'Apyest' variables to the no-mouth model resulted in a slight increase in correct classification, AUC = .704 ($CI_{95\%}$ = [.693, .715], σ_M = .0050), accuracy = 65%.

The two most prominent emotions demonstrated in the dataset were 'happy' and 'neutral', with over 83% of the sample being in either one of these groups. The point coordinate data were reduced down to those subjects the emotion classifier deemed as being 'happy', thus attempting to isolate on images demonstrating similar facial expressions across the white male gun domain. When constrained to just 'happy' subjects, model classification still proved to be successful, AUC = .696 ($CI_{95\%}$ = [.683, .709], σ_M = .0059). Model accuracy for only facial points on just happy subjects was 65%. Adding the 'Apyest' data generated a modest increase in classification efficacy, AUC = .707 ($CI_{95\%}$ = [.698, .717], σ_M = .0042), accuracy = 65%. See Figure 13a – 13b.

Figure 13

ROC Plot – Happy – White Male Gun

Figure 13a - Face Points - Happy	Figure 13b - Face Points - Neutral
	

Images with neutral faces were also isolated on in such a manner. Both the neutral point only model and the point model with 'Apyest' data proved to be accurate classifiers, AUC = .691 ($CI_{95\%}$ = [.675, .706], σ_M = .0068) and AUC = .699 ($CI_{95\%}$ = [.684, .714], σ_M = .0066), respectively. Accuracy of the neutral point model was 64%, while the neutral point model with 'Apyest' data had an accuracy of 65%. First and fourth quartile facial meshes were taken for both 'happy' and 'neutral' models and are displayed in Figure 14.

Figure 14

Facial Quartile Points Plot - Happy and Neutral - White Male Gun

Figure 14a - Happy - Left	Figure 14b - Happy - Combined	Figure 14c - Happy - Right

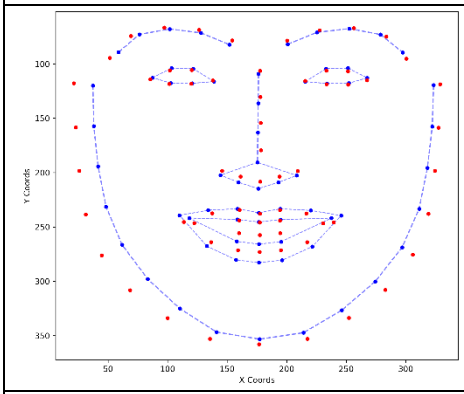
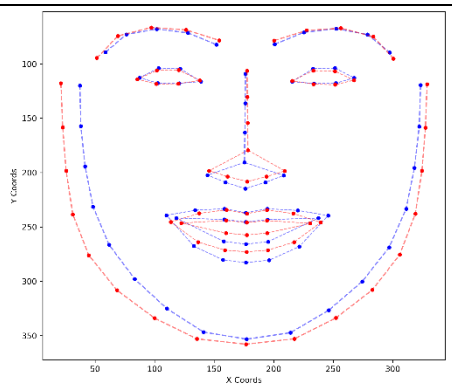
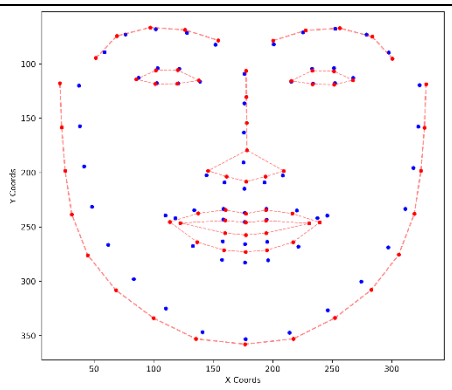
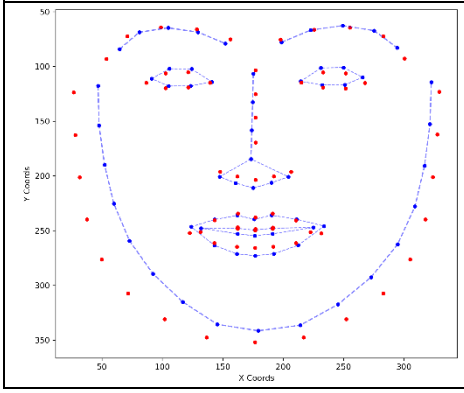
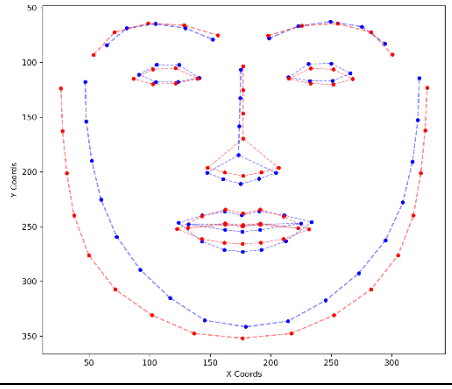
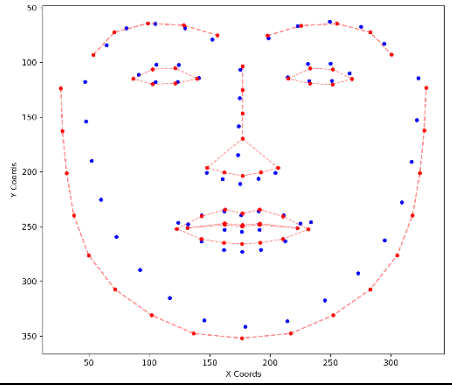

		

Figure 14d - Neutral - Left	**Figure 14e - Neutral - Combined**	**Figure 14f - Neutral - Right**
		

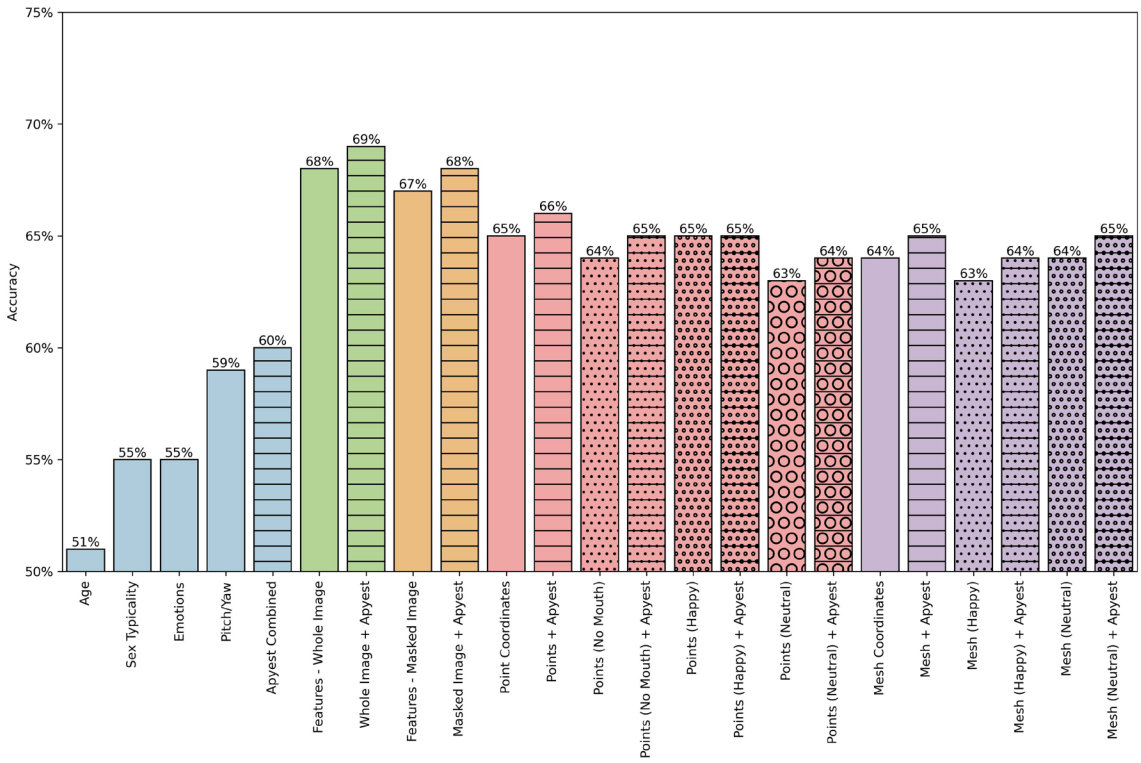
These findings confirm Hypothesis 5. Point models constrained by facial expression performed far better than chance, proving that images can be categorized by facial morphology alone and that facial expression cannot be the primary determinant of these models' success.

Mesh Coordinates. In addition to the point coordinate data, mesh coordinate data were also collected using the mediapipe library. In contrast to the dlib library which stores 68 facial points, media pipe stores 468 facial points to make its lattice. The X and Y coordinates for these 468 points were used as predictors in a logistic regression model. This model comprised only of mesh face coordinates was enough to accurately classify images 64% of the time, AUC = .691 ($CI_{95\%}$ = [.679, .703], σ_M = .0052), confirming Hypothesis 6. Adding the 'Apyest' variables improved classification slightly but not substantially, AUC = .704 ($CI_{95\%}$ = [.691, .716], σ_M = .0055), accuracy = 65%.

The mesh coordinates were also narrowed by facial expression, similar to the point coordinate data. With a sample of only happy subjects, the mesh coordinates were suitable predictors, accurately classifying 63% of subjects, AUC = .685 ($CI_{95\%}$ = [.677, .692], σ_M = .0034). Results for the same model with 'Apyest' variables were also positive, AUC = .698 ($CI_{95\%}$ = [.690, .707], σ_M = .0039), accuracy = 64%. Neutral subjects were also accurately classified at a rate of 64%, AUC = .692 ($CI_{95\%}$ = [.675, .710], σ_M = .0078), and the model with the 'Apyest' variables improved upon classification slightly, AUC = .705 ($CI_{95\%}$ = [.687, .723], σ_M = .0078), accuracy = 65%. These findings confirm Hypothesis 7, and prove that images can be classified when only utilizing mesh coordinates, even when controlling facial expression. Accuracies and AUCs are plotted for the all image, white male gun models in Figures 15 and 16. Accuracy and AUC plots are available for all other models in Appendix M.

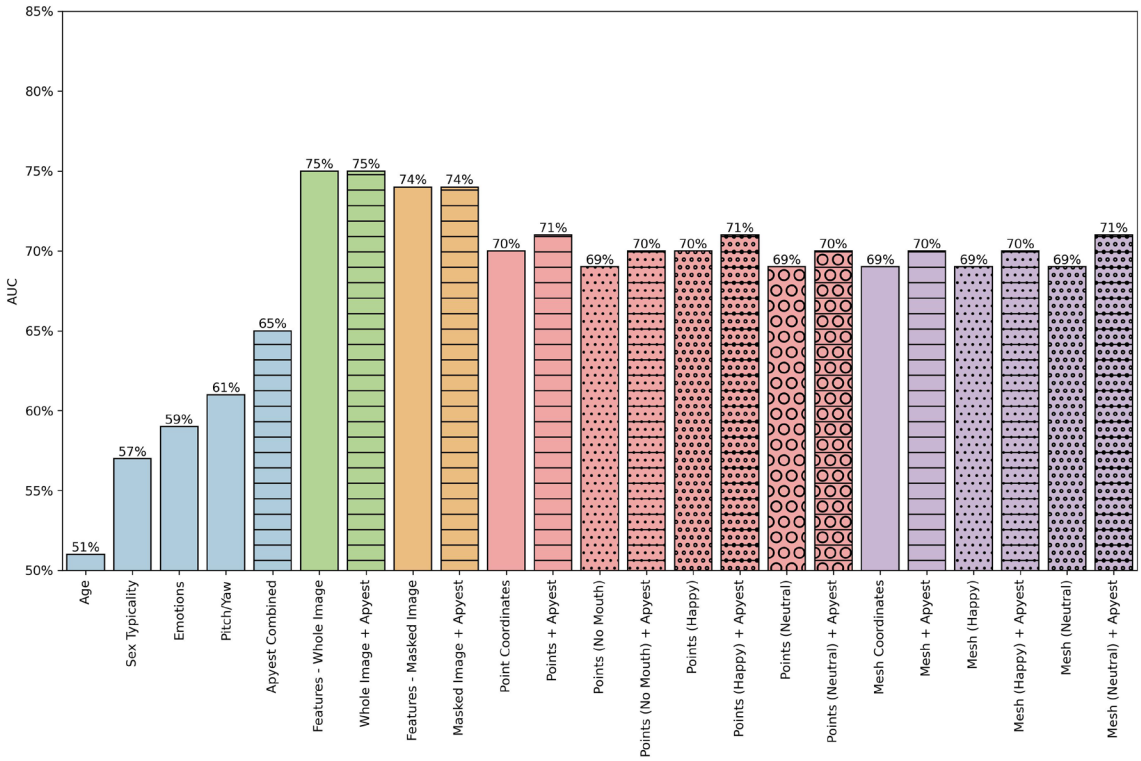
Figure 15

Accuracy – All Images – White Male Gun



Note: Blue bars represent the control variables and the combined 'Apyest' model. Green bars represent features of the whole image, while orange bars represent features of the masked image. Red bars represent point coordinates and purple bars represent mesh coordinates.

Figure 16
AUC – All Images – White Male Gun



Note: Blue bars represent the control variables and the combined 'Apyest' model. Green bars represent features of the whole image, while orange bars represent features of the masked image. Red bars represent point coordinates and purple bars represent mesh coordinates.

Summary – White Males – Gun. Results from this set of analyses confirm all of the hypotheses. First, the original conceptual effect from Wang and Kosinski (2018) and Kosinski (2021) was replicated, namely, being able to correctly classify images from their features alone. Features were a more powerful classifier than any of the other variables recorded. Further, removing the background from the images appeared to have little effect on classification power.

There were also several strong indicators that the classifier was utilizing facial morphology. Utilizing only facial point or facial mesh coordinates resulted in a decrease in classification power; however, the classifier was still able to categorize images well

above chance levels. This was also the case across all of the analyzed subgroups (the no-mouth subgroups, the happy subgroups, and the neutral subgroups). These analyses provide the strongest support yet that feature only models are almost certainly utilizing facial morphology when classifying images.

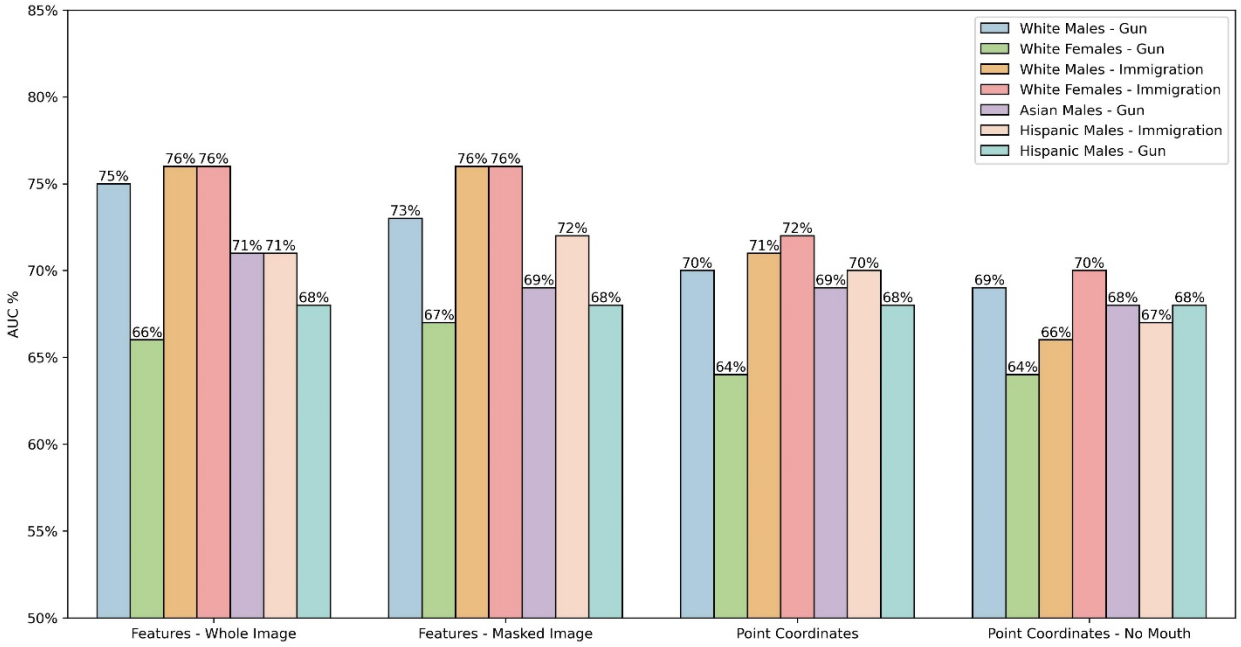
All Groups

It is possible that the positive findings for the white male gun group are an aberration, and that the effects would not translate to the other groups of analysis.

In an attempt to determine if this methodology is consistent across all comparisons, the same analyses were performed on the remaining six groups of analysis. If the methodology is sound, we should expect a similar ability to classify subjects' political orientation across the rest of the comparisons using only facial morphology. If, however, the methodology is dependent upon the specific comparison in question, the negative results from these analyses should also be revealing.

Results for the models across all comparisons was confirmatory. Of particular interest is the comparison of the feature models to the masked models as well as to the point and mesh coordinates. AUCs are presented for all comparisons in Figure 17. Metrics for all models are presented in Appendix N.

Figure 17
AUC – Features and Point Coordinates – All Groups

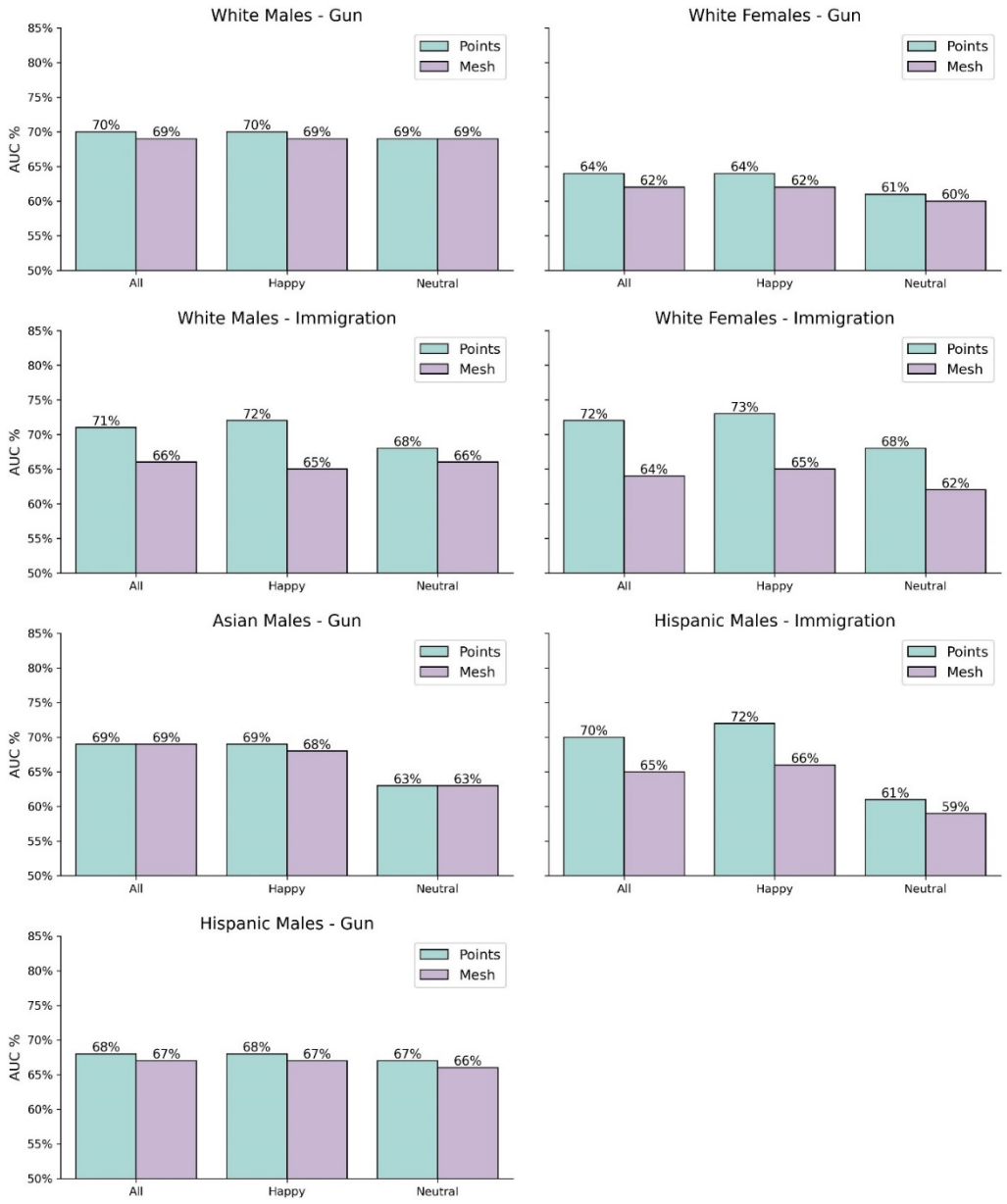


Several things are clear from the bar plot. First, every group of interest was able to be classified successfully using only features, replicating the principal finding across all conditions and confirming Hypothesis 1 for all groups of analysis. Second, removing the background from the image had little effect on classification power in any of the groups of analysis, proving that the background is not particularly informative in regards to classification and confirming Hypothesis 2 for each comparison. Third, the point coordinate models performed well above chance across every comparison, proving that images can be classified using only facial morphology, and that the initial confirmatory results were not exclusive to the white male gun comparison. This confirms Hypotheses 3 for all groups of analysis. Finally, even when removing the points related to the mouth, the classifiers were still able to accurately categorize images at rates well above chance, confirming Hypothesis 4 for all groups of analysis.

Regarding classifier quality, utilizing computer vision to attain point coordinates allowed for better classification than using the mesh coordinates, although both performed far better than chance. Across every analysis, the point coordinates performed similarly or better than the mesh coordinates. See Figure 18.

Figure 18

AUC – Point vs. Mesh Comparison – All Groups of Analysis



Examining Figure 18, we can draw several conclusions. First, point models constrained by happy and neutral facial expressions were solidly predictive, confirming Hypothesis 5 for all groups of analysis. Constraining the images by facial expression sometimes reduced model accuracy slightly, in particular for neutral facial expressions and only in some comparisons. However, each model demonstrated an AUC well above chance, and many of the comparisons demonstrated relatively consistent AUCs across conditions.

Figure 18 also illustrates that using only the mesh models was sufficient to categorize images, confirming Hypothesis 6 for all groups of analysis. Mesh models constrained by facial expression were also able to be classified, confirming Hypothesis 7 for all groups of analysis.

From these results, it is evident that the classifier was effective at classifying followers across all groups of analysis. However, it is still unclear if the effect is similar across groups. For example, it was previously discovered that, for the white male gun comparison, the facial area for conservative subjects was larger than the facial area for liberal subjects.

This effect was duplicated across all male groups of analysis. Males across all five of the comparisons demonstrated the same effect in regards to facial area. Conservative male subjects have a larger facial area, in general, than do liberal male subjects, independent of either the topic or of the racial background of the subject.

However, this effect was inverted across the female samples, with conservative women demonstrating, on average, less facial area than their liberal counterparts. The

effect is more difficult to observe in the female immigration sample; nevertheless, liberal female faces remain longer and wider on average than conservative female faces.

Neutral facial point images for all groups are presented in Figures 19a – 19d and 20. All facial point images are presented in Appendix O.

Figure 19

Facial Quartile Points Plot - Neutral – White Males and Females

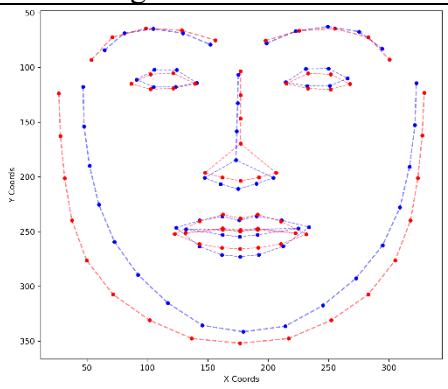
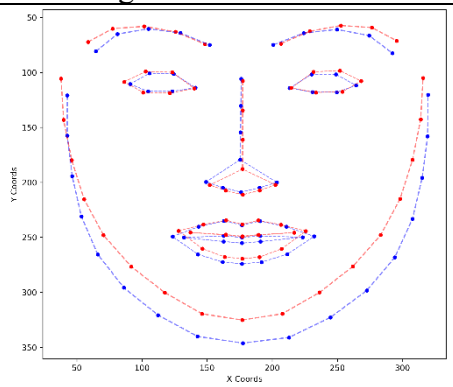
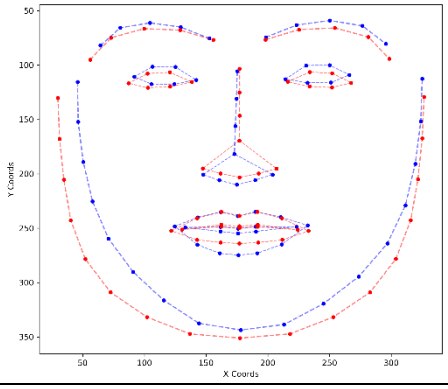
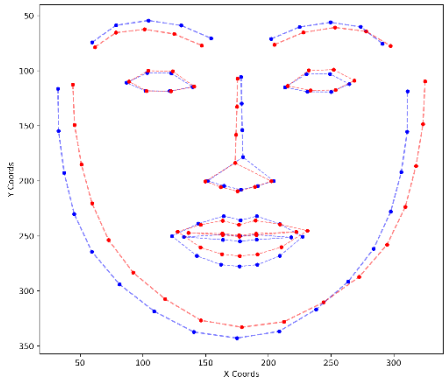
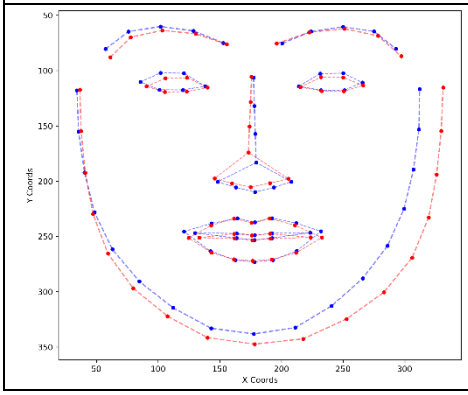
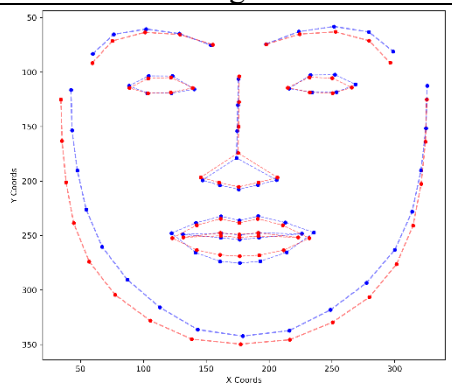
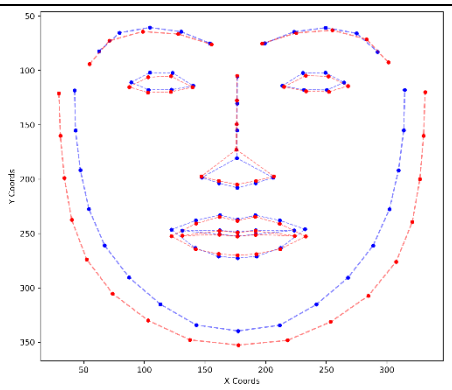
<table>
<tr><td colspan="2">White – Gun – Neutral</td></tr>
<tr><td>Figure 19a - Males</td><td>Figure 19b - Females</td></tr>
<tr><td> 
</td><td> 
</td></tr>
<tr><td colspan="2">White – Immigration – Neutral</td></tr>
<tr><td>Figure 19c - Males</td><td>Figure 19d - Females</td></tr>
<tr><td> 
</td><td> 
</td></tr>
</table>

Figure 20

Facial Quartile Points Plot - Neutral – Asian and Hispanic Males

	Males – Neutral	
Figure 20a - Asian – Gun	Figure 20b - Hispanic – Immmigration	Figure 20c - Hispanic – Gun
		

Finally, classification ability was markedly lower in the white female – gun comparison in comparison to the other comparisons. While it is unclear as to why this is, it is possible the sample for this particular subgroup was less 'clean' than the other groups, resulting in more noise in the model. Contrastingly, it is possible that the subjects in the white female gun groups simply demonstrate less variation between left and right groups. Further research is required to make the determination.

Chapter 10: Discussion, Implications, and Limitations

This dissertation contributes to a small but burgeoning line of research that explores the role that features play in image categorization. Several conclusions can be drawn from the research presented. First, features served as adequate predictors across each of the seven groups of analysis, replicating the main finding from previous research of this nature and confirming Hypothesis 1. Subjects were able to be sorted into social categories at rates far exceeding chance merely by using features extracted from their facial images. More narrowly, this study replicates the central finding of Kosinski (2021), which found that subjects could be classified by their political ideology utilizing only features extracted from images. As in Kosinski (2021), models created on extracted features had the highest accuracy ratings of any of the models run.

Second, removal of the background did not have much effect on model accuracy across any of the groups of analysis. Previous research has found that images can be correctly categorized through only the background of the image (Wang, 2022). However, the accuracy of the no background feature models were only slightly reduced in comparison to the whole image feature models, indicating that the background had little to do with the latter's classification accuracies. These findings confirmed Hypothesis 2 for all groups of analysis.

Third, models that used only facial points, and thus only facial morphology, were still able to classify images well above chance for each of the seven groups of analysis. Similarly, models that relied upon facial morphology but without the mouth points were still predictive at rates above chance, as were models that controlled for happy and neutral facial expressions. These findings make certain that facial morphology can be used to

classify subjects into ideological subgroups, confirming Hypotheses 3, 4, and 5 for each group of interest.

Fourth, similar to the point models, models that utilized only the mesh points were accurate classifiers at rates well above chance. Importantly, this replicates the finding for point coordinates, providing incontrovertible support for the idea that images can be categorized into political subgroups utilizing only the facial morphology of the subject in the photograph. When controlling for facial expression, mesh models were still accurate predictors at rates above chance for all groups, although the predictive power was slightly reduced in comparison to the point coordinates. These findings confirmed Hypotheses 6 and 7 for all groups of analysis.

Group Differences

Despite the relative consistency in these findings, the groups analyzed demonstrated some differences as well. For example, age was an excellent predictor across the immigration subgroups but a mediocre one for subgroups related to the gun topic. Sex typicality was a strong predictor in male gun subgroups, a modest predictor in male immigration subgroups, and no better than chance in female subgroups. Also, in spite of having the second largest sample size, the accuracy rate for the white female gun group was substantially lower than that of the other groups.

These findings suggest that groups with similar political orientations might differ in dramatic and unforeseen ways from one another. This should evoke a modicum of skepticism in the reader in regards to attempting to apply previously trained models to novel images, in spite of the positive results demonstrated here. Researchers attempting to

categorize images in such a way should anticipate modest differences between common subgroups, even if the presumption is the subjects are from the same underlying population.

Sample Size

Two topics in regards to sample size bear discussion. First, as stated previously, this methodology was expected to be subject to sample size in making its predictions. In the same way this research took guidance from previous research in order to assess the minimum sample size for inclusion, future researchers might look to this manuscript for guidance regarding sample size in performing their own physiognomic research.

Second, an examination of the models' sample sizes might reveal some characteristics about the nature of this data that might be illuminating in other capacities. Specifically, critics of this type of analysis have warned of dystopian futures resembling that of "Black Mirror" or "Minority Report" (Resnick, 2018; Levin, 2017). The tacit implication behind these foreboding predictions is that technological gains over time will result in faster processing speeds and better methodologies for classifying images in the future. In the same way that neural networks were once theoretical but are now commonplace, some might have concerns that future leaps in technology will be better at interpreting these modest effects and classifying people into social categories without their consent.

To facilitate discussion of both of these topics, see Figures 22a-b. The below plots represent the AUCs of the previously presented models by their sample size. Regression lines were plotted for models of the same type. Typically what one might expect to see is that the classification accuracies improve as sample sizes increase. This occurs because models will tend to become better classifiers as more data is provided, broadly speaking.

Figure 21

Model Type by AUC by Sample Size

Model Type by AUC by Sample Size

Figure 21a - Features, Point and Mesh Coords

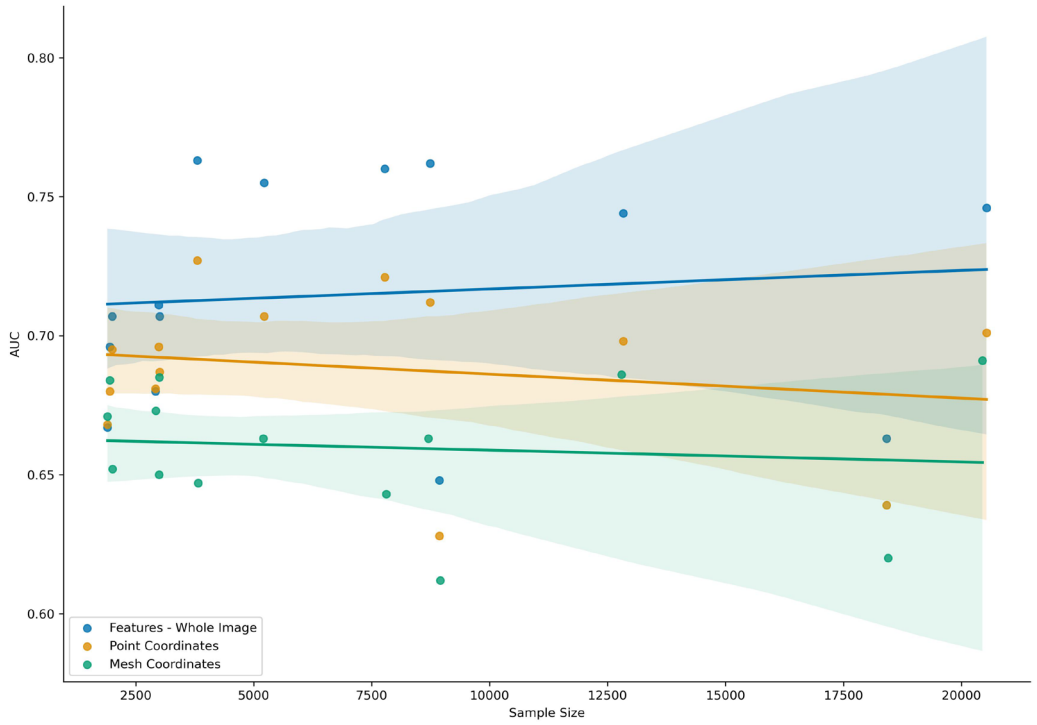
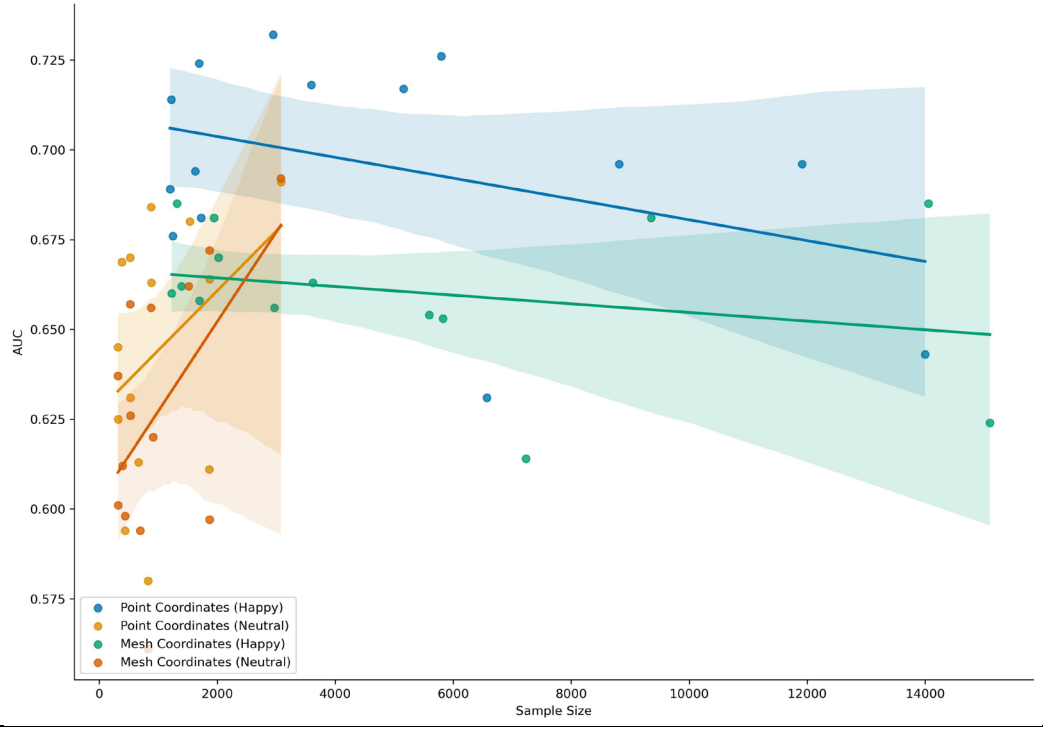


Figure 21b - Happy/Neutral Point and Mesh Coords



Three conclusions can be drawn from the figures presented. First, the models utilizing features, point coordinates, mesh coordinates, and happy subjects did not appear to suffer much in regards to their classification as their samples decreased in size, as indicated by the regression lines for these models. Conversely neutral models took the biggest hit in classification accuracy from reductions in sample size, but these findings occurred in conditions where sample size was most constrained in this dataset. This could offer some explanation as to why the error for neutral models is a fair amount larger in comparison to the other models. Despite this limitation, it is clear from this analysis that the classifiers were utilizing facial morphology as demonstrated by the facial point models. Regardless, as a general rule, it appears that for most of the models, a sample size of 2,000 images per group was sufficient to achieve success in image categorization. Researchers attempting this methodology in the future should aim for this figure in order to achieve the statistical power necessary for the best classification rates possible.

Second, it appears as though increasing sample size above 2,000 did not result in increases in classification accuracy, even as those sample sizes increased by an order of magnitude. Indeed, some model types even decreased in classification accuracy as sample size increased. This observation implies that the presented mean lines for model AUC might not differ very much from the mean lines for the broader population. In other words, there is evidence to suggest that the morphological effects presented in this analysis appear to be quite small *in reality*. If true, this necessarily means that it is unlikely that increases in technology will provide for more accurate classifications in the future, because the data itself only provides a small effect with which to classify images.

Finally, researchers attempting to ascertain the true population mean might require a very large sample size to see their model errors reduced to a number approaching zero. See Figure 23. Such scholars would likely require a sample size approaching 30 or 40 thousand images for feature, point, or mesh models in order to expect very low error rates. These same scholars might require a sample size of around 20 thousand images when utilizing models constrained by facial expression, although more research is required to be certain, in particular for neutral point and mesh models.

Figure 22

Model Type by Error by Sample Size

Model Type by Error by Sample Size

Figure 22a - Features, Point and Mesh Coords

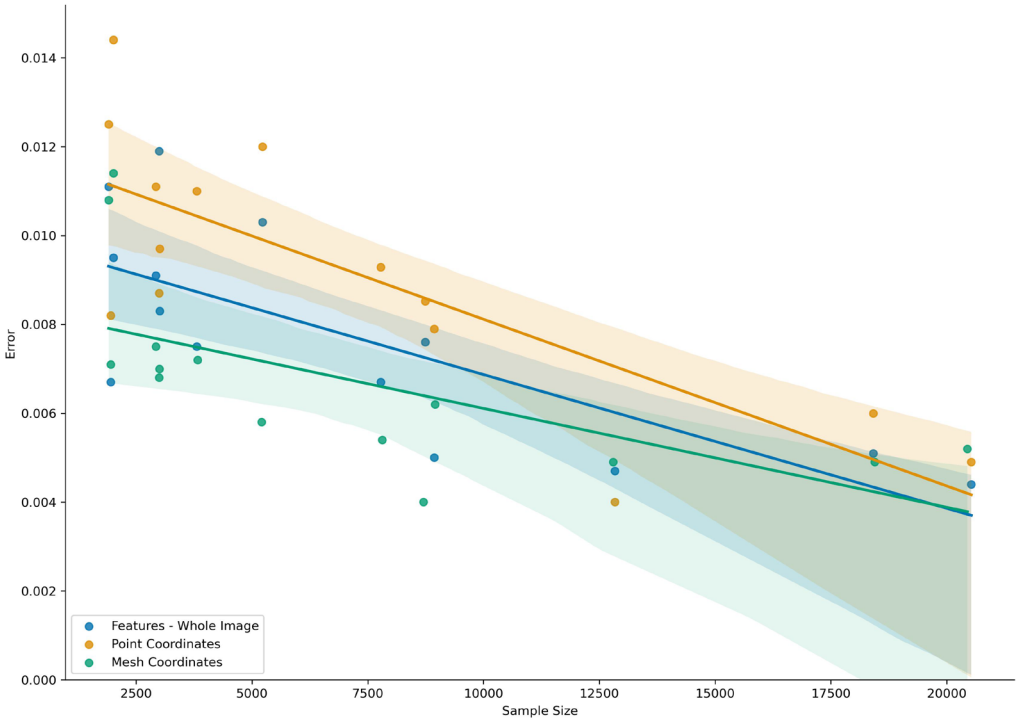
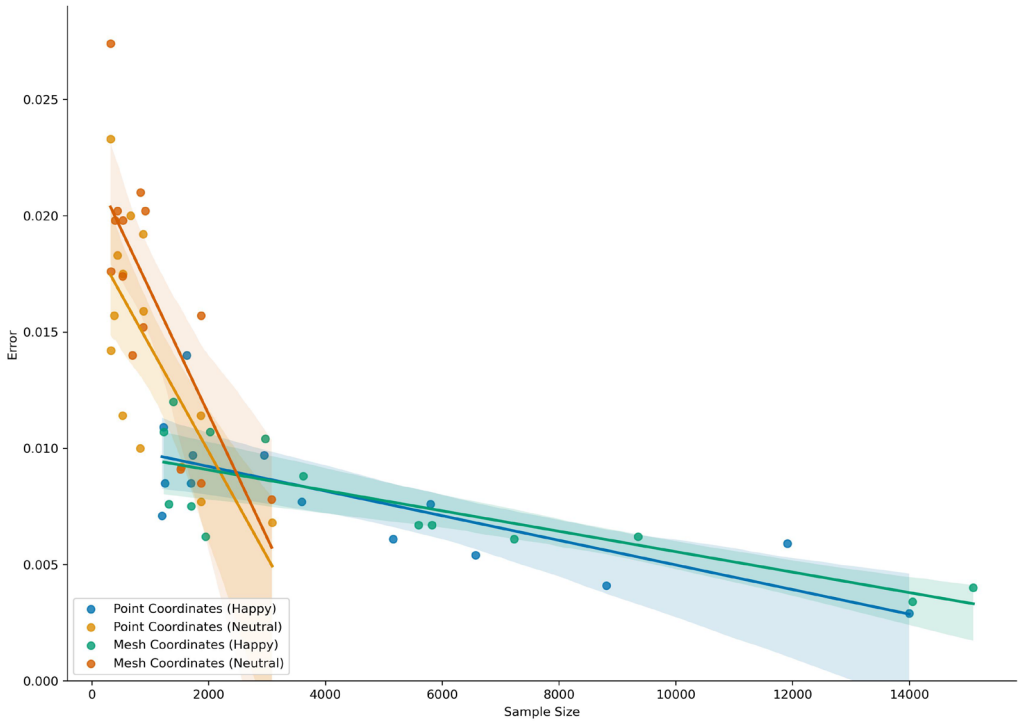


Figure 22b - Happy/Neutral Point and Mesh Coords



Benefits

This research offers several benefits in comparison to previous research performed in this domain. Because all of the information used in these analyses was from the photographs themselves, this study mimics the conditions for how this technology might be employed on novel data in the future. For example, neural networks might be employed to process a variety of facial information in order to make classifiers more accurate, including determining the sex, race, or age of the subject in the photograph. Self-updating programs might incorporate this information into their algorithms for classification, fine-tuning models slightly over time as more subjects are introduced to the algorithm.

Additionally, unlike previous research, this paradigm provided a clear demarcation between facial morphology and transient facial features. Through the use of facial point and facial mesh coordinates, this research unambiguously eliminated virtually all of the

influence of transient facial features, providing the strongest evidence to date that subjects can be classified into social categories utilizing only their facial morphology. While Kosinski (2021) attempted to control for many of the transient features in his images, he failed to isolate on facial morphology. Prior claims that facial morphology was implicated in image categorization were largely speculative. In contrast, this work provides clear evidence that facial morphology is implicated in image categorization.

This study is also more inclusive than any prior work in this field, a criticism of previous research (Anderson, 2017). By including two topics of analysis as well as the incorporation of multiple racial groups, this research represents the broadest application of the technique yet performed.

Limitations

Perhaps the most important limitation regarding this study is that the sample undoubtedly contains some amount of error. For example, most images containing multiple people in the frame were eliminated automatically by the algorithm, incorporating an element of selection bias into the sample. To illustrate one such potentiality, it is possible that extraverted individuals are more likely to take pictures in groups than introverted individuals. If true, extraverted individuals were disproportionately removed from the sample. More broadly, it is possible that the images removed in the categorization process differed from the remaining sample in many ways that are opaque. Future analyses that are more inclusive of these removed participants could differ in their results.

At the same time, despite our respectable ICC score, qualitative image removal was somewhat subjective. While it is relatively easy to remove images of the former or current president, of a famous celebrity, or of a historical figure, some images were much more

ambiguous. As such, it is very likely that this sample contains some images that are not of the account holder. Despite these sources of potential bias, these analyses appear to be quite resilient, even in the face of some sampling error.

In terms of the methodology, success in image categorization tacitly implies that the concerns about minority persecutions presented in Wang and Kosinski (2018) are both legitimate and prescient. However, it remains true that the effects uncovered in these analyses are quite small and certainly unreliable in regards to attributing characteristics to any particular individual. While Wang and Kosinski (2018) laud classification rates higher than 90%, those scores occurred while using a very specific set of images that was well labeled as well as utilizing a forced choice paradigm where each image had a 50% probability of belonging to one category or the other. No other credible research has claimed classification accuracies so high utilizing this methodology. Researchers taking images 'from the wild' should anticipate a sizeable reduction in classification ability, especially when applying previously trained models to novel images.

Additionally, this process was quite computationally expensive and intellectually challenging to implement. While the accuracies are impressive in a theoretical capacity, it is important to note that these analyses could have 100% accuracy by simply recording the information about which account a subject followed on Twitter. Companies attempting to use this technology as a means to advertise to potential customers could almost certainly achieve better 'hit' rates through less complex means. Any classification information that might be gleaned by such a process could result in a sufficiently high false positive rate that motivated marketing could have the opposite effect intended. For example, if discovered and brought to public attention, companies targeting individuals in such a way

could face severe backlash from consumers in the form of negative publicity, boycotts, and protests. It is unlikely that business owners will engage in such a risky practice just to identify potential customers at rates modestly better than a coin flip, especially when directed ad platforms such as Google and FaceBook appear to be quite successful in their targeting by comparison. As such, this is a very meandering way to achieve an accuracy rate far below the simpler method of just observing which group subjects came from.

A further limitation of this research is that it relies heavily upon the DeepFace algorithm to retrieve information regarding the subject in the image. Despite reaching accuracy levels that mimic that of human beings in face matching tasks, previous research has shown that classifier algorithms like DeepFace demonstrate considerable variation in classification accuracy depending upon the race and sex of the target (Buolamwini & Gebru, 2018). Further, these types of systems have been shown to vary in their classification decisions based upon transient properties of the images such as brightness or contrast (Cavazos et al., 2021). As such, we can anticipate some error in our classifications, and that error is likely exacerbated in minority conditions.

Explanations for the Effect

Taken in the aggregate, this research represents the most comprehensive scrutiny physiognomy has seen to date. Given the strong confirmatory results, any scholar might be tempted to posit theoretical reasons as to why this research uncovered the findings that it did. For example, it is possible from an evolutionary standpoint that "wearing" one's political ideology on one's face might assist the broader population in selecting leadership during peace-time or war-time conditions (Little, 2012). If this is true, those populations that were more adaptive in determining their leadership would have been more successful

in passing their genes along to future generations. Additionally, it is possible that morphological indicators of underlying attributes are beneficial in terms of mating strategy, with individuals demonstrating shared political or personality characteristics providing for more fitness in their offspring across the aggregate.

An alternative explanation might be that conservatives tend to espouse traditional gender roles, and that those individuals demonstrating the greatest sex typicality in terms of gender roles are genetically predisposed towards conservative ideology (Duncan et al., 1997). In other words, it is possible there is some common underlying genetic link between the morphological differences we have observed and attitudes towards gender role adoption more broadly.

Further still, it is possible that individuals vary in their morphological sex typicality, and that those individuals demonstrating more traditional forms of sex typicality are more likely to espouse system justifying ideals simply because they are the beneficiaries of that system (Jost et al., 2004). Males with more masculine features and females with more feminine features could gravitate towards conservative forms of ideology specifically because endorsement of system change could threaten their respective positions within current social hierarchies.

One could continue in this effort, proposing a comprehensive slew of post hoc rationales that each fit the pattern of these findings. Locations might differ by political orientation but also by regional eating habits, with the morphological differences we are seeing being due to regional weight differences rather than political ideology (Morland et al., 2002). Individuals exhibiting similar facial structures might conform to one anothers' behavior or attitudes beginning in early life, creating clusters of individuals demonstrating

similar political position taking and also having similar facial structures (Morgan et al., 2015). Conservative ideology might be more likely to value traditionally masculine or feminine facial features in individuals, with individuals demonstrating those qualities being motivated to subscribe to the ideology through some type of operant conditioning during interactions with conservatives (Duncan et al., 1997; Staddon & Cerutti, 2003). Because liberal people demonstrate more openness to experience in terms of Big Five personality characteristics, they may be more inclusive in their group membership, creating a sample morphologically different from conservatives who are potentially more exclusionary (Gerber et al., 2011). Liberals and conservatives could have received different quantities of hormones in the womb, a potential common precursor to both morphological and attitudinal divides between these groups (Wang & Kosinski, 2018). Individuals with certain types of facial features might observe the behaviors and attitudes of others with those same facial features and adopt liberal or conservative ideology vicariously through modeling (Bandura, 1965). As one can see, there are an abundance of potential explanations for the findings, and any particular researcher might feel more persuaded by the theory that matches the field of study that they are most engaged in.

However, not all of these theories can possibly be true concurrently, and the preponderance of potential rationales for explaining these findings should invoke caution in the endorsement of some to the exclusion of others that might have equivalent plausibility, including those left unmentioned here.

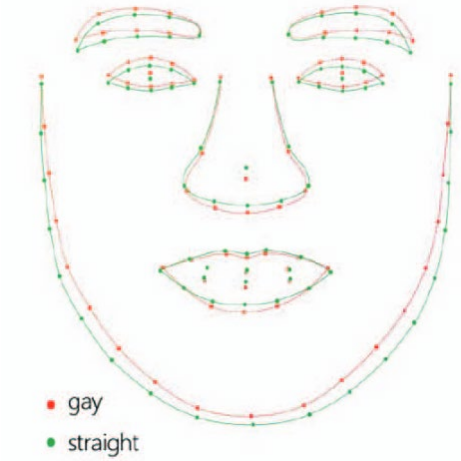
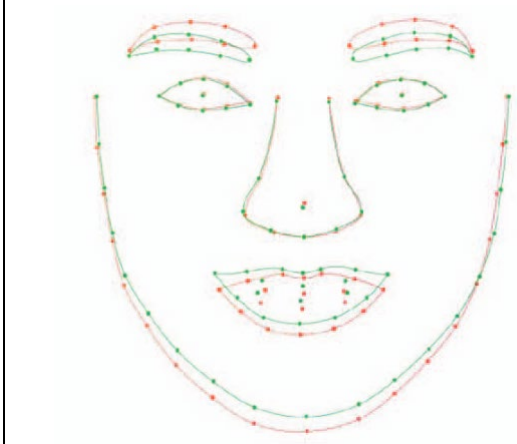
Putting that aside, there are several difficulties in attempting to explain these findings from a theoretical perspective. First, the analysis demonstrates that facial morphology is different between liberals and conservatives; however, it reveals nothing

about *why* facial morphology differs between liberals and conservatives. To propose any explanation would be to project our own belief systems onto data that is otherwise ambiguous in this regard. To illustrate this clearly, one only need recall that historical physiognomists also observed differences in people's facial morphology. These differences were then used as evidence to reinforce existing social hierarchies, namely that some groups of people are genetically inferior to other groups or that some groups of people are predisposed to moral turpitude. Deriving concrete conclusions from ambiguous data should be considered dubious, even when proposing explanations that have nothing to do with eugenics or racial superiority.

Second, although this research utilized different categories of classification, the findings demonstrate a striking similarity to those found in Wang and Kosinski's (2018) analysis of heterosexuals and homosexuals. From their 'Average facial landmark' plot (p. 251), one can observe that homosexual men tended to have less facial area than heterosexual men, while the inverse was true for women. See Figure 21a and 21b. Given such results and based upon the standards set by Wang and Kosinski (2018), this paper too could claim that the prenatal hormone theory of development is implicated in the differences between liberals and conservatives in regards to their facial morphology, despite this research demonstrating no positive or negative support for such a conclusion. In other words, because these findings match that of Wang and Kosinski (2018), the only thing preventing the application of their explanation to these findings is that the theory lacks any sort of face validity in regards to this sample. However, the fact that a theory demonstrates some amount of face validity is, in itself, insufficient evidence to support that theory's veracity.

Figure 23

Average Facial Landmarks – Wang and Kosinski (2018), p. 251

Figure 23a - Gay vs. Straight Males	Figure 23b - Gay vs. Straight Females
	

Third, physiognomy remains a highly controversial topic, and for good reasons. Modern people tend to be highly suspicious of anything that even hints of genetic determinism, despite the fact that genetics undoubtedly influence our trajectories in a myriad of ways (Funk et al., 2013; Alford et al., 2005; Hatemi & McDermott, 2012). For this reason specifically, great lengths have been undertaken to remove as much ambiguity and interpretation from this analysis as possible. Despite the rather cogent analysis performed by Wang and Kosinski (2018), the authors were roundly criticized by news media and activist organizations. Much of this criticism centered around their use of PHT as an explanation for their positive findings (e.g., Murphy, 2017; Resnick, 2018; Agüera y Arcas et al., 2018, as a start). By overextending themselves in their conclusions, the authors opened themselves up to criticism unrelated to the scientific veracity of their methodology, a mistake left unreplicated here.

Finally, researchers in this arena should exercise some caution in drawing firm conclusions from their results, especially given reports of the replication crisis in social

psychology specifically and in academia more broadly (Earp & Trafimow, 2015). Had the results of these analyses been non-confirmatory, this manuscript would have reflected those findings without hesitation. As such, it seems unfair to now propose a theory claiming explanatory power regarding a phenomenon that is quite poorly understood, even by those most familiar with it. Further, drawing seemingly innocuous deductions from ambiguous data could result in incorrect conclusions, a legitimate threat with this type of research in particular. The difference between 'acceptable' explanations and 'non-acceptable' explanations for these morphological differences are, at the current time, entirely subjective due to a lack of supporting evidence. This subjectivity is precisely what made previous physiognomic work so unjust, at least in comparison to the methodologically rigorous observation of modest differences in people's facial morphology.

Government Intervention

Previous researchers in this arena have warned of a dystopian future where facial features are used as a means to classify people in order to persecute certain populations. While this could potentially occur, it seems unlikely for several reasons. First and most importantly, the false positive rates should be sufficiently high so as to discourage classification in such a manner. From a pragmatic perspective, targets of government persecution can be found much more easily and with greater accuracy (Fox, 2020). Second, if any government is threatening its citizens with violence, the problem is the violence, rather than the means by which that violence occurs. Third, there are much easier ways to get more reliable information on individuals. Fourth, the effect sizes in this study appear to be relatively small, suggesting that increases in processing power in the future will not

necessarily result in greater model accuracy. Fifth, even if a model classifies its subjects correctly 100% of the time in regards to previously labeled images, there is no confirmatory evidence that future subjects applied to this same model would be accurate in their predictions. In other words, even if a perfect classifier is somehow eventually created, it would not mean that the classifier will be perfect when applied to novel data where group belonging is in doubt.

In spite of these limitations, several calls have been made for government intervention, oversight, and regulation in order to curb the potential negative effects of this type of research. For example, Wang and Kosinski (2018) write "Delaying or abandoning the publication of these findings could deprive individuals of the chance to take preventive measures and policymakers the ability to introduce legislation to protect people" (p. 255). Wang (2022) argues "that the burden of privacy protection should not be shifted to the consumers, but must be initiated by governments and companies" (p. 48). However, as of this writing, no one has attempted to describe exactly what this potential legislation might look like.

It must be acknowledged that any hypothetical legislation proposed to curb research of this type in the future would have an abundance of downsides. First, limiting access to information through APIs would cause a lot of applications to stop working. While a typical end user might not need access to hundreds or thousands of profile images, developers do require unfettered access to this information to make their applications work. Having the government distinguish between 'good' and 'bad' developers leaves enough latitude that one might anticipate unfair treatment or legislative favoritism.

Second, restricting the ability to upload or download images freely borders dangerously on restricting freedom of speech as well as freedom of association. Given that this technology has not been shown to actively threaten anyone at this time, proposing such restrictions on individuals or companies might be a case where the cure is more damaging than the illness.

Third, Americans have often been shown to be skeptical of their own leadership, particularly at the national level. One need only look at the favorability ratings of the two presidential candidates in 2016 to demonstrate that people often, broadly speaking, lack faith in their national representation (Saad, 2016). Regardless of any individual political leader, there should be some skepticism as to whether politicians could or should be restricting speech and association in such a manner. While citizens might desire the government to work in the best interest of the people, it seems at least plausible that such legislation could be passed to the benefit of both authorities and corporations while ignoring the concerns of the population more broadly. At the same time, it is well documented that the American government routinely violates the privacy of its own citizens (Greene, 2017; Hvistendahl & Biddle, 2022). Given these conclusions, we suggest caution and restraint in demanding that speech be curtailed in order to remedy a problem where harm has yet to be demonstrated.

That having been said, the potential threats surrounding computer vision and its uses are very real. While it is this researcher's opinion that these types of analyses specifically do not represent a threat to the average person at this time, computer vision use more broadly could certainly endanger the liberty of individuals in the future. Many locales have banned the use of computer vision by the government entirely (Simonite, 2021), and

activist groups should remain vigilant in regards to the use of computer vision, particularly by government organizations, and particularly in those cases where breaches of individual privacy might be concerned.

Future Research

Because of the inherent ambiguity in interpreting these results, as well as the controversial nature of the research being performed, future scholarship in this domain should seek to uncover findings that eliminate plausible explanations for the effects uncovered here and elsewhere. Such scholarship should apply critical research questions that attempt to exclude explanations for these effects rather than confirming them. Specifically, future researchers might attempt to eliminate some explanations for this effect in order to progress theoretical rationales as to where this effect is stemming from.

For example, while it is clear that there is some type of biological mechanism at play here, it is unclear as to whether the effects uncovered in this analysis are due primarily to genetic or environmental factors. It is possible that our groups differ in a variety of ways, including coming from different regions. There is evidence, for instance, that individuals from different regions demonstrate, in the aggregate, different personality characteristics as well as different political, economic, social, and health indicators from one another (Rentfrow et al., 2013). If true, this would suggest a potential alternative explanation for our findings. It could be that our samples of liberals and conservatives differ by region, and that the effects we are seeing are due to regional factors rather than ideological ones. In other words, liberals and conservatives drawn from the same general location might demonstrate less facial variation between groups.

It is further possible that liberals and conservatives differ by eating habits, and that the facial variation effects observed are due to elective variation between groups. For example, if conservatives demonstrate greater gender typicality in comparison to liberals, we might expect conservative men and women to demonstrate greater sexual dimorphism in regards to weight (Duncan et al., 1997). If true, conservative men might be heavier on average than liberal men, and the difference between facial area for conservative men and conservative women might be larger than the difference between liberal men and liberal women. Future research in this domain might attempt to constrain subjects by weight or by location in order to remove the influence of these environmental effects.

Contrastingly, these effects might be due to genetic influences. Previous research in this domain has demonstrated that political position taking is highly heritable (Funk et al., 2013; Alford et al., 2005; Hatemi & McDermott, 2012). Thus, it is possible that the morphological differences observed are due, at least in part, to common genetic markers overrepresented in conservative and liberal populations. These genetic markers might coincide with markers related to the development and appearance of the face. If true, facial morphology could be indicative of personality attributes broadly as well as political ideology specifically, and as such, there might be evolutionary reasons as to why people 'wear' their political ideology on their face.

One way to examine the genetic influence of morphological indicators would be to utilize twin samples. While it appears that political ideology is heritable, some sets of twins will inevitably demonstrate greater differentiation in ideological position taking than other sets of twins. One might imagine that twins demonstrating greater differentiation between each other might be more difficult to classify for the algorithm. In other words,

if the effects presented here are due to heritability, then those twins demonstrating 'less' heritability of political factors should be more ambiguous in regards to their group membership than those demonstrating 'greater' heritability, at least hypothetically. Research demonstrating a correlation between the propensity score for the correct classification of an image and the heritability of political factors could be supportive of a genetic explanation for these effects. However, at the current stage this research provides *no confirmatory evidence* that both facial morphology and political orientation stem from the same genetic origin, or that these findings are due to any genetic precursors.

Further, as of now, there is no evidence that the classifiers used to differentiate between liberals and conservatives (or heterosexuals and homosexuals) are utilizing the same facial information that human beings use to classify such individuals. It is possible that the classifiers employed in research of this type are categorizing images based on entirely different facial morphology than human beings use to perform the same task. Future researchers could attempt to isolate on facial morphology and utilize the predictions made by the classifier to provide a sample of archetype liberal and conservative face points according to the algorithm. They could then have human participants classify these sample images in an attempt to determine if human beings can classify these images at a rate above chance. If so, it would be a strong indicator that the classifier is utilizing the same morphological information as human beings when making its predictions.

Future researchers might also attempt to stress this methodology in order to derive conclusions about how the classifier is coming to make its decisions. For example, as previously stated, the effect found here is astonishingly similar to the effect uncovered in Wang and Kosinski (2018). It is possible that the classifiers employed in Wang and

Kosinski (2018) and this study are utilizing the same strategy despite the different topics of interest (sexual orientation vs. political ideology). In other words, it is possible that the topics of sexual orientation and political ideology demonstrate similar differences between groups in regards to facial area and that selecting images based on facial area provides a point of differentiation that is effective in classification across multiple domains but lacks specificity in regards to any particular topic. If true, then classifiers demonstrating success in differentiating between subjects in one arena might also work in differentiating subjects in a different, entirely unrelated arena. To examine this, future researchers might attempt to utilize their classifiers to differentiate images under different conditions from which the classifier was trained. For example, one might wish to train their classifier on subjects in the sexual orientation domain and apply their classifier to a novel sample of political images. If the classifier correctly categorizes the novel images, it seems likely that the classifier is not differentiating images based upon the sexual orientation of the person in the image specifically, but rather some confounding variable that correlates with indicators related to both sexual orientation as well as political ideology. Further, future research might attempt to control for facial area in their analyses in order to determine if facial area is the main driver of these effects or if there are other factors at play such as eyebrow positioning, nose shape, or the size of the eyes in proportion to the face. If researchers can accurately classify images when controlling for facial area, one might believe that these predictions are more nuanced than originally imagined.

Finally, researchers might utilize simulations in order to generate data with which to test classifiers. For example, researchers might generate simulated points from left or right leaning images that are constrained in a variety of aspects. By simulating points from

conservative and liberal faces constrained by facial area, we might attempt to classify images into left and right categories while holding facial area constant. This might help determine if classifiers will still be successful when controlling for obvious points of differentiation. If classifiers are still able to work in these conditions, it would suggest morphological differences between groups that are unrelated to the jawline. This would indicate that there are more points of morphological diversion than originally imagined between these groups, and that predictions made in regards to this topic might differ in measurable ways from predictions made in other domains.

Conclusion

Because physiognomy has such an ignoble history, many people might question the wisdom of engaging in a practice that has caused a great deal of suffering in the past.

While this hesitation is understandable, this type of research needs to occur more often, rather than less. Bad actors have been able to play in the sandbox of eugenics, racism, and bigotry precisely because physiognomy is so poorly understood. Examining these effects in detail reduces their power to harm, rather than increasing it. By repeatedly illustrating the modest effects and dramatic limitations of these types of analyses, we relegate physiognomy to what it truly is: small effects over an aggregate population.

Further, it is this researcher's opinion that there is an inconsistency in believing that physiognomy is a monster in our closet and concurrently being unwilling to turn on the light. Avoiding physiognomic research implies both that the effects are more real than they are and that the power to classify is more reliable than it is, lending tacit legitimacy to its potential misuse. More research in this arena will have the opposite effect that critics predict, taking something opaque and alarming and transforming it into something

transparent and measurable. By measuring the effect, we are able to demonstrate both how modest it is as well as how inadequate classifying people in such a manner would be.

References

ACLU. (2014). *An ethical framework for facial recognition.* ACLU.

ACLU. (2018, July 27). *ACLU comment on new Amazon statement responding to face recognition technology* test. ACLU. https://www.aclunc.org/news/aclu-comment-new-amazon-statement-responding-face-recognition-technology-test

Agüera y Arcas, B., Mitchell, M., Todorov, A. (2017, May 6). *Physiognomy's new clothes.* Medium. https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a

Agüera y Arcas, B., Todorov, A., & Mitchell, M. (2018, January 11). *Do algorithms reveal sexual orientation or just expose our stereotypes?*. Medium. https://medium.com/@blaisea/do-algorithms-reveal-sexual-orientation-or-just-expose-our-stereotypes-d998fafdf477.

Alford, J. R., Funk, C. L., & Hibbing, J. R. (2005). Are political orientations genetically transmitted?. *American Political Science Review, 99*(2), 153-167.

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, *111*(2), 256-274.

Anderson, D. (2017, September 8). *GLAAD and HRC call on Stanford University & responsible media to debunk dangerous & flawed report claiming to identify LGBTQ people through facial recognition technology*. GLAAD. https://www.glaad.org/blog/glaad-and-hrc-call-stanford-university-responsible-media-debunk-dangerous-flawed-report

Antonakis, J., & Eubanks, D. L. (2017). Looking leadership in the face. *Current Directions in Psychological Science, 26*(3), 270-275.

Babich, N. (2020, July 28). *What is computer vision and how does it work? An introduction.* Adobe. https://xd.adobe.com/ideas/principles/emerging-technology/what-is-computer-vision-how-does-it-work/

Bagheri, R. (2020, Jan 9). *Understanding Singular Value Decomposition and its application in data science.* TowardsDataScience. https://towardsdatascience.com/understanding-singular-value-decomposition-and-its-application-in-data-science-388a54be95d

Baker, K. (2005). Singular value decomposition tutorial. The Ohio State University, 24.

Ballew, C. C., & Todorov, A. (2007). Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences, 104*(46), 17948-17953.

Bandura, A. (1965). Vicarious processes: A case of no-trial learning. In *Advances in Experimental Social Psychology* (Vol. 2, pp. 1-55). Academic Press.

Bendel, O. (2018, March). The uncanny return of physiognomy. *In 2018 AAAI Spring Symposium Series.*

Berggren, N., Jordahl, H., & Poutvaara, P. (2017). The right look: Conservative politicians look better and voters reward it. *Journal of Public Economics, 146*, 79-86.

Berry, D. S., & Brownlow, S. (1989). Were the physiognomists right? Personality correlates of facial babyishness. *Personality and Social Psychology Bulletin, 15*(2), 266-279.

Berry, D. S., & McArthur, L. Z. (1985). Some components and consequences of a babyface. *Journal of Personality and Social Psychology, 48*(2), 312.

Bhandari, P. (2021). *Correlation Coefficient | Types, Formulas, & Examples.* Scribbr. https://www.scribbr.com/statistics/correlation-coefficient/.

Biehl, M., Matsumoto, D., Ekman, P., Hearn, V., Heider, K., Kudoh, T., & Ton, V. (1997). Matsumoto and'Ekman's Japanese and Caucasian Facial Expressions of Emotion (JACFEE): Reliability data and cross-national differences. *Journal of Nonverbal Behavior, 21*(1), 3-21.

Biel, J. I., Teijeiro-Mosquera, L., & Gatica-Perez, D. (2012, October). Facetube: predicting personality from facial expressions of emotion in online conversational video. In *Proceedings of the 14th ACM international conference on Multimodal interaction* (pp. 53-56). ACM.

Bjornsdottir, R. T., & Rule, N. O. (2017). The visibility of social class from facial cues. *Journal of Personality and Social Psychology, 113*(4), 530-546.

Block, J., & Block, J. H. (2006). Nursery school personality and political orientation two decades later. *Journal of Research in Personality*, *40*(5), 734-749.

Boothroyd, L. G., Cross, C. P., Gray, A. W., Coombes, C., & Gregson-Curtis, K. (2011). Perceiving the facial correlates of sociosexuality: Further evidence. *Personality and Individual Differences, 50*(3), 422-425.

Breedlove, S. M. (2017). Prenatal influences on human sexual orientation: Expectations versus data. *Archives of Sexual Behavior, 46*(6), 1583-1592.

Brewer, M. B. (1988). A dual process model of impression formation. In R. S. Wyer, Jr. & T. K. Srull (Eds.), *Advances in social cognition* (Vol. 1, pp. 1-36). Erlbaum.

Bull, R., & Hawkes, C. (1982). Judging politicians by their faces. *Political Studies, 30*(1),

95-101.

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency* (pp. 77-91). PMLR.

Burke, D., & Sulikowski, D. (2010). A new viewpoint on the evolution of sexually dimorphic human faces. *Evolutionary Psychology*, 8(4), 147470491000800404.

Carbon, C. C., Schweinberger, S. R., Kaufmann, J. M., & Leder, H. (2005). The Thatcher illusion seen by the brain: an event-related brain potentials study. *Cognitive Brain Research, 24(*3), 544-555.

Carney, D. R., Colvin, C. R., & Hall, J. A. (2007). A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality, 41*(5), 1054-1072.

Carney D. R., Jost J. T., Gosling S. D., & Potter J. (2008). The secret lives of liberals and conservatives: personality profiles, interaction styles, and the things they leave behind. *Political Psychology, 29*, 807–840.

Carpinella, C. M., & Johnson, K. L. (2013). Appearance-based politics: Sex-typed facial cues communicate political party affiliation. *Journal of Experimental Social Psychology, 49*(1), 156-160.

Cavazos, J. G., Phillips, P. J., Castillo, C. D., & O'Toole, A. J. (2020). Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?. *IEEE transactions on biometrics, behavior, and identity science*, *3*(1), 101-111.

Chen, E. E., & Wojcik, S. P. (2016). A practical guide to big data research in psychology. *Psychological Methods*, 21(4), 458.

Cheung, C. H., Rutherford, H. J., Mayes, L. C., & McPartland, J. C. (2010). Neural responses to faces reflect social personality traits. *Social Neuroscience, 5*(4), 351-359.

Civile, C., McLaren, R. P., & McLaren, I. P. (2014). The face inversion effect—Parts and wholes: Individual features and their configuration. *Quarterly Journal of Experimental Psychology, 67*(4), 728-746.

Costa, P. T., & McCrae, R. R. (1992). Multiple uses for longitudinal personality data. *European Journal of Personality, 6*, 85–102.

Costa, P. T., Jr., & McCrae, R. R. (2002). Looking backward: Changes in the mean levels of personality traits from 80 to 12. In D. Cervone & W. Mischel (Eds.), *Advances in personality science* (pp. 219 –237). Guilford.

Daniller, A. (2019, November 12). *Americans' immigration policy priorities: Division between – and within – the two parties.* Pew. https://www.pewresearch.org/fact-tank/2019/11/12/americans-immigration-policy-priorities-divisions-between-and-within-the-two-parties/

Darwin, C., & Prodger, P. (1998). *The expression of the emotions in man and animals*. Oxford University Press.

Defazio, A., Bach, F., & Lacoste-Julien, S. (2014). SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, *27*.

DeLisi, M. (2013). Revisiting Lombroso. In Cullen, F. & Vilcox, P. (Eds.), *The Oxford Handbook of Criminological Theory* (pp. 5-21). New York: Routledge.

Dixon, S. (2021). *Distribution of Twitter users worldwide as of April 2021, by age group*. Statista. https://www.statista.com/statistics/283119/age-distribution-of-global-twitter-users/

Dixon, S. (2022). *Distribution of Twitter users worldwide as of January 2022, by gender*. Statista. https://www.statista.com/statistics/828092/distribution-of-users-on-twitter-worldwide-gender/

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, *4*(1), eaao5580.

Duncan, L. E., Peterson, B. E., & Winter, D. G. (1997). Authoritarianism and gender roles: Toward a psychological analysis of hegemonic relationships. *Personality and Social Psychology Bulletin, 23*(1), 41-49.

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*, 621.

Eaves, L., Heath, A., Martin, N., Maes, H., Neale, M., Kendler, K., Kirk, K., & Corey, L. (1999). Comparing the biological and cultural inheritance of personality and social attitudes in the Virginia 30 000 study of twins and their relatives. *Twin Research and Human Genetics, 2*(2), 62-80.

Edmondson, C. (2019, July 7). *ICE Used Facial Recognition to Mine State Driver's License Databases*. The New York Times. https://www.nytimes.com/2019/07/07/us/politics/ice-drivers-licenses-facial-recognition.html

Eidelman, S., Crandall, C., Goodman, J., & Blanchar, J. (2012). Low effort thought promotes political conservatism. *Personality and Social Psychology Bulletin, 38*, 808–820.

Ekman, P. (2003). Darwin, deception, and facial expression. *Annals of the New York Academy of Sciences, 1000*(1), 205-221.

Ellis, L., & Ames, M. A. (1987). Neurohormonal functioning and sexual orientation: A theory of homosexuality–heterosexuality. *Psychological Bulletin*, 101(2), 233-258.

Farah, M. J., Wilson, K. D., Drain, H. M., & Tanaka, J. R. (1995). The inverted face inversion effect in prosopagnosia: Evidence for mandatory, face-specific perceptual mechanisms. *Vision Research*, 35(14), 2089-2093.

Ferwerda, B., Schedl, M., & Tkalcic, M. (2016, January). Using Instagram picture features to predict users' personality. In *International Conference on Multimedia Modeling* (pp. 850-861). Springer.

Fink, B., Grammer, K., & Matts, P. J. (2006). Visible skin color distribution plays a role in the perception of age, attractiveness, and health in female faces. *Evolution and Human Behavior, 27*(6), 433-442.

Flach, P. A. (2016). ROC analysis. In *Encyclopedia of Machine Learning and Data Mining* (pp. 1-8). Springer.

Fox, G. (2020, October 1). Egypt police 'using dating apps' to find and imprison LGBT+ people. The Independent. https://www.independent.co.uk/news/world/middle-east/egypt-lgbt-gay-facebook-grindr-jail-torture-police-hrw-b742231.html

Funk, C. L., Smith, K. B., Alford, J. R., Hibbing, M. V., Eaton, N. R., Krueger, R. F., Eaves, L.J., & Hibbing, J. R. (2013). Genetic and environmental transmission of political orientations. *Political Psychology, 34*(6), 805-819.

Furnham, A., & Fenton-O'Creevy, M. (2018). Personality and political orientation. *Personality and Individual Differences*, *129*, 88-91.

Gallup. (n.d.) Guns. Gallup. Retrieved from: https://news.gallup.com/poll/1645/guns.aspx

GeeksforGeeks. (2021). Deep Face Recognition. GeeksforGeeks.com. https://www.geeksforgeeks.org/deep-face-recognition/

Gerber, A. S., Huber, G. A., Doherty, D., & Dowling, C. M. (2011). Personality traits and the consumption of political information. *American Politics Research, 39*(1), 32-84.

Goodman, R. (2018, October 12). *Why Amazon's automated hiring tool discriminated against women.* ACLU. https://www.aclu.org/blog/womens-rights/womens-rights-workplace/why-amazons-automated-hiring-tool-discriminated-against

Government Europa. (2019, June 4). *Facial recognition ethics framework for policing in*

London issued. Government Europa. https://www.governmenteuropa.eu/facial-recognition-ethics-framework/93476/

Greco, A., Saggese, A., Vento, M., & Vigilante, V. (2020). A convolutional neural network for gender recognition optimizing the accuracy/speed tradeoff. *IEEE Access, 8,* 130771-130781.

Greene, R. (2017, June 22). How the government can read your email. *Politico.* https://www.politico.com/agenda/story/2017/06/22/section-702-surveillance-program-national-security-000463/

Hall, J. A., Pennington, N., & Lueders, A. (2014). Impression management and formation on Facebook: A lens model approach. *New Media & Society*, *16*(6), 958-982.

Hardesty, L. (2017, April 4). *Explained: Neural networks.* MIT News. https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414

Hashemi, M., & Hall, M. (2020). Criminal tendency detection from facial images and the gender bias effect. *Journal of Big Data, 7*(1), 1-16.

Hassin, R., & Trope, Y. (2000). Facing faces: studies on the cognitive aspects of physiognomy. *Journal of Personality and Social Psychology*, *78*(5), 837-852.

Hatemi, P. K., & McDermott, R. (2012). The genetics of politics: Discovery, challenges, and progress. *Trends in Genetics, 28*(10), 525-533.

Hatemi, P. K., Medland, S. E., Klemmensen, R., Oskarsson, S., Littvay, L., Dawes, C. T., … & Christensen, K. (2014). Genetic influences on political ideologies: Twin analyses of 19 measures of political ideologies from five democracies and genome-wide findings from three populations. *Behavior Genetics, 44*(3), 282-294.

Hellman, M. (2019, June 12). *Amazon speaks out in favor of U.S. regulating facial-recognition technology*. The Seattle Times. Retrieved from https://www.seattletimes.com/business/amazon/amazon-speaks-out-in-favor-of-regulating-facial-recognition/

Hibbing, J. R., Smith, K. B., & Alford, J. R. (2013). *Predisposed: Liberals, conservatives, and the biology of political differences.* Routledge.

Hill, K. (2012, Feb 16). *How Target figured out a teen girl was pregnant before her father did.* Forbes. https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#4ac58eef6668

Hvistendahl, M. & Biddle, S. (2022, Feb 8). Use of controversial phone-cracking tool is spreading across federal government. *The Intercept. https://theintercept.com/2022/02/08/cellebrite-phone-hacking-government-agencies/*

Hyde, J. S. (2014). Gender similarities and differences. *Annual Review of Psychology*, 65, 373-398.

Jenkinson, J. (1997). Face facts: a history of physiognomy from ancient Mesopotamia to the end of the 19th century. *Journal of Biocommunication, 24*(3), 2-7.

Johnson, S. (2020, February 22). *These are the 7 most partisan issues in America right now.* BigThink.com. https://bigthink.com/the-present/political-polarization/

Jones, J. (2021, July 23). *American remain divided on preferred immigration levels.* Gallup. https://news.gallup.com/poll/352664/americans-remain-divided-preferred-immigration-levels.aspx

Jost, J. T., Banaji, M. R., & Nosek, B. A. (2004). A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political Psychology*, *25*(6), 881-919.

Kaminski, G., Dridi, S., Graff, C., & Gentaz, E. (2009). Human ability to detect kinship in strangers' faces: effects of the degree of relatedness. *Proceedings of the Royal Society B: Biological Sciences, 276*(1670), 3193-3200.

Kandler, C., Bleidorn, W., & Riemann, R. (2012). Left or right? Sources of political orientation: the roles of genetic factors, cultural transmission, assortative mating, and personality. *Journal of Personality and Social Psychology*, *102*(3), 633.

Kazem, A. J., & Widdig, A. (2013). Visual phenotype matching: cues to paternity are present in rhesus macaque faces. *PLoS One*, 8(2), e55846.

Kelion, L. (2019, May 22). *Amazon heads off facial recognition rebellion.* BBC News. https://www.bbc.com/news/technology-48339142

Kenny, D. A., Albright, L., Malloy, T. E., & Kashy, D. A. (1994). Consensus in interpersonal perception: Acquaintance and the big five. *Psychological Bulletin, 116*(2), 245-258.

King, D. (2009). Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research, 10*, 1755-1758.

Kobayashi, H., & Kohshima, S. (1997). Unique morphology of the human eye. *Nature*, *387*(6635), 767-768.

Kosinski, M. (2021). Facial recognition technology can expose political orientation from naturalistic facial images. *Scientific Reports*, *11*(1), 1-7.

Kraus, M. W., & Keltner, D. (2009). Signs of socioeconomic status: A thin-slicing approach. *Psychological Science, 20*(1), 99-106.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

Leopold, D. A., & Rhodes, G. (2010). A comparative view of face perception. *Journal of Comparative Psychology, 124*(3), 233.

Leuner, J. (2019). A replication Study: Machine learning models are capable of predicting sexual orientation from facial images. *arXiv preprint arXiv*:1902.10739.

Levi, G., & Hassner, T. (2015). Age and gender classification using convolutional neural networks. *In Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 34-42).

Levin, S. (2017, September 12). *Face-reading AI will be able to detect your politics and IQ, professor says.* The Guardian. https://www.theguardian.com/technology/2017/sep/12/artificial-intelligence-face-recognition-michal-kosinski

Little, A. C., & Perrett, D. I. (2007). Using composite images to assess accuracy in personality attribution to faces. *British Journal of Psychology, 98*(1), 111-126.

Little, A. C., Burriss, R. P., Jones, B. C., & Roberts, S. C. (2007). Facial appearance affects voting decisions. *Evolution and Human Behavior, 28*(1), 18-27.

Little, A. C., Roberts, S. C., Jones, B. C., & DeBruine, L. M. (2012). The perception of attractiveness and trustworthiness in male faces affects hypothetical voting decisions differently in wartime and peacetime scenarios. *Quarterly Journal of Experimental Psychology, 65*(10), 2018-2032.

Lönnqvist, J. E. (2017). Just because you look good, doesn't mean you're right. *Personality and Individual Differences, 108*, 133-135.

Maina, S. (2021, Feb 25). *Preventing Overfitting with Lasso, Ridge, and Elastic-net Regularization in Machine Learning.* TowardsDataScience.com https://towardsdatascience.com/preventing-overfitting-with-lasso-ridge-and-elastic-net-regularization-in-machine-learning-d1799b05d382

Martin, Z. (2014, May 27). *ACLU: Facial recognition needs ethical framework.* SecureIDNews.com. https://www.secureidnews.com/news-item/aclu-facial-recognition-needs-ethical-framework/

McNeish, D. M. (2015). Using lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences. *Multivariate Behavioral Research, 50*(5), 471-484.

Mitchell, A., & Diamond, L. (2018, February 2). *China's surveillance state should scare everyone.* The Atlantic.

https://www.theatlantic.com/international/archive/2018/02/china-surveillance/552203/

Montoya, E. R., Terburg, D., Bos, P. A., & Van Honk, J. (2012). Testosterone, cortisol, and serotonin as key regulators of social aggression: A review and theoretical perspective. *Motivation and Emotion, 36*(1), 65-73.

Morgan, T. J., Laland, K. N., & Harris, P. L. (2015). The development of adaptive conformity in young children: effects of uncertainty and consensus. *Developmental Science*, *18*(4), 511-524.

Morland, K., Wing, S., Roux, A. D., & Poole, C. (2002). Neighborhood characteristics associated with the location of food stores and food service places. *American Journal of Preventive Medicine*, *22*(1), 23-29.

Mueller, U., & Mazur, A. (1996). Facial dominance of West Point cadets as a predictor of later military rank. *Social Forces, 74*(3), 823-850.

Murphy, H. (2017, October 9). *Why Stanford researchers tried to create a 'gaydar' machine.* The New York Times. https://www.nytimes.com/2017/10/09/science/stanford-sexual-orientation-study.html

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research.* https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html

Narkhede, S. (2018, June 26). Understanding AUC – ROC Curve. *TowardsDataScience.com. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5*

Naumann, L. P., Vazire, S., Rentfrow, P. J., & Gosling, S. D. (2009). Personality judgments based on physical appearance. *Personality and Social Psychology Bulletin, 35*(12), 1661-1671.

Newport, F. (2018, May 22). *In U.S., estimate of LGBT population rises to 4.5%.* Gallup. https://news.gallup.com/poll/234863/estimate-lgbt-population-rises.aspx

Oliphant, J. B. (2017, June 23). *Bipartisan support for some gun proposals, stark partisan divisions on many others.* Pew. https://www.pewresearch.org/fact-tank/2017/06/23/bipartisan-support-for-some-gun-proposals-stark-partisan-divisions-on-many-others/

Olivola, C. Y., Funk, F., & Todorov, A. (2014). Social attributions from faces bias human choices. *Trends in Cognitive Sciences, 18*(11), 566-570.

Paxton, A. (accepted; anticipated publication date: 2020). *The Belmont Report in the age of big data: Ethics at the intersection of psychological science and data science*. To appear in S. E. Woo, L. Tay, & R. Proctor (Eds.), *Big data methods for psychological research: New horizons and challenges*. American Psychological Association. https://alexandrapaxton.com/files/paxton-belmont_report_in_big_data-accepted.pdf

Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research, 96*(1), 3-14.

Penton-Voak, I. S., & Chen, J. Y. (2004). High salivary testosterone is linked to masculine male facial appearance in humans. *Evolution and Human Behavior, 25*(4), 229-241.

Peterson, R. D., & Palmer, C. L. (2017). Effects of physical attractiveness on political beliefs. *Politics and the Life Sciences, 36*(2), 3-16.

Read, S. J., Droutman, V, & Miller, L. C. (2017) Virtual personalities: A neural network model of the structure and dynamics of personality. In Vallacher, R. R., Read, S. J., & Nowak, A. (Eds.), *Computational social psychology.* Routledge.

Reinhart, R. J. (2018, May 18). *In the news: Key insights into Americans' views on guns.* Gallup. https://news.gallup.com/poll/234800/news-key-insights-americans-views-guns.aspx

Rentfrow, P. J., Gosling, S. D., Jokela, M., Stillwell, D. J., Kosinski, M., & Potter, J. (2013). Divided we stand: Three psychological regions of the United States and their political, economic, social, and health correlates. *Journal of Personality and Social Psychology*, *105*(6), 996-xxxx.

Resnick, B. (2018, January 9). *This psychologist's 'gaydar' research makes us uncomfortable. That's the point.* Vox. https://www.vox.com/science-and-health/2018/1/29/16571684/michal-kosinski-artificial-intelligence-faces

Rezlescu, C., Duchaine, B., Olivola, C. Y., & Chater, N. (2012). Unfakeable facial configurations affect strategic choices in trust games with or without information about past behavior. *PloS One, 7*(3), e34293.

Rhodes, G., & Simmons, L. W. (2007). Symmetry, attractiveness and sexual selection. In R. I. M. Dunbar & L. Barrett (Eds.), *The Oxford handbook of evolutionary psychology* (pp. 333–364). Oxford University Press.

Roulette, J. (2019, July 18). *Orlando cancels Amazon Rekognition program, capping 15 months of glitches and controversy*. Orlando Weekly. https://www.orlandoweekly.com/Blogs/archives/2019/07/18/orlando-cancels-

amazon-rekognition-capping-15-months-of-glitches-and-controversy

Rule, N. O., & Ambady, N. (2008). The face of success: Inferences from chief executive officers' appearance predict company profits. *Psychological Science*, *19*(2), 109-111.

Rule, N. O., & Ambady, N. (2010). Democrats and Republicans can be differentiated from their faces. *PloS One, 5*(1), e8733. https://tspace.library.utoronto.ca/bitstream/1807/33130/1/Rule%26Ambady%282008_PsychSci%29.pdf

Rule, N. O., Ambady, N., Adams Jr., R. B., & Macrae, C. N. (2008). Accuracy and awareness in the perception and categorization of male sexual orientation. *Journal of Personality and Social Psychology, 95*(5), 1019-1028.

Saad, L. (2016, Nov 8). Trump and Clinton finish with historically poor images. *Gallup.* https://news.gallup.com/poll/197231/trump-clinton-finish-historically-poor-images.aspx

Samochowiec, J., Wänke, M., & Fiedler, K. (2010). Political ideology at face value. *Social Psychological and Personality Science*, *1*(3), 206-213.

SAS. (n.d.) *Computer vision: What it is and why it matters.* SAS. https://www.sas.com/en_us/insights/analytics/computer-vision.html

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61, 85-117.

Sedgewick, J. R., Flath, M. E., & Elias, L. J. (2017). Presenting your best self (ie): the influence of gender on vertical orientation of selfies on tinder. *Frontiers in Psychology, 8*, 604.

Segalin, C., Celli, F., Polonio, L., Kosinski, M., Stillwell, D., Sebe, N., Cristani, M., & Lepri, B. (2017, October). What your Facebook profile picture reveals about your personality. In *Proceedings of the 25th ACM International Conference on Multimedia* (pp. 460-468).

Segalin, C., Cheng, D. S., & Cristani, M. (2017). Social profiling through image understanding: Personality inference using convolutional neural networks. *Computer Vision and Image Understanding*, *156*, 34-50.

Serengil, S. (n.d.) *DeepFace – The Most Popular Open Source Facial Recognition Library.* Viso.ai. https://viso.ai/computer-vision/deepface/

Serengil, S. I. & Ozpinar, A. (2020, Sep 26). *deepface*. PyPi.org. https://pypi.org/project/deepface/#description

Shaikh, F. (2017, May 18). *Why are GPUs necessary for training Deep Learning models?* Analytics Vidhya. https://www.analyticsvidhya.com/blog/2017/05/gpus-necessary-for-deep-learning/

Sham, A. H., Aktas, K., Rizhinashvili, D., Kuklianov, D., Alisinanoglu, F., Ofodile, I., ... & Anbarjafari, G. (2022). Ethical AI in facial expression analysis: Racial bias. *Signal, Image and Video Processing*, 1-8.

Sheth, K. (2017, April 25). *Did you know only 66 years separated the first successful plane flights and moon landings?* World Atlas. https://www.worldatlas.com/articles/did-you-know-only-66-years-separated-the-first-successful-plane-flights-and-moon-landings.html

Skorska, M. N., Geniole, S. N., Vrysen, B. M., McCormick, C. M., & Bogaert, A. F. (2015). Facial structure predicts sexual orientation in both men and women. *Archives of Sexual Behavior*, *44*(5), 1377-1394.

Smith, B. (2018, July 13). *Facial recognition technology: the need for public regulation and corporate responsibility.* Microsoft. https://blogs.microsoft.com/on-the-issues/2018/07/13/facial-recognition-technology-the-need-for-public-regulation-and-corporate-responsibility/

Snow, J. (2018, July 26). *Amazon's face recognition falsely matched 28 members of Congress with mugshots*. ACLU. https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28

Spirina, K. (2019, April 29). *Ethics of facial recognition: How to make business uses fair and transparent.* Medium. https://towardsdatascience.com/ethics-of-facial-recognition-how-to-make-business-uses-fair-and-transparent-98e3878db08d

Staddon, J. E., & Cerutti, D. T. (2003). Operant conditioning. *Annual review of Psychology*, *54*, 115.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, 31(1), 137-149.

Stanley, J. (2018, April 16). *A looming implication of face recognition: Private photo blacklists.* ACLU. https://www.aclu.org/blog/privacy-technology/surveillance-technologies/looming-implication-face-recognition-private-photo

Stenner, K. (2009). Three kinds of "conservatism". *Psychological Inquiry, 20*(2-3), 142-159.

Subramanian, S. V., & Perkins, J. M. (2009). Are Republicans healthier than Democrats?. *International Journal of Epidemiology*, 39(3), 930-931.

Swets, J. A. (2014). *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers.* Psychology Press.

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology, 46*(2), 225-245.

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology, 66*, 519-545.

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*(12), 455-460.

Tomasello, M., Hare, B., Lehmann, H., & Call, J. (2007). Reliance on head versus eyes in the gaze following of great apes and human infants: the cooperative eye hypothesis. *Journal of Human Evolution, 52*(3), 314-320.

Truett, K. R. (1993). Age differences in conservatism. *Personality and Individual Differences, 14*(3), 405-411.

Tskhay, K. O., & Rule, N. O. (2013). Accuracy in categorizing perceptually ambiguous groups: A review and meta-analysis. *Personality and Social Psychology Review, 17*(1), 72-86.

Vinciarelli, A., & Mohammadi, G. (2014). A survey of personality computing. IEEE *Transactions on Affective Computing*, *5*(3), 273-291.

Wang, D. (2021, October 28). Presentation in Self-Posted Facial Images Can Expose Sexual Orientation: Implications for Research and Privacy. *PsyArXiv.* https://doi.org/10.31234/osf.io/u7vcd

Wang, Y., & Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology, 114*(2), 246-257.

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science, 17*(7), 592-598.

Winter, N. J. (2010). Masculine republicans and feminine democrats: Gender and Americans' explicit and implicit images of the political parties. *Political Behavior, 32*(4), 587-618.

Wolffhechel, K., Fagertun, J., Jacobsen, U. P., Majewski, W., Hemmingsen, A. S., Larsen, C. L., ... & Jarmer, H. (2014). Interpretation of appearance: The effect of facial features on first impressions and personality. *PloS one*, *9*(9), e107721.

Wu, X., & Zhang, X. (2016). Automated inference on criminality using face images. *arXiv preprint arXiv*:1611.04135, 4038-4052.

Wu, Y. C. J., Chang, W. H., & Yuan, C. H. (2015). Do Facebook profile pictures reflect user's personality?. *Computers in Human Behavior*, *51*, 880-889.

Yildiz, D., Munson, J., Vitali, A., Tinati, R., & Holland, J. A. (2017). Using Twitter data for demographic research. *Demographic Research*, *37*, 1477-1514.

Zaiontz, C. (n.d.) *Effect Size for Chi-square Test.* Real-statistics.com. https://www.real-statistics.com/chi-square-and-f-distributions/effect-size-chi-square/

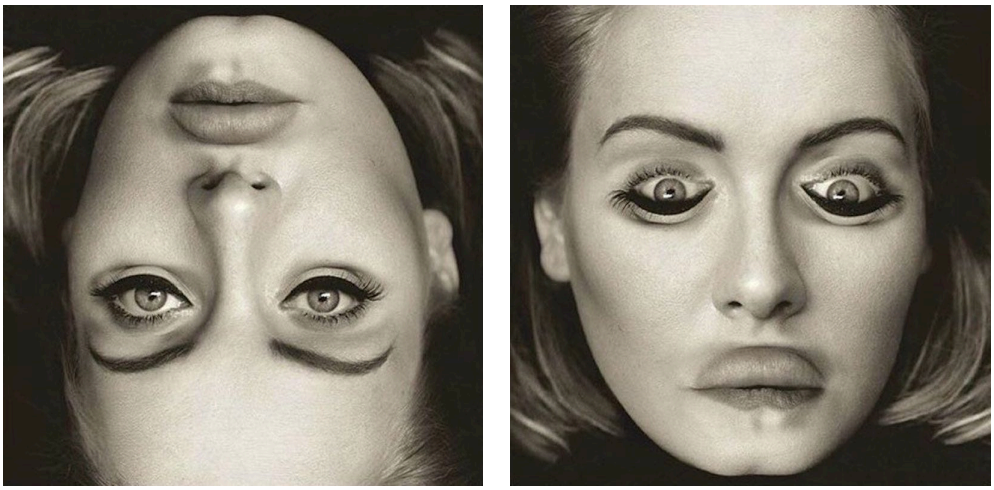
Zebrowitz, L. A., & Montepare, J. M. (1992). Impressions of babyfaced individuals across the life span. *Developmental Psychology*, *28*(6), 1143-1152.

Zhang, T., Qin, R. Z., Dong, Q. L., Gao, W., Xu, H. R., & Hu, Z. Y. (2017). Physiognomy: Personality traits prediction by learning. *International Journal of Automation and Computing*, *14*(4), 386-395.

Zou, J., Han, Y., & So, S. S. (2008). Overview of artificial neural networks. In Livingstone, D.J. (Ed.) *Artificial Neural Networks* (pp. 14-22). Humana Press. 10:1007/978-1-60327-101-1

Appendix A

Figure 24
Thatcher Effect

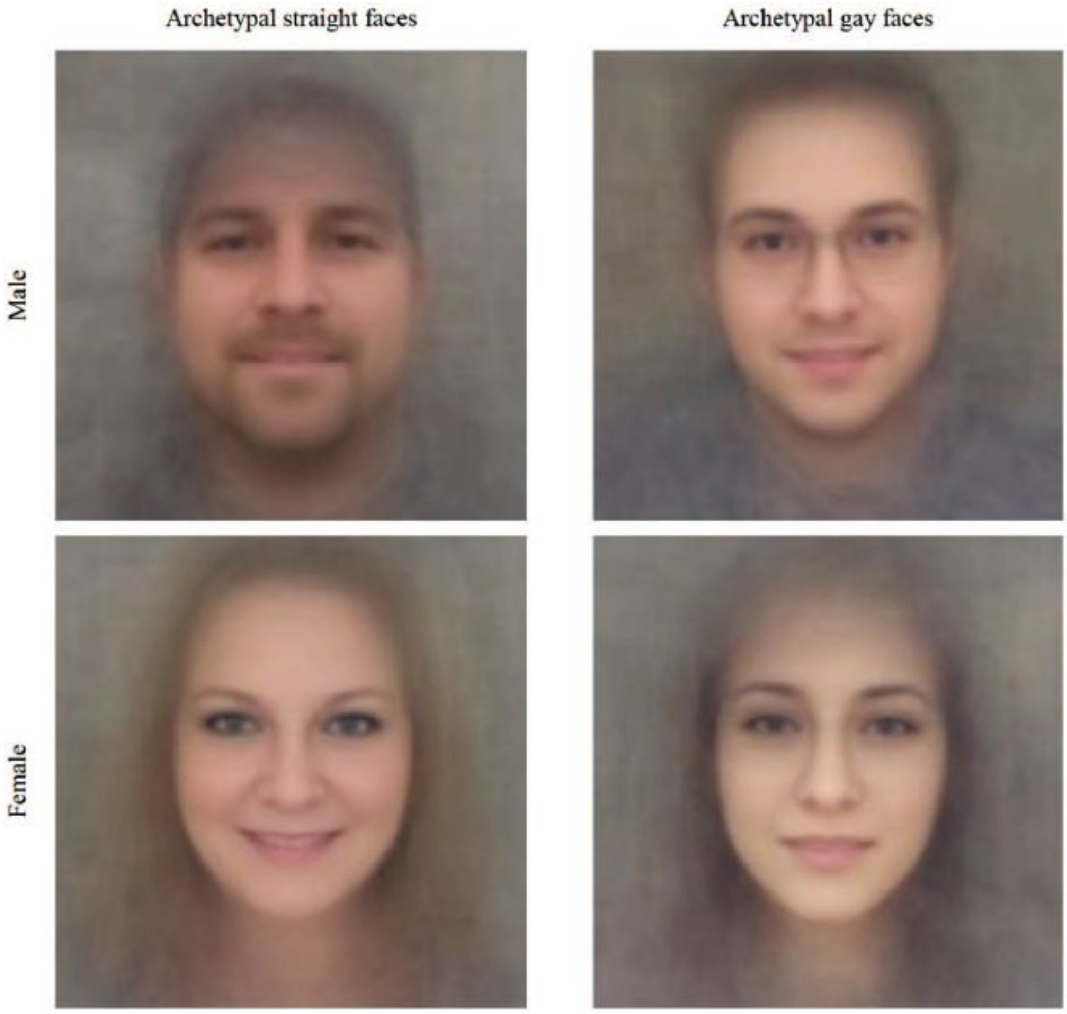


The “Thatcher effect”. Inverted faces are processed correctly despite the incorrect placement of the component features.

Return to relevant section

Appendix B

Figure 25
Composite Images from Wang and Kosinski (2018)



Critics of Wang and Kosinski (2018) argue that head pose might be inferred from the dating site images rather than sexual orientation. The straight male and gay female composites appear to have flatter eyebrows and larger nostrils, perhaps indicating that these photographs were more likely to be shot from straight ahead or below.

Return to relevant section

Appendix C

Table 4
Organizations of Interest, # of followers

Gun Policy		
Organization	Position	# of Followers
Everytown for Gun Safety	Pro-Gun Control (left)	~250,000
The National Rifle Association	Anti-Gun Control (right)	~890,000
Immigration Policy		
Organization	Position	# of Followers
Federation for American Immigration Reform (FAIR)	Anti-Immigration (right)	~470,000
United We Dream	Pro-immigration (left)	~130,000

Return to relevant section

Appendix D

Figure 26
Facial Points Mapped to an Image

Return to relevant section

Appendix E

Figure 27
Example of a Classification Table

Table 4.—The Observed and the Predicted Frequencies for Remedial Reading Instructions by Logistic Regression With the Cutoff of 0.50

Observed	Predicted		% Correct
	Yes	No	
Yes	2	57	3.39
No	1	129	99.23
Overall % correct			69.31

Note. Sensitivity = 2/(2+57)% = 3.39%. Specificity = 129/(1+129)% = 99.23%. False positive = 1/(1+2)% = 33.33%. False negative = 57/(57+129)% = 30.65%.

An example of a classification table. True positives occur when observed and predicted are both "Yes", while true negatives occur when both are "No". A false positive occurs when observed is "no" but predicted is "yes", while a false negative would be the inverse. (Peng et al., 2002, p. 8).

Return to relevant section

Appendix F

Table 5

Linear Model Results for Sex Typicality

OLS Regression Results

Dep. Variable:	SexTypicality	R-squared:	0.036
Model:	OLS	Adj. R-squared:	0.036
Method:	Least Squares	F-statistic:	1335.
Date:	Wed, 09 Nov 2022	Prob (F-statistic):	0.00
Time:	14:40:35	Log-Likelihood:	-9.0243e+05
No. Observations:	247515	AIC:	1.805e+06
Df Residuals:	247507	BIC:	1.805e+06
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	94.5220	0.053	1785.077	0.000	94.418	94.626
C(DomSex)[T.Females]	-0.7188	0.081	-8.855	0.000	-0.878	-0.560
C(Topic)[T.Immigration]	0.0292	0.087	0.336	0.737	-0.141	0.199
C(orientation)[T.Right]	3.4520	0.060	57.107	0.000	3.333	3.570
C(DomSex)[T.Females]:C(Topic)[T.Immigration]	-0.0497	0.133	-0.373	0.709	-0.311	0.211
C(orientation)[T.Right]:C(Topic)[T.Immigration]	-1.7177	0.108	-15.858	0.000	-1.930	-1.505
C(DomSex)[T.Females]:C(orientation)[T.Right]	-2.8565	0.104	-27.387	0.000	-3.061	-2.652
C(DomSex)[T.Females]:C(orientation)[T.Right]:C(Topic)[T.Immigration]	0.9184	0.183	5.029	0.000	0.560	1.276

Omnibus:	144928.787	Durbin-Watson:	2.004
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1034205.033
Skew:	-2.870	Prob(JB):	0.00
Kurtosis:	11.205	Cond. No.	17.6

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

	sum_sq	df	F	PR(>F)
Intercept	2.739739e+08	1.0	3.186499e+06	0.000000e+00
C(DomSex)	6.741865e+03	1.0	7.841238e+01	8.416070e-19
C(Topic)	9.713818e+00	1.0	1.129782e-01	7.367790e-01
C(orientation)	2.803928e+05	1.0	3.261155e+03	0.000000e+00
C(DomSex):C(Topic)	1.196746e+01	1.0	1.391896e-01	7.090884e-01
C(orientation):C(Topic)	2.162187e+04	1.0	2.514768e+02	1.319594e-56
C(DomSex):C(orientation)	6.449039e+04	1.0	7.500661e+02	6.854493e-165
C(DomSex):C(orientation):C(Topic)	2.174350e+03	1.0	2.528914e+01	4.938237e-07
Residual	2.128055e+07	247507.0	NaN	NaN

Return to relevant section

Appendix G

Table 6

Linear Model Results for Age

OLS Regression Results

Dep. Variable:	Age	R-squared:	0.023
Model:	OLS	Adj. R-squared:	0.023
Method:	Least Squares	F-statistic:	841.1
Date:	Thu, 10 Nov 2022	Prob (F-statistic):	0.00
Time:	23:40:36	Log-Likelihood:	-7.6914e+05
No. Observations:	247515	AIC:	1.538e+06
Df Residuals:	247507	BIC:	1.538e+06
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	30.9883	0.031	1002.785	0.000	30.928	31.049
C(DomSex)[T.Females]	1.1976	0.047	25.278	0.000	1.105	1.290
C(orientation)[T.Right]	0.2010	0.035	5.697	0.000	0.132	0.270
C(Topic)[T.Immigration]	-1.0445	0.051	-20.627	0.000	-1.144	-0.945
C(DomSex)[T.Females]:C(orientation)[T.Right]	-0.1172	0.061	-1.926	0.054	-0.237	0.002
C(DomSex)[T.Females]:C(Topic)[T.Immigration]	0.0226	0.078	0.291	0.771	-0.130	0.175
C(Topic)[T.Immigration]:C(orientation)[T.Right]	2.6192	0.063	41.435	0.000	2.495	2.743
C(DomSex)[T.Females]:C(Topic)[T.Immigration]:C(orientation)[T.Right]	-0.3745	0.107	-3.514	0.000	-0.583	-0.166

Omnibus:	16828.880	Durbin-Watson:	1.997
Prob(Omnibus):	0.000	Jarque-Bera (JB):	20857.753
Skew:	0.657	Prob(JB):	0.00
Kurtosis:	3.544	Cond. No.	17.6

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

	sum_sq	df	F	PR(>F)
Intercept	2.944675e+07	1.0	1.005578e+06	0.000000e+00
C(DomSex)	1.871188e+04	1.0	6.389929e+02	8.362466e-141
C(orientation)	9.504021e+02	1.0	3.245532e+01	1.221023e-08
C(Topic)	1.245959e+04	1.0	4.254830e+02	1.878133e-94
C(DomSex):C(orientation)	1.086094e+02	1.0	3.708906e+00	5.412399e-02
C(DomSex):C(Topic)	2.484113e+00	1.0	8.483008e-02	7.708558e-01
C(Topic):C(orientation)	5.027627e+04	1.0	1.716886e+03	0.000000e+00
C(DomSex):C(Topic):C(orientation)	3.616038e+02	1.0	1.234842e+01	4.414378e-04
Residual	7.247846e+06	247507.0	NaN	NaN

Return to relevant section

Appendix H

Table 7

Linear Model Results for Happy

OLS Regression Results

Dep. Variable:	Happy	R-squared:	0.013
Model:	OLS	Adj. R-squared:	0.013
Method:	Least Squares	F-statistic:	307.4
Date:	Thu, 10 Nov 2022	Prob (F-statistic):	0.00
Time:	23:44:29	Log-Likelihood:	-6.1141e+05
No. Observations:	158479	AIC:	1.223e+06
Df Residuals:	158471	BIC:	1.223e+06
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	94.8832	0.080	1188.077	0.000	94.727	95.040
C(orientation)[T.Right]	-1.6271	0.093	-17.475	0.000	-1.810	-1.445
C(DomSex)[T.Females]	1.7543	0.116	15.088	0.000	1.526	1.982
C(Topic)[T.Immigration]	-0.8293	0.135	-6.158	0.000	-1.093	-0.565
C(orientation)[T.Right]:C(DomSex)[T.Females]	1.0406	0.151	6.890	0.000	0.745	1.337
C(orientation)[T.Right]:C(Topic)[T.Immigration]	0.5968	0.172	3.471	0.001	0.260	0.934
C(Topic)[T.Immigration]:C(DomSex)[T.Females]	-0.2035	0.197	-1.032	0.302	-0.590	0.183
C(orientation)[T.Right]:C(DomSex)[T.Females]:C(Topic)[T.Immigration]	0.3006	0.271	1.110	0.267	-0.230	0.831

Omnibus:	87326.733	Durbin-Watson:	2.007
Prob(Omnibus):	0.000	Jarque-Bera (JB):	546190.231
Skew:	-2.712	Prob(JB):	0.00
Kurtosis:	10.300	Cond. No.	17.2

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

	sum_sq	df	F	PR(>F)
Intercept	1.854582e+08	1.0	1.411527e+06	0.000000e+00
C(orientation)	4.012097e+04	1.0	3.053617e+02	2.593566e-68
C(DomSex)	2.991166e+04	1.0	2.276584e+02	2.098035e-51
C(Topic)	4.982844e+03	1.0	3.792455e+01	7.370990e-10
C(orientation):C(DomSex)	6.237700e+03	1.0	4.747529e+01	5.590848e-12
C(orientation):C(Topic)	1.582535e+03	1.0	1.204471e+01	5.195342e-04
C(Topic):C(DomSex)	1.398523e+02	1.0	1.064419e+00	3.022110e-01
C(orientation):C(DomSex):C(Topic)	1.619014e+02	1.0	1.232235e+00	2.669744e-01
Residual	2.082125e+07	158471.0	NaN	NaN

Return to relevant section

Appendix I

Table 8

Linear Model Results for Sad

OLS Regression Results

Dep. Variable:	Sad	R-squared:	0.001
Model:	OLS	Adj. R-squared:	0.000
Method:	Least Squares	F-statistic:	2.290
Date:	Thu, 10 Nov 2022	Prob (F-statistic):	0.0249
Time:	23:45:52	Log-Likelihood:	-84223.
No. Observations:	19073	AIC:	1.685e+05
Df Residuals:	19065	BIC:	1.685e+05
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	71.3273	0.426	167.278	0.000	70.492	72.163
C(orientation)[T.Right]	-1.2851	0.472	-2.720	0.007	-2.211	-0.359
C(DomSex)[T.Females]	-1.2194	0.842	-1.449	0.147	-2.869	0.430
C(Topic)[T.Immigration]	0.2287	0.672	0.340	0.734	-1.089	1.547
C(orientation)[T.Right]:C(DomSex)[T.Females]	1.1420	1.062	1.075	0.282	-0.940	3.224
C(orientation)[T.Right]:C(Topic)[T.Immigration]	-0.0057	0.800	-0.007	0.994	-1.575	1.563
C(Topic)[T.Immigration]:C(DomSex)[T.Females]	-0.3959	1.278	-0.310	0.757	-2.901	2.110
C(orientation)[T.Right]:C(DomSex)[T.Females]:C(Topic)[T.Immigration]	-0.7926	1.722	-0.460	0.645	-4.169	2.583

Omnibus:	21365.081	Durbin-Watson:	2.018
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1196.737
Skew:	-0.040	Prob(JB):	1.35e-260
Kurtosis:	1.776	Cond. No.	21.1

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

	sum_sq	df	F	PR(>F)
Intercept	1.122322e+07	1.0	27981.808061	0.000000
C(orientation)	2.968459e+03	1.0	7.400982	0.006525
C(DomSex)	8.421076e+02	1.0	2.099549	0.147359
C(Topic)	4.641734e+01	1.0	0.115728	0.733719
C(orientation):C(DomSex)	4.636101e+02	1.0	1.155876	0.282336
C(orientation):C(Topic)	2.039252e-02	1.0	0.000051	0.994311
C(Topic):C(DomSex)	3.847621e+01	1.0	0.095929	0.756774
C(orientation):C(DomSex):C(Topic)	8.492717e+01	1.0	0.211741	0.645411
Residual	7.646778e+06	19065.0	NaN	NaN

Return to relevant section

Appendix J

Table 9

Linear Model Results for Pitch

OLS Regression Results

Dep. Variable:	Pitch	R-squared:	0.049
Model:	OLS	Adj. R-squared:	0.049
Method:	Least Squares	F-statistic:	1820.
Date:	Thu, 10 Nov 2022	Prob (F-statistic):	0.00
Time:	23:47:55	Log-Likelihood:	-9.6362e+05
No. Observations:	247515	AIC:	1.927e+06
Df Residuals:	247507	BIC:	1.927e+06
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.8500	0.068	42.035	0.000	2.717	2.983
C(orientation)[T.Right]	-2.1803	0.077	-28.170	0.000	-2.332	-2.029
C(DomSex)[T.Females]	3.8962	0.104	37.483	0.000	3.692	4.100
C(Topic)[T.Immigration]	-0.6332	0.111	-5.700	0.000	-0.851	-0.415
C(orientation)[T.Right]:C(DomSex)[T.Females]	2.9622	0.134	22.181	0.000	2.700	3.224
C(orientation)[T.Right]:C(Topic)[T.Immigration]	0.6556	0.139	4.727	0.000	0.384	0.927
C(Topic)[T.Immigration]:C(DomSex)[T.Females]	-0.1211	0.171	-0.710	0.478	-0.456	0.213
C(orientation)[T.Right]:C(DomSex)[T.Females]:C(Topic)[T.Immigration]	-0.6756	0.234	-2.889	0.004	-1.134	-0.217

Omnibus:	45929.219	Durbin-Watson:	2.002
Prob(Omnibus):	0.000	Jarque-Bera (JB):	878843.282
Skew:	-0.359	Prob(JB):	0.00
Kurtosis:	12.203	Cond. No.	17.6

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

	sum_sq	df	F	PR(>F)
Intercept	2.490750e+05	1.0	1766.915408	0.000000e+00
C(orientation)	1.118616e+05	1.0	793.535586	2.593480e-174
C(DomSex)	1.980591e+05	1.0	1405.012768	1.250460e-306
C(Topic)	4.579251e+03	1.0	32.484788	1.202651e-08
C(orientation):C(DomSex)	6.935230e+04	1.0	491.978847	6.756262e-109
C(orientation):C(Topic)	3.150159e+03	1.0	22.346935	2.277009e-06
C(Topic):C(DomSex)	7.107283e+01	1.0	0.504184	4.776681e-01
C(orientation):C(DomSex):C(Topic)	1.176526e+03	1.0	8.346170	3.865347e-03
Residual	3.489008e+07	247507.0	NaN	NaN

Return to relevant section

Appendix K

Table 10

Linear Model Results for Yaw

OLS Regression Results

Dep. Variable:	Yaw	R-squared:	0.001
Model:	OLS	Adj. R-squared:	0.001
Method:	Least Squares	F-statistic:	45.84
Date:	Thu, 10 Nov 2022	Prob (F-statistic):	2.34e-65
Time:	23:49:49	Log-Likelihood:	-9.3469e+05
No. Observations:	247515	AIC:	1.869e+06
Df Residuals:	247507	BIC:	1.869e+06
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.4230	0.060	7.012	0.000	0.305	0.541
C(orientation)[T.Right]	-0.3984	0.069	-5.785	0.000	-0.533	-0.263
C(DomSex)[T.Females]	-0.8938	0.092	-9.665	0.000	-1.075	-0.713
C(Topic)[T.Immigration]	0.1866	0.099	1.888	0.059	-0.007	0.380
C(orientation)[T.Right]:C(DomSex)[T.Females]	0.2765	0.119	2.327	0.020	0.044	0.509
C(orientation)[T.Right]:C(Topic)[T.Immigration]	0.0091	0.123	0.073	0.942	-0.233	0.251
C(Topic)[T.Immigration]:C(DomSex)[T.Females]	-0.3802	0.152	-2.504	0.012	-0.678	-0.083
C(orientation)[T.Right]:C(DomSex)[T.Females]:C(Topic)[T.Immigration]	0.2702	0.208	1.299	0.194	-0.138	0.678

Omnibus:	43551.297	Durbin-Watson:	2.003
Prob(Omnibus):	0.000	Jarque-Bera (JB):	856779.260
Skew:	-0.270	Prob(JB):	0.00
Kurtosis:	12.099	Cond. No.	17.6

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

	sum_sq	df	F	PR(>F)
Intercept	5.486608e+03	1.0	49.172137	2.350539e-12
C(orientation)	3.734118e+03	1.0	33.465947	7.260961e-09
C(DomSex)	1.042213e+04	1.0	93.405307	4.299580e-22
C(Topic)	3.978399e+02	1.0	3.565525	5.899225e-02
C(orientation):C(DomSex)	6.043008e+02	1.0	5.415871	1.995530e-02
C(orientation):C(Topic)	6.008543e-01	1.0	0.005385	9.415018e-01
C(Topic):C(DomSex)	6.998853e+02	1.0	6.272519	1.226310e-02
C(orientation):C(DomSex):C(Topic)	1.881589e+02	1.0	1.686319	1.940881e-01
Residual	2.761674e+07	247507.0	NaN	NaN

Return to relevant section

Appendix L
Figure 28
ROC Plots for White Males – Gun – All Images

Figure 28a - Age	Figure 28b - Pitch, Yaw	Figure 28c - Emotions

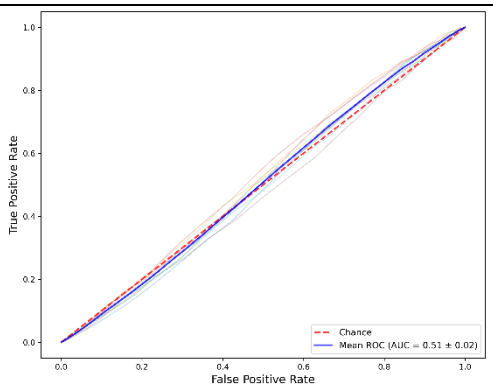
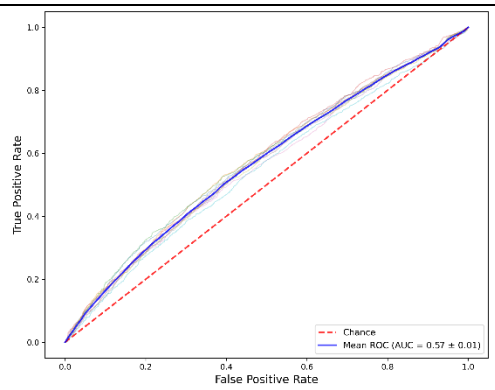
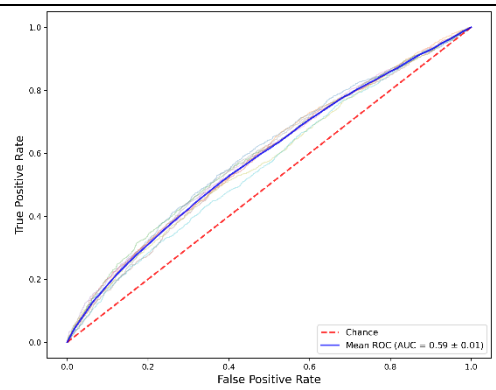

		

Figure 28d - Sex Typicality	**Figure 28e - Combined Apyest**	**Figure 28f - Features**

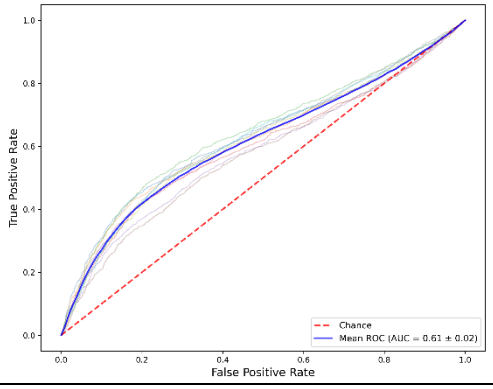
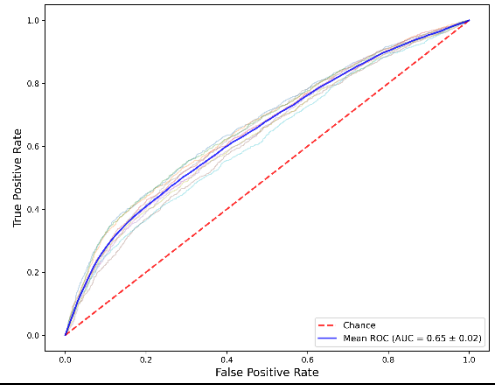
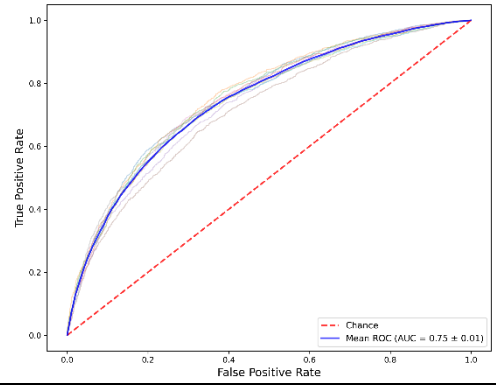

		

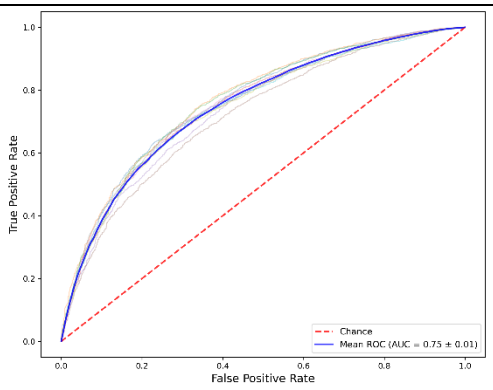
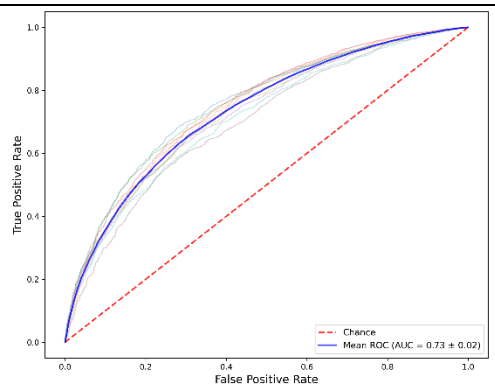
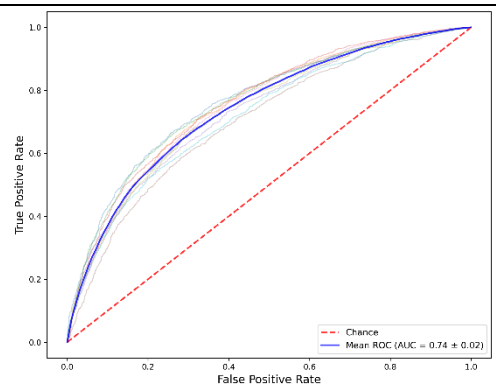
Figure 28g - Features + Apyest	**Figure 28h - Features (Masked)**	**Figure 28i - Features (M) + Apyest**
		

Figure 28j - Point Coordinates	Figure 28k - Points + Apyest	Figure 28l - Points (No Mouth)

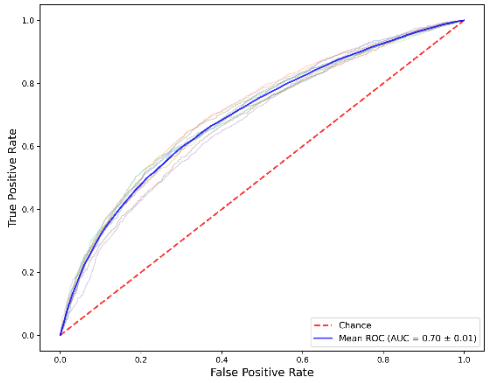
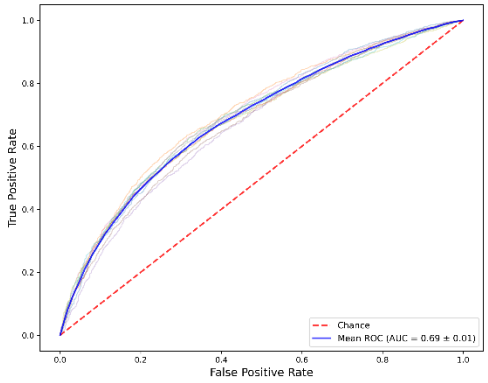
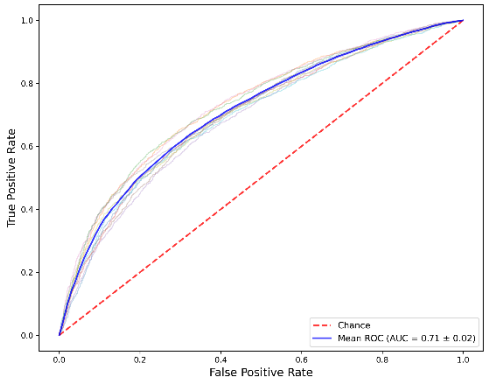

		

Figure 28m - Points (NM) + Apyest	**Figure 28n - Points (Happy)**	**Figure 28o - Points (H) + Apyest**

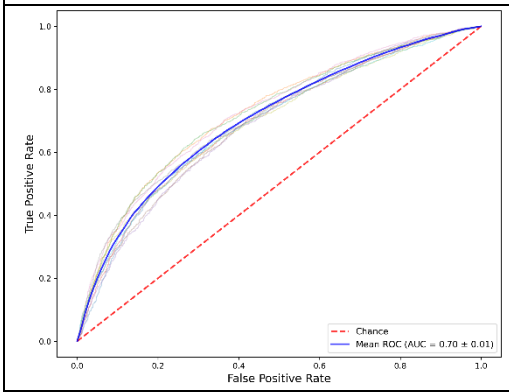
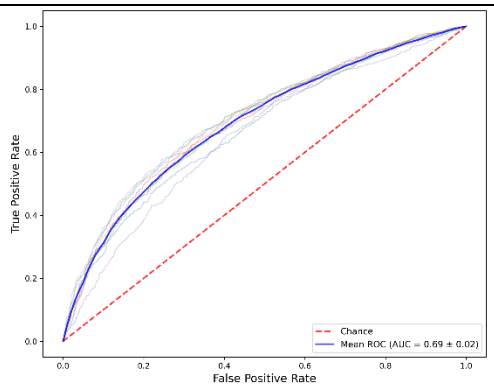
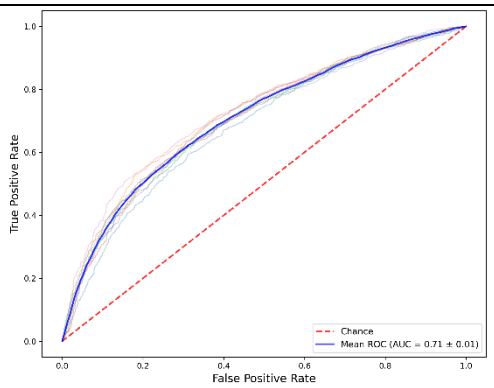

		

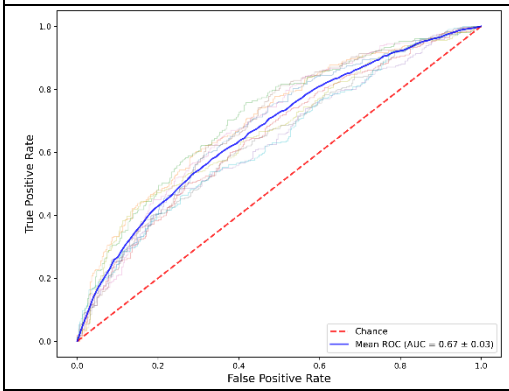
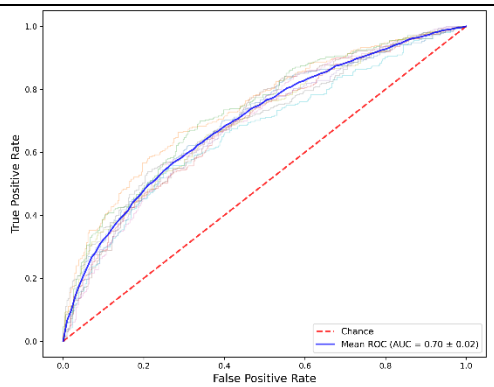
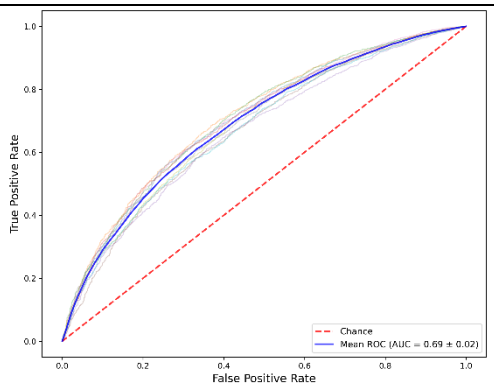
Figure 28p - Points (Neutral)	**Figure 28q - Points (N) + Apyest**	**Figure 28r - Mesh Coordinates**
		

Figure 28s - Mesh + Apyest	Figure 28t - Mesh (Happy)	Figure 28u - Mesh (H) + Apyest

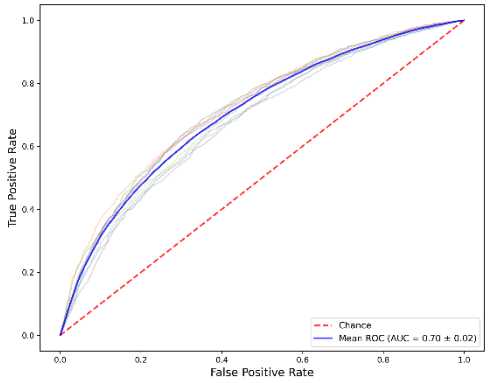
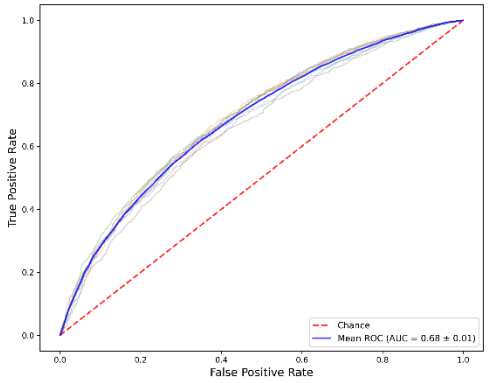
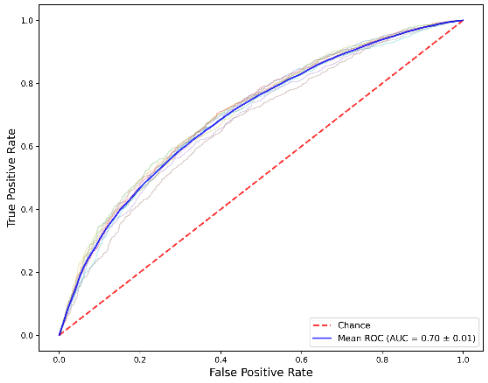
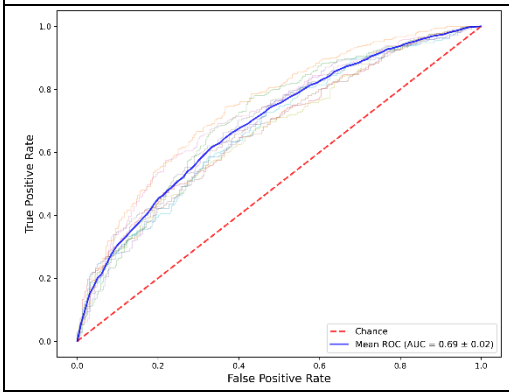
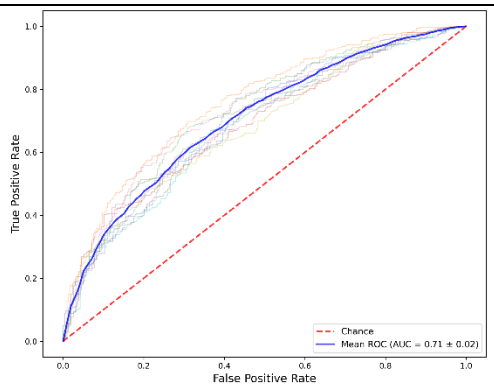

		

Figure 28v - Mesh (Neutral)	**Figure 28w - Mesh (N) + Apyest**	
		

Return to relevant section

Figure 29

ROC Plots for White Males – Gun – Reduced Images

Figure 29a - Age	Figure 29b - Pitch, Yaw	Figure 29c - Emotions

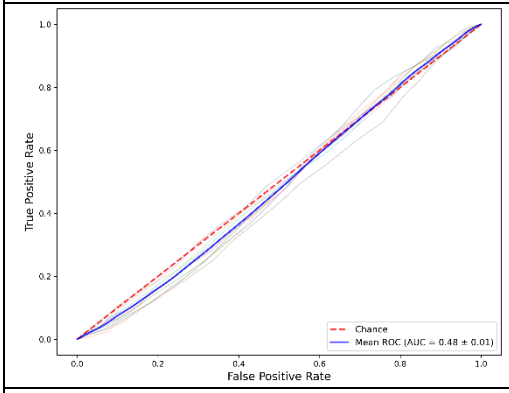
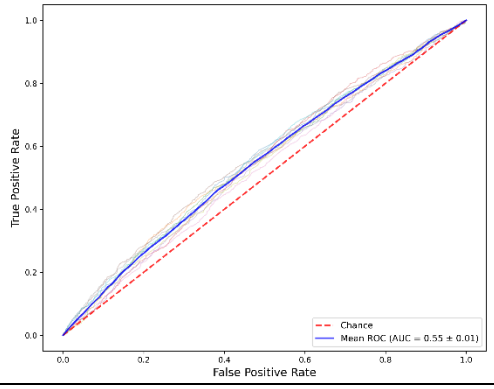
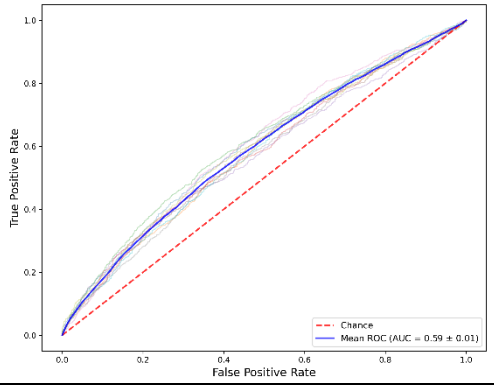

		

Figure 29d - Sex Typicality	**Figure 29e - Combined Apyest**	**Figure 29f - Features**

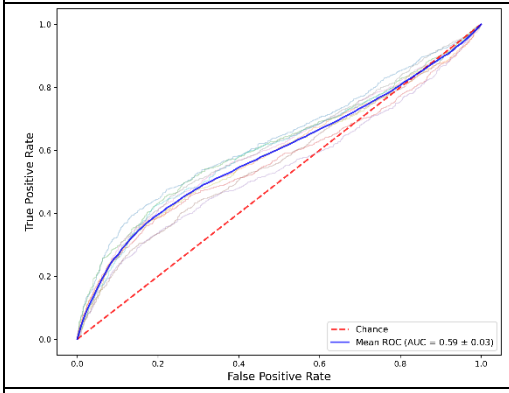
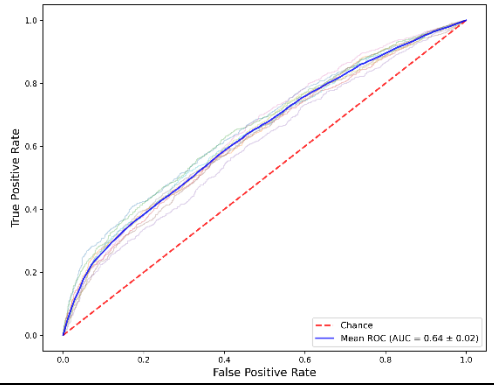
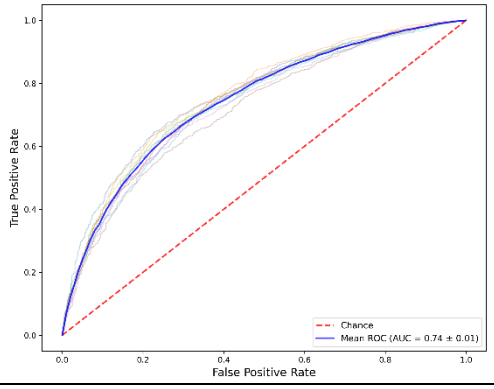

		

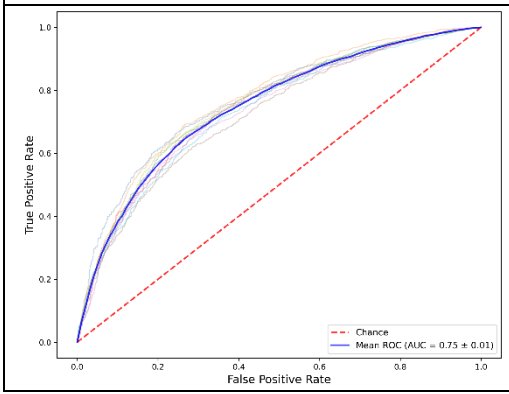
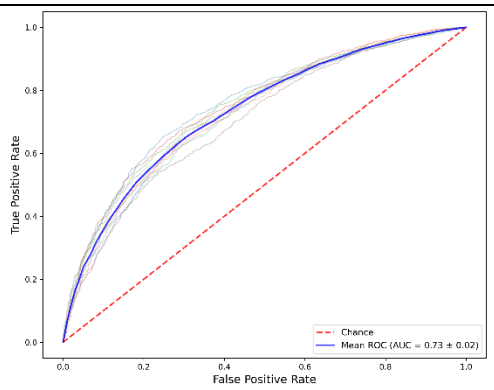
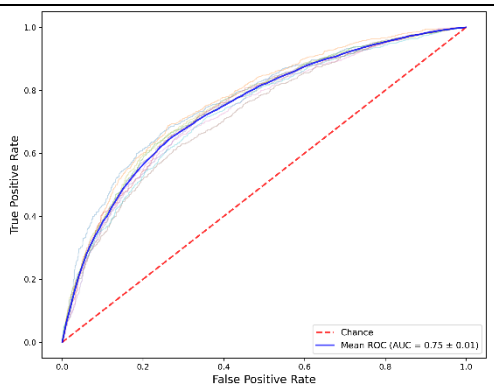
Figure 29g - Features + Apyest	**Figure 29h - Features (Masked)**	**Figure 29i - Features (M) + Apyest**
		

Figure 29j - Point Coordinates	Figure 29k - Points + Apyest	Figure 29l - Points (No Mouth)

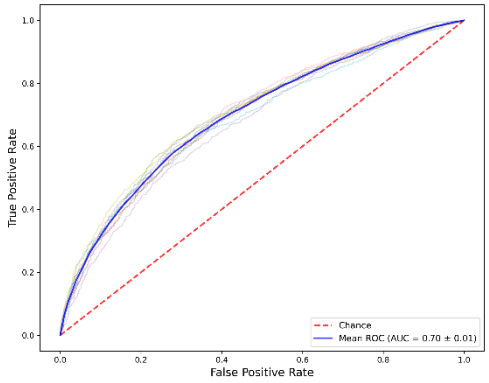
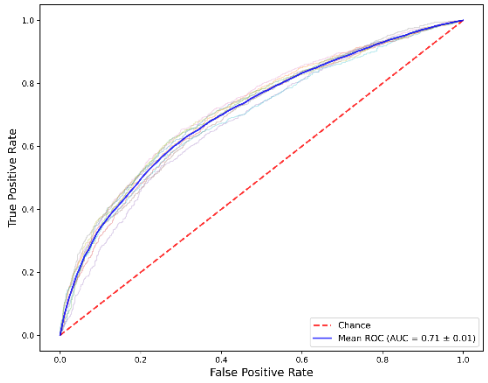
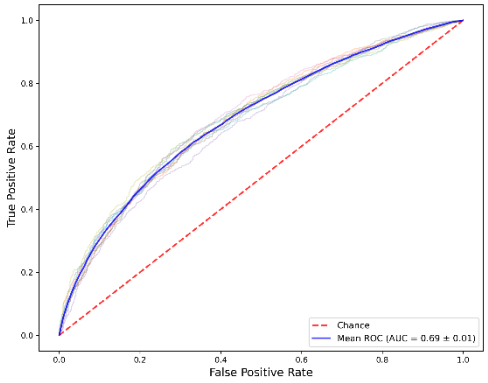

		

Figure 29m - Points (NM) + Apyest	**Figure 29n - Points (Happy)**	**Figure 29o - Points (H) + Apyest**

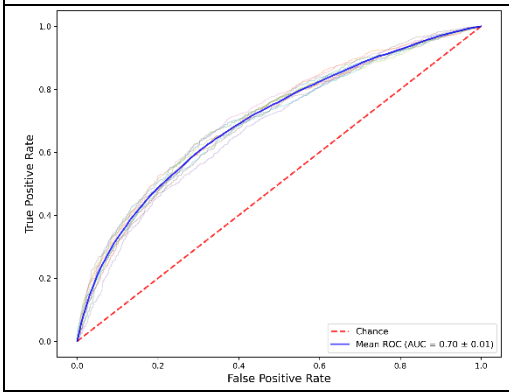
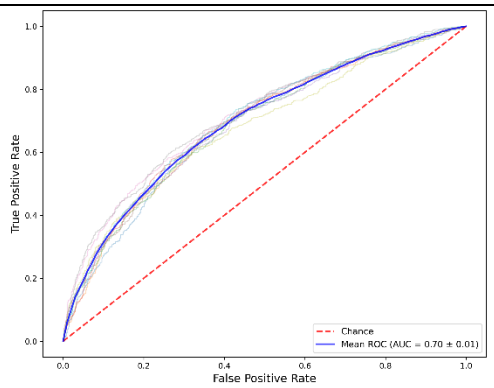
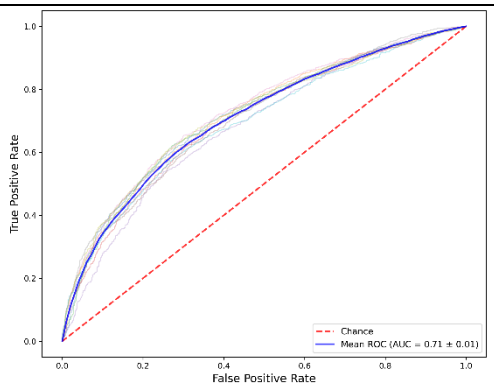

		

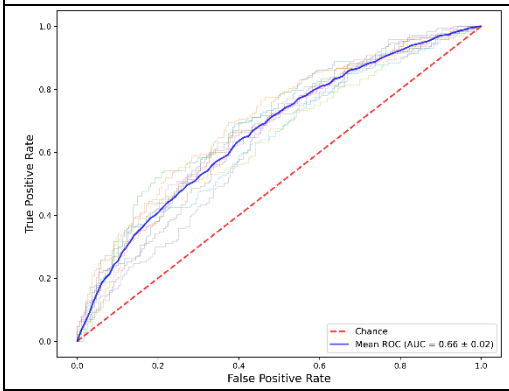
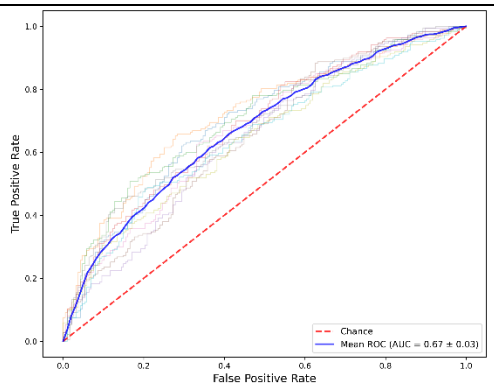
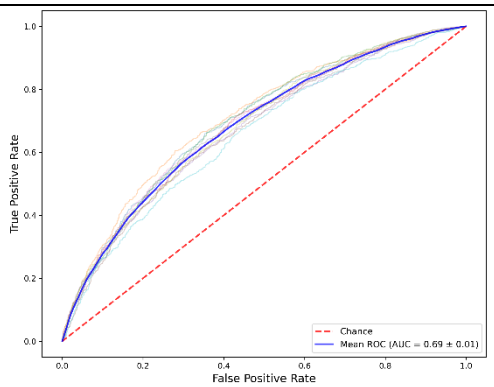
Figure 29p - Points (Neutral)	**Figure 29q - Points (N) + Apyest**	**Figure 29r - Mesh Coordinates**
		

Figure 29s - Mesh + Apyest	Figure 29t - Mesh (Happy)	Figure 29u - Mesh (H) + Apyest

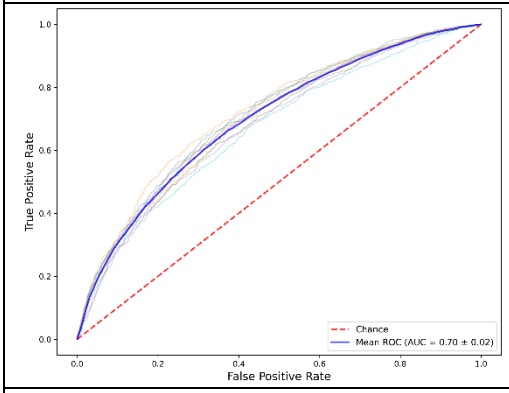
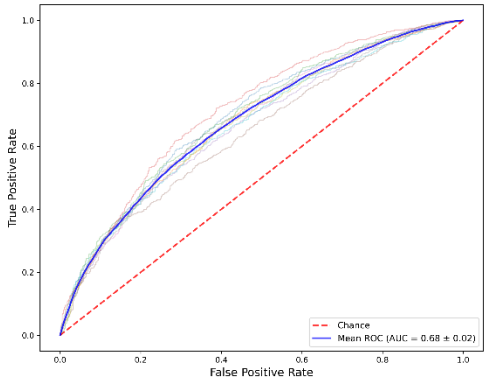
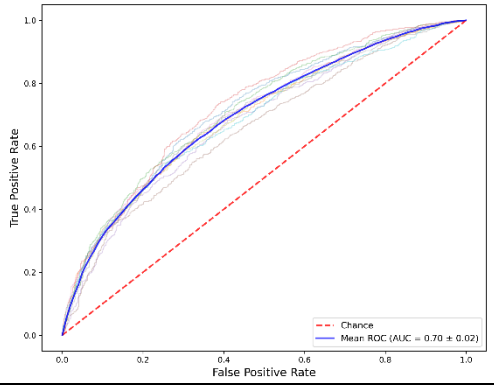

		

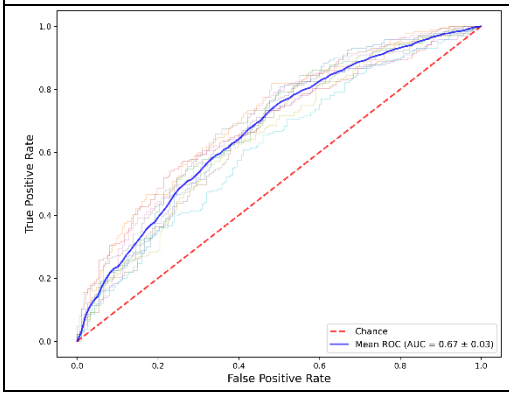
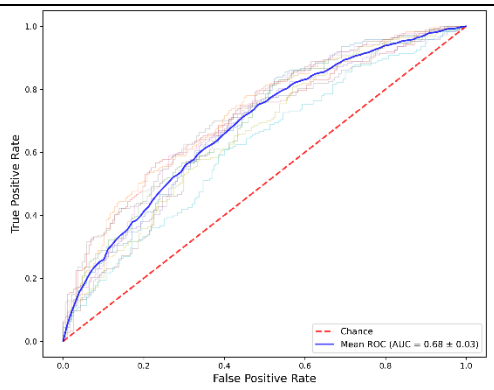
Figure 29v - Mesh (Neutral)	**Figure 29w - Mesh (N) + Apyest**	
		

Figure 30

ROC Plots for White Females – Gun – All Images

Figure 30a - Age	Figure 30b - Pitch, Yaw	Figure 30c - Emotions

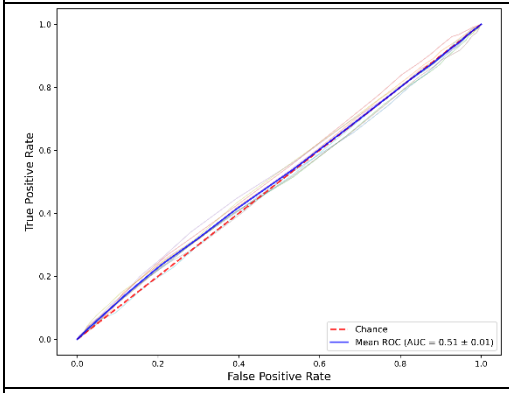
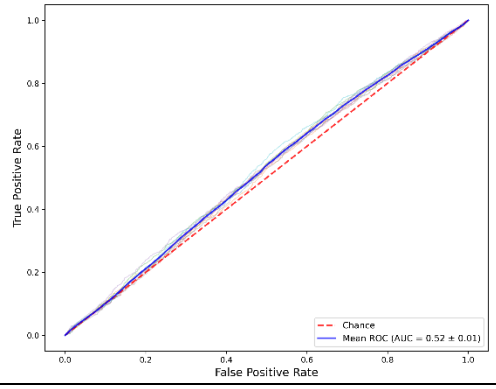
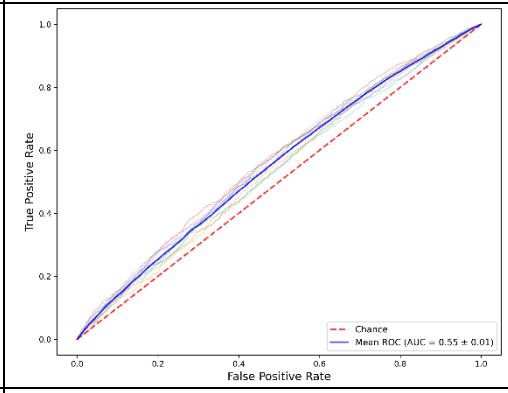

		

Figure 30d - Sex Typicality	**Figure 30e - Combined Apyest**	**Figure 30f - Features**

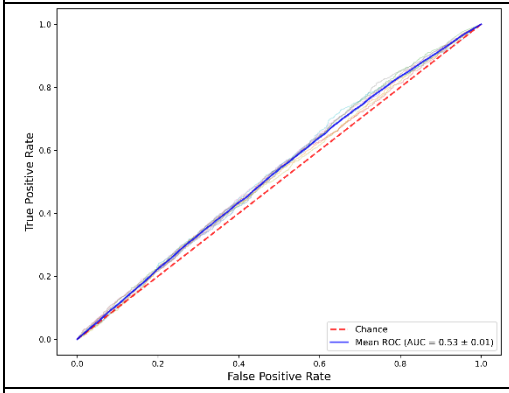
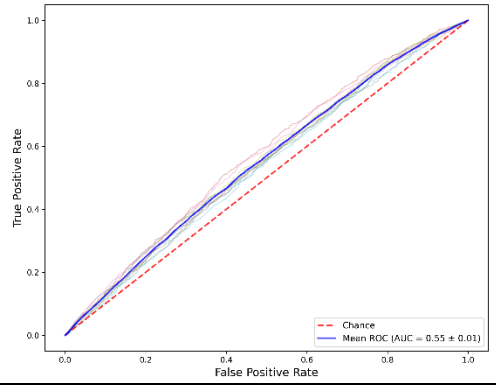
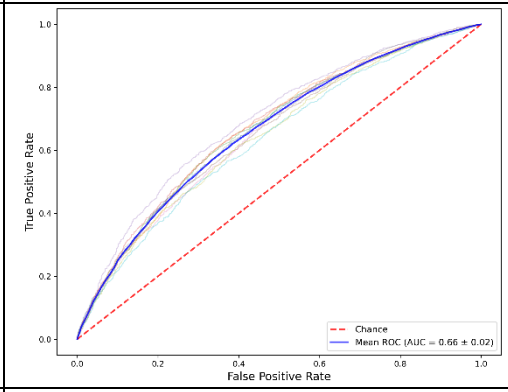

		

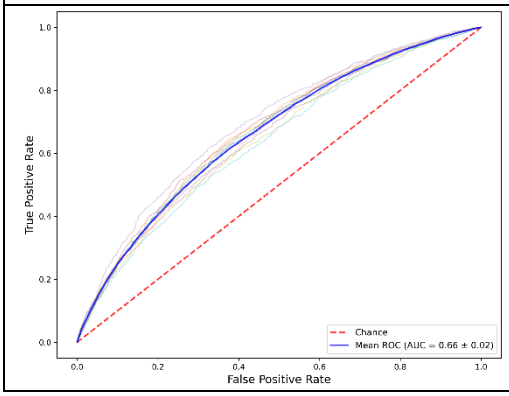
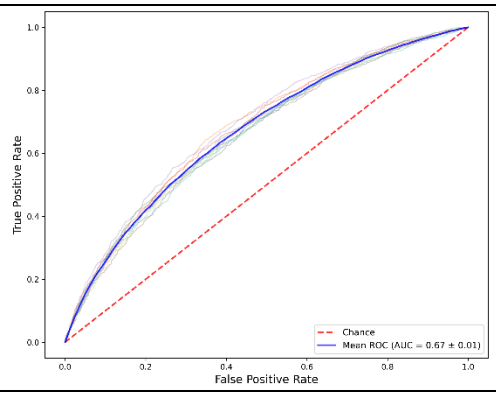
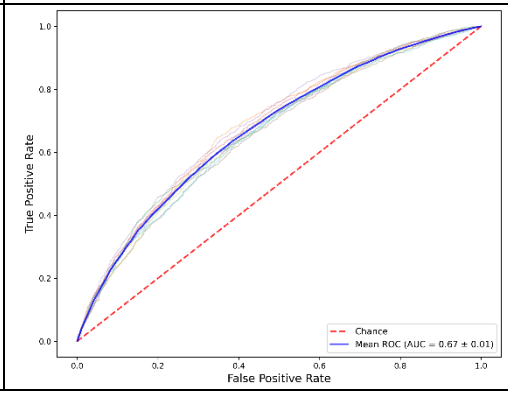
Figure 30g - Features + Apyest	**Figure 30h - Features (Masked)**	**Figure 30i - Features (M) + Apyest**
		

Figure 30j - Point Coordinates	Figure 30k - Points + Apyest	Figure 30l - Points (No Mouth)

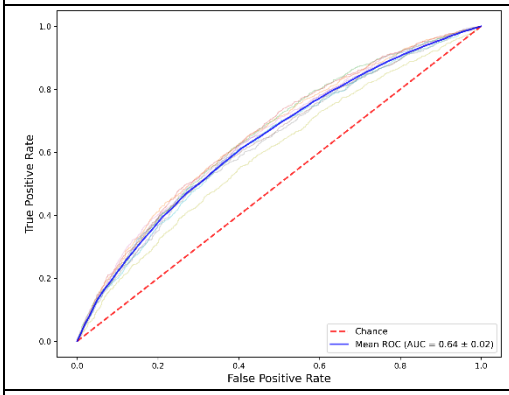
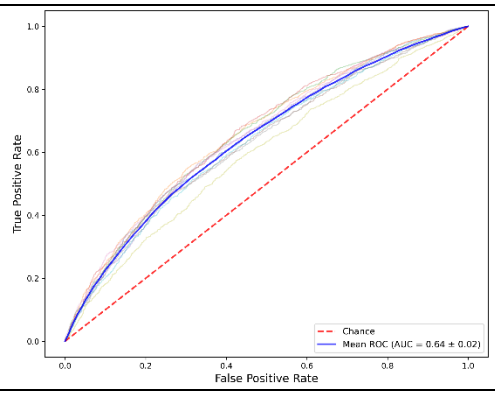
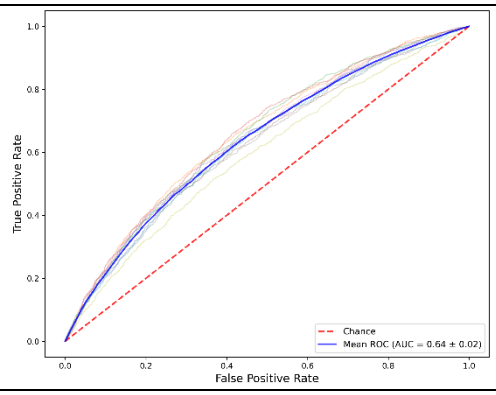

		

Figure 30m - Points (NM) + Apyest	**Figure 30n - Points (Happy)**	**Figure 30o - Points (H) + Apyest**

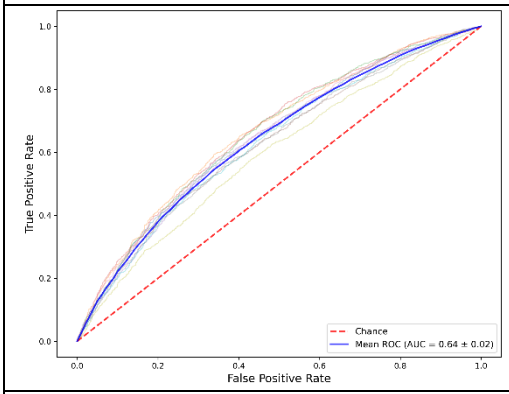
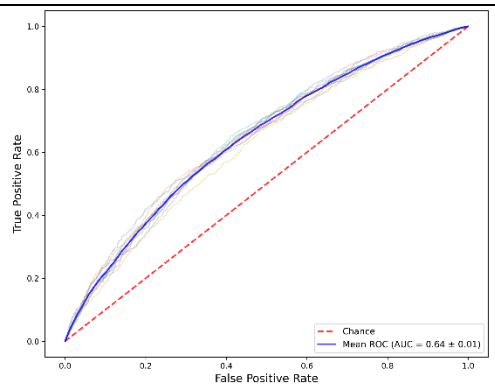
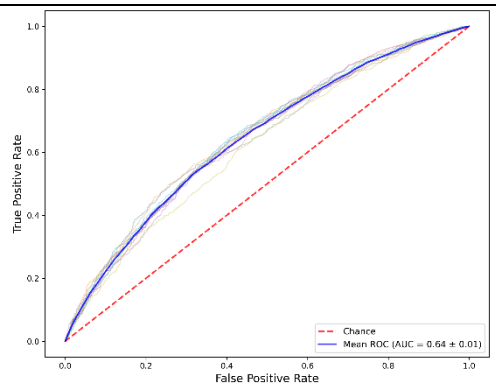

		

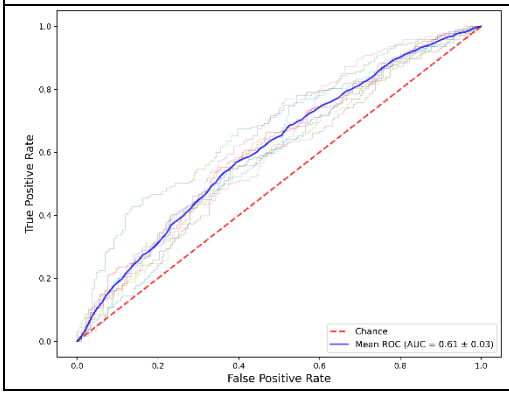
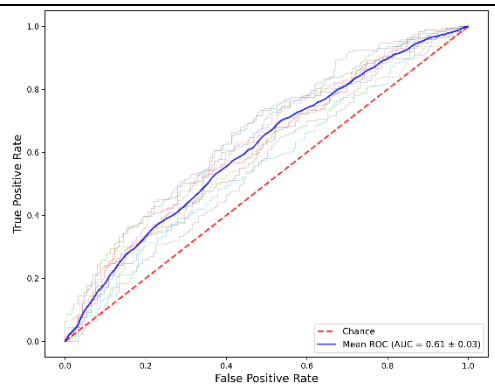
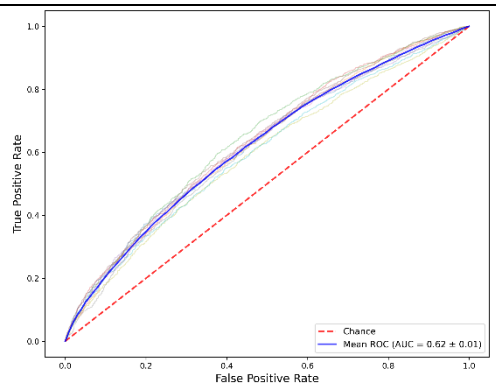
Figure 30p - Points (Neutral)	**Figure 30q - Points (N) + Apyest**	**Figure 30r - Mesh Coordinates**
		

Figure 30s - Mesh + Apyest	Figure 30t - Mesh (Happy)	Figure 30u - Mesh (H) + Apyest

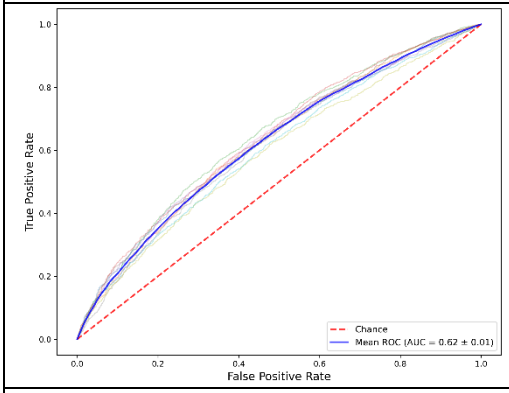
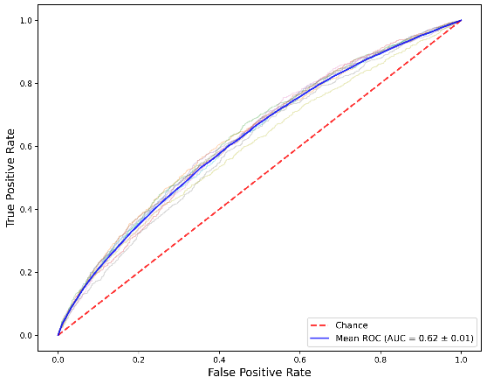
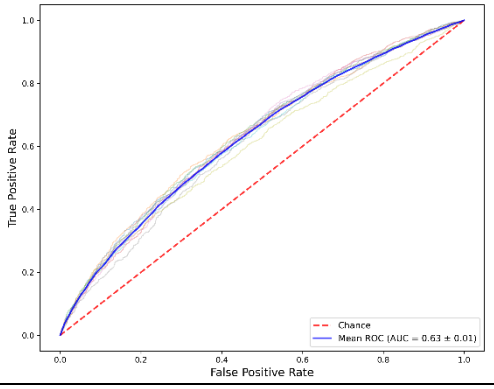

		

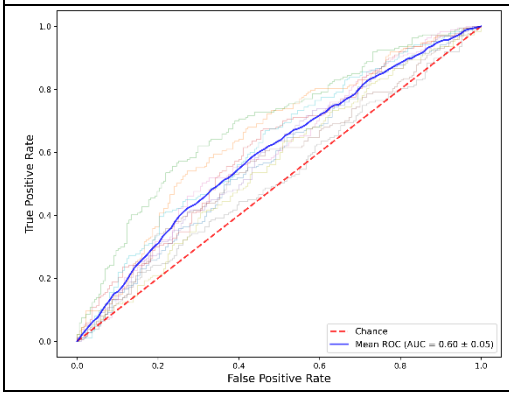
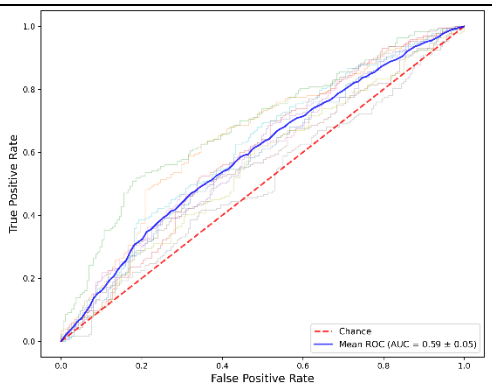
Figure 30v - Mesh (Neutral)	Figure 30w - Mesh (N) + Apyest	
		

Figure 31

ROC Plots for White Females – Gun – Reduced Images

Figure 31a - Age	Figure 31b - Pitch, Yaw	Figure 31c - Emotions

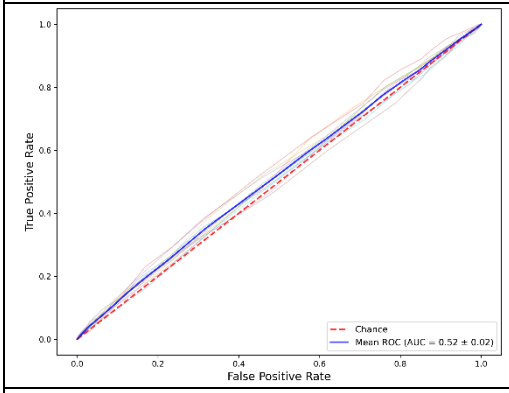
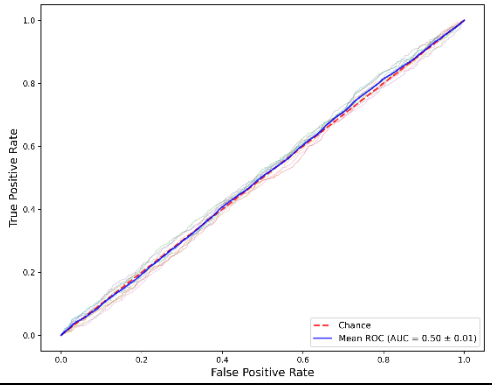
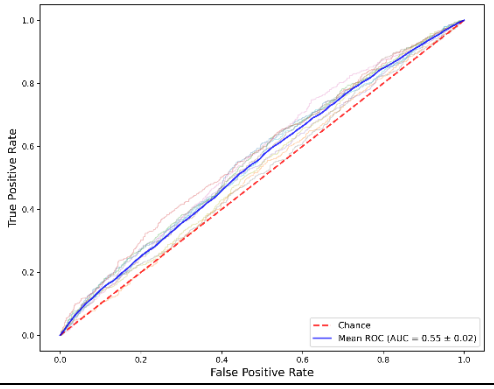

		

Figure 31d - Sex Typicality	**Figure 31e - Combined Apyest**	**Figure 31f - Features**

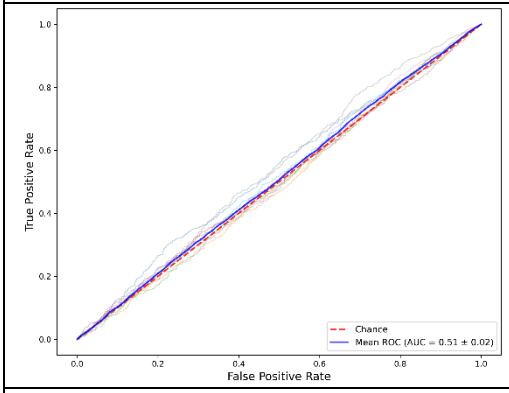
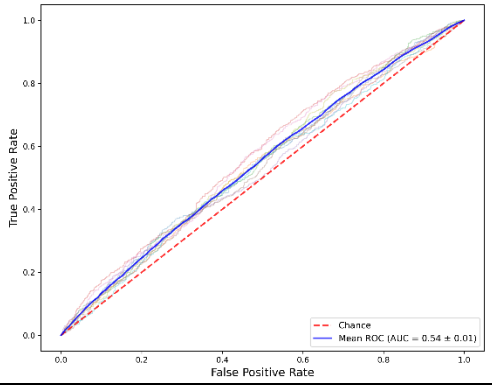
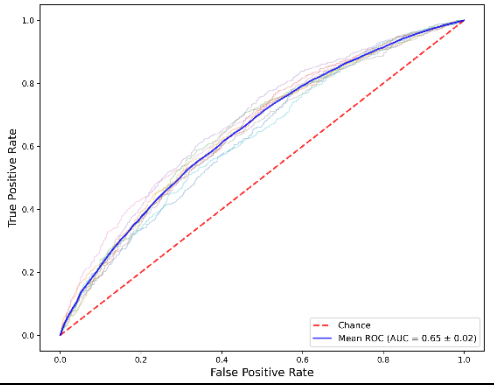

		

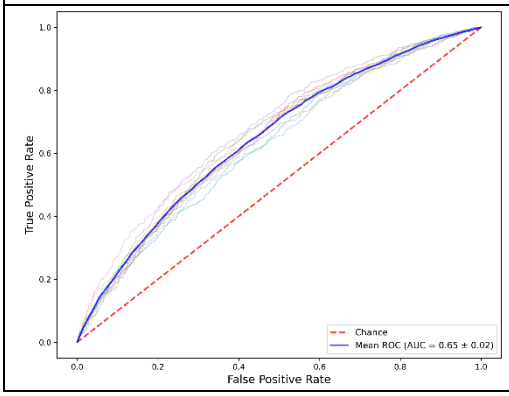
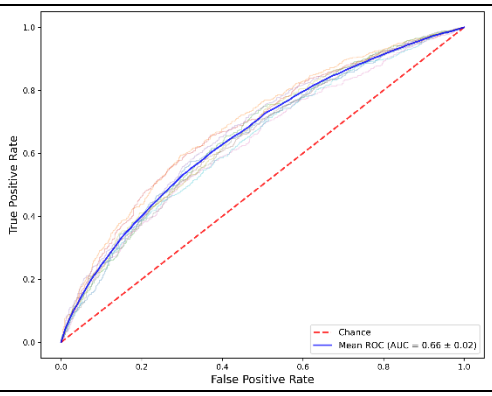
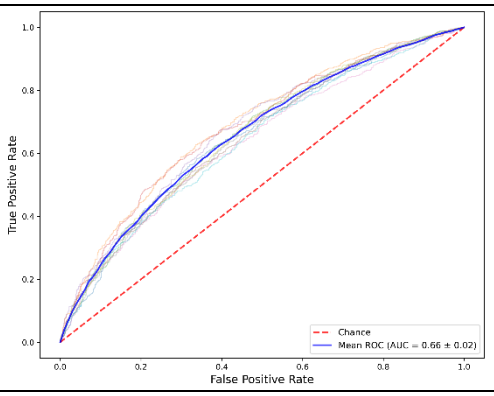
Figure 31g - Features + Apyest	**Figure 31h - Features (Masked)**	**Figure 31i - Features (M) + Apyest**
		

Figure 31j - Point Coordinates	Figure 31k - Points + Apyest	Figure 31l - Points (No Mouth)

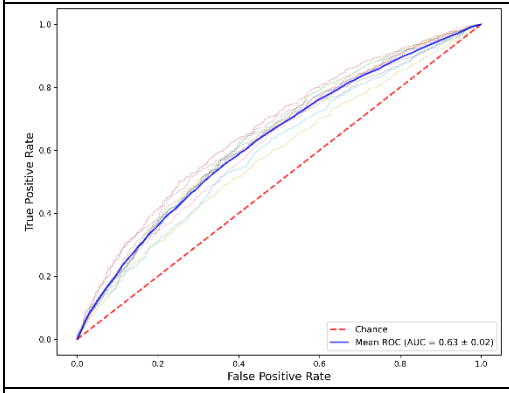
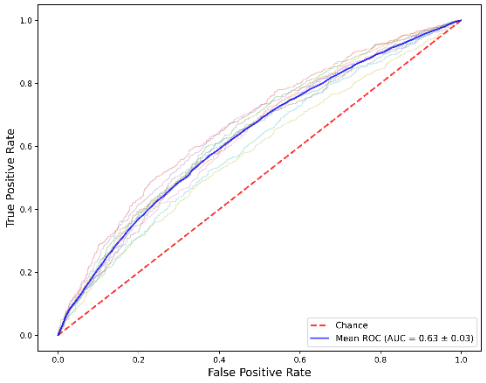
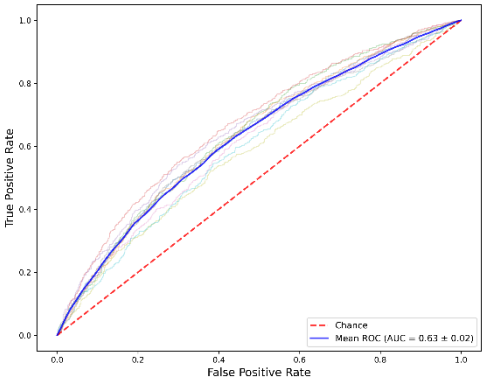

		

Figure 31m - Points (NM) + Apyest	**Figure 31n - Points (Happy)**	**Figure 31o - Points (H) + Apyest**

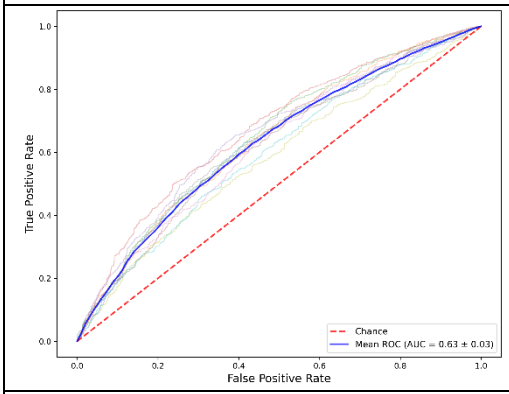
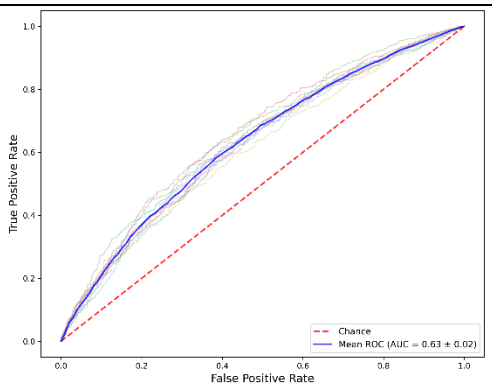
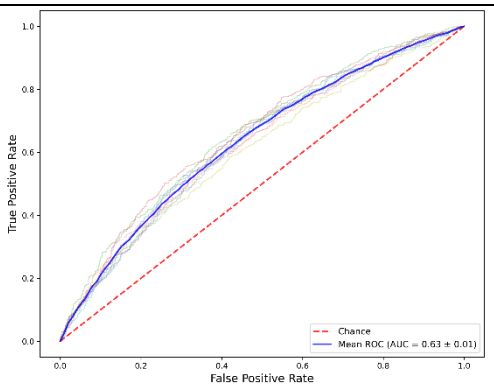

		

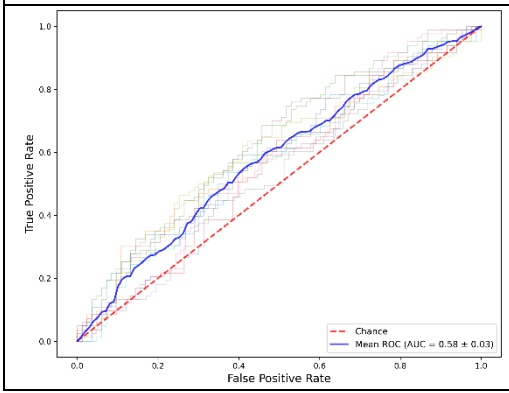
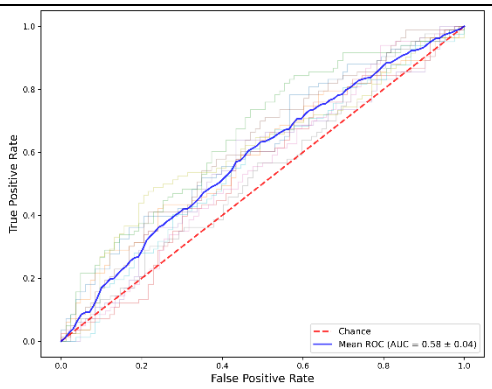
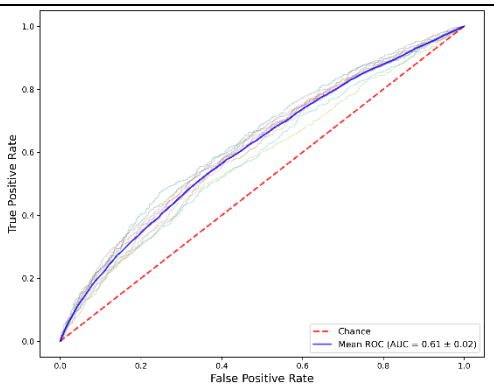
Figure 31p - Points (Neutral)	**Figure 31q - Points (N) + Apyest**	**Figure 31r - Mesh Coordinates**
		

Figure 31s - Mesh + Apyest	Figure 31t - Mesh (Happy)	Figure 31u - Mesh (H) + Apyest

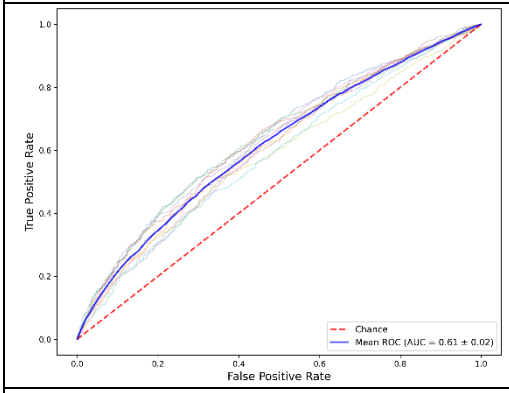
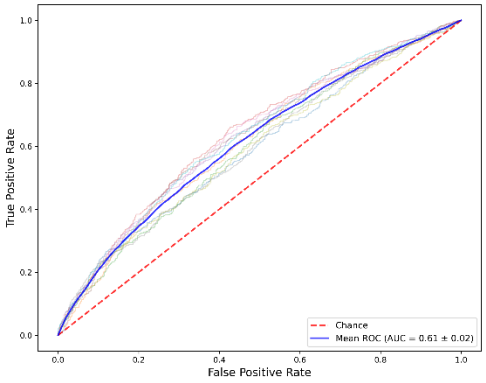
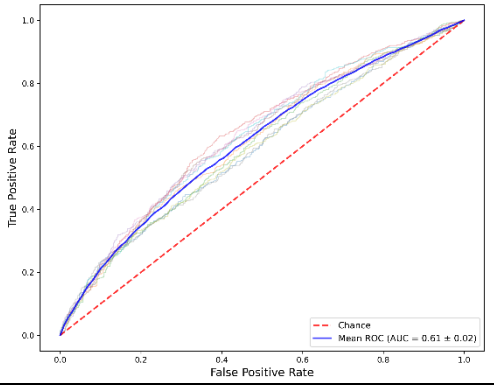

		

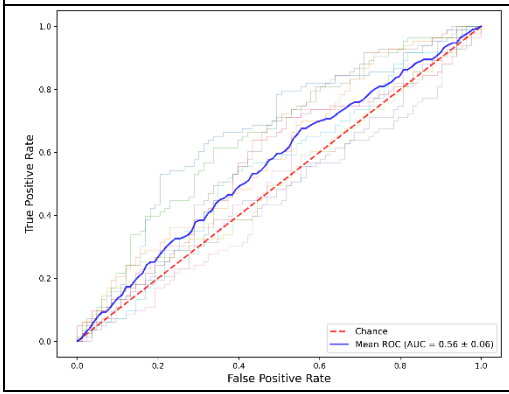
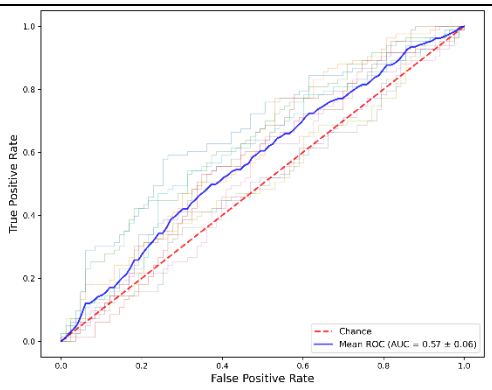
Figure 31v - Mesh (Neutral)	**Figure 31w - Mesh (N) + Apyest**	
		

Figure 32

ROC Plots for White Males – Immigration – All Images

Figure 32a - Age	Figure 32b - Pitch, Yaw	Figure 32c - Emotions

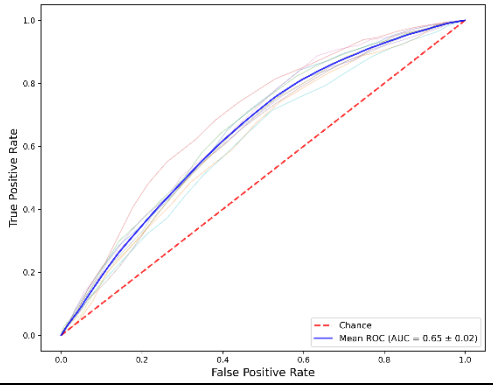
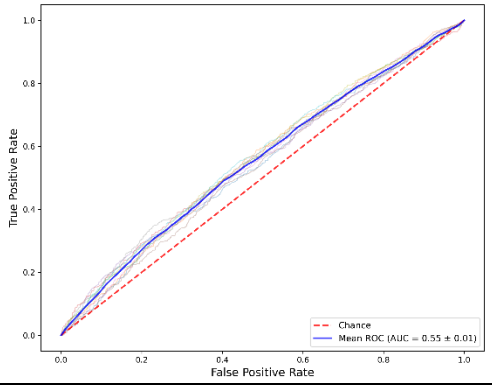
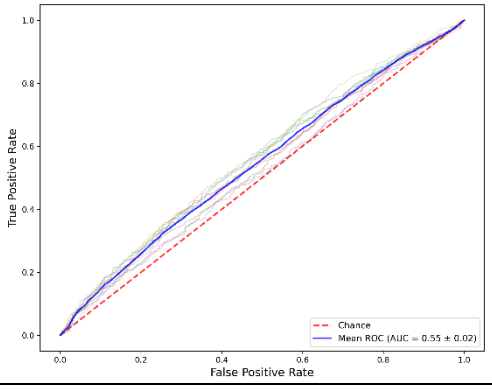

		

Figure 32d - Sex Typicality	**Figure 32e - Combined Apyest**	**Figure 32f - Features**

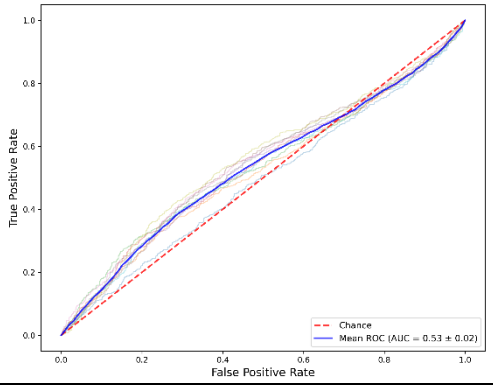
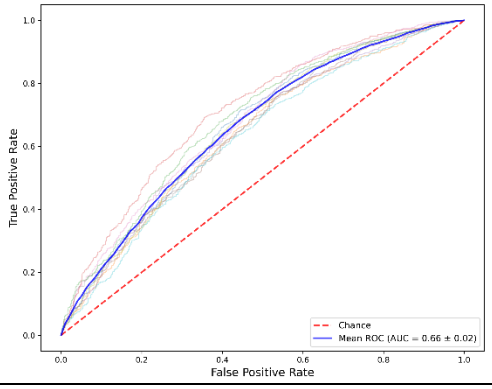
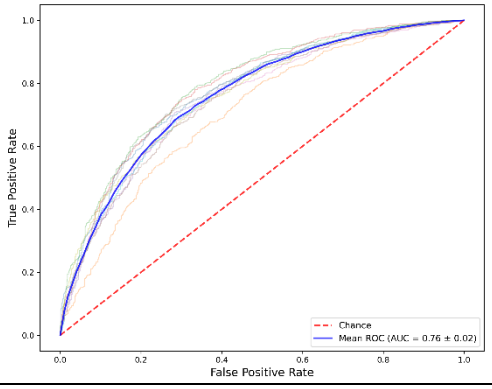

		

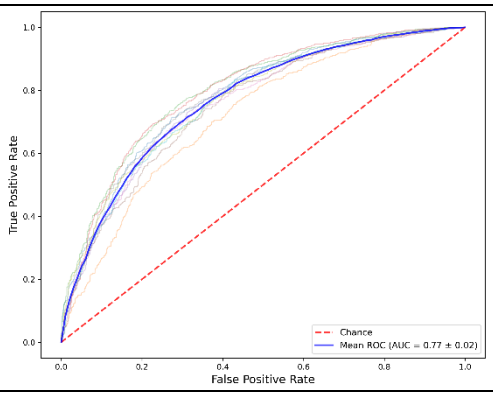
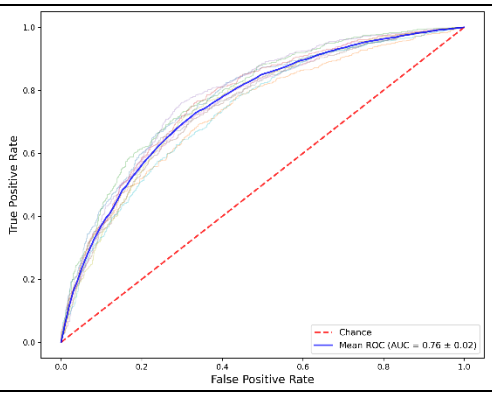
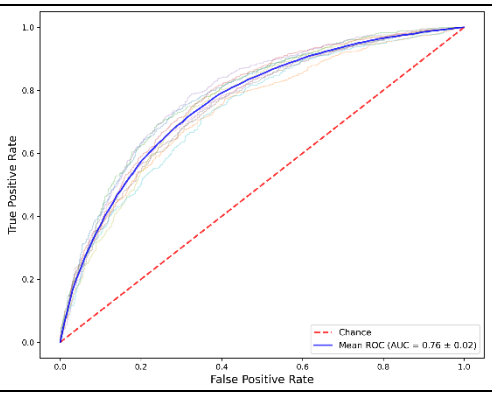
Figure 32g - Features + Apyest	**Figure 32h - Features (Masked)**	**Figure 32i - Features (M) + Apyest**
		

Figure 32j - Point Coordinates	Figure 32k - Points + Apyest	Figure 32l - Points (No Mouth)

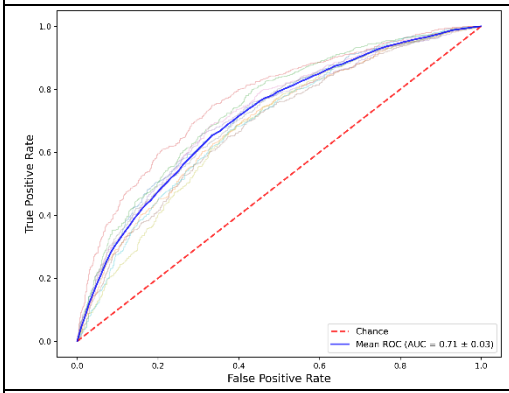
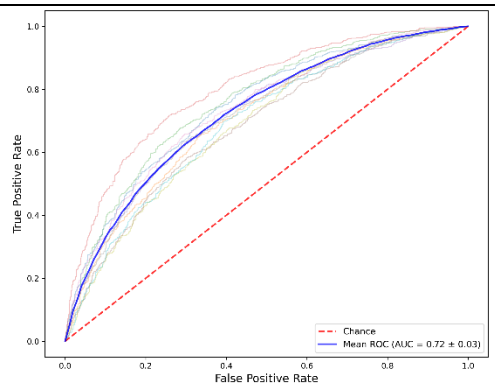
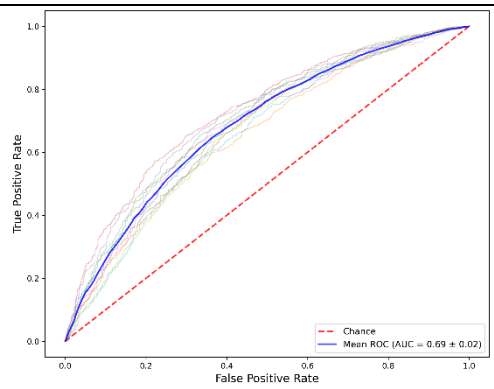

		

Figure 32m - Points (NM) + Apyest	**Figure 32n - Points (Happy)**	**Figure 32o - Points (H) + Apyest**

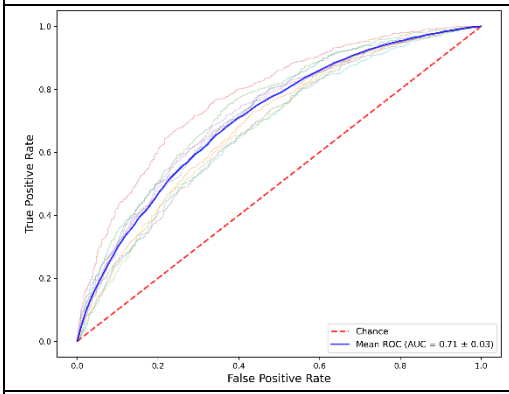
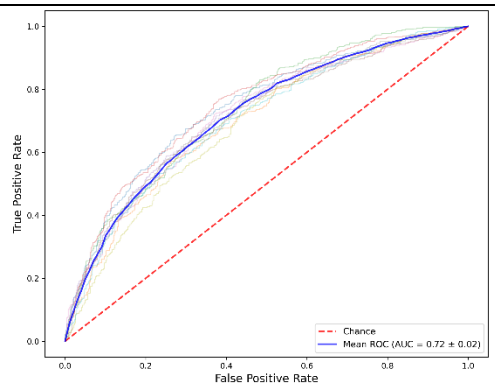
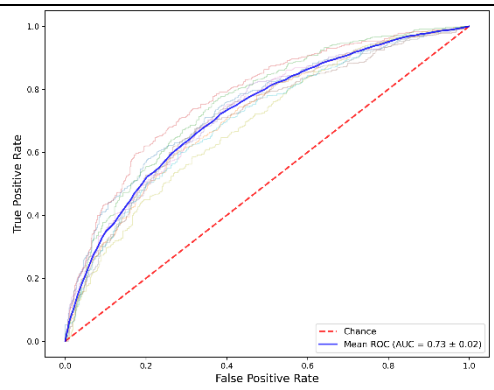

		

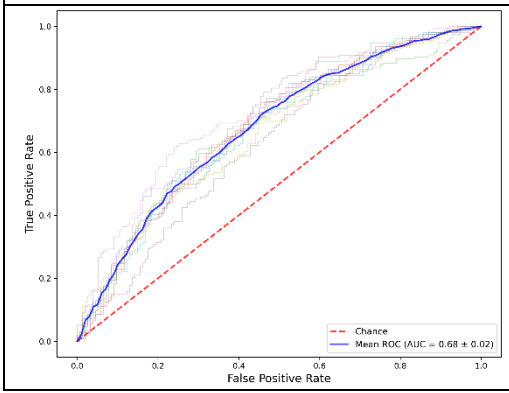
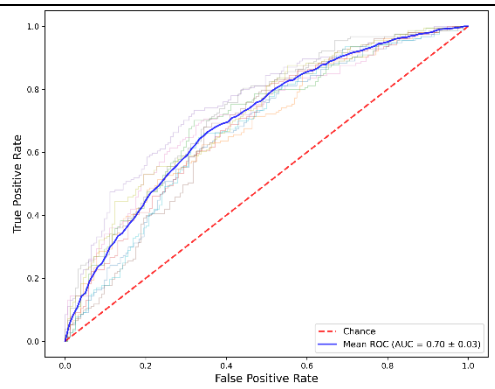
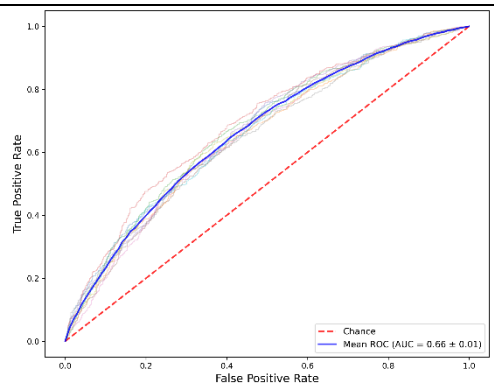
Figure 32p - Points (Neutral)	**Figure 32q - Points (N) + Apyest**	**Figure 32r - Mesh Coordinates**
		

Figure 32s - Mesh + Apyest	Figure 32t - Mesh (Happy)	Figure 32u - Mesh (H) + Apyest

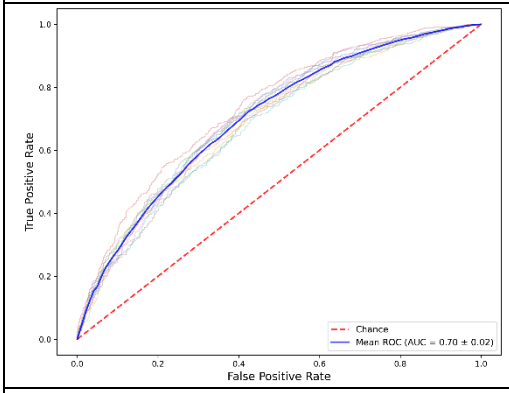
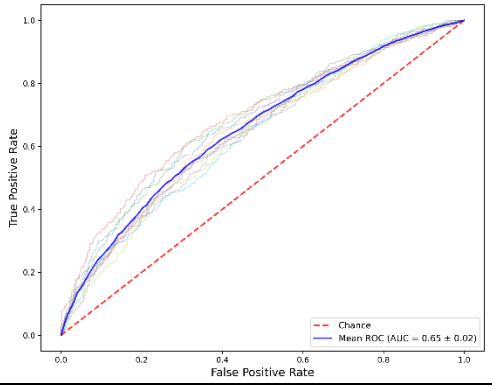
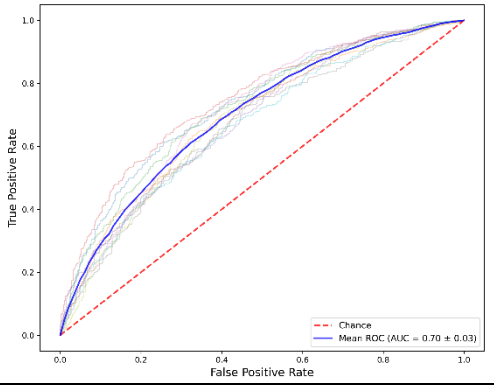

		

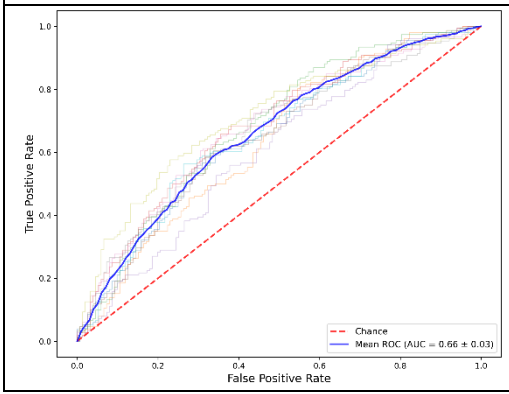
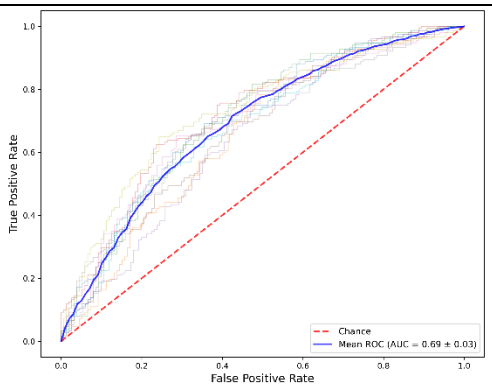
Figure 32v - Mesh (Neutral)	**Figure 32w - Mesh (N) + Apyest**	
		

Figure 33

ROC Plots for White Males – Immigration – Reduced Images

Figure 33a - Age	Figure 33b - Pitch, Yaw	Figure 33c - Emotions

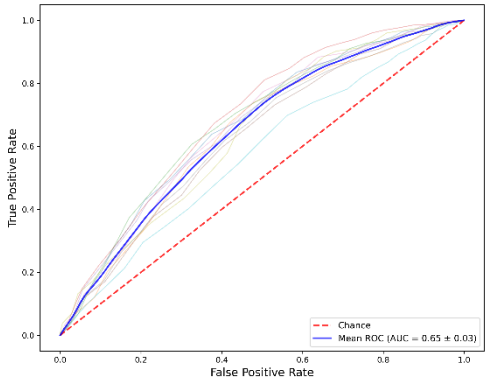
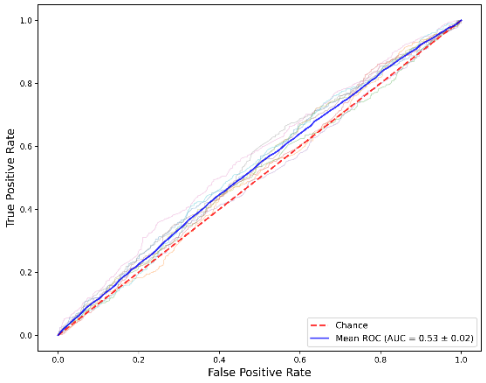
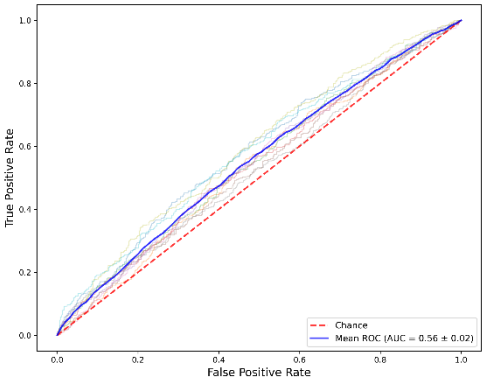

		

Figure 33d - Sex Typicality	**Figure 33e - Combined Apyest**	**Figure 33f - Features**

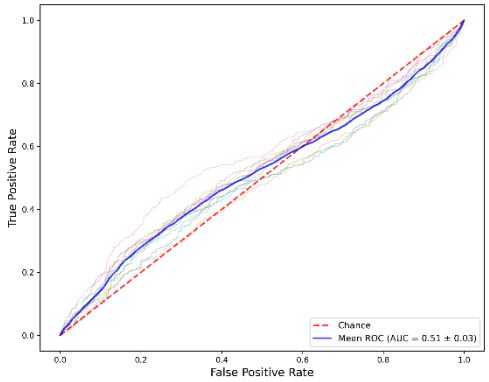
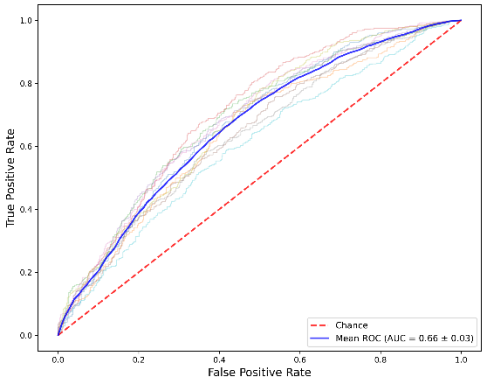
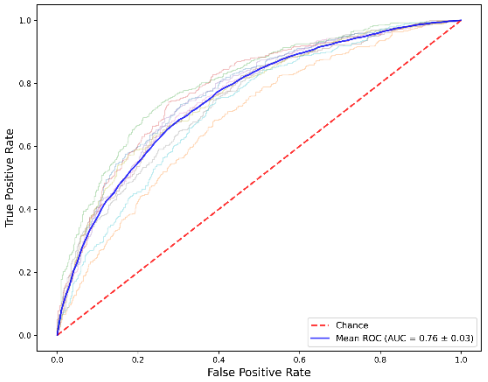

		

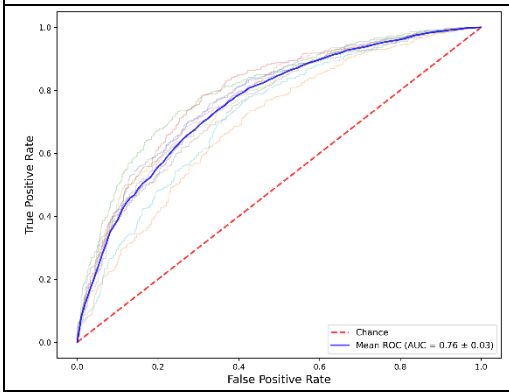
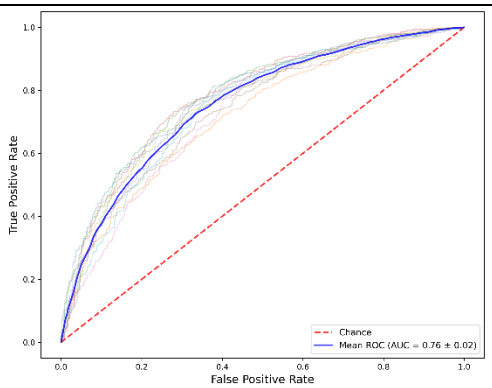
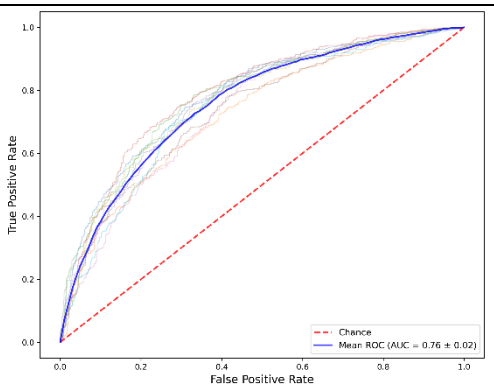
Figure 33g - Features + Apyest	**Figure 33h - Features (Masked)**	**Figure 33i - Features (M) + Apyest**
		

Figure 33j - Point Coordinates	Figure 33k - Points + Apyest	Figure 33l - Points (No Mouth)

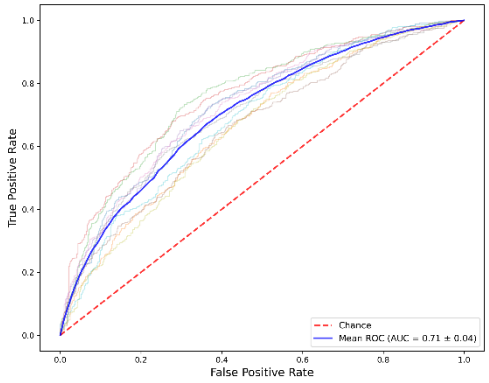
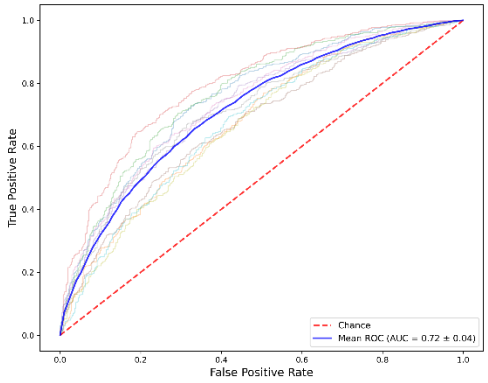
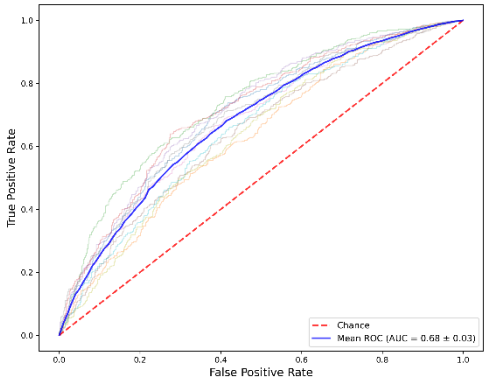

		

Figure 33m - Points (NM) + Apyest	**Figure 33n - Points (Happy)**	**Figure 33o - Points (H) + Apyest**

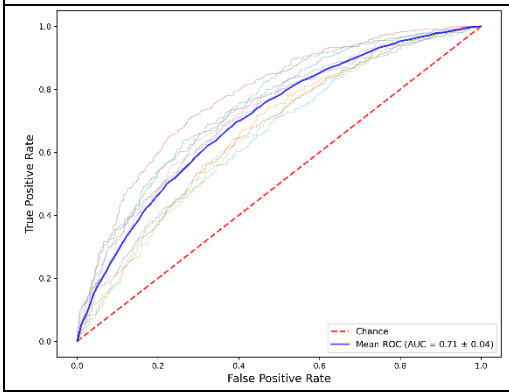
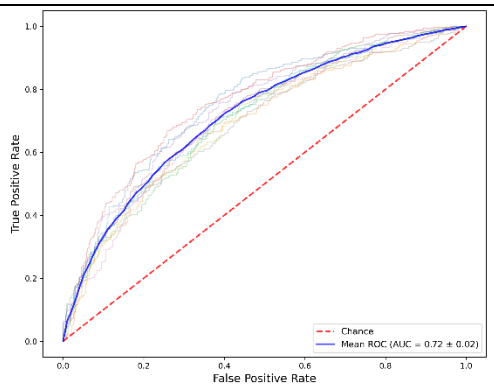
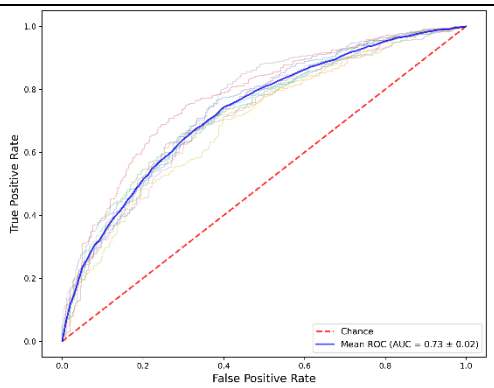

		

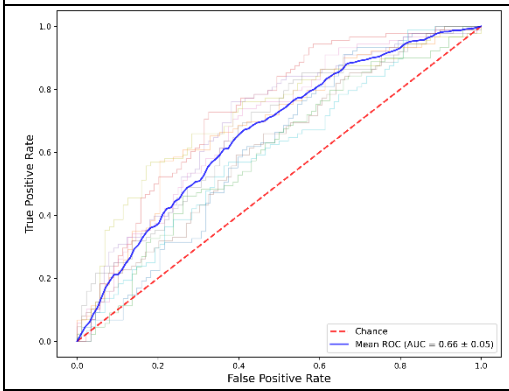
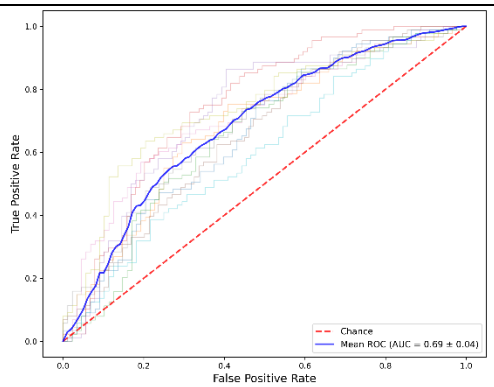
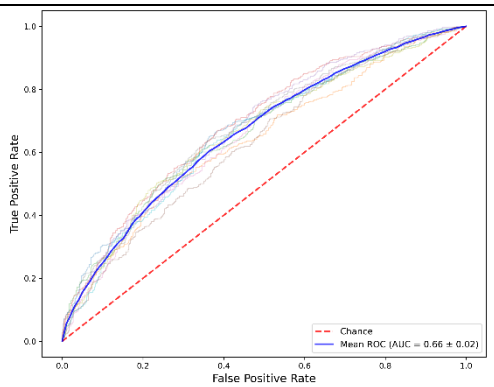
Figure 33p - Points (Neutral)	**Figure 33q - Points (N) + Apyest**	**Figure 33r - Mesh Coordinates**
		

Figure 33s - Mesh + Apyest	Figure 33t - Mesh (Happy)	Figure 33u - Mesh (H) + Apyest

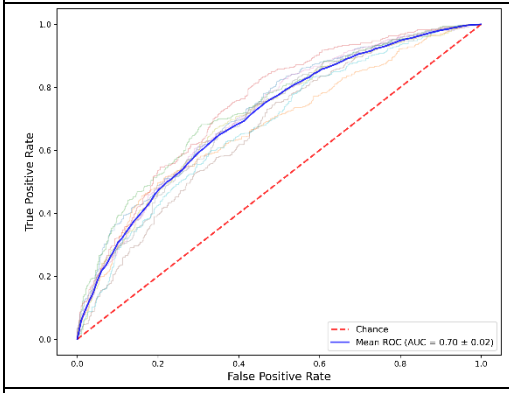
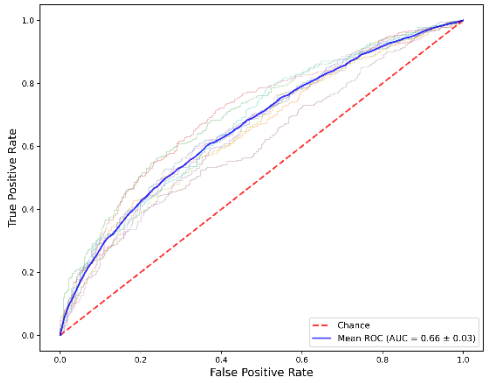
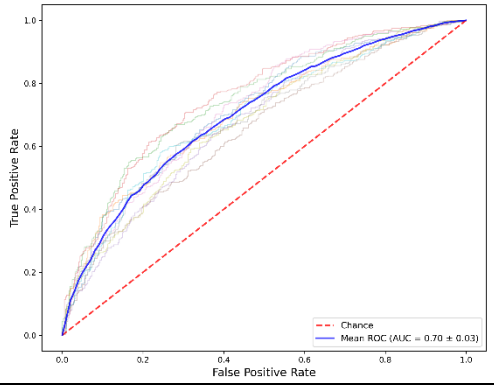

		

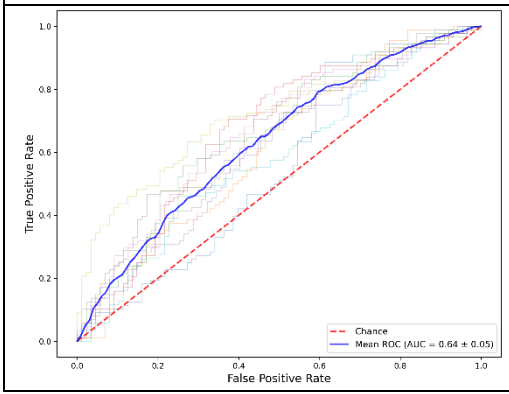
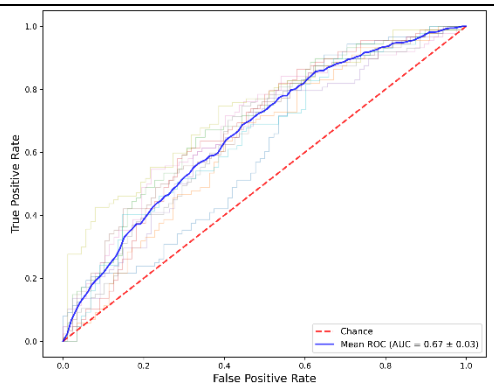
Figure 33v - Mesh (Neutral)	**Figure 33w - Mesh (N) + Apyest**	
		

Figure 34

ROC Plots for White Females – Immigration – All Images

Figure 34a - Age	Figure 34b - Pitch, Yaw	Figure 34c - Emotions

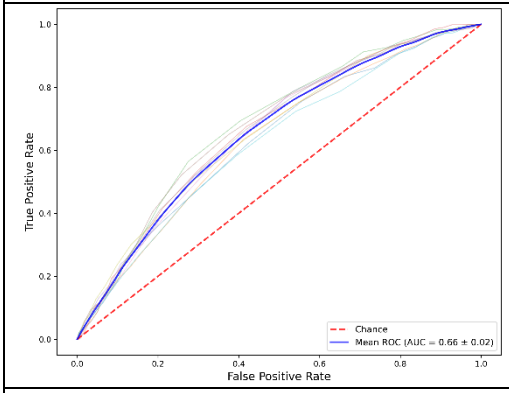
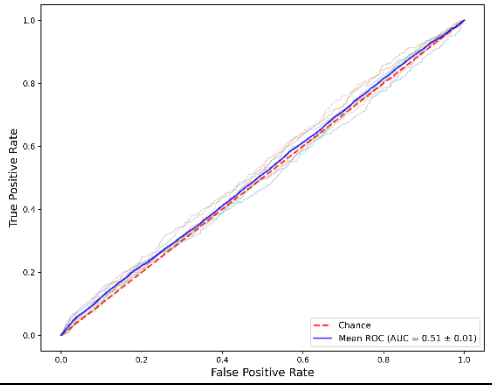
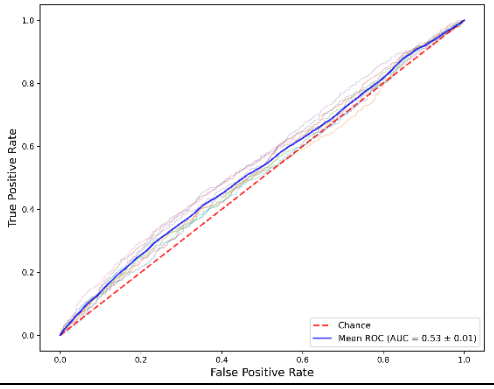

		

Figure 34d - Sex Typicality	**Figure 34e - Combined Apyest**	**Figure 34f - Features**

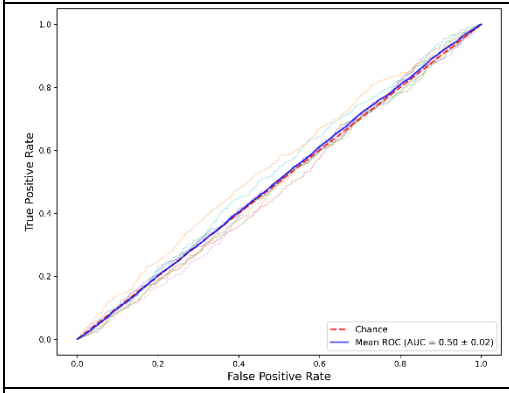
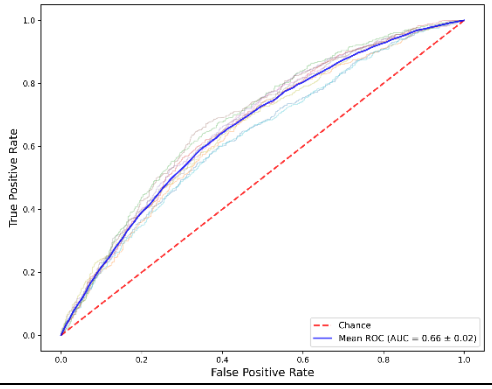
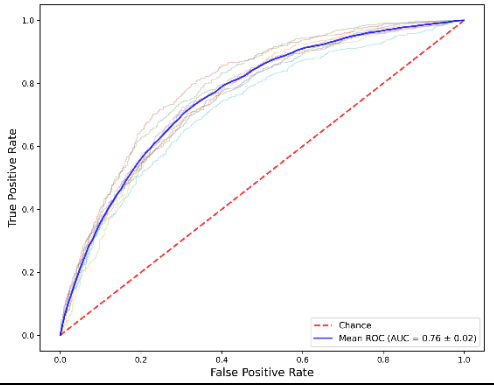

		

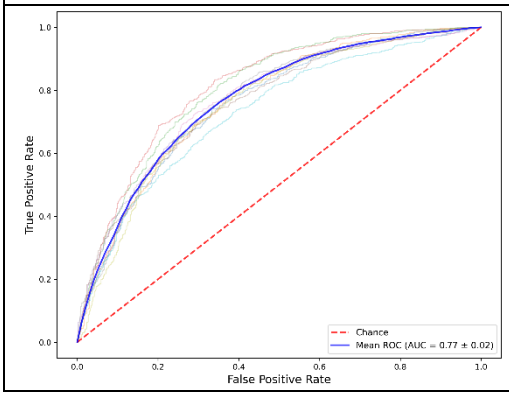
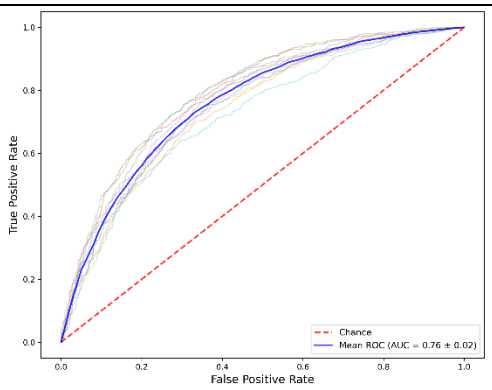
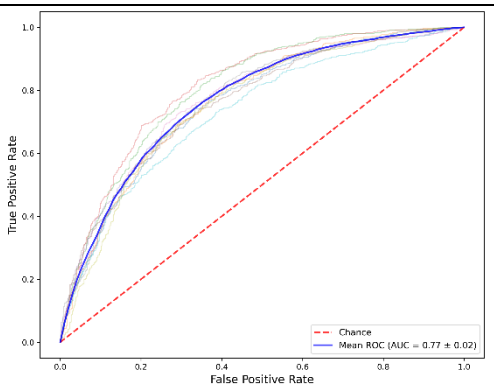
Figure 34g - Features + Apyest	**Figure 34h - Features (Masked)**	**Figure 34i - Features (M) + Apyest**
		

Figure 34j - Point Coordinates	Figure 34k - Points + Apyest	Figure 34l - Points (No Mouth)

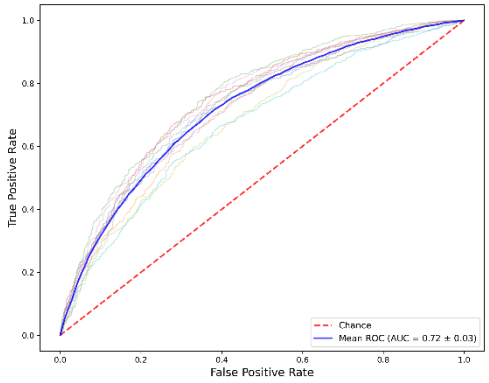
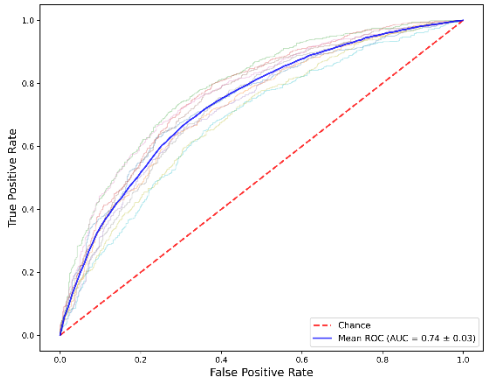
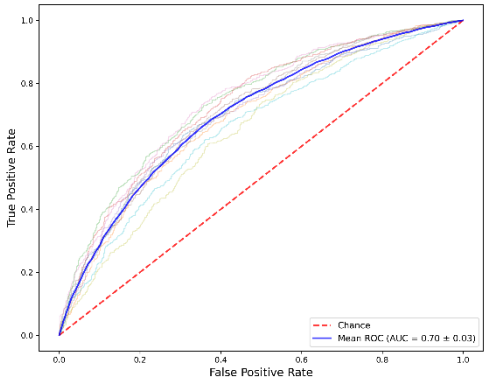

		

Figure 34m - Points (NM) + Apyest	**Figure 34n - Points (Happy)**	**Figure 34o - Points (H) + Apyest**

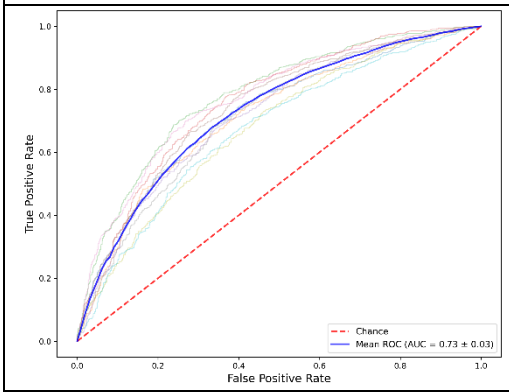
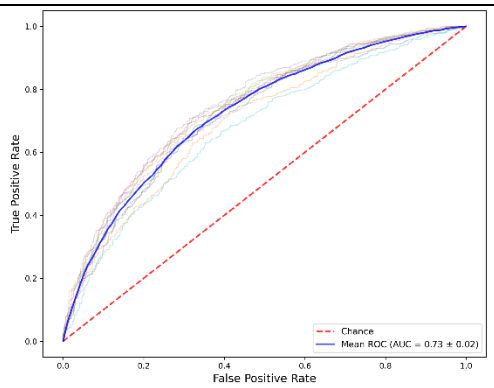
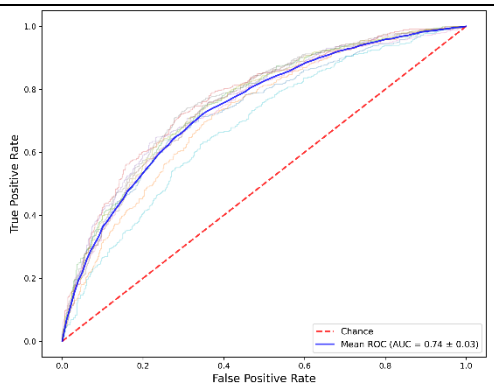

		

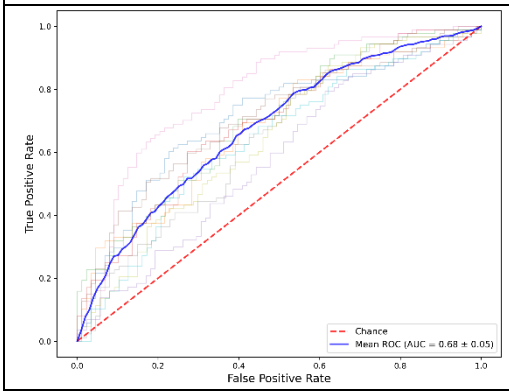
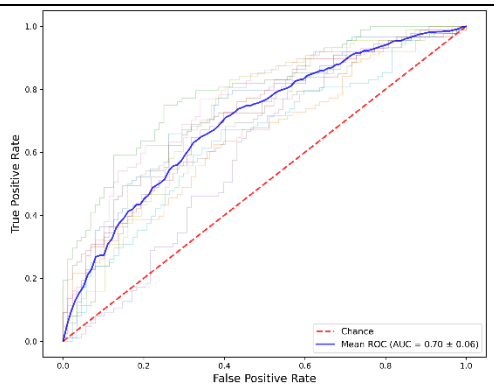
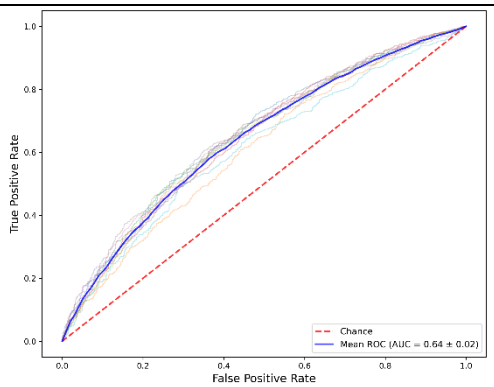
Figure 34p - Points (Neutral)	**Figure 34q - Points (N) + Apyest**	**Figure 34r - Mesh Coordinates**
		

Figure 34s - Mesh + Apyest	Figure 34t - Mesh (Happy)	Figure 34u - Mesh (H) + Apyest

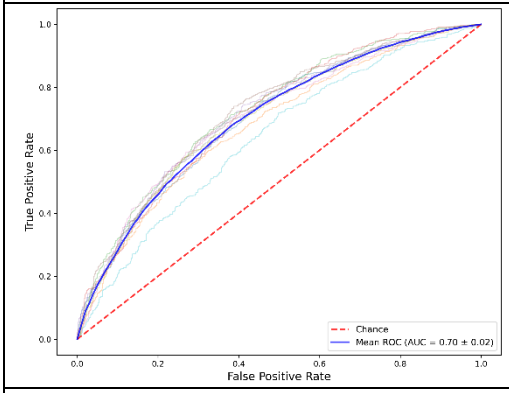
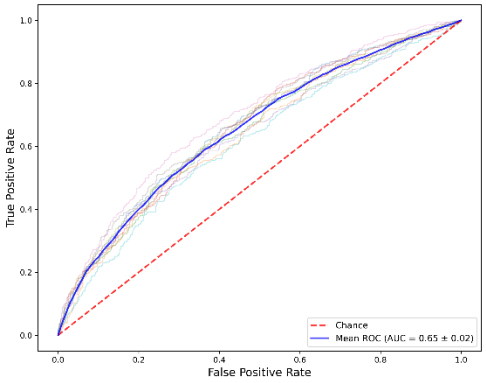
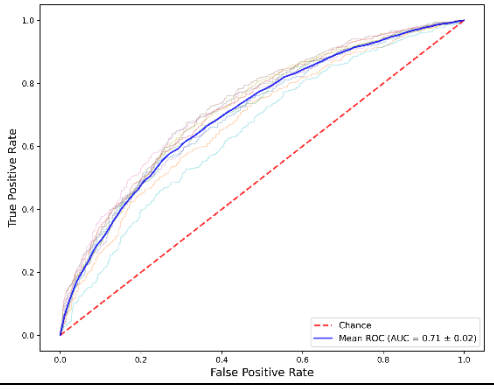

		

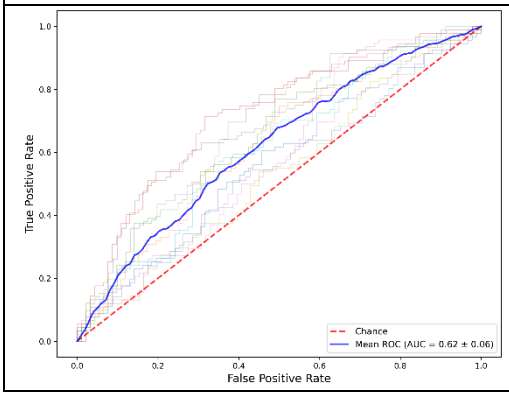
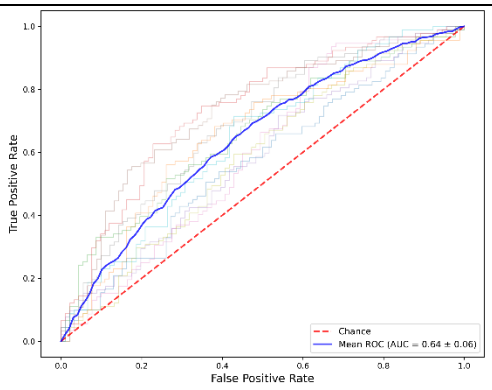
Figure 34v - Mesh (Neutral)	**Figure 34w - Mesh (N) + Apyest**	
		

Figure 35

ROC Plots for White Females – Immigration – Reduced Images

Figure 35a - Age	Figure 35b - Pitch, Yaw	Figure 35c - Emotions

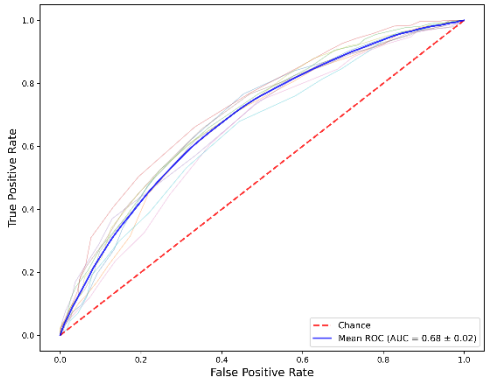
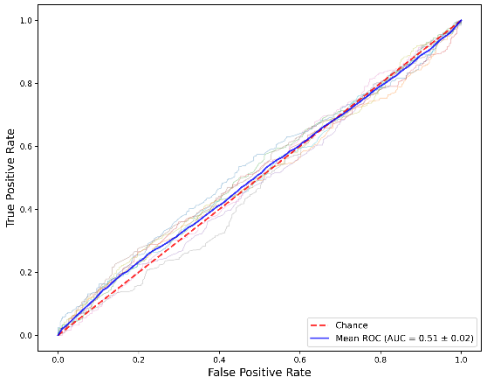
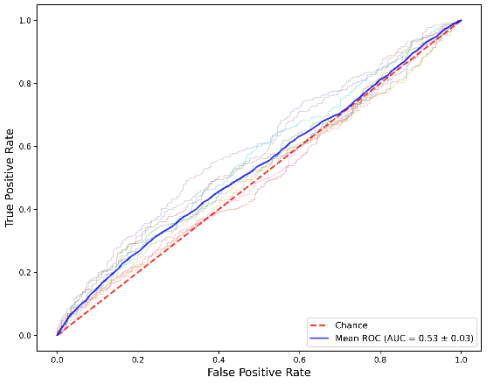

		

Figure 35d - Sex Typicality	**Figure 35e - Combined Apyest**	**Figure 35f - Features**

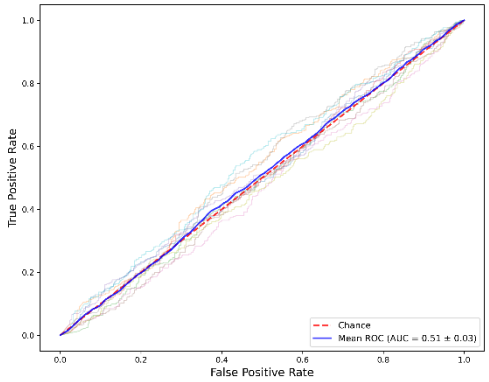
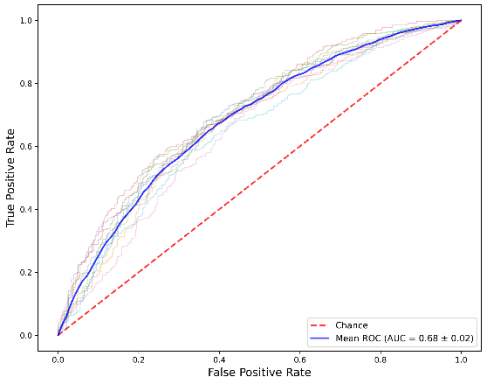
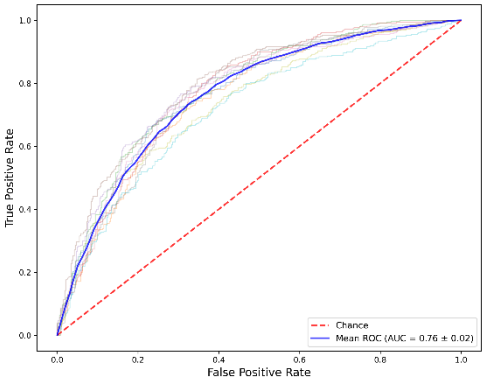

		

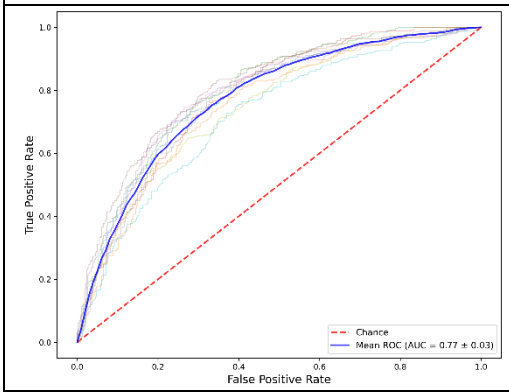
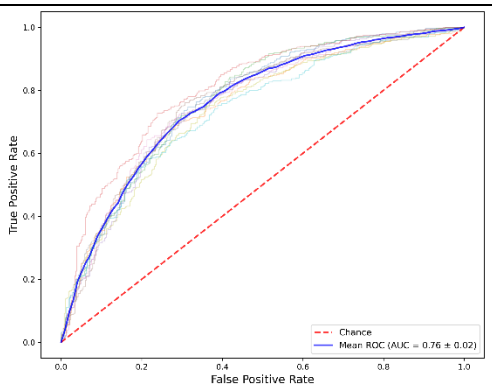
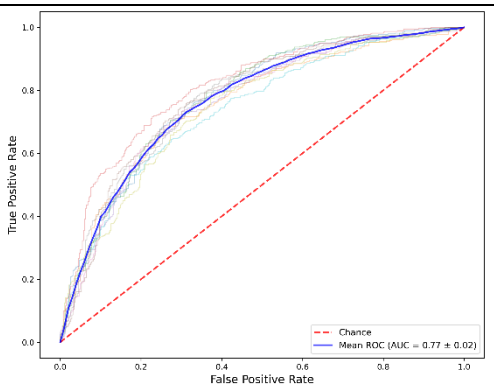
Figure 35g - Features + Apyest	**Figure 35h - Features (Masked)**	**Figure 35i - Features (M) + Apyest**
		

Figure 35j - Point Coordinates	Figure 35k - Points + Apyest	Figure 35l - Points (No Mouth)

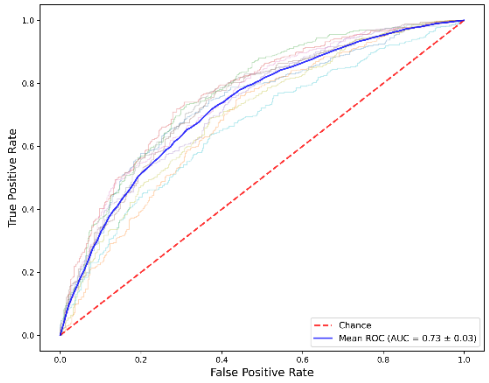
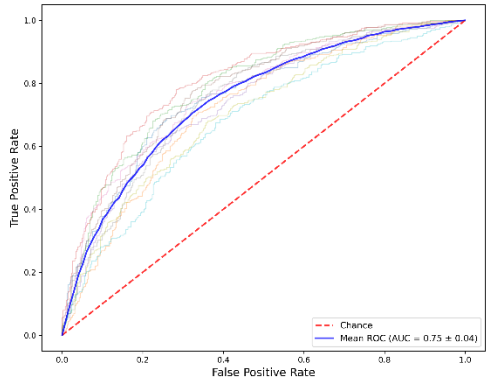
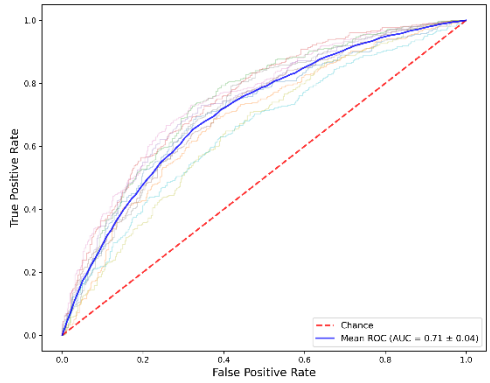

		

Figure 35m - Points (NM) + Apyest	**Figure 35n - Points (Happy)**	**Figure 35o - Points (H) + Apyest**

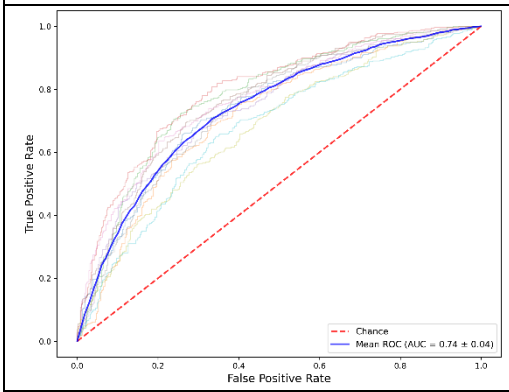
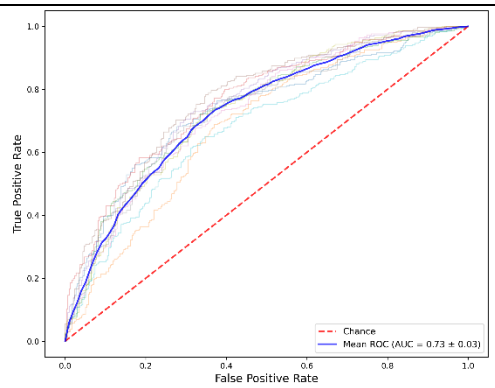
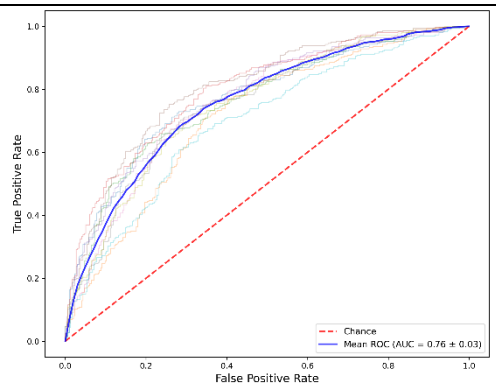

		

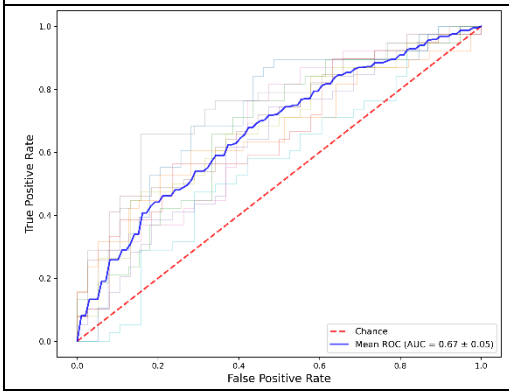
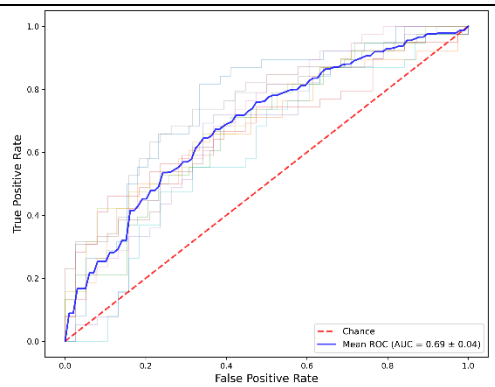
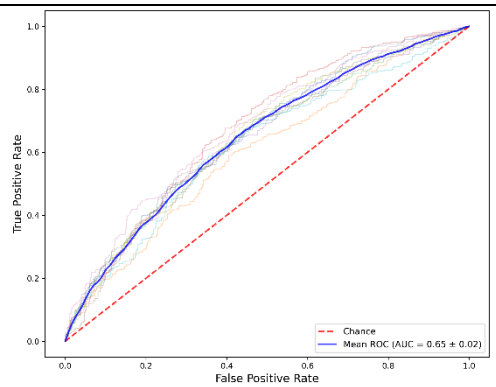
Figure 35p - Points (Neutral)	**Figure 35q - Points (N) + Apyest**	**Figure 35r - Mesh Coordinates**
		

Figure 35s - Mesh + Apyest	Figure 35t - Mesh (Happy)	Figure 35u - Mesh (H) + Apyest

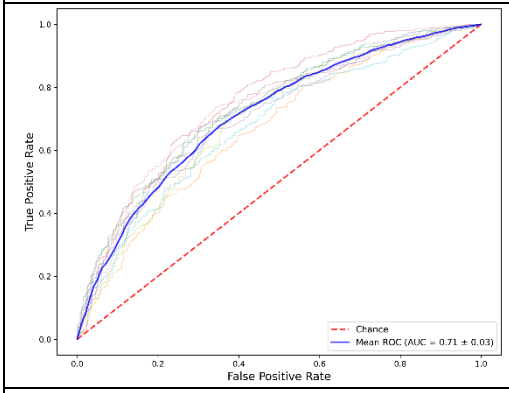
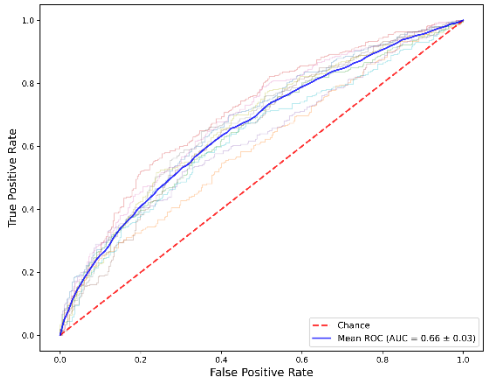
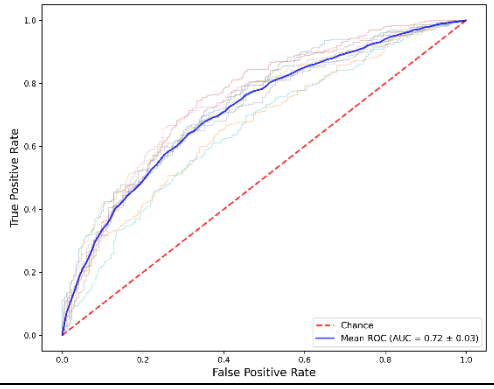

		

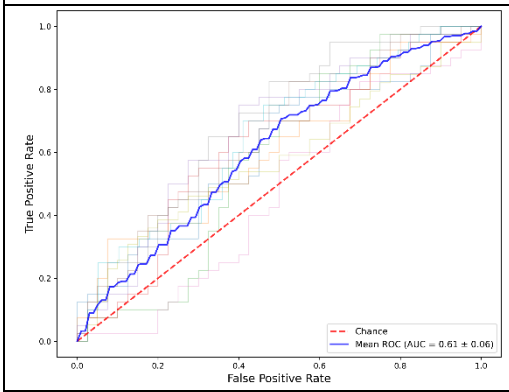
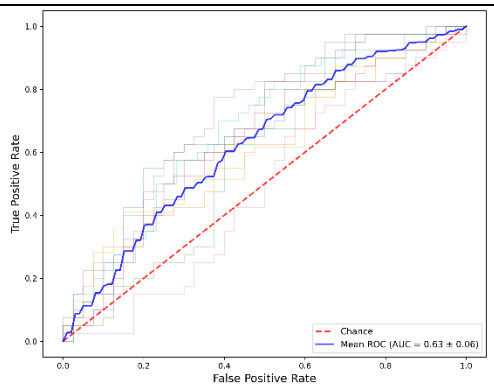
Figure 35v - Mesh (Neutral)	**Figure 35w - Mesh (N) + Apyest**	
		

Figure 36

ROC Plots for Asian Males – Gun – All Images

Figure 36a - Age	Figure 36b - Pitch, Yaw	Figure 36c - Emotions

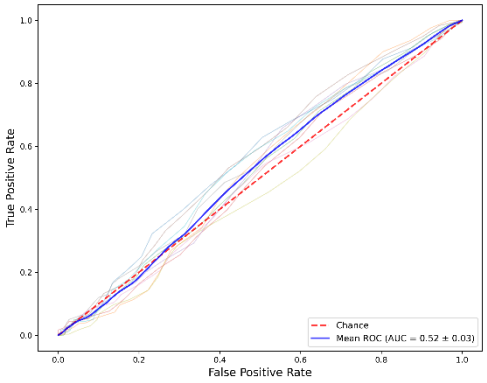
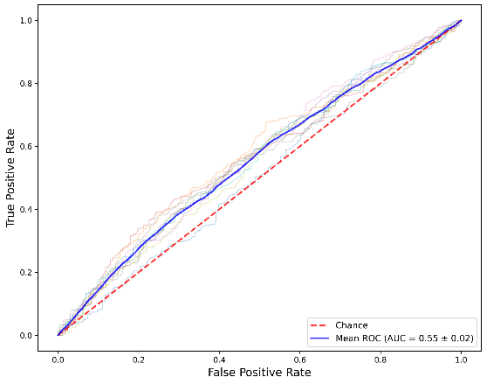
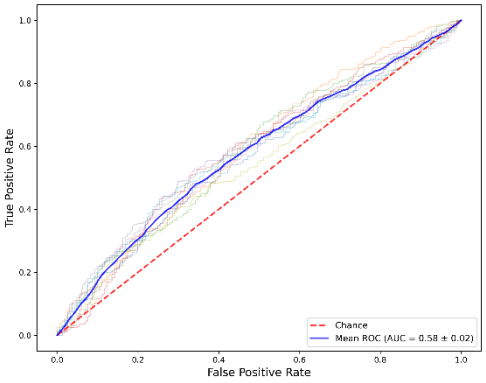

		

Figure 36d - Sex Typicality	**Figure 36e - Combined Apyest**	**Figure 36f - Features**

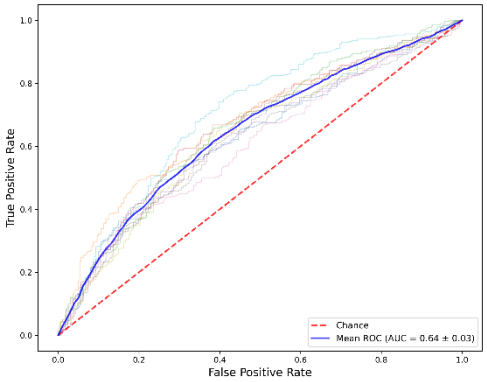
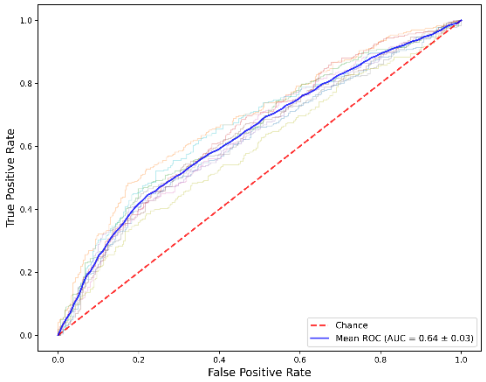
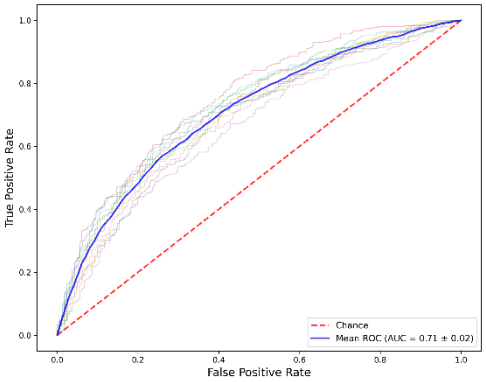

		

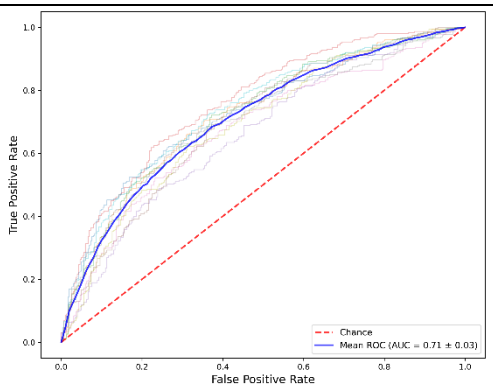
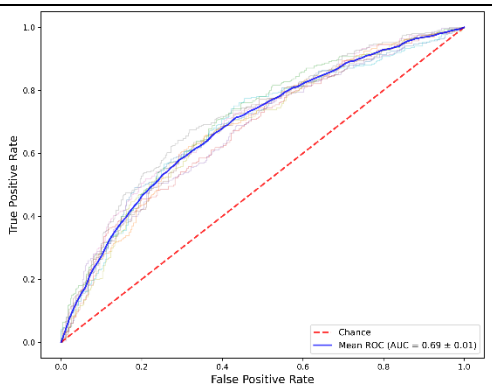
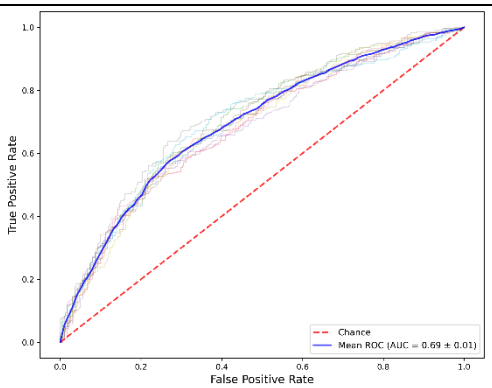
Figure 36g - Features + Apyest	**Figure 36h - Features (Masked)**	**Figure 36i - Features (M) + Apyest**
		

Figure 36j - Point Coordinates	Figure 36k - Points + Apyest	Figure 36l - Points (No Mouth)

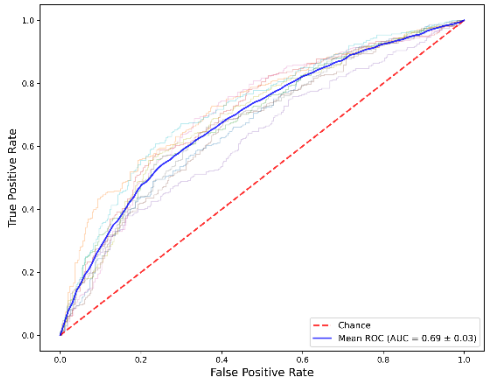
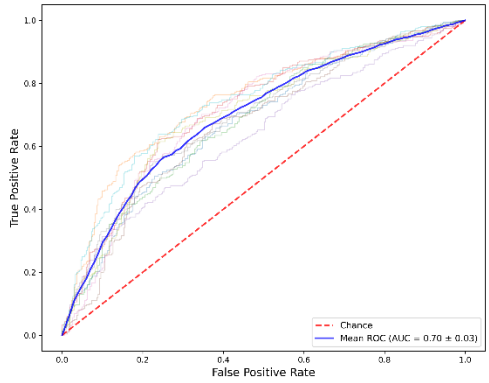
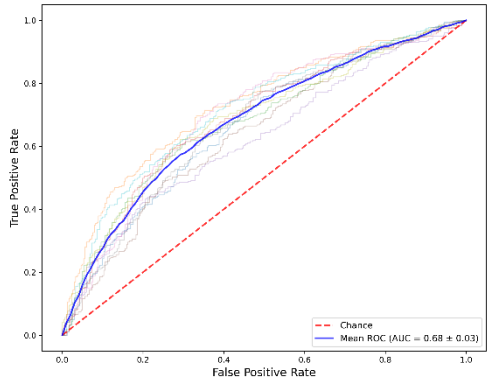

		

Figure 36m - Points (NM) + Apyest	**Figure 36n - Points (Happy)**	**Figure 36o - Points (H) + Apyest**

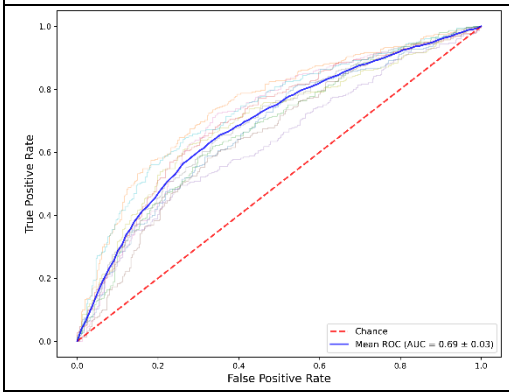
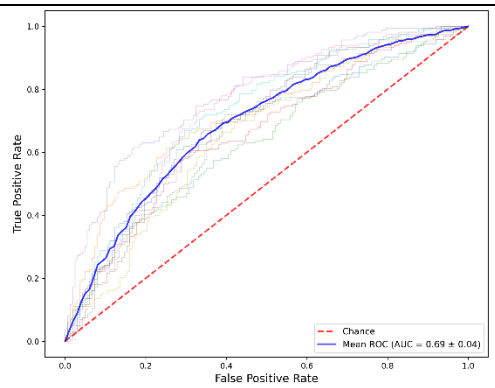
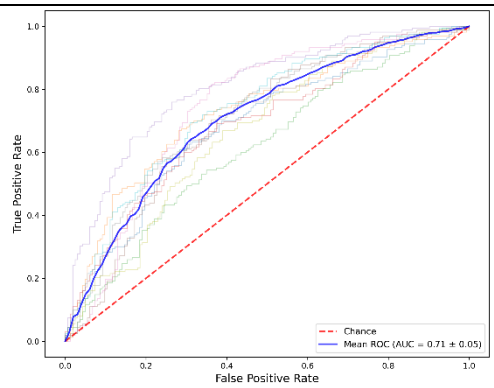

		

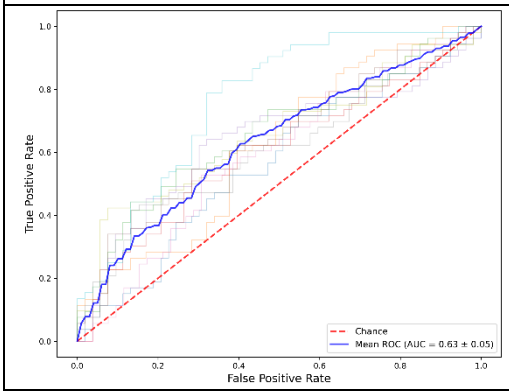
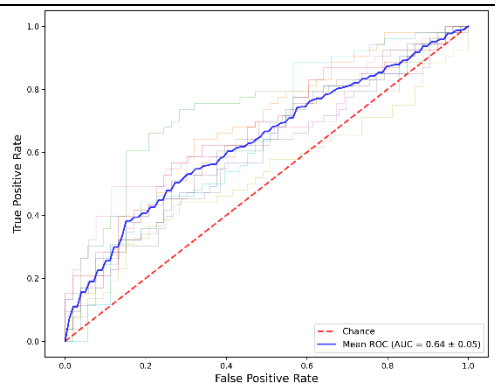
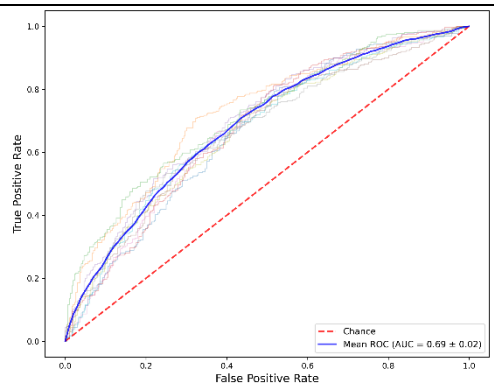
Figure 36p - Points (Neutral)	**Figure 36q - Points (N) + Apyest**	**Figure 36r - Mesh Coordinates**
		

Figure 36s - Mesh + Apyest	Figure 36t - Mesh (Happy)	Figure 36u - Mesh (H) + Apyest

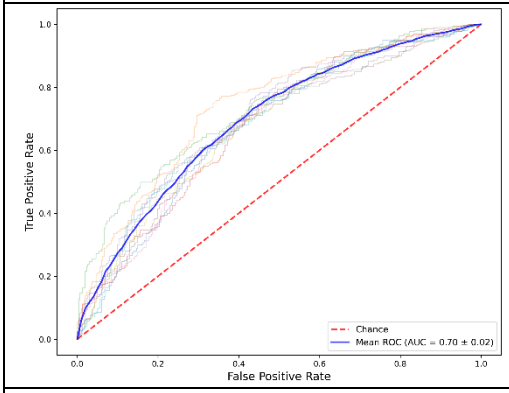
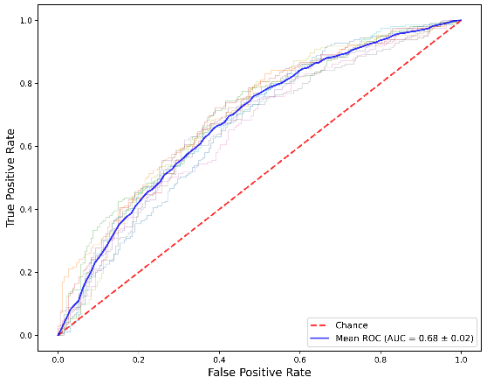
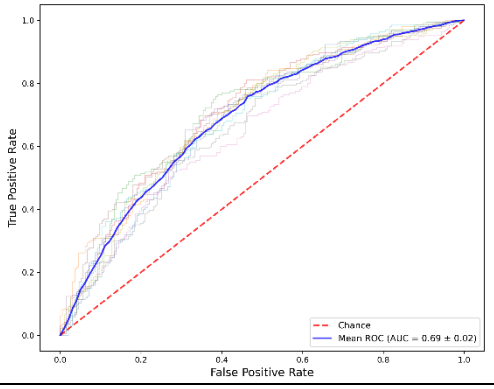

		

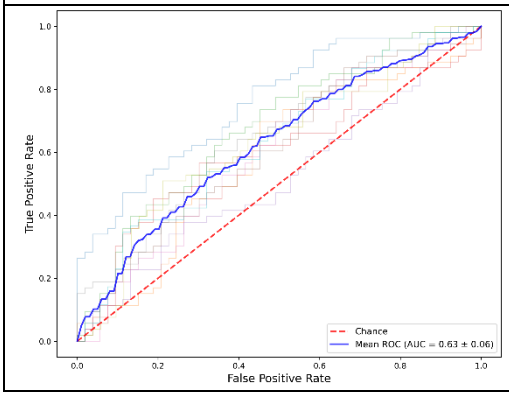
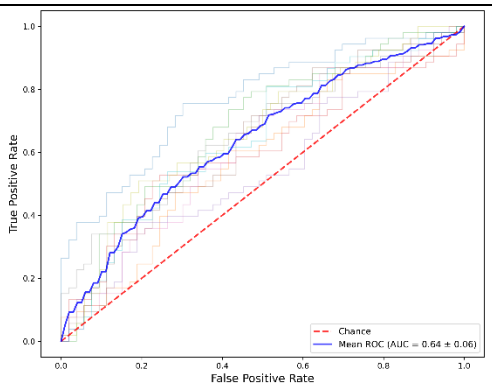
Figure 36v - Mesh (Neutral)	**Figure 36w - Mesh (N) + Apyest**	
		

Figure 37

ROC Plots for Asian Males – Gun – Reduced Images

Figure 37a - Age	Figure 37b - Pitch, Yaw	Figure 37c - Emotions

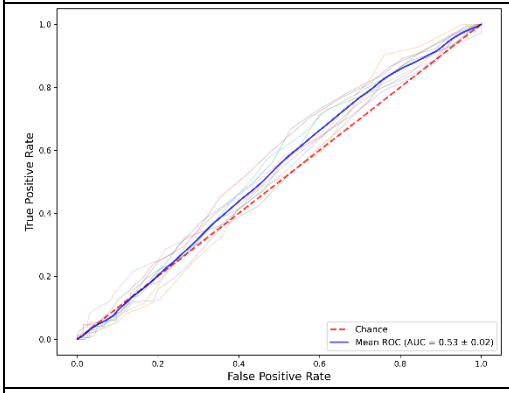
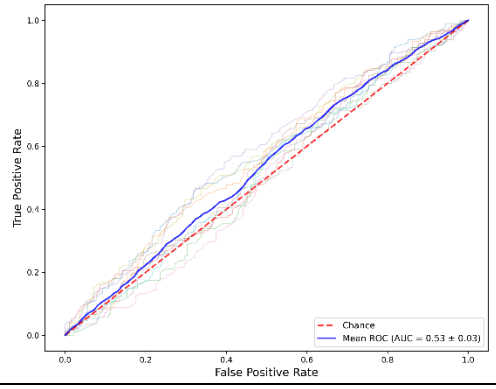
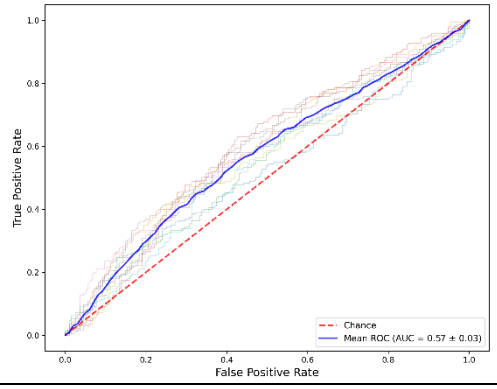

		

Figure 37d - Sex Typicality	**Figure 37e - Combined Apyest**	**Figure 37f - Features**

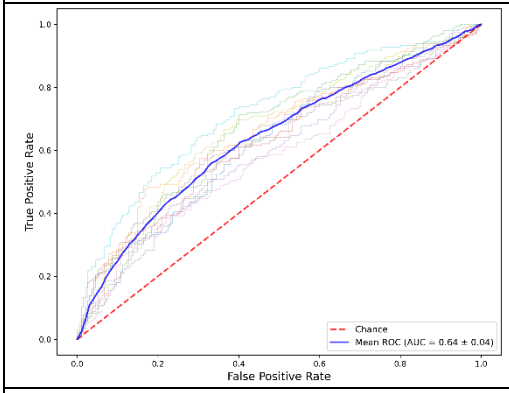
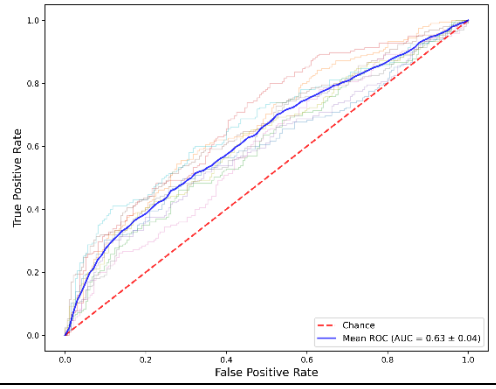
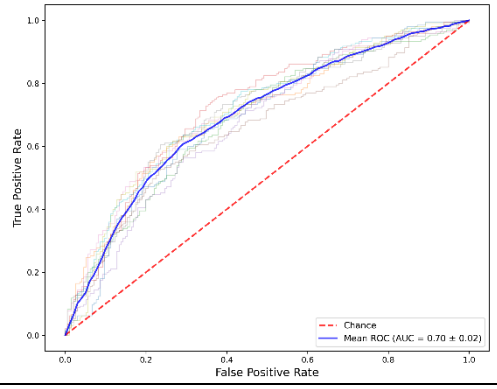

		

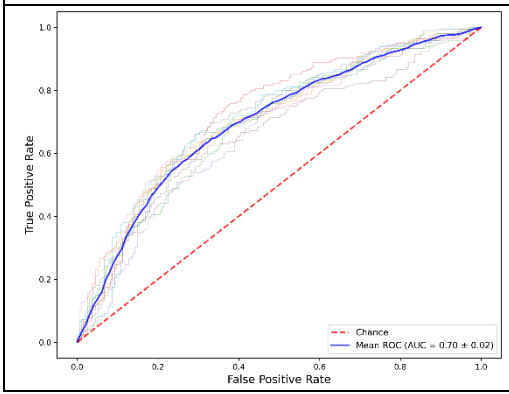
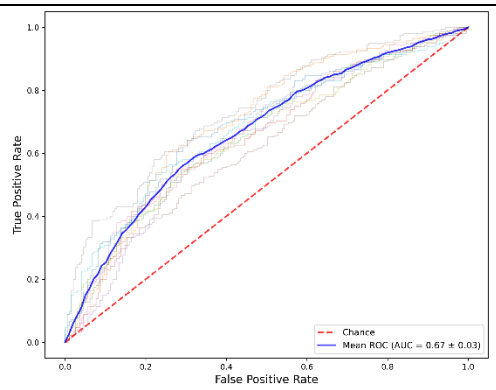
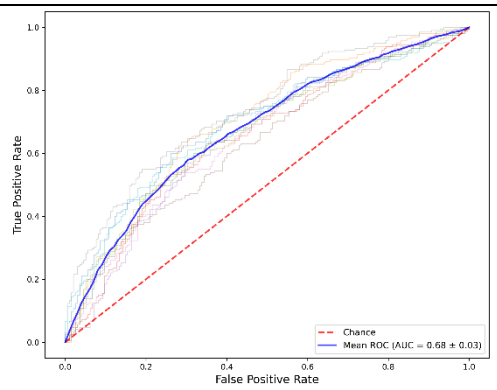
Figure 37g - Features + Apyest	**Figure 37h - Features (Masked)**	**Figure 37i - Features (M) + Apyest**
		

Figure 37j - Point Coordinates	Figure 37k - Points + Apyest	Figure 37l - Points (No Mouth)

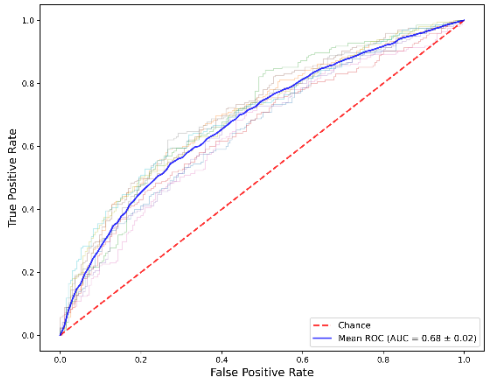
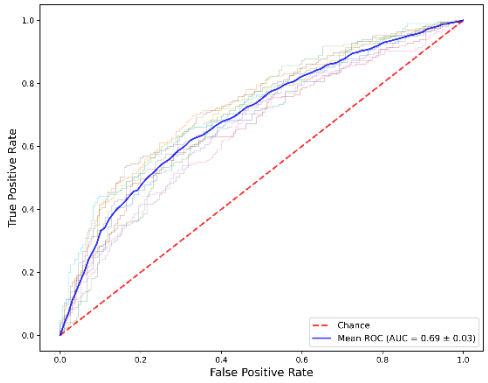
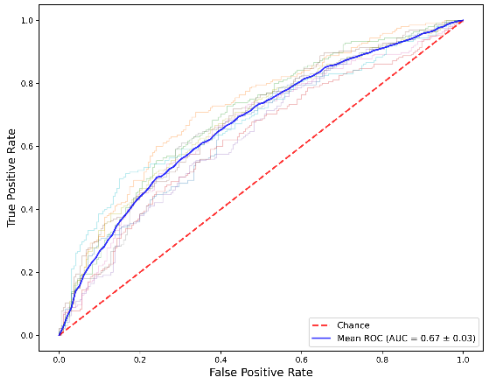

		

Figure 37m - Points (NM) + Apyest	**Figure 37n - Points (Happy)**	**Figure 37o - Points (H) + Apyest**

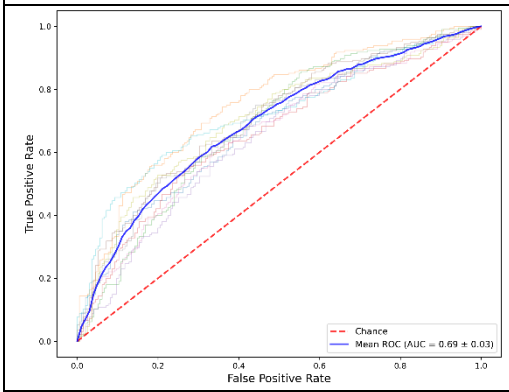
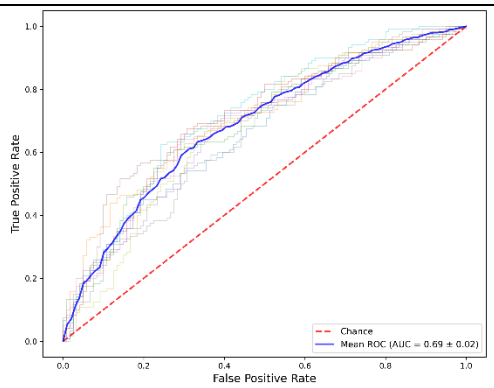
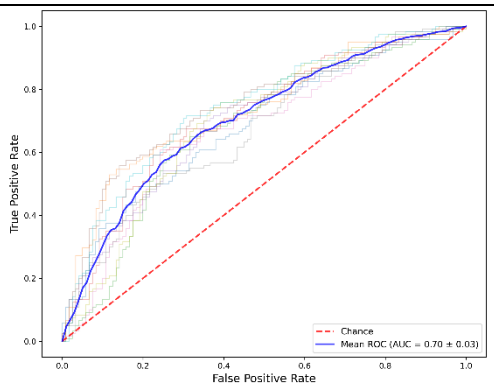

		

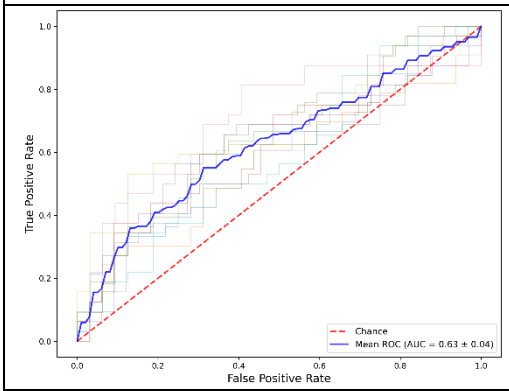
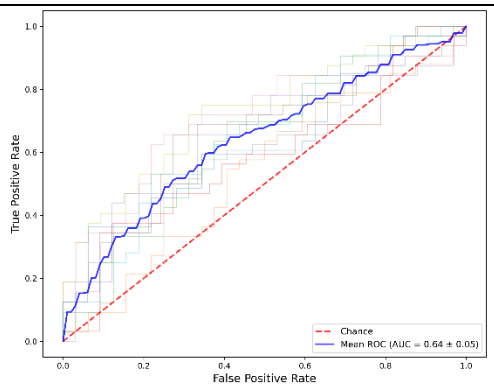
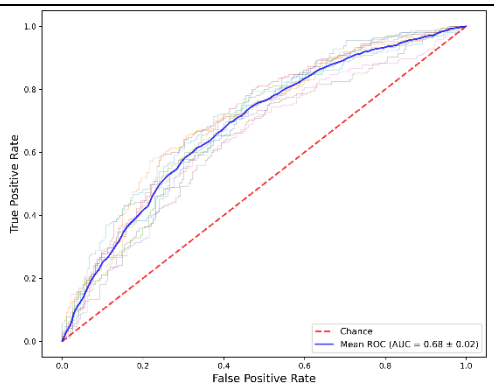
Figure 37p - Points (Neutral)	**Figure 37q - Points (N) + Apyest**	**Figure 37r - Mesh Coordinates**
		

Figure 37s - Mesh + Apyest	Figure 37t - Mesh (Happy)	Figure 37u - Mesh (H) + Apyest

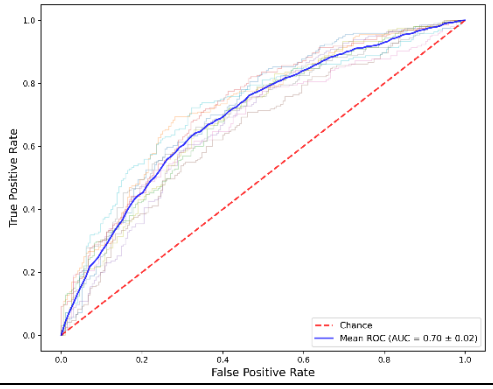
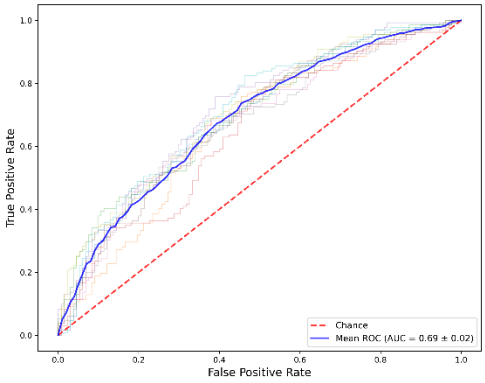
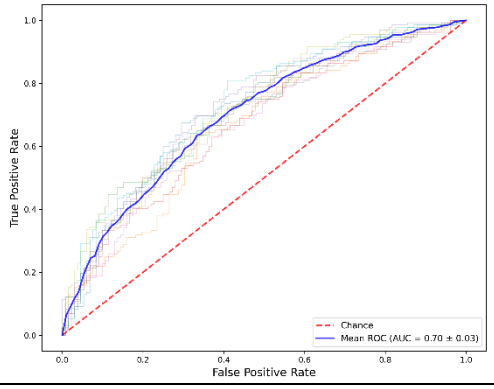

		

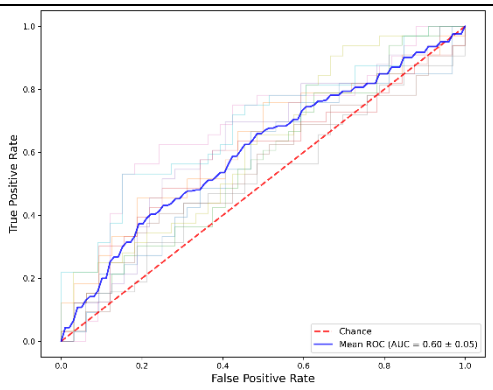
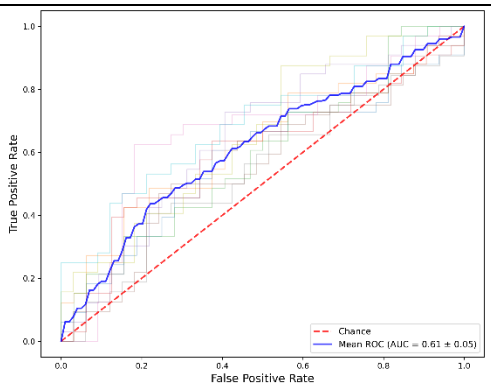
Figure 37v - Mesh (Neutral)	**Figure 37w - Mesh (N) + Apyest**	
		

Figure 38

ROC Plots for Hispanic Males – Immigration – All Images

Figure 38a - Age	Figure 38b - Pitch, Yaw	Figure 38c - Emotions

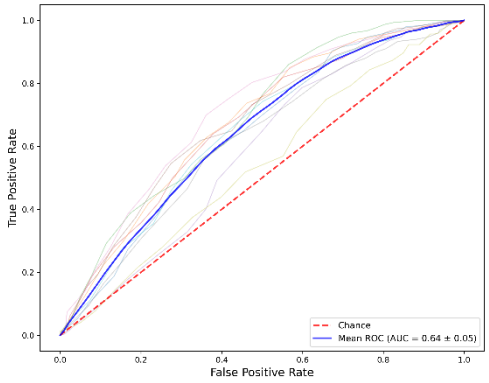
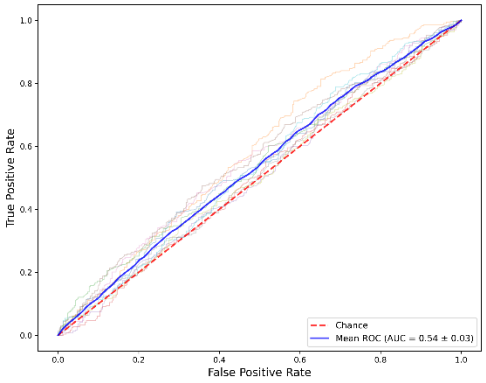
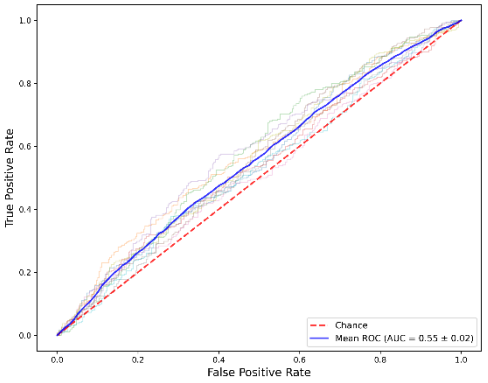

		

Figure 38d - Sex Typicality	**Figure 38e - Combined Apyest**	**Figure 38f - Features**

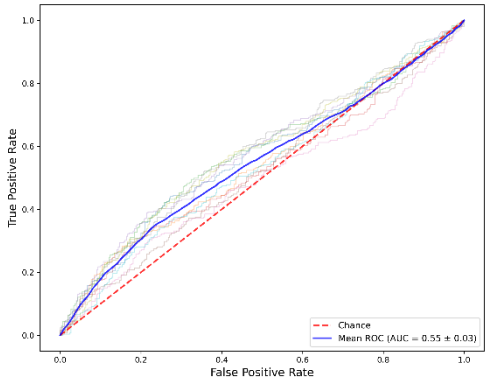
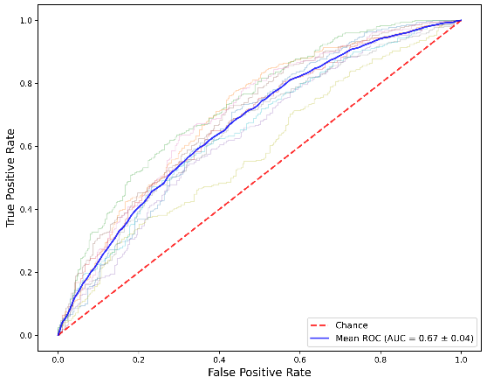
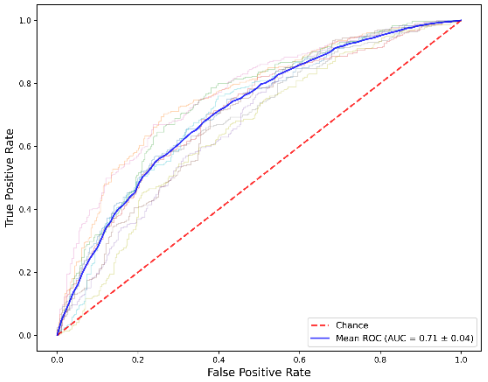

		

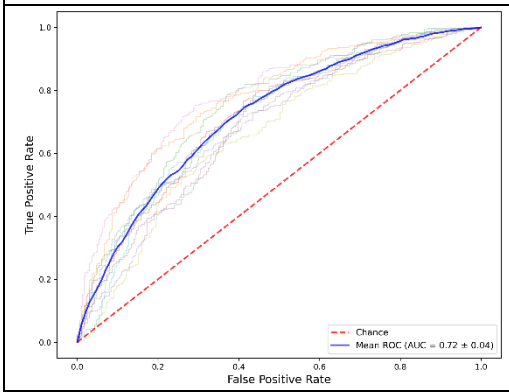
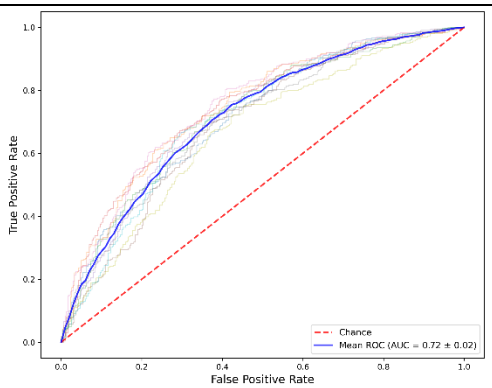
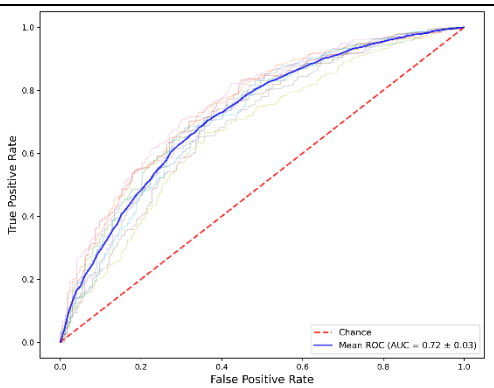
Figure 38g - Features + Apyest	**Figure 38h - Features (Masked)**	**Figure 38i - Features (M) + Apyest**
		

Figure 38j - Point Coordinates	Figure 38k - Points + Apyest	Figure 38l - Points (No Mouth)

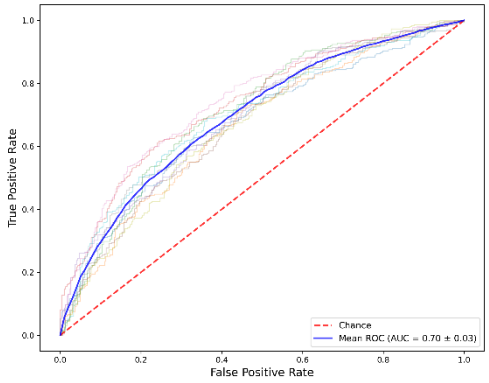
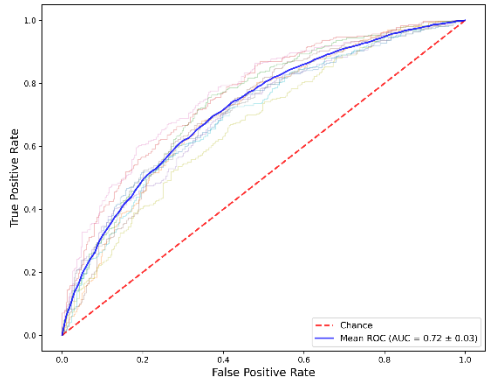
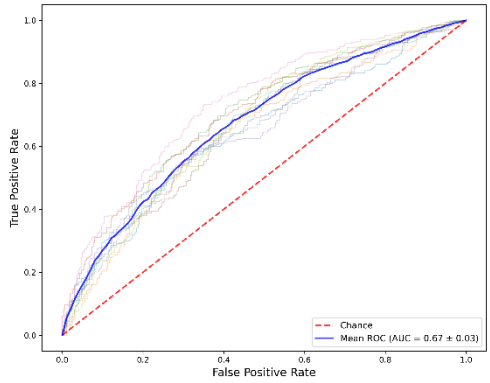

		

Figure 38m - Points (NM) + Apyest	**Figure 38n - Points (Happy)**	**Figure 38o - Points (H) + Apyest**

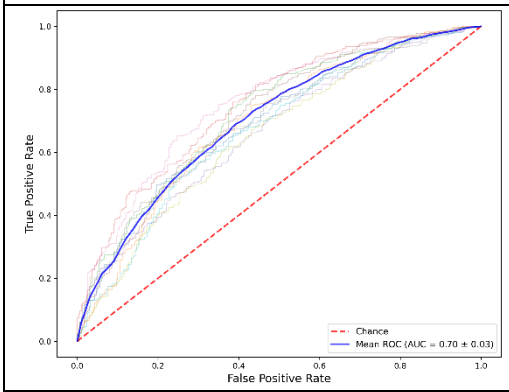
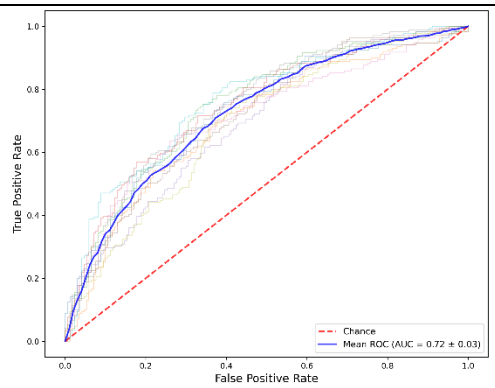
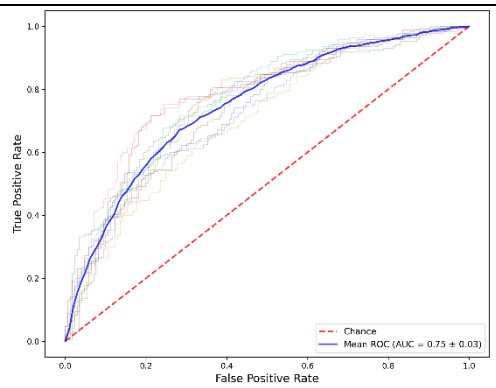

		

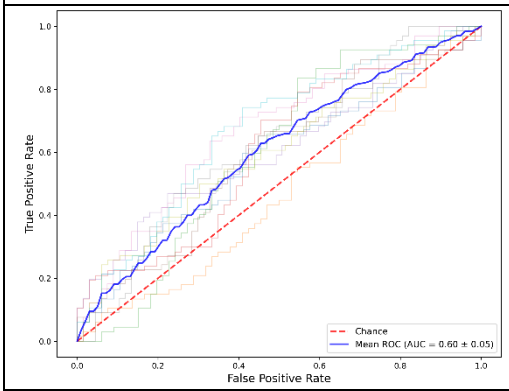
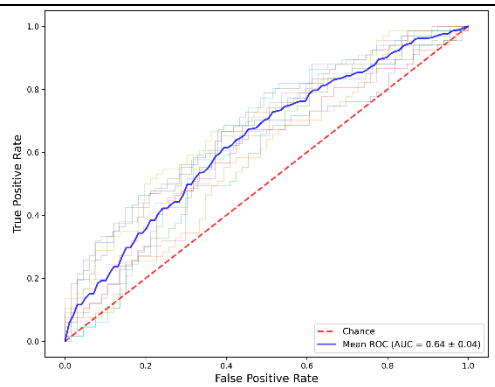
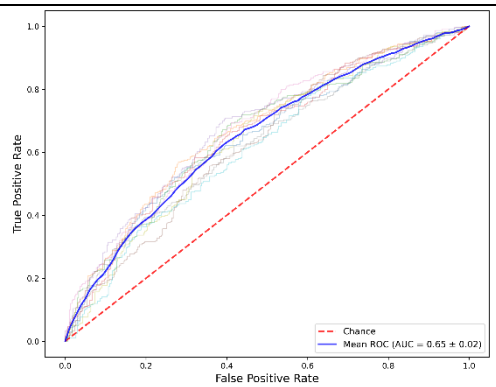
Figure 38p - Points (Neutral)	**Figure 38q - Points (N) + Apyest**	**Figure 38r - Mesh Coordinates**
		

Figure 38s - Mesh + Apyest	Figure 38t - Mesh (Happy)	Figure 38u - Mesh (H) + Apyest

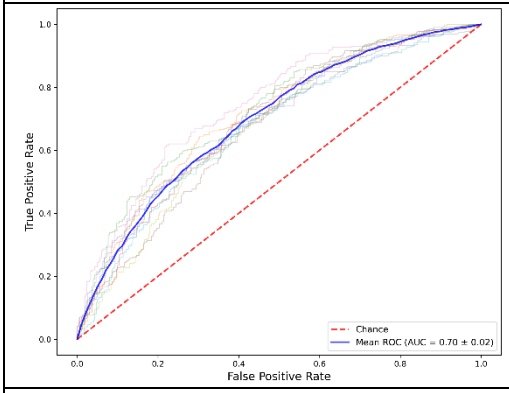
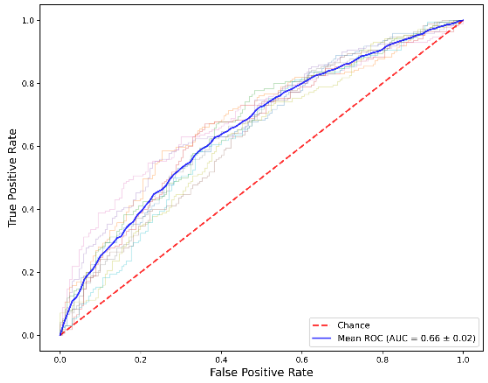
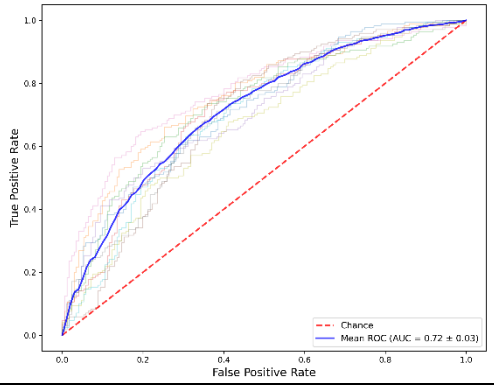

		

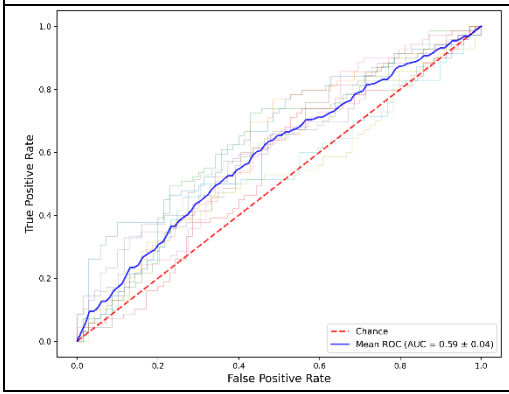
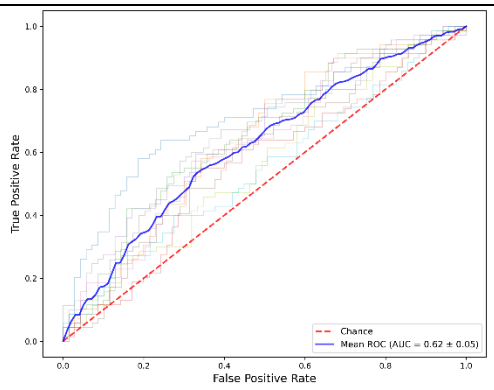
Figure 38v - Mesh (Neutral)	**Figure 38w - Mesh (N) + Apyest**	
		

Figure 39

ROC Plots for Hispanic Males – Immigration – Reduced Images

Figure 39a - Age	Figure 39b - Pitch, Yaw	Figure 39c - Emotions

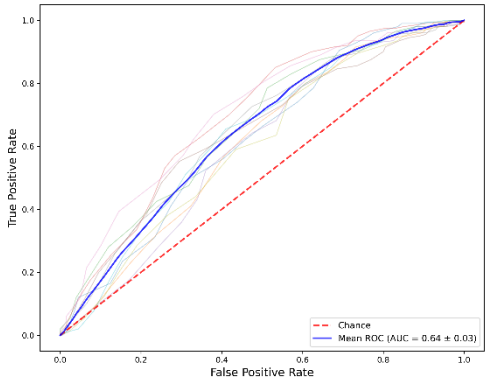
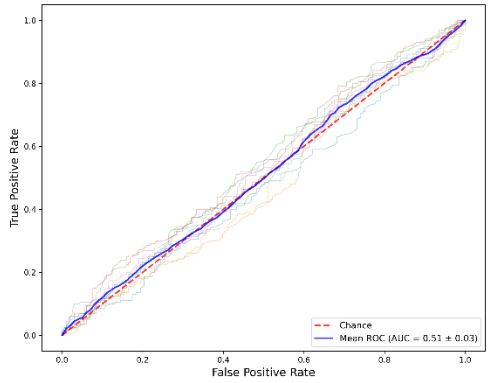
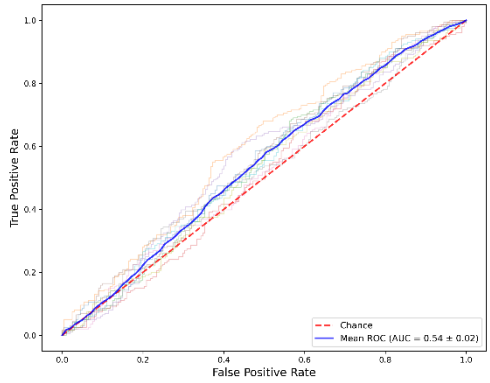

		

Figure 39d - Sex Typicality	**Figure 39e - Combined Apyest**	**Figure 39f - Features**

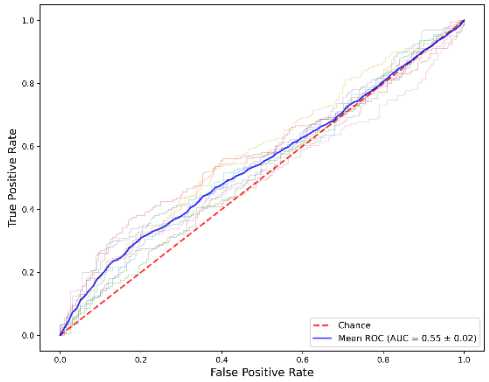
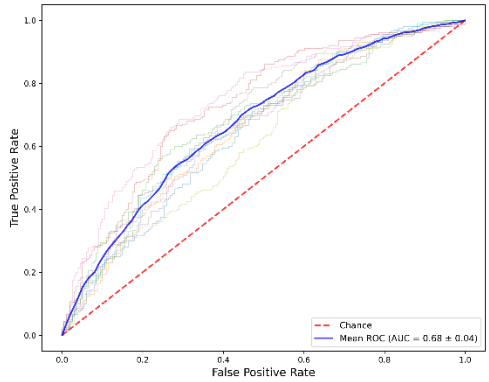
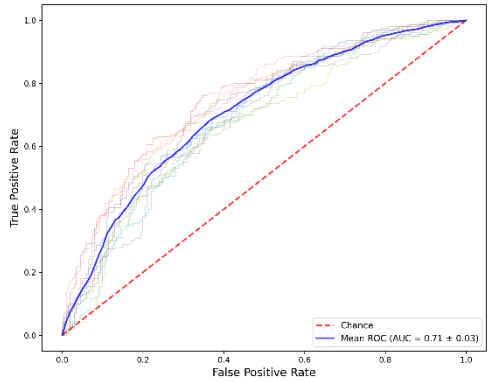

		

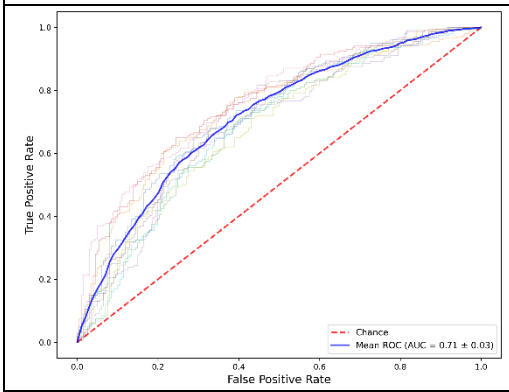
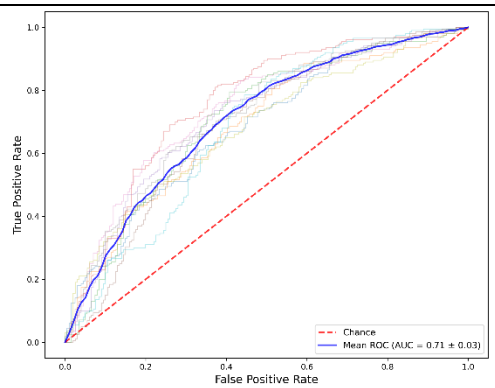
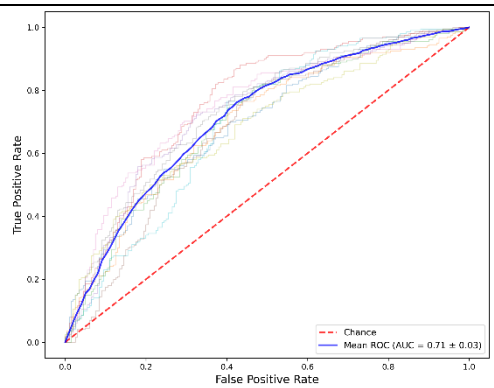
Figure 39g - Features + Apyest	**Figure 39h - Features (Masked)**	**Figure 39i - Features (M) + Apyest**
		

Figure 39j - Point Coordinates	Figure 39k - Points + Apyest	Figure 39l - Points (No Mouth)

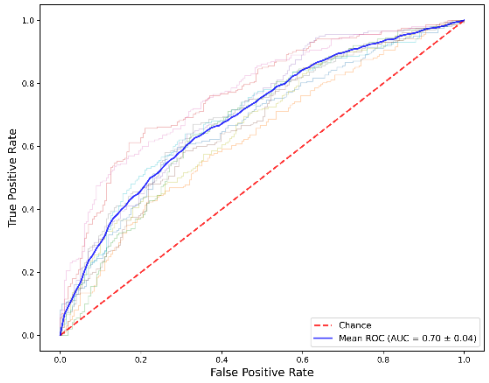
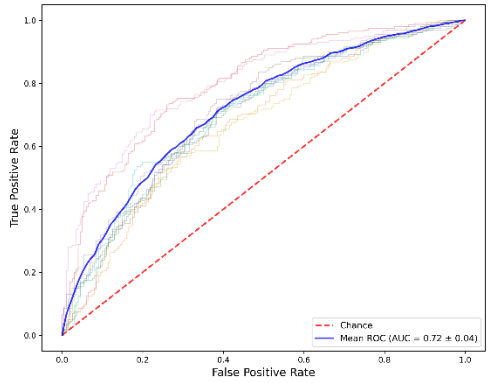
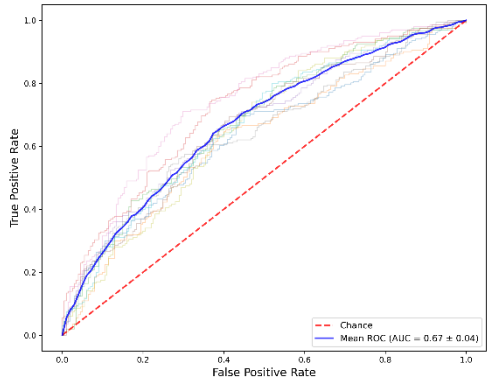

		

Figure 39m - Points (NM) + Apyest	**Figure 39n - Points (Happy)**	**Figure 39o - Points (H) + Apyest**

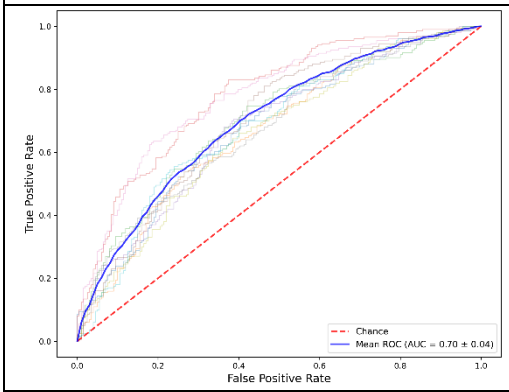
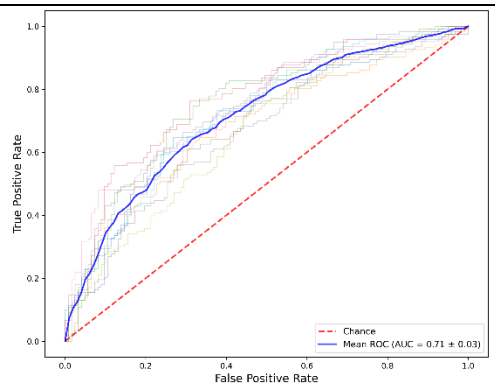
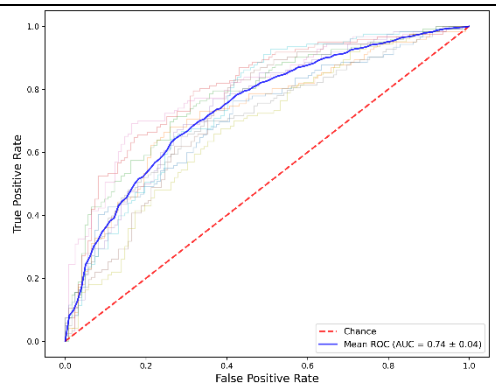

		

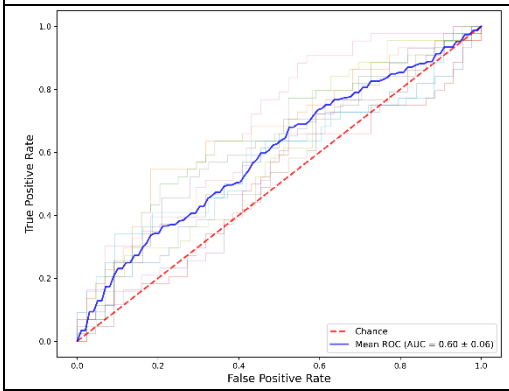
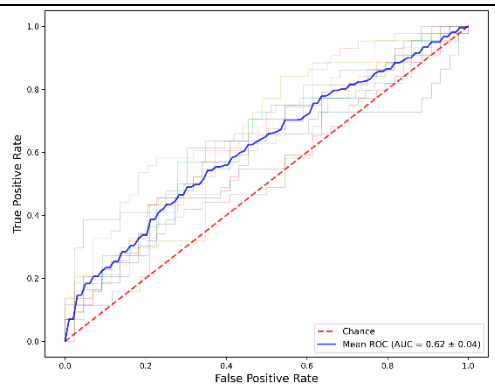
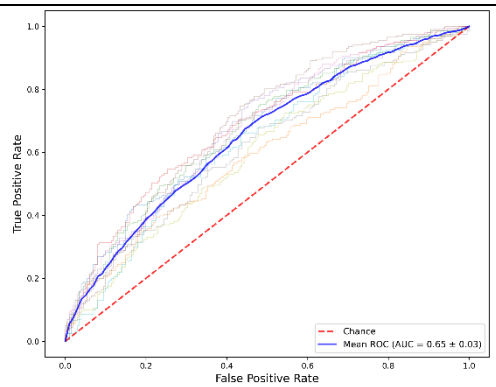
Figure 39p - Points (Neutral)	**Figure 39q - Points (N) + Apyest**	**Figure 39r - Mesh Coordinates**
		

Figure 39s - Mesh + Apyest	Figure 39t - Mesh (Happy)	Figure 39u - Mesh (H) + Apyest

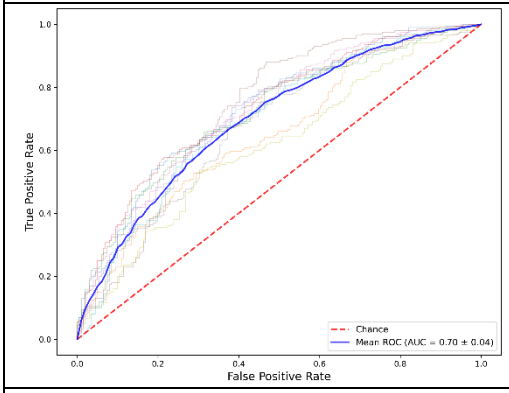
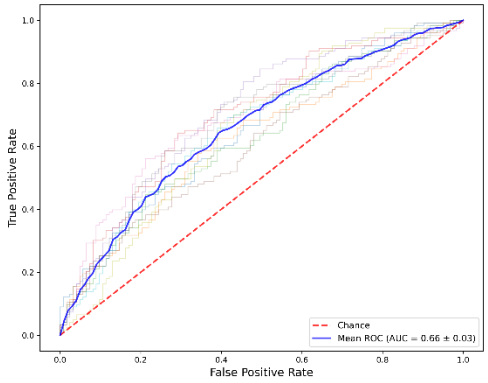
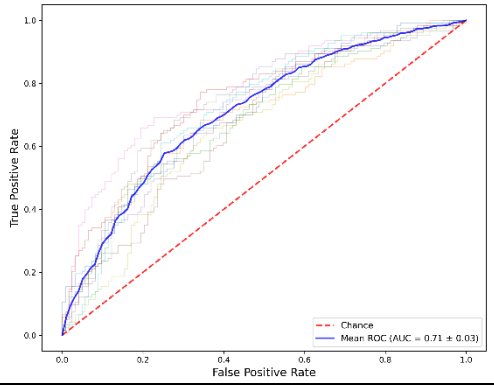

		

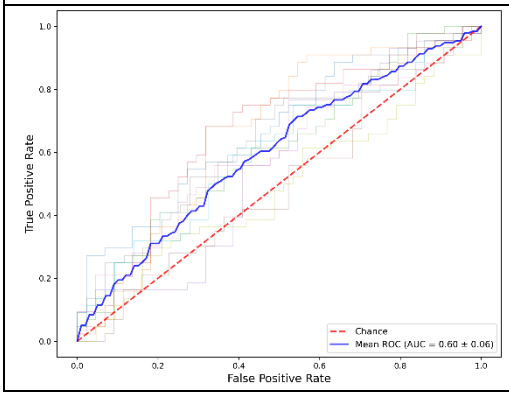
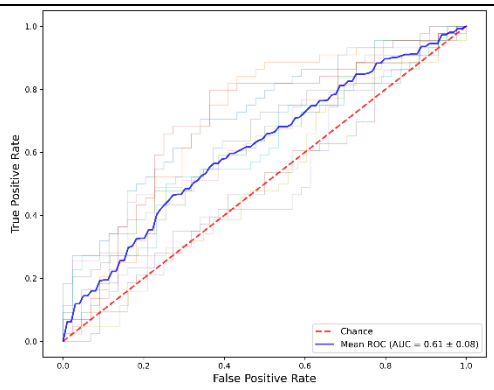
Figure 39v - Mesh (Neutral)	**Figure 39w - Mesh (N) + Apyest**	
		

Figure 40

ROC Plots for Hispanic Males – Gun – All Images

Figure 40a - Age	Figure 40b - Pitch, Yaw	Figure 40c - Emotions

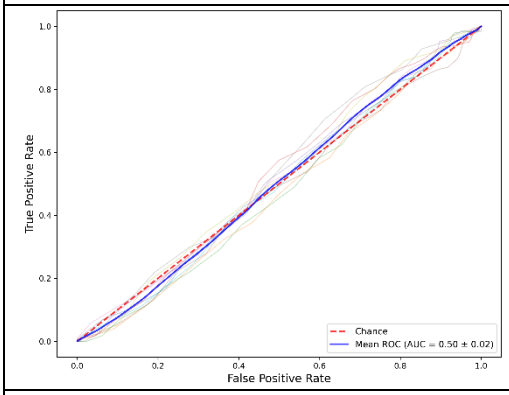
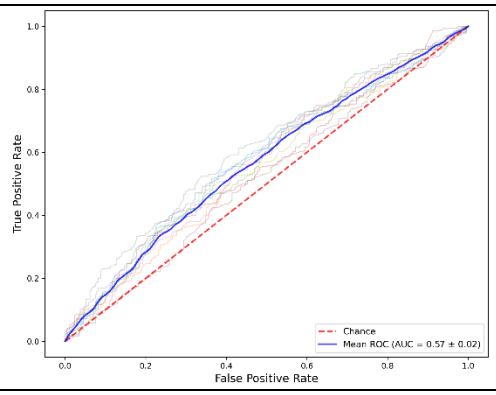
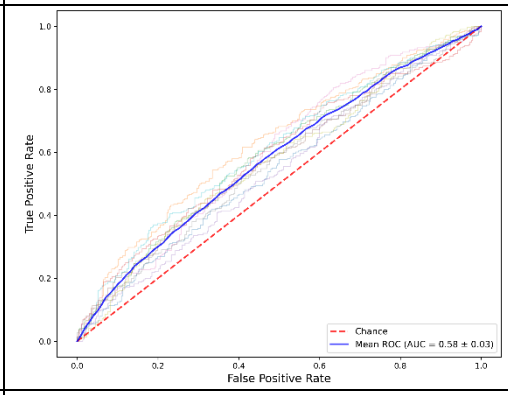

		

Figure 40d - Sex Typicality	**Figure 40e - Combined Apyest**	**Figure 40f - Features**

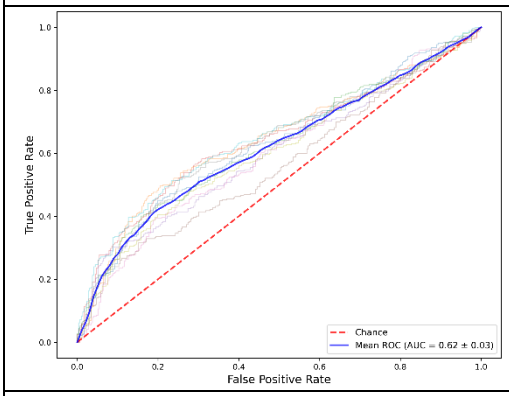
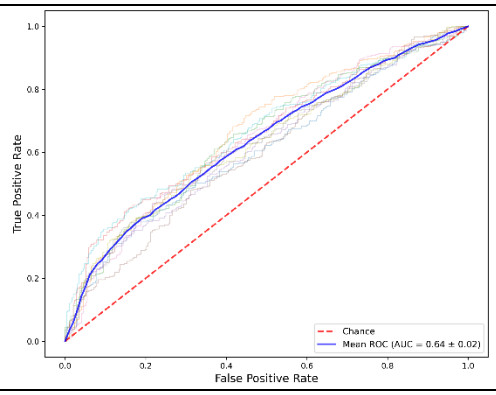
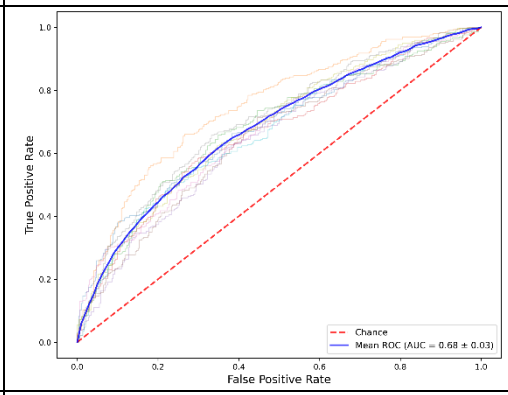

		

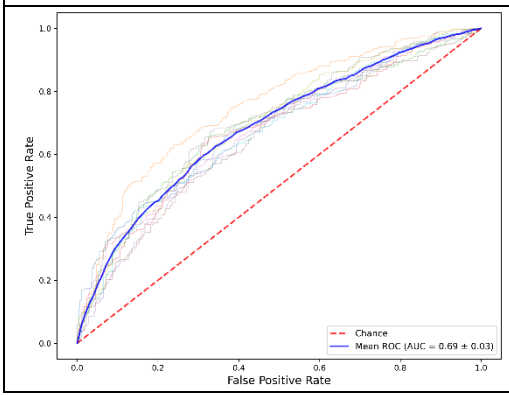
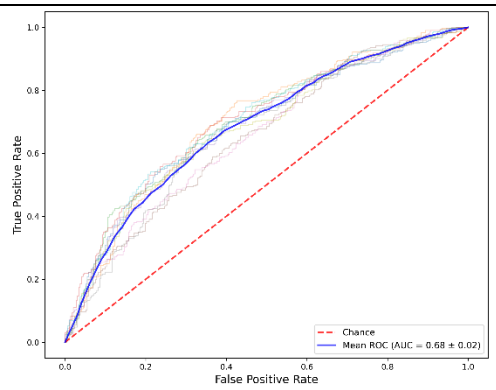
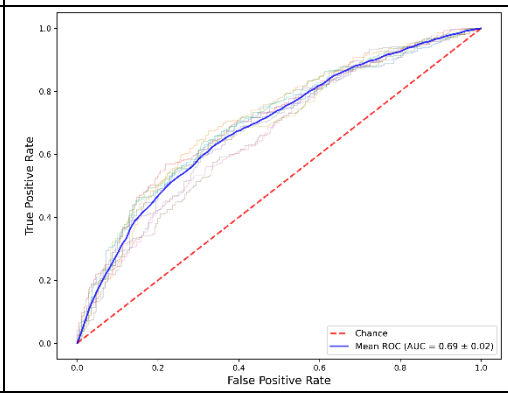
Figure 40g - Features + Apyest	**Figure 40h - Features (Masked)**	**Figure 40i - Features (M) + Apyest**
		

Figure 40j - Point Coordinates	Figure 40k - Points + Apyest	Figure 40l - Points (No Mouth)

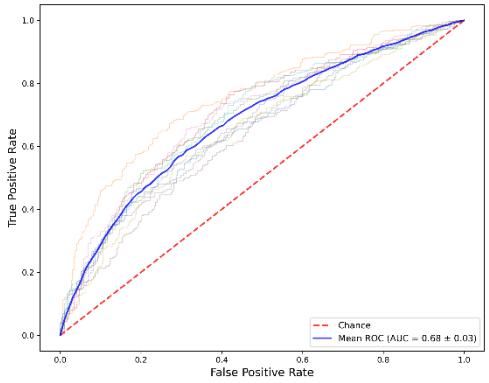
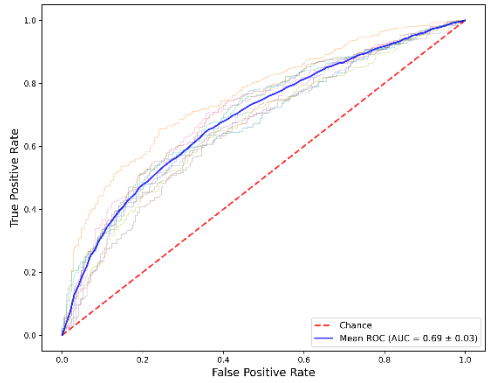
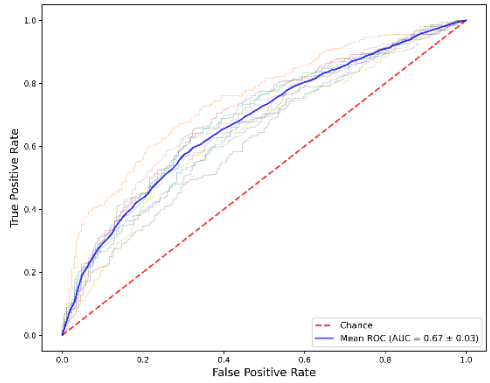

		

Figure 40m - Points (NM) + Apyest	**Figure 40n - Points (Happy)**	**Figure 40o - Points (H) + Apyest**

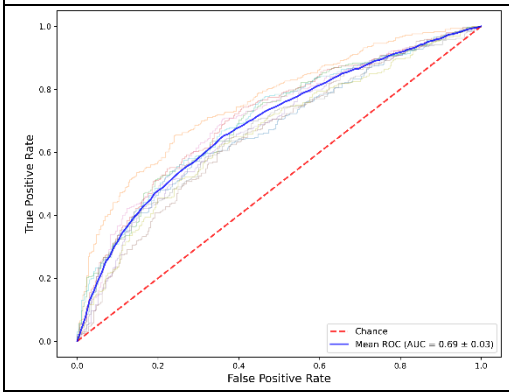
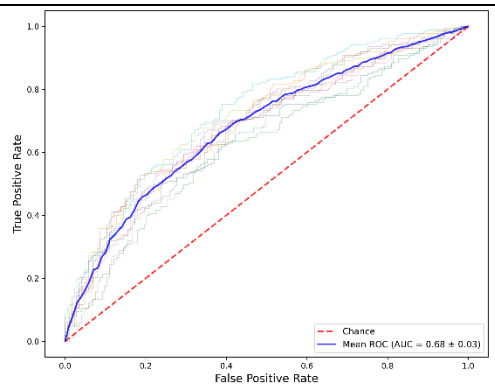
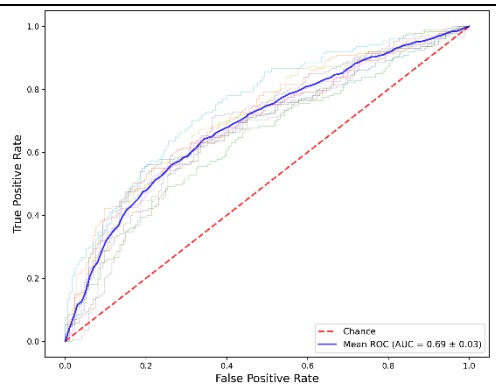

		

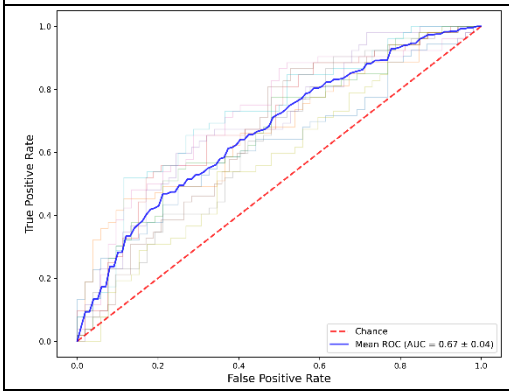
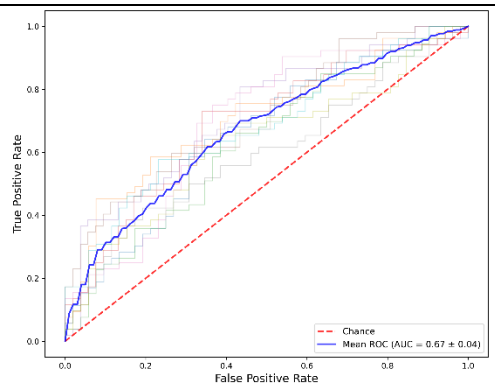
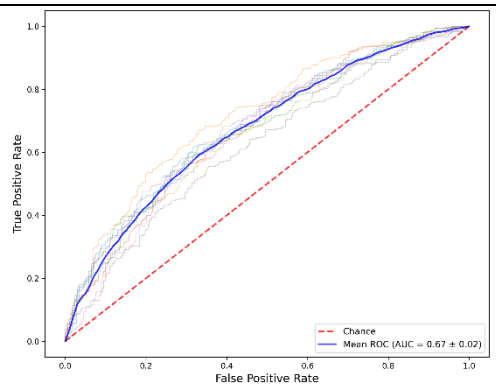
Figure 40p - Points (Neutral)	**Figure 40q - Points (N) + Apyest**	**Figure 40r - Mesh Coordinates**
		

Figure 40s - Mesh + Apyest	Figure 40t - Mesh (Happy)	Figure 40u - Mesh (H) + Apyest

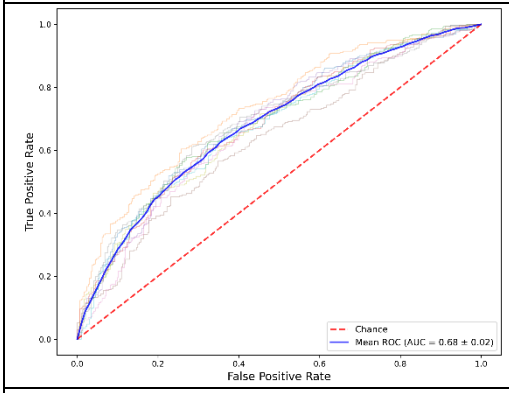
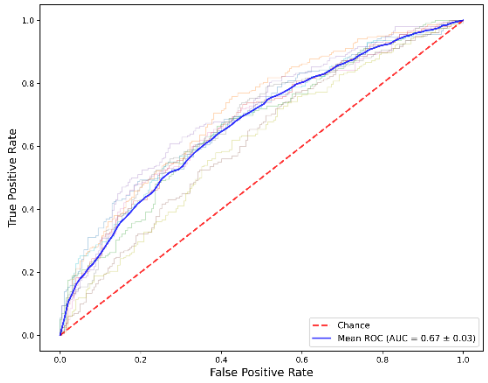
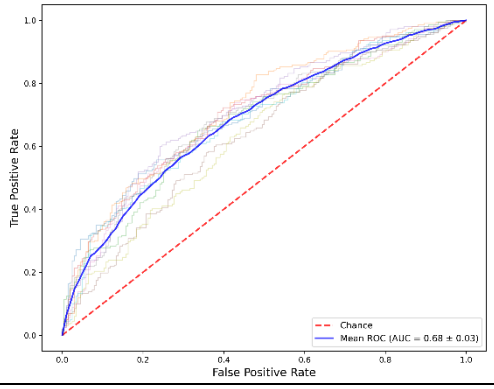

		

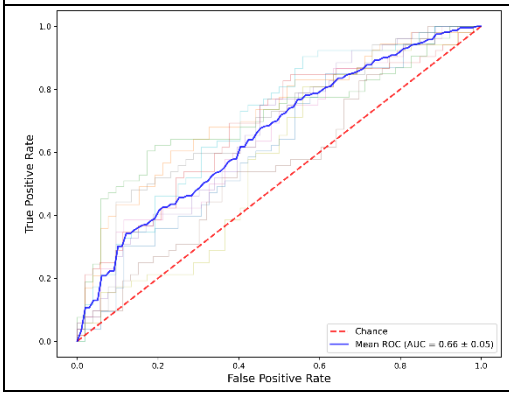
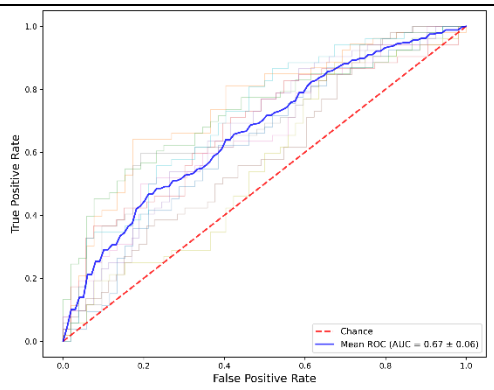
Figure 40v - Mesh (Neutral)	**Figure 40w - Mesh (N) + Apyest**	
		

Figure 41

ROC Plots for Hispanic Males – Gun – Reduced Images

Figure 41a - Age	Figure 41b - Pitch, Yaw	Figure 41c - Emotions

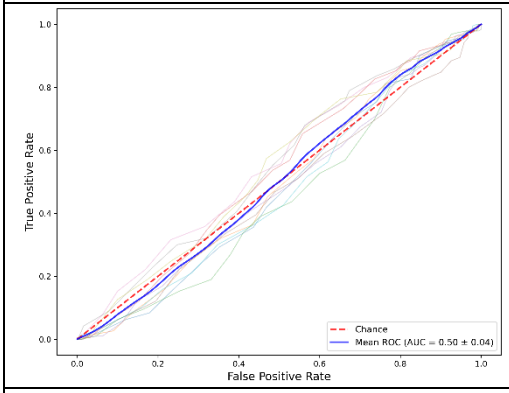
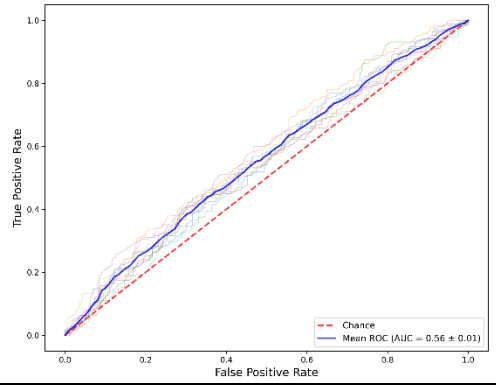
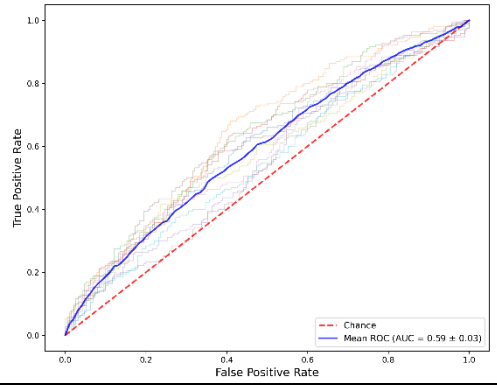

		

Figure 41d - Sex Typicality	**Figure 41e - Combined Apyest**	**Figure 41f - Features**

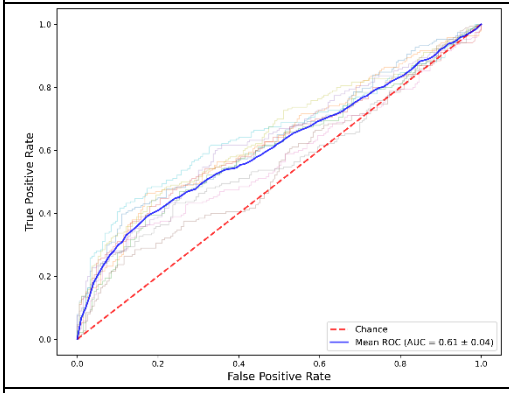
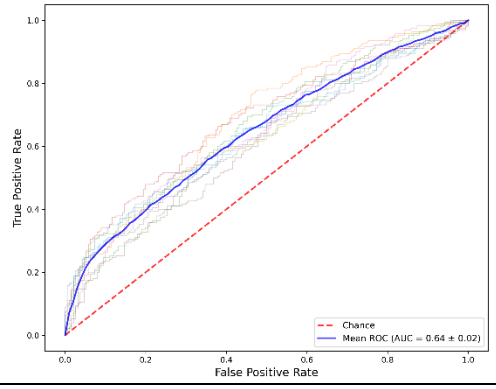
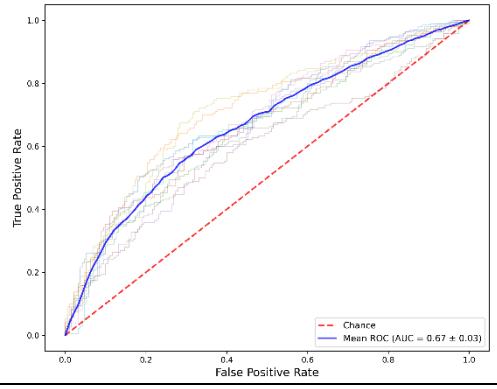

		

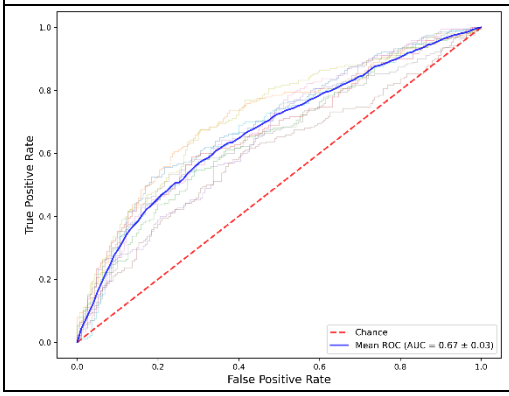
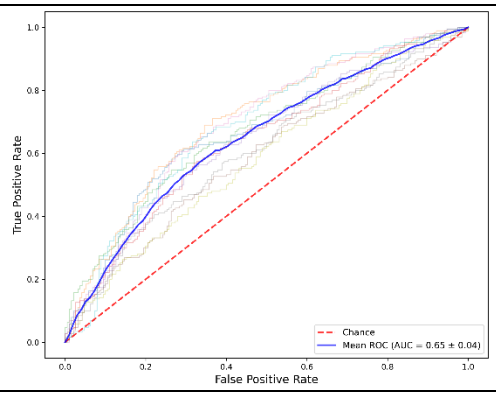
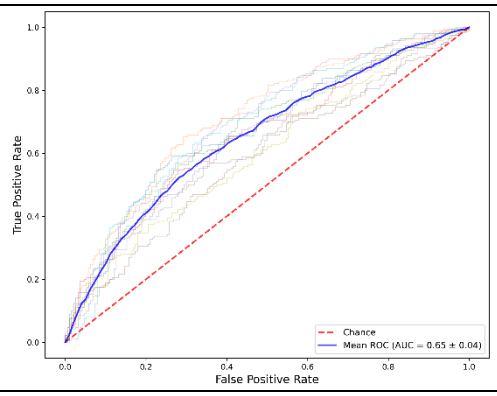
Figure 41g - Features + Apyest	**Figure 41h - Features (Masked)**	**Figure 41i - Features (M) + Apyest**
		

Figure 41j - Point Coordinates	Figure 41k - Points + Apyest	Figure 41l - Points (No Mouth)

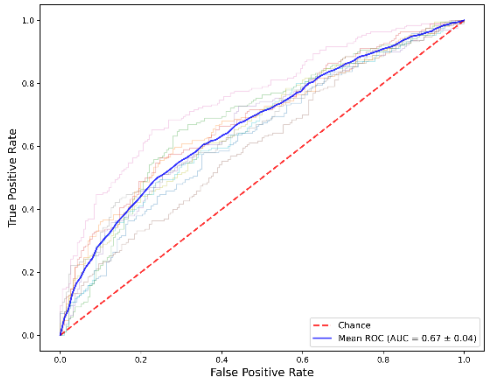
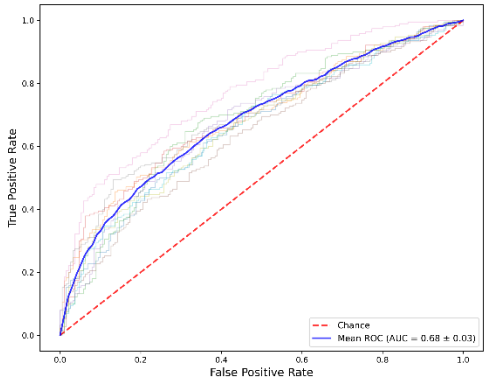
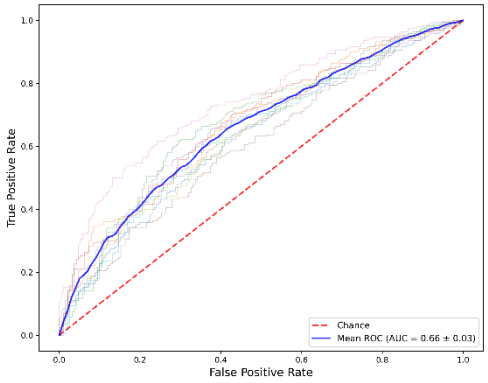

		

Figure 41m - Points (NM) + Apyest	**Figure 41n - Points (Happy)**	**Figure 41o - Points (H) + Apyest**

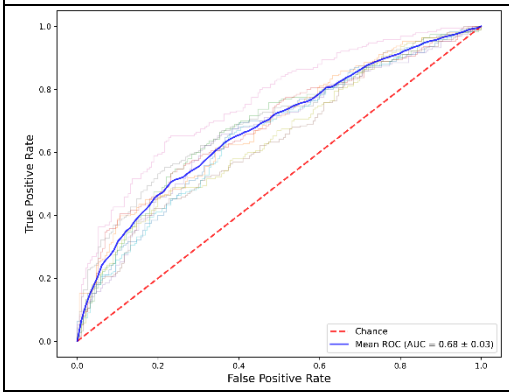
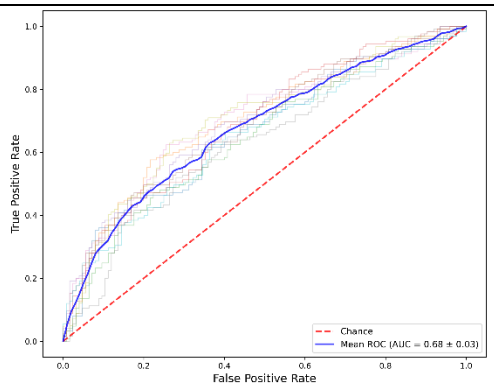
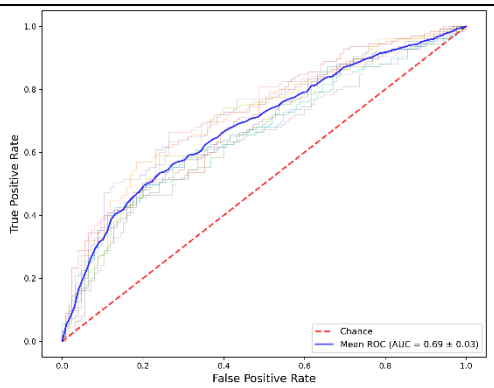

		

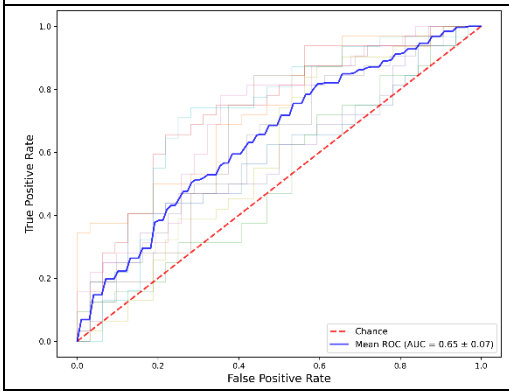
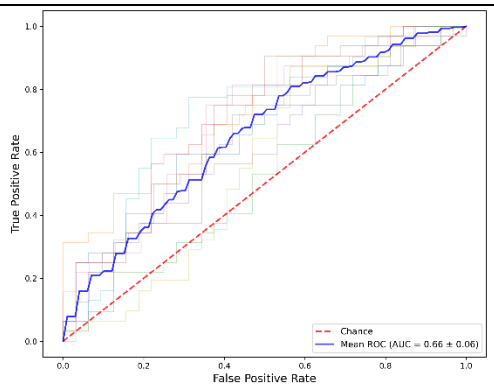
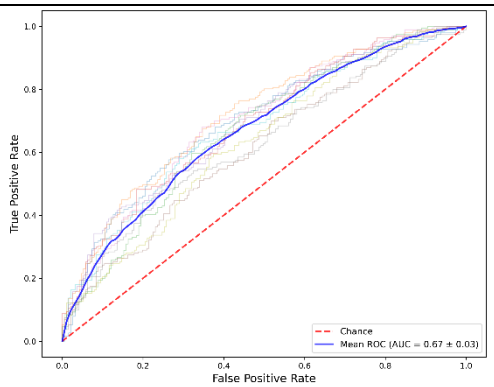
Figure 41p - Points (Neutral)	**Figure 41q - Points (N) + Apyest**	**Figure 41r - Mesh Coordinates**
		

Figure 41s - Mesh + Apyest	Figure 41t - Mesh (Happy)	Figure 41u - Mesh (H) + Apyest

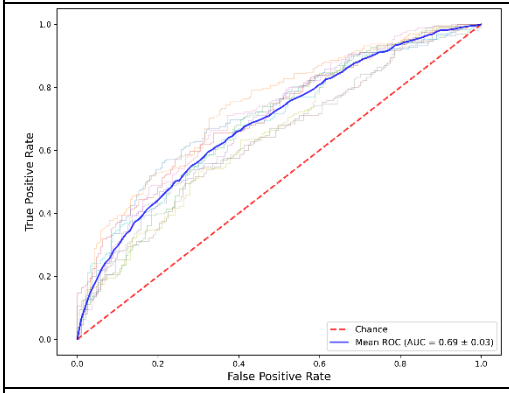
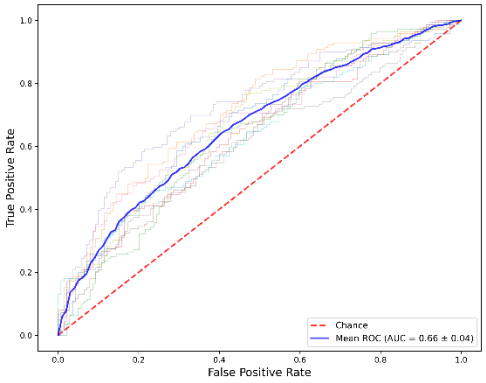
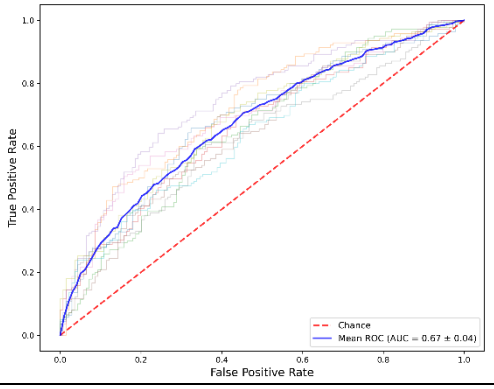
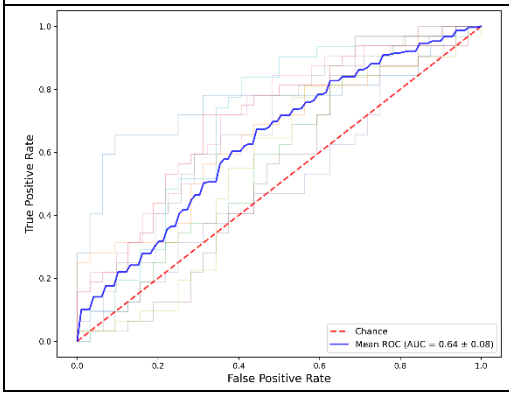
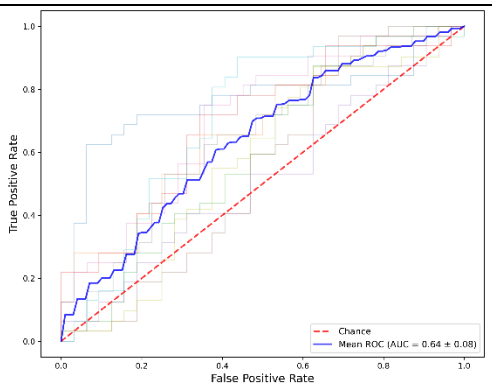

		

Figure 41v - Mesh (Neutral)	**Figure 41w - Mesh (N) + Apyest**	
		

Appendix M
Figure 42
Accuracy – White Males – Gun – All Images

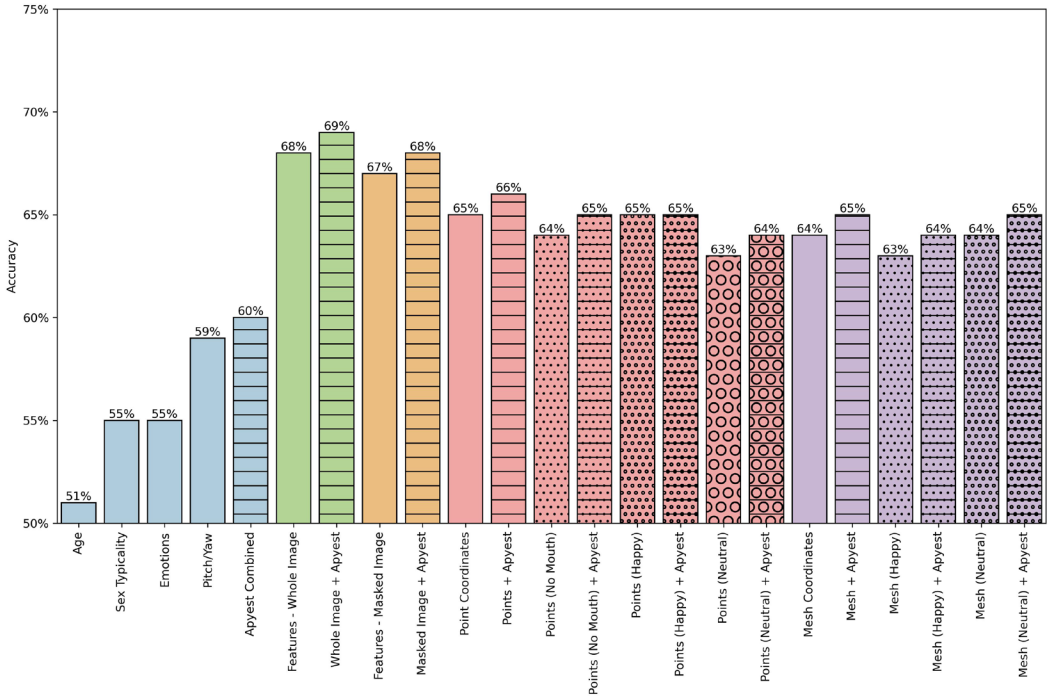
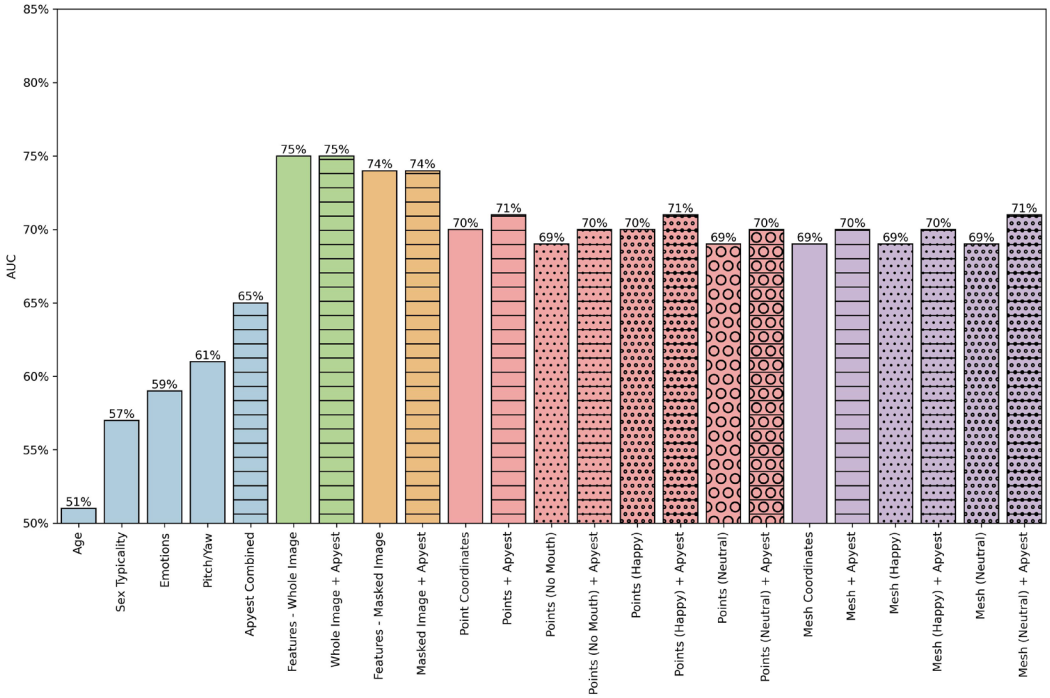


Figure 43
AUC – White Males – Gun – All Images



Return to relevant section

Figure 44

Accuracy – White Males – Gun – Reduced Images

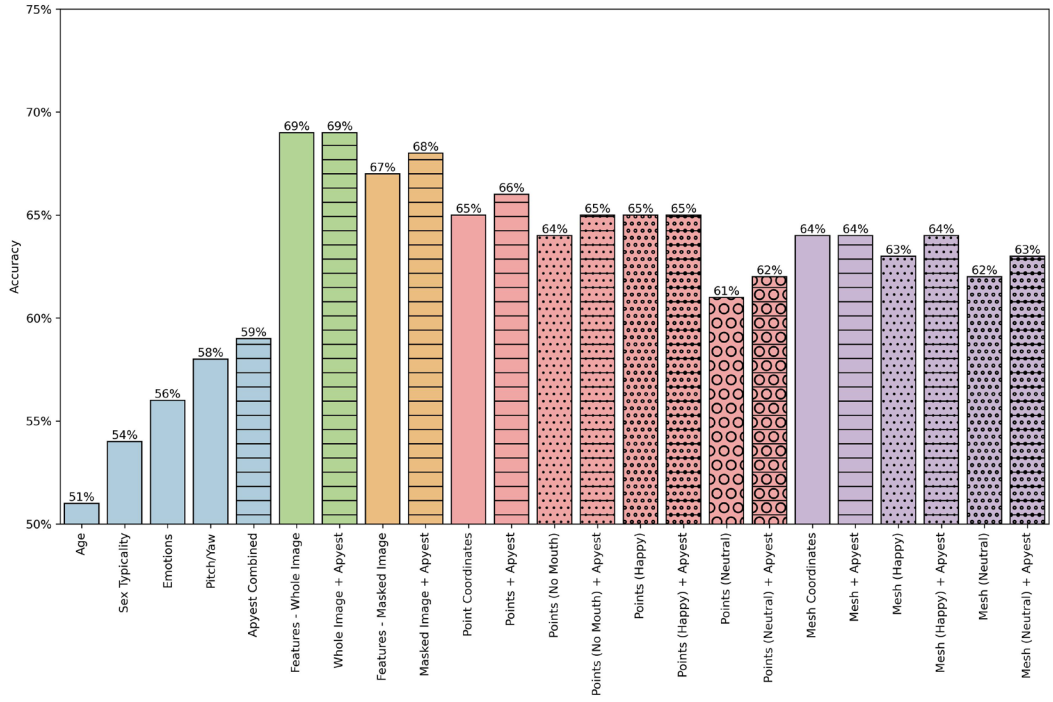


Figure 45

AUC – White Males – Gun – Reduced Images

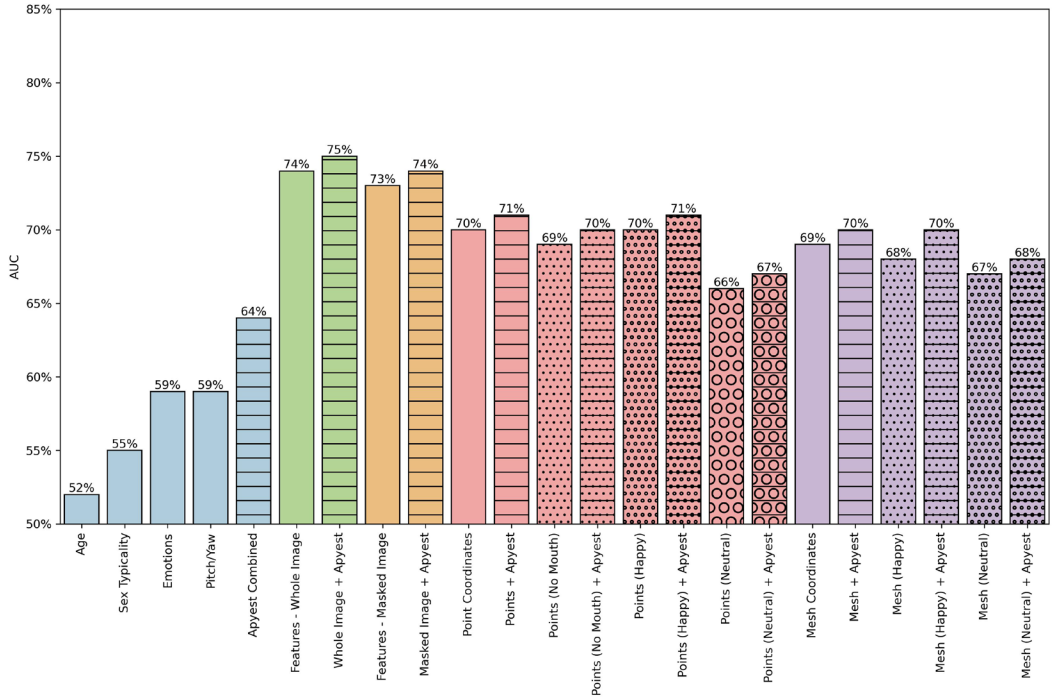


Figure 46

Accuracy – White Females – Gun – All Images

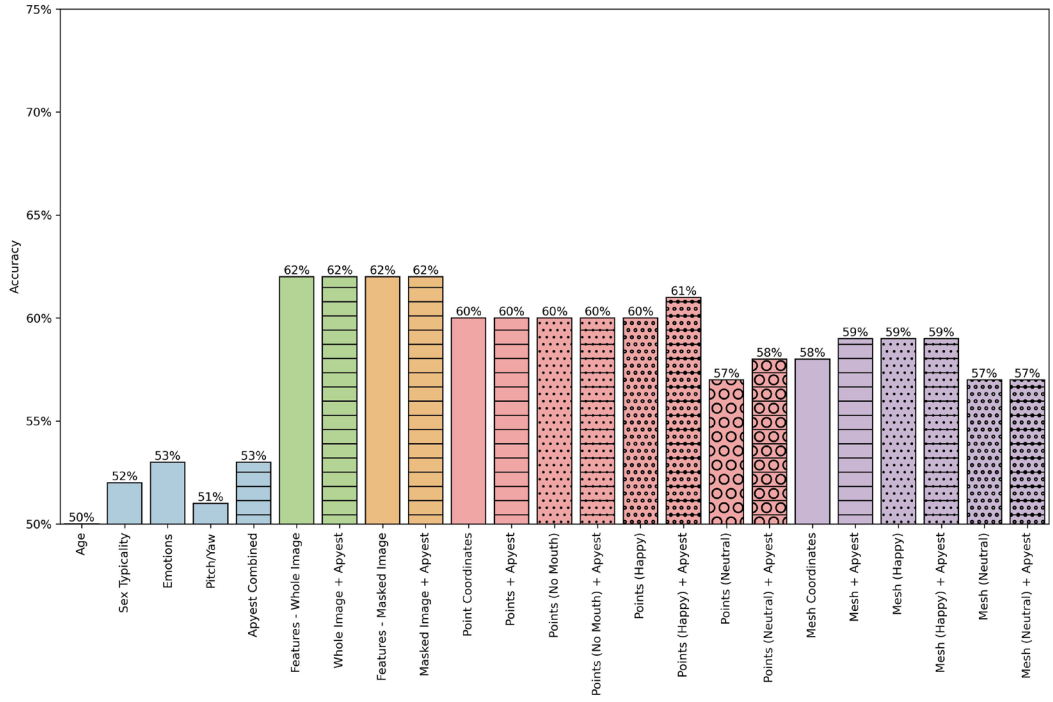


Figure 47

AUC – White Females – Gun – All Images

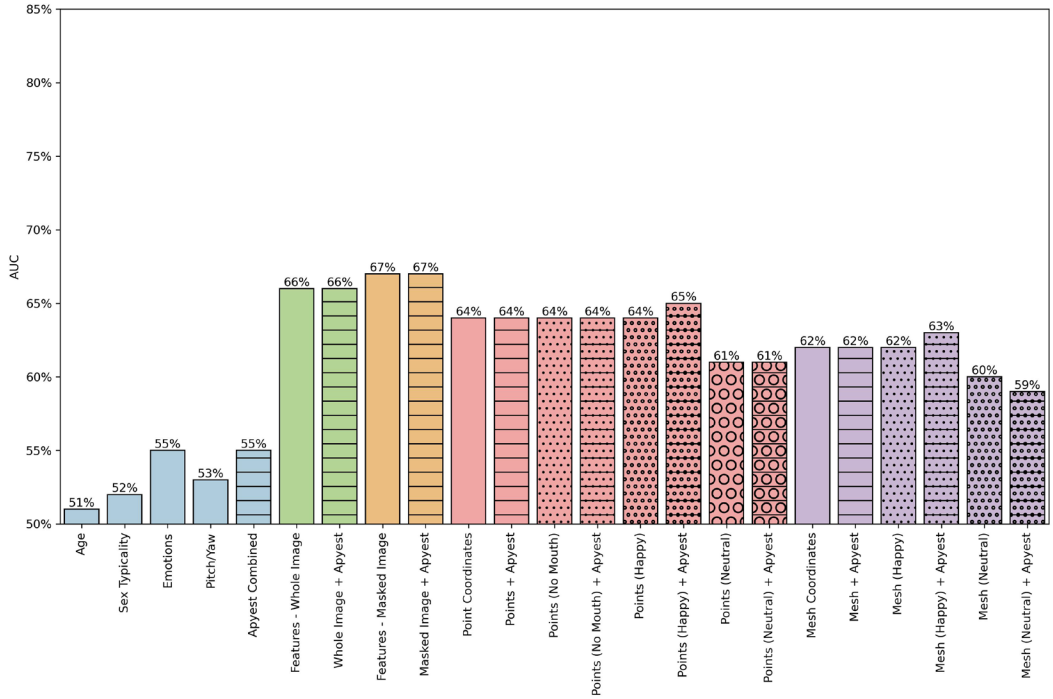


Figure 48

Accuracy – White Females – Gun – Reduced Images

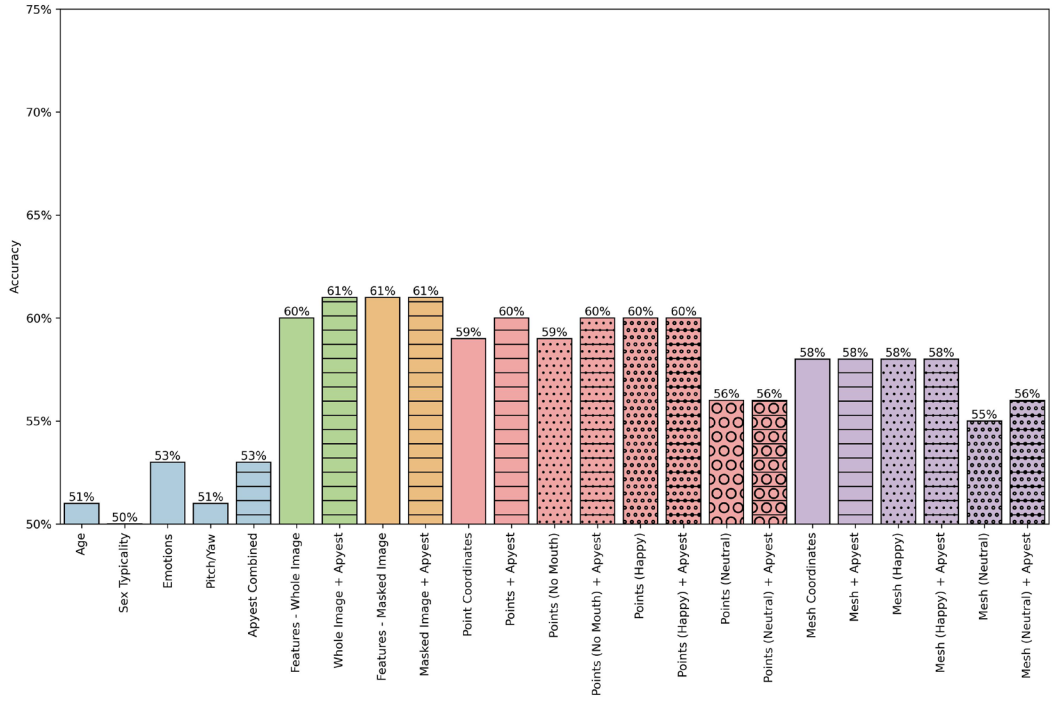


Figure 49

AUC – White Females – Gun – Reduced Images

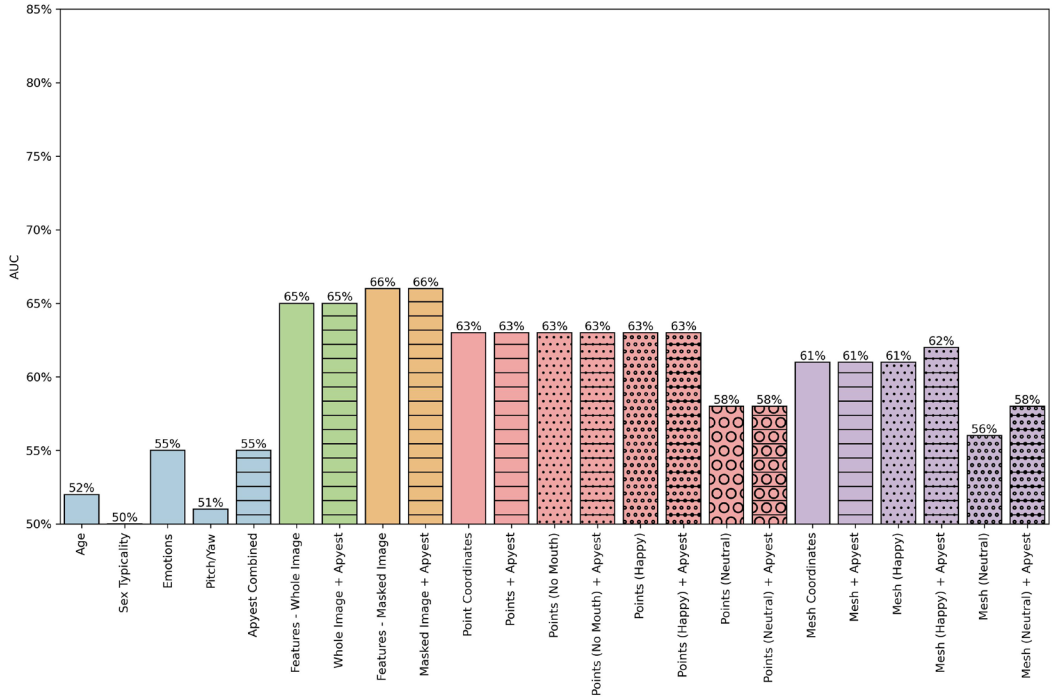


Figure 50

Accuracy – White Males – Immigration – All Images

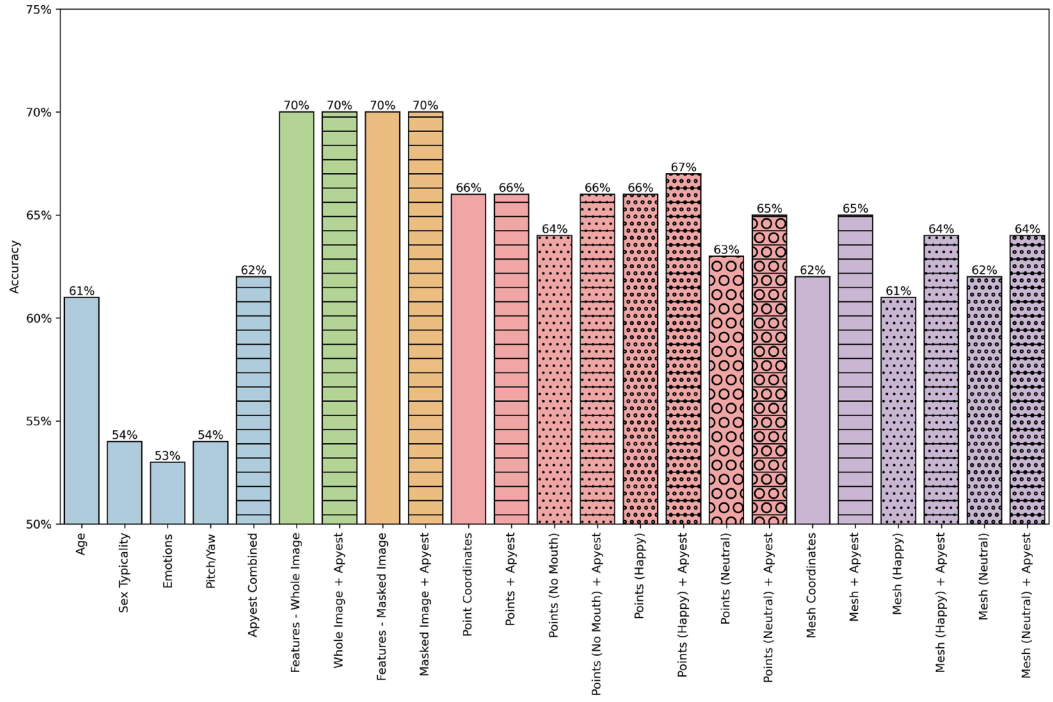


Figure 51

AUC – White Males – Immigration – All Images

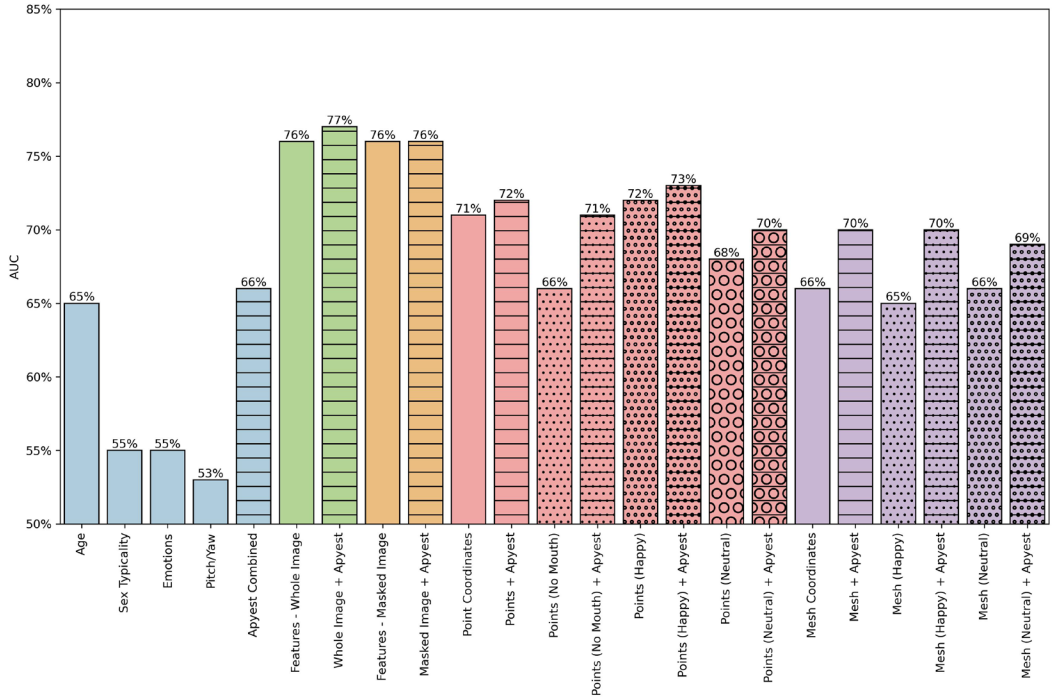


Figure 52

Accuracy – White Males – Immigration – Reduced Images

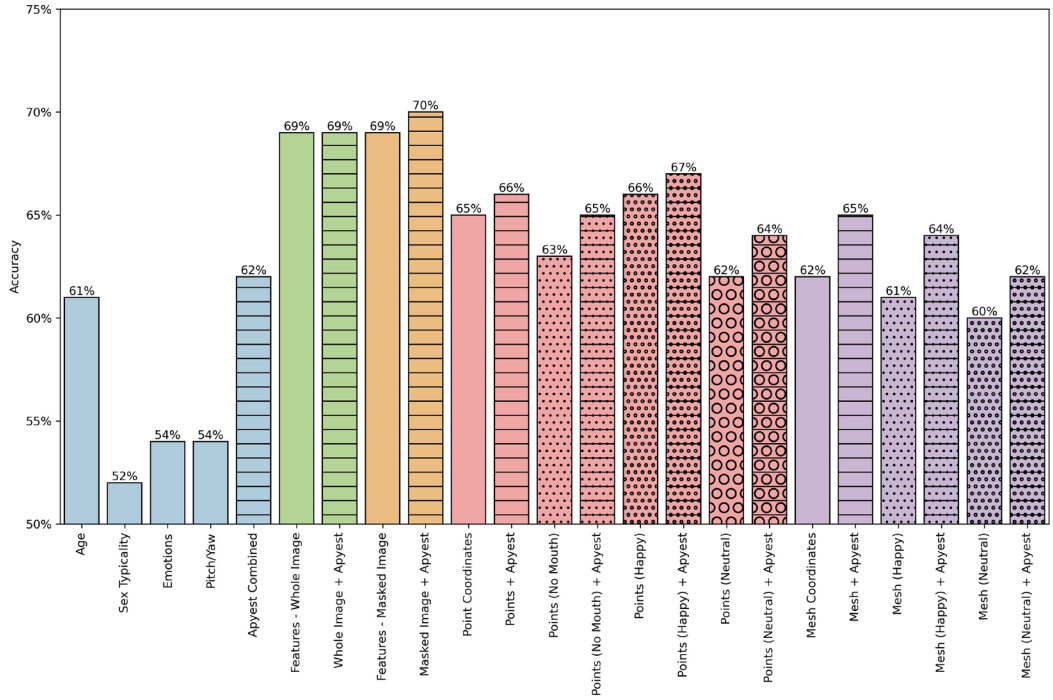


Figure 53

AUC – White Males – Immigration – Reduced Images

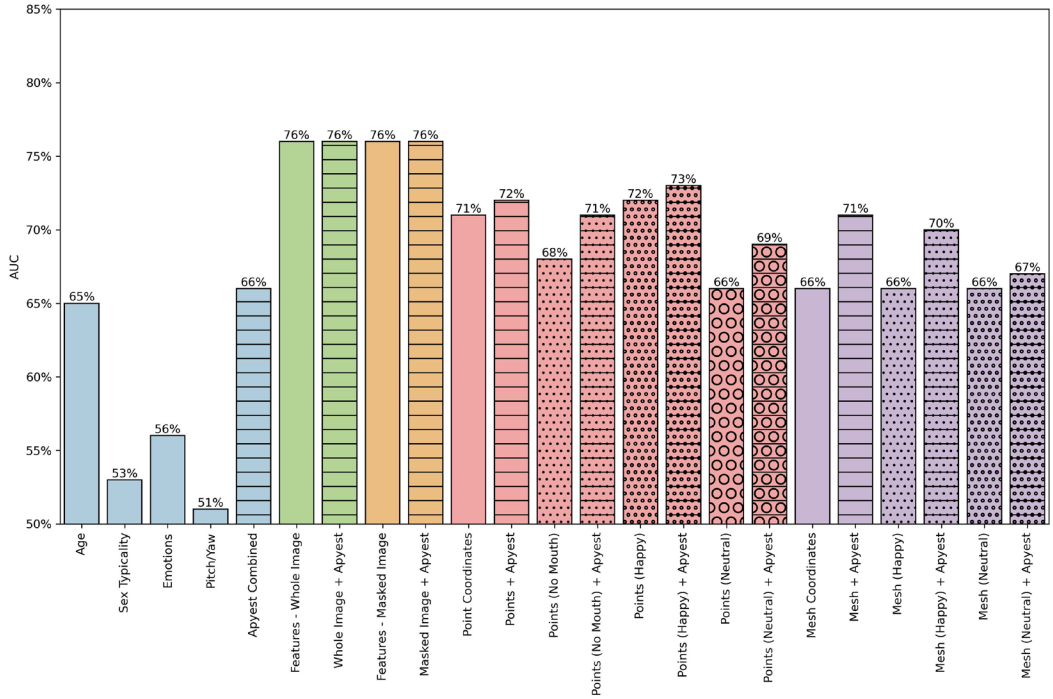


Figure 54

Accuracy – White Females – Immigration – All Images

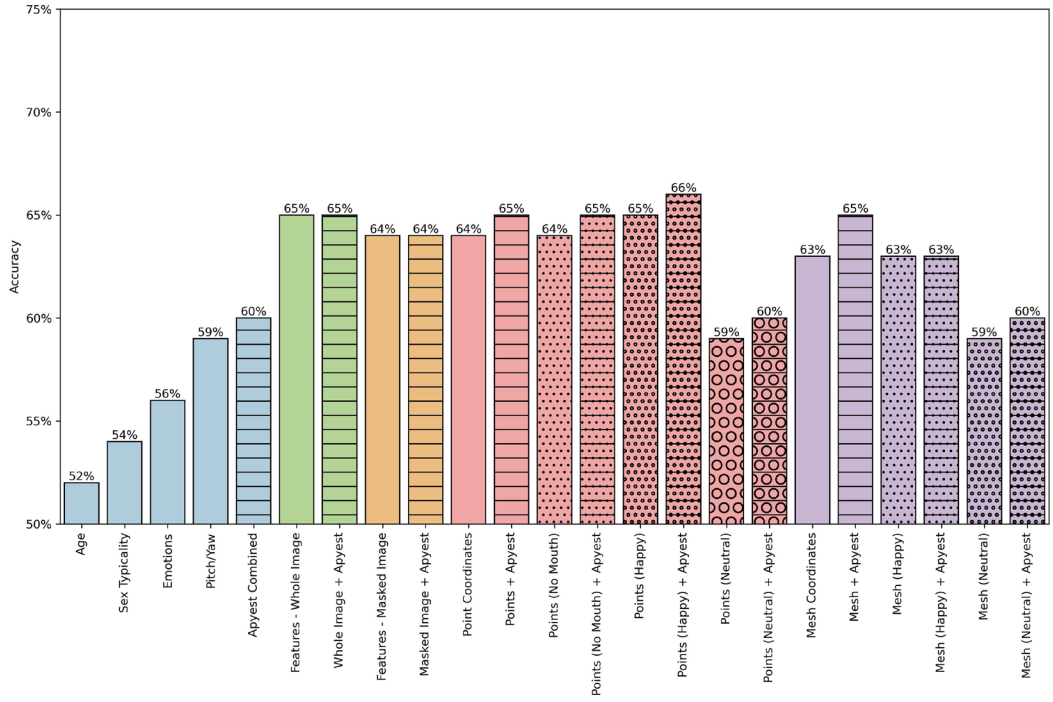


Figure 55

AUC – White Females – Immigration – All Images

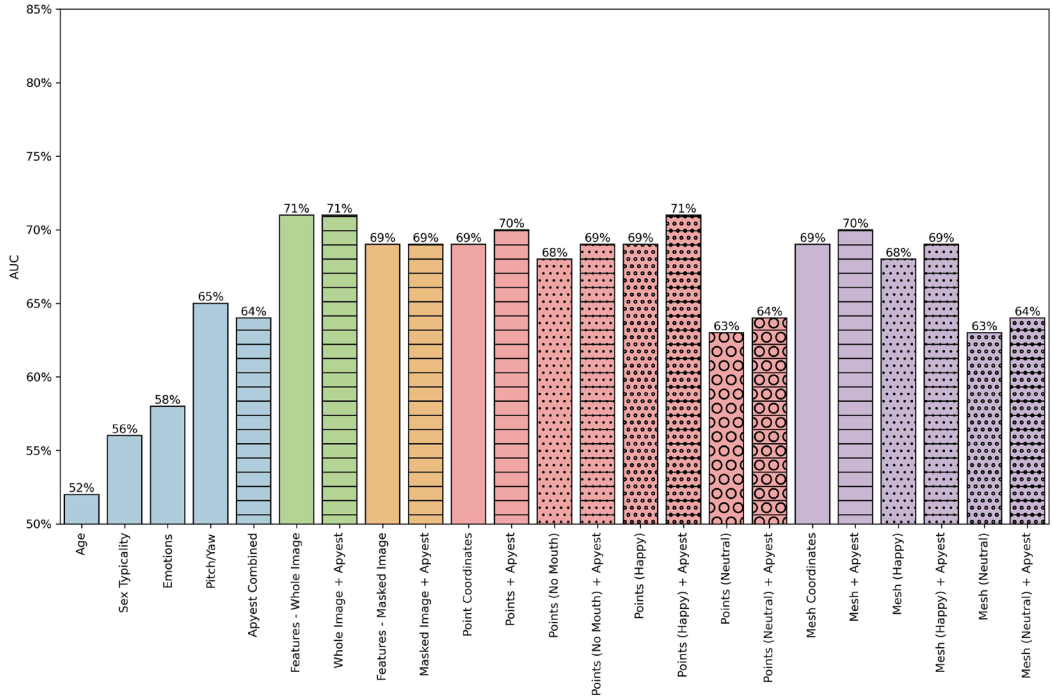


Figure 56

Accuracy – White Females – Immigration – Reduced Images

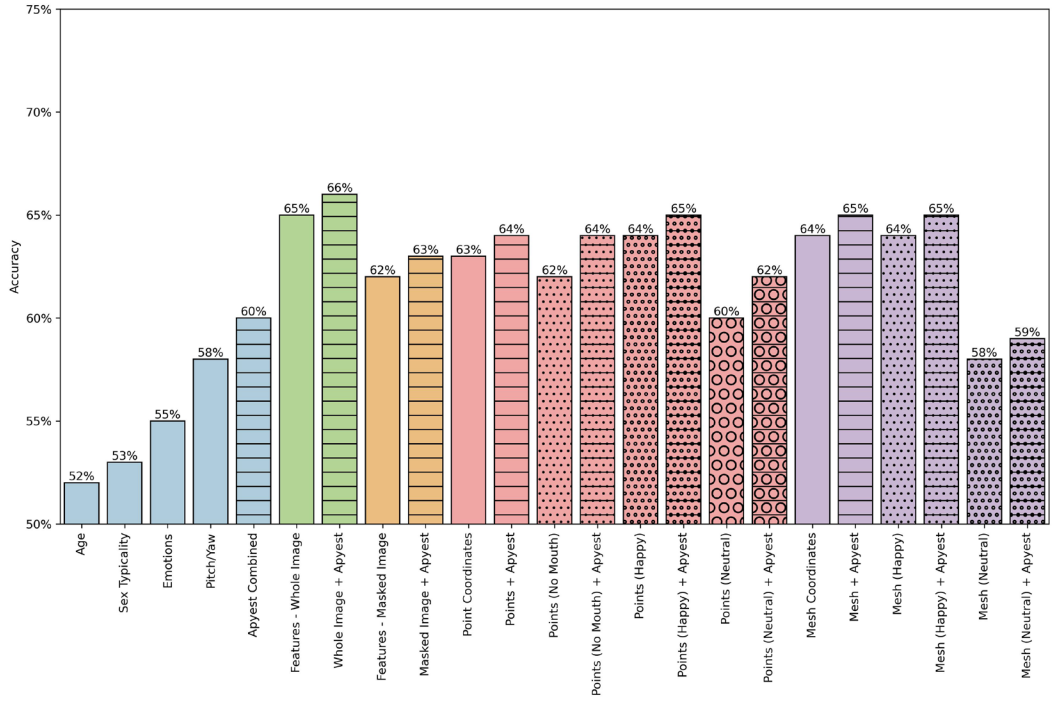


Figure 57

AUC – White Females – Immigration – Reduced Images

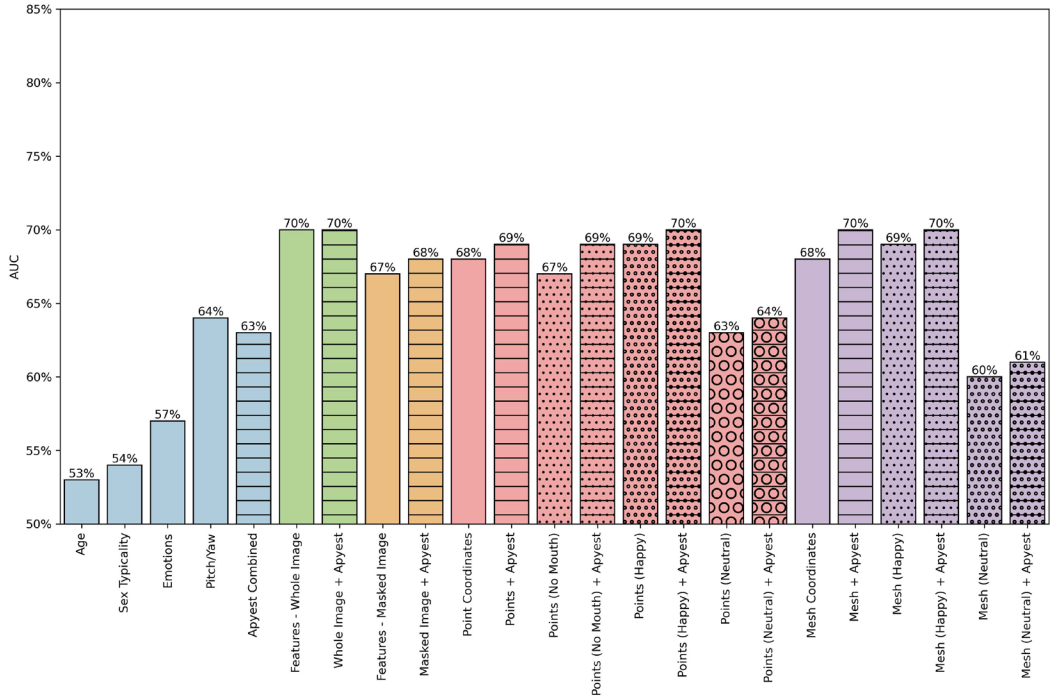


Figure 58

Accuracy – Asian Males – Gun – All Images

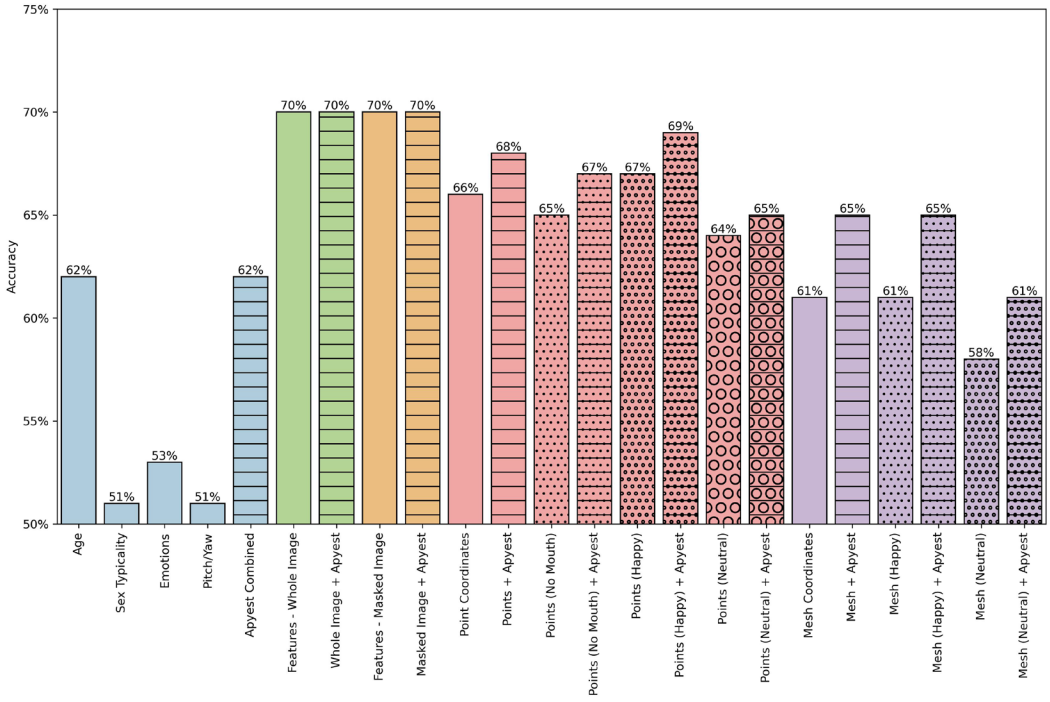


Figure 59

AUC – Asian Males – Gun – All Images

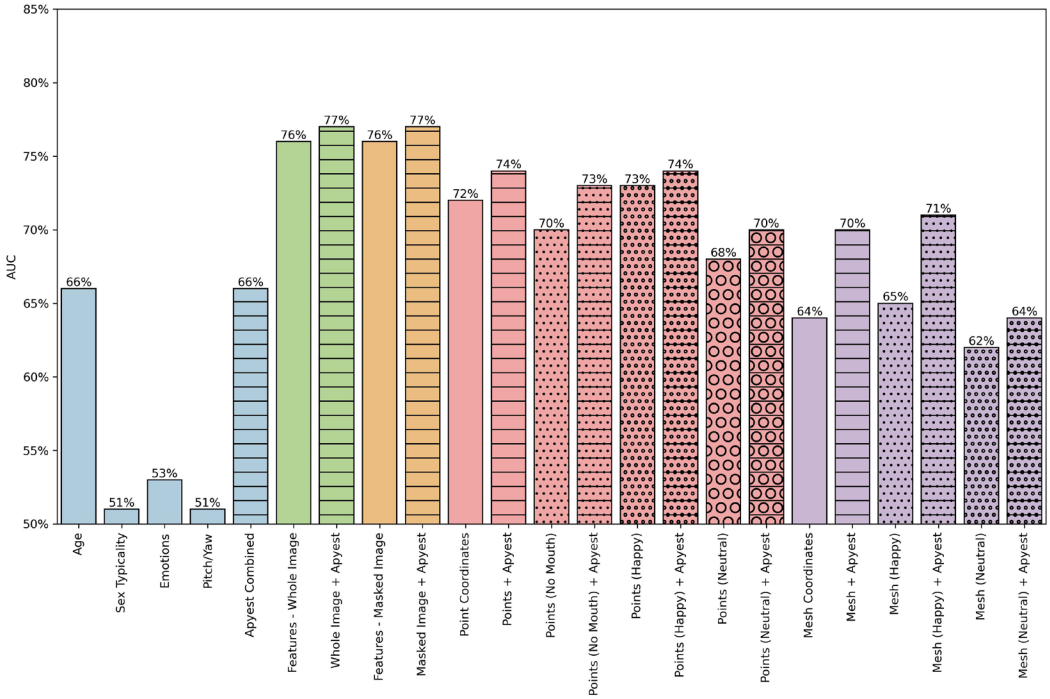


Figure 60

Accuracy – Asian Males – Gun – Reduced Images

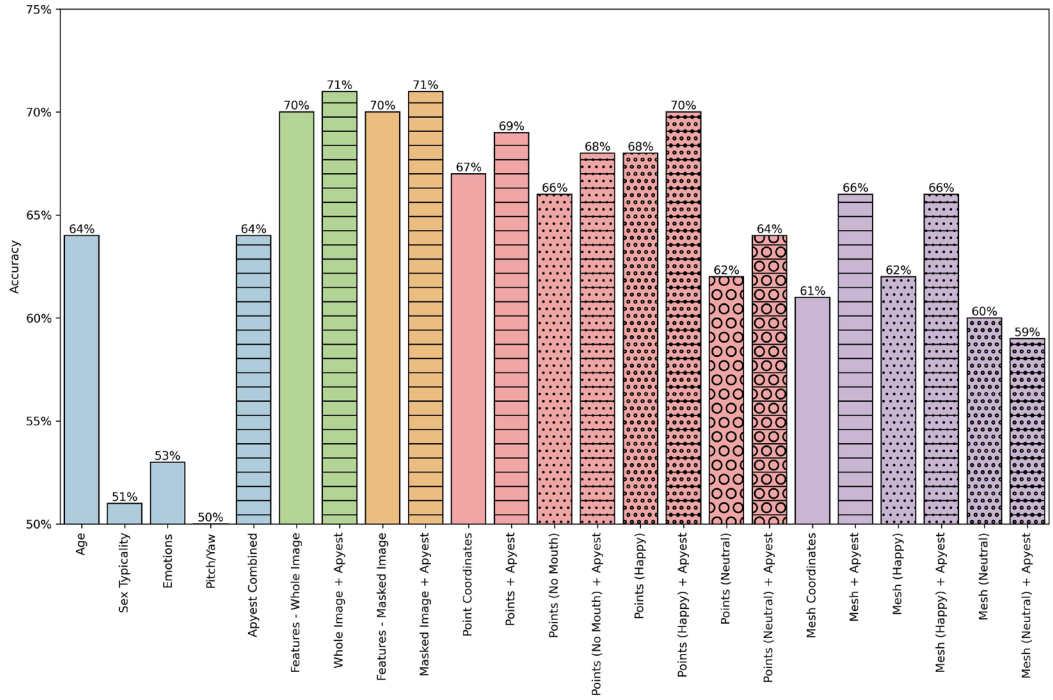


Figure 61

AUC – Asian Males – Gun – Reduced Images

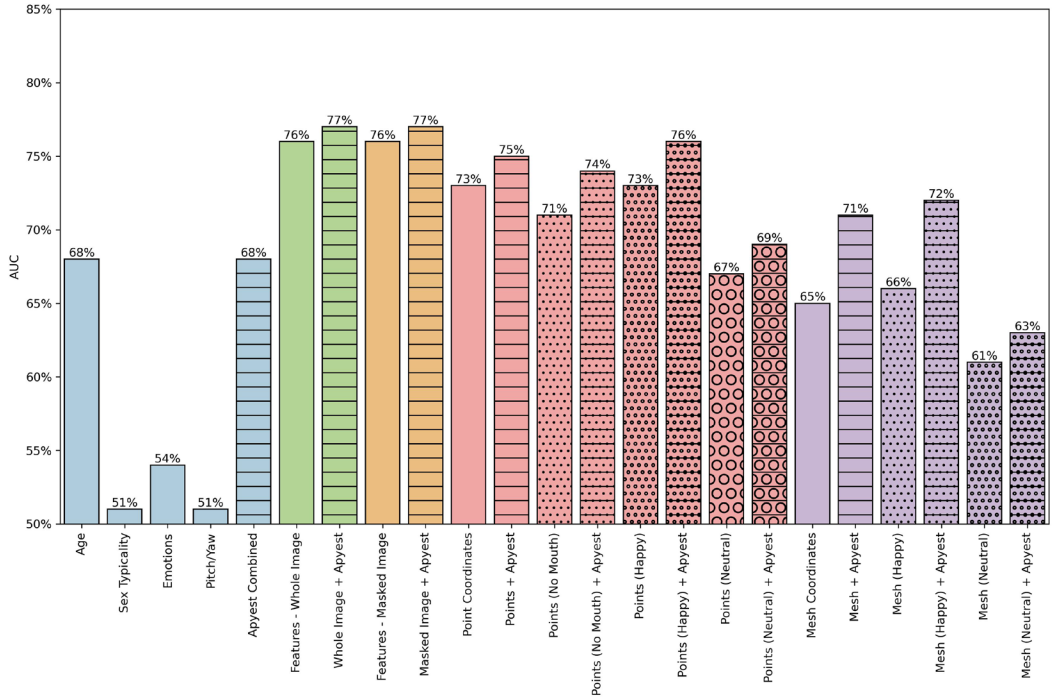


Figure 62

Accuracy – Hispanic Males – Immigration – All Images

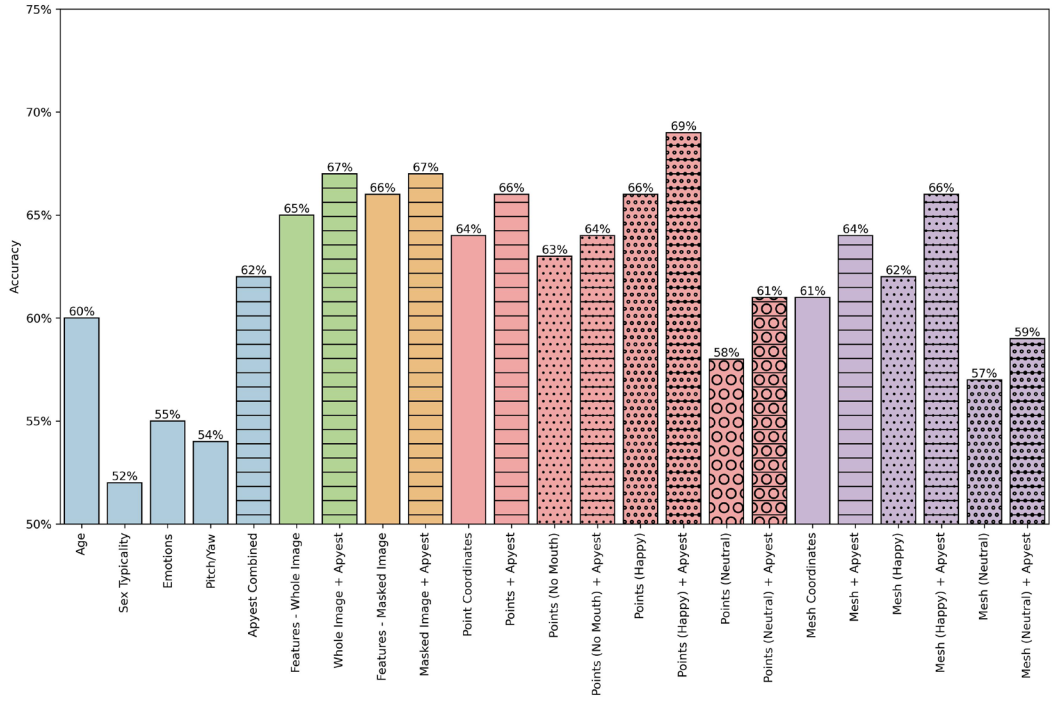


Figure 63

AUC – Hispanic Males – Immigration – All Images

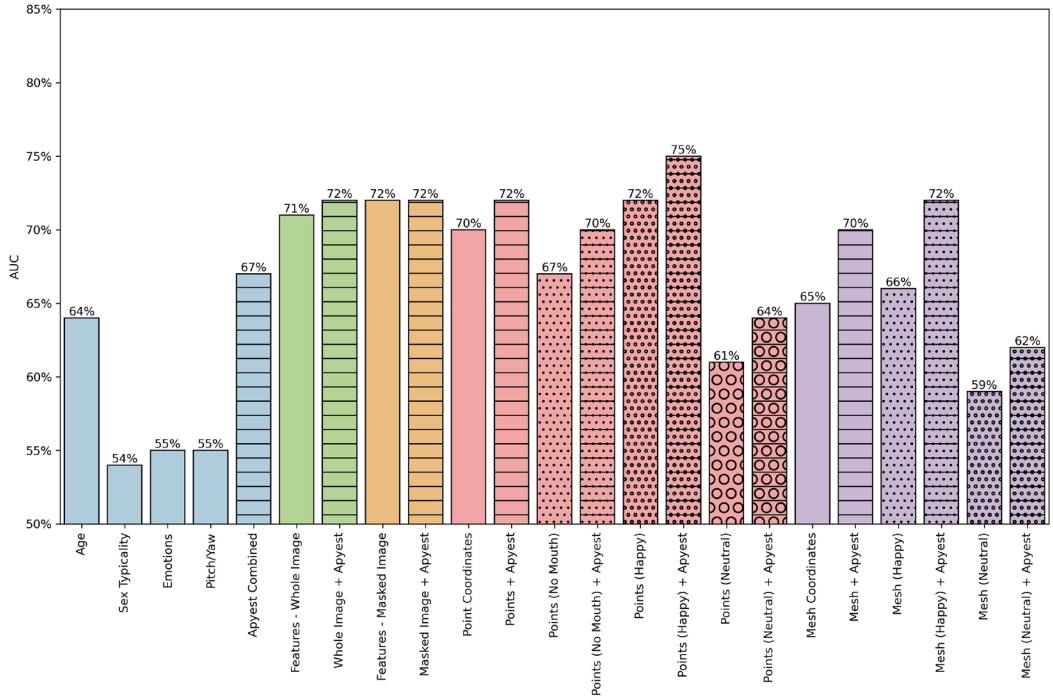


Figure 64

Accuracy – Hispanic Males – Immigration – Reduced Images

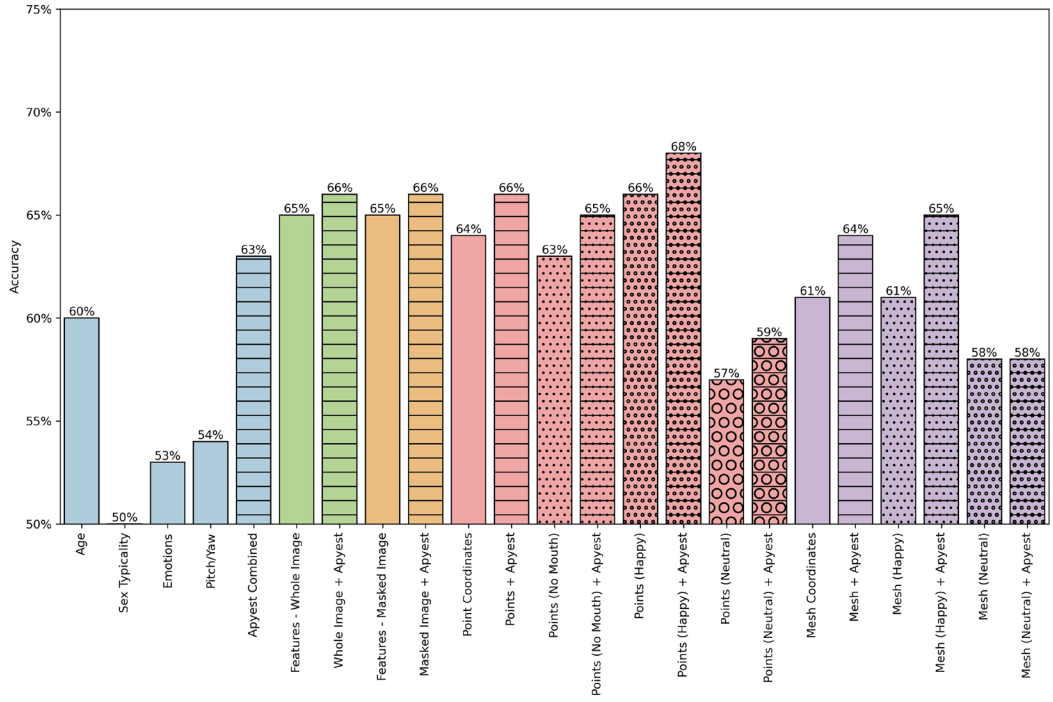


Figure 65

AUC – Hispanic Males – Immigration – Reduced Images

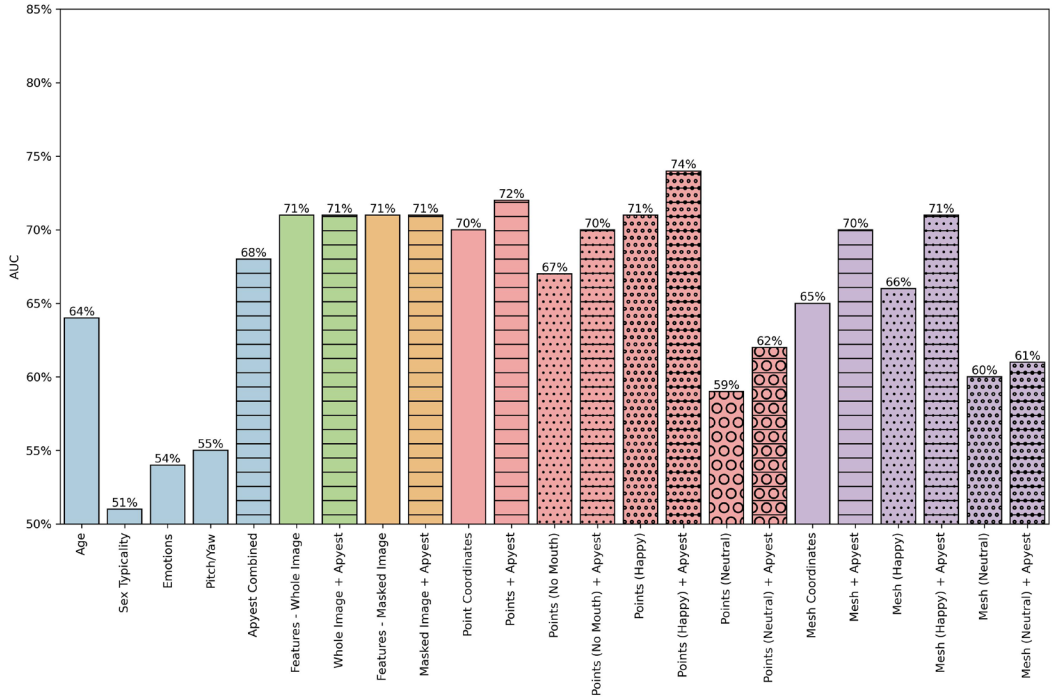


Figure 66

Accuracy – Hispanic Males – Gun – All Images

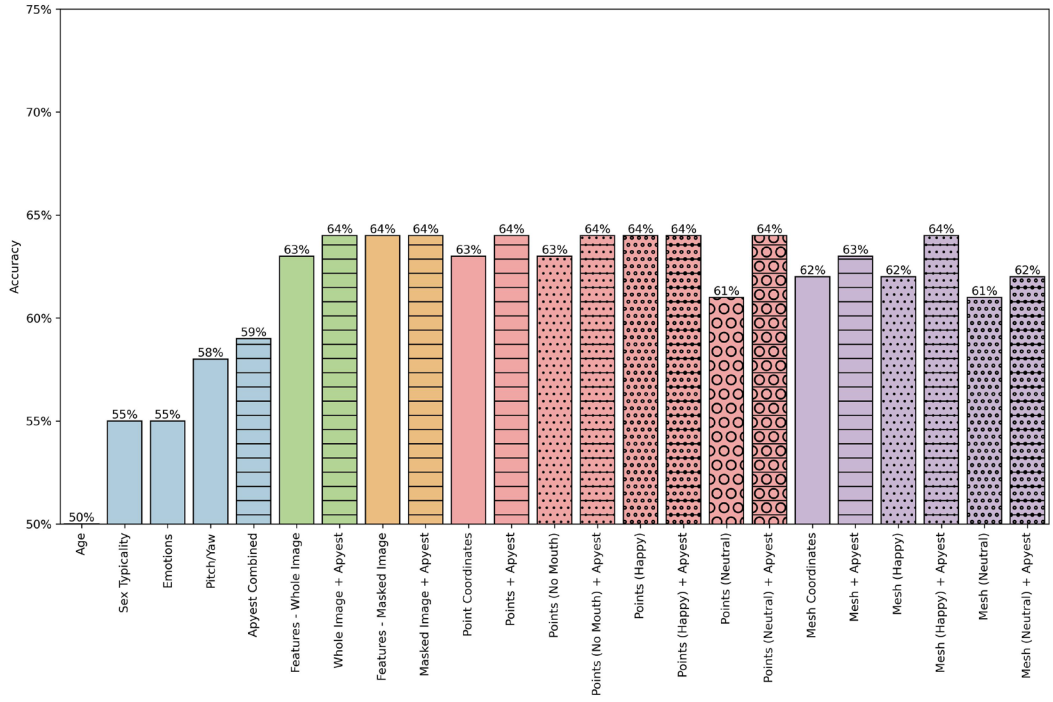


Figure 67

AUC – Hispanic Males – Gun – All Images

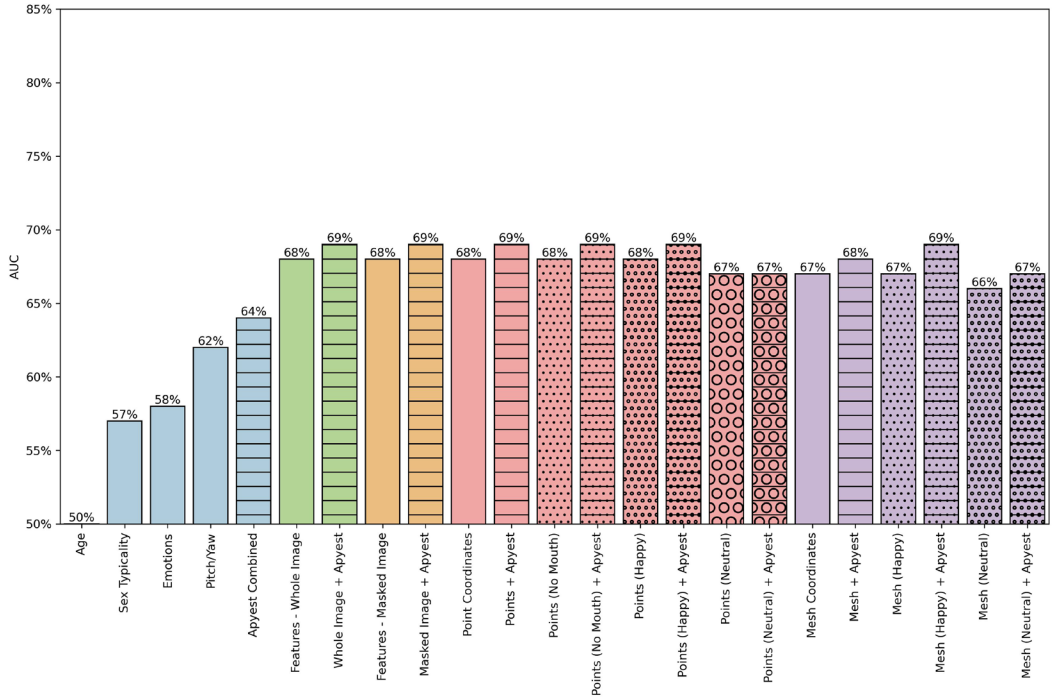


Figure 68

Accuracy – Hispanic Males – Gun – Reduced Images

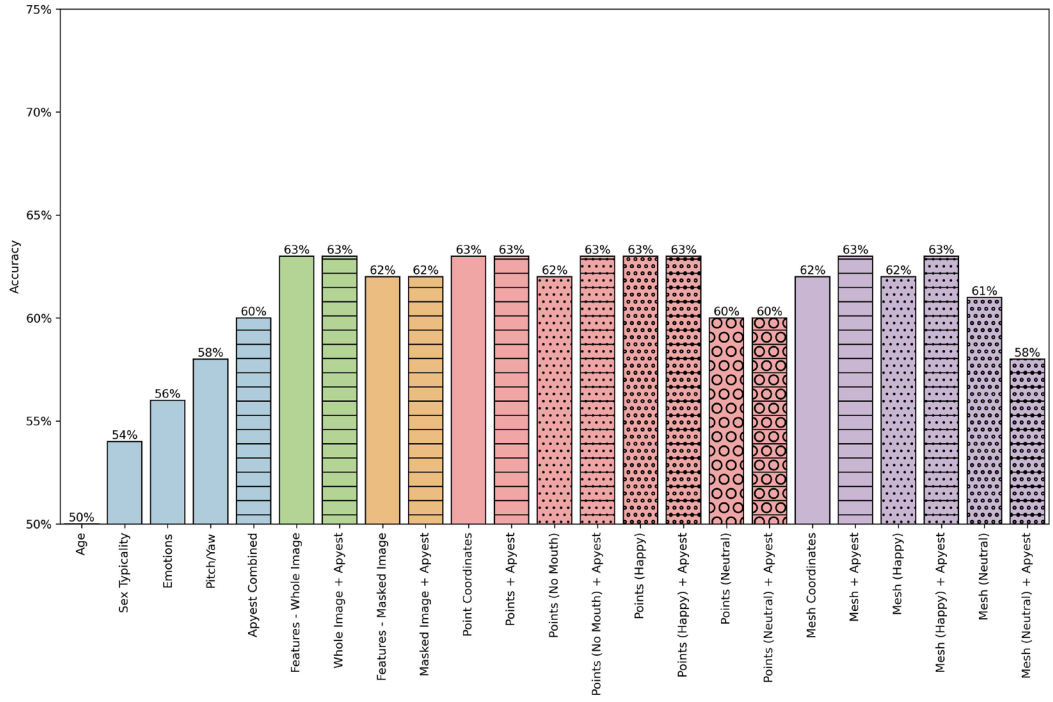
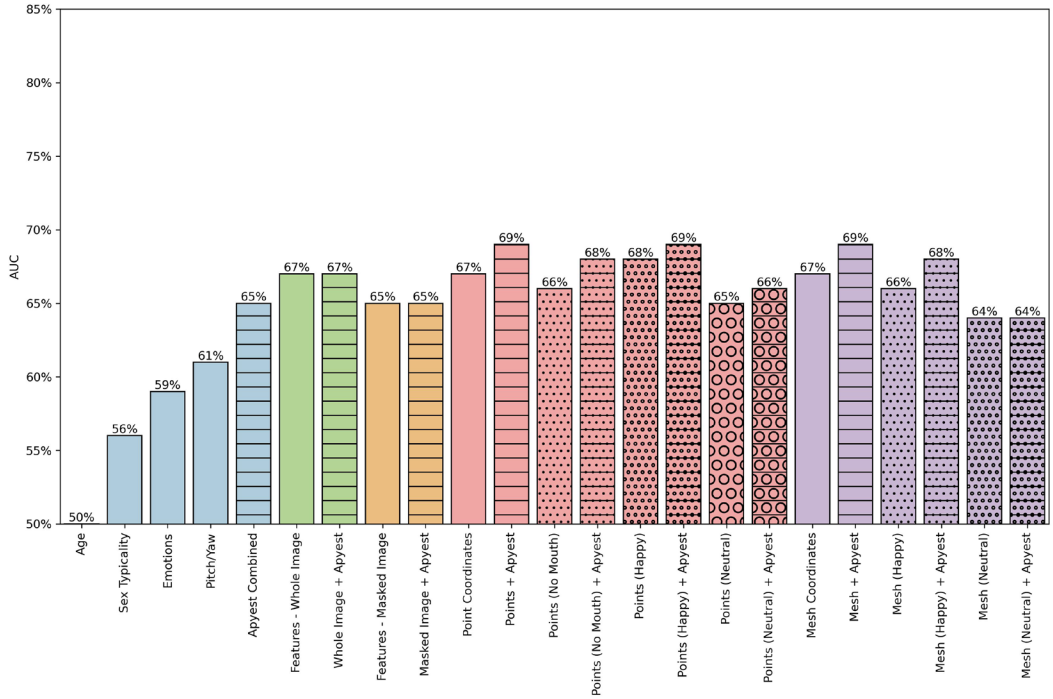


Figure 69

AUC – Hispanic Males – Gun – Reduced Images



Appendix N
Table 11
Model Metrics – White Males – Gun – All Images

	White Males Gun					
	n (of smaller group)	Area Under Curve (AUC)	CI Lower	CI Upper	Standard Error	Accuracy
Age	20,531	0.505	0.494	0.517	0.0051	51%
Pitch/Yaw	20,531	0.569	0.562	0.577	0.0033	55%
Emotions	20,531	0.586	0.578	0.594	0.0036	55%
Sex Typicality	20,531	0.613	0.598	0.629	0.0067	59%
Combined Apyest Model	20,531	0.646	0.632	0.659	0.0060	60%
Features - Whole Image	20,531	0.746	0.736	0.756	0.0044	68%
Features (WI) + Apyest	20,531	0.750	0.740	0.761	0.0045	69%
Features - Masked Image	20,531	0.735	0.721	0.748	0.0059	67%
Features (MI) + Apyest	20,531	0.741	0.727	0.755	0.0063	68%
Point Coordinates	20,531	0.701	0.689	0.712	0.0049	65%
Point Coordinates + Apyest	20,531	0.710	0.698	0.721	0.0051	66%
Point Coordinates - No Mouth	20,531	0.692	0.681	0.701	0.0045	64%
Point Coordinates - No Mouth + Apyest	20,531	0.704	0.693	0.715	0.0050	65%
Point Coordinates (Happy)	11,913	0.696	0.683	0.709	0.0059	65%
Point Coordinates (Happy) + Apyest	11,913	0.707	0.698	0.717	0.0042	65%
Point Coordinates (Neutral)	3,087	0.691	0.675	0.706	0.0068	63%
Point Coordinates (Neutral) + Apyest	3,087	0.699	0.684	0.714	0.0066	64%
Mesh Coordinates	20,446	0.691	0.679	0.703	0.0052	64%
Mesh Coordinates + Apyest	20,446	0.704	0.691	0.716	0.0055	65%
Mesh Coordinates (Happy)	14,052	0.685	0.677	0.692	0.0034	63%
Mesh Coordinates (Happy) + Apyest	14,052	0.698	0.690	0.707	0.0039	64%
Mesh Coordinates (Neutral)	3,078	0.692	0.675	0.710	0.0078	64%
Mesh Coordinates (Neutral) + Apyest	3,078	0.705	0.687	0.723	0.0078	65%

Return to relevant section

Table 12

Model Metrics – White Males – Gun – Reduced Images

	White Males Gun - Reduced					
	n (of smaller group)	Area Under Curve (AUC)	CI Lower	CI Upper	Standard Error	Accuracy
Age	12,836	0.515	0.505	0.526	0.0047	51%
Pitch/Yaw	12,836	0.550	0.542	0.558	0.0034	54%
Emotions	12,836	0.589	0.580	0.598	0.0038	56%
Sex Typicality	12,836	0.592	0.572	0.611	0.0087	58%
Combined Apyest Model	12,836	0.637	0.624	0.649	0.0056	59%
Features - Whole Image	12,836	0.744	0.734	0.755	0.0047	69%
Features (WI) + Apyest	12,836	0.747	0.737	0.758	0.0047	69%
Features - Masked Image	12,836	0.732	0.721	0.743	0.0051	67%
Features (MI) + Apyest	12,836	0.738	0.727	0.750	0.0051	68%
Point Coordinates	12,836	0.698	0.689	0.707	0.0040	65%
Point Coordinates + Apyest	12,836	0.708	0.698	0.718	0.0044	66%
Point Coordinates - No Mouth	12,836	0.688	0.682	0.695	0.0030	64%
Point Coordinates - No Mouth + Apyest	12,836	0.702	0.693	0.711	0.0039	65%
Point Coordinates (Happy)	8,813	0.696	0.686	0.715	0.0041	65%
Point Coordinates (Happy) + Apyest	8,813	0.706	0.696	0.716	0.0044	65%
Point Coordinates (Neutral)	1,869	0.664	0.647	0.682	0.0077	61%
Point Coordinates (Neutral) + Apyest	1,869	0.674	0.654	0.694	0.0087	62%
Mesh Coordinates	12,796	0.686	0.675	0.697	0.0049	64%
Mesh Coordinates + Apyest	12,796	0.698	0.686	0.710	0.0052	64%
Mesh Coordinates (Happy)	9,355	0.681	0.666	0.695	0.0062	63%
Mesh Coordinates (Happy) + Apyest	9,355	0.696	0.682	0.711	0.0065	64%
Mesh Coordinates (Neutral)	1,869	0.672	0.653	0.692	0.0085	62%
Mesh Coordinates (Neutral) + Apyest	1,869	0.683	0.662	0.704	0.0093	63%

Table 13

Model Metrics – White Females – Gun – All Images

	White Females Gun					
	n (of smaller group)	Area Under Curve (AUC)	CI Lower	CI Upper	σM	Accuracy
Age	18,416	0.510	0.501	0.518	0.0040	50%
Pitch/Yaw	18,416	0.522	0.518	0.526	0.0019	52%
Emotions	18,416	0.553	0.543	0.562	0.0044	53%
Sex Typicality	18,416	0.527	0.523	0.532	0.0019	51%
Combined Apyest Model	18,416	0.551	0.542	0.560	0.0039	53%
Features - Whole Image	18,416	0.663	0.651	0.675	0.0051	62%
Features (WI) + Apyest	18,416	0.663	0.651	0.675	0.0052	62%
Features - Masked Image	18,416	0.671	0.662	0.680	0.0038	62%
Features (MI) + Apyest	18,416	0.672	0.663	0.680	0.0038	62%
Point Coordinates	18,416	0.639	0.626	0.653	0.0060	60%
Point Coordinates + Apyest	18,416	0.641	0.626	0.656	0.0066	60%
Point Coordinates - No Mouth	18,416	0.638	0.624	0.652	0.0063	60%
Point Coordinates - No Mouth + Apyest	18,416	0.640	0.625	0.656	0.0069	60%
Point Coordinates (Happy)	13,996	0.643	0.636	0.650	0.0029	60%
Point Coordinates (Happy) + Apyest	13,996	0.645	0.637	0.652	0.0035	61%
Point Coordinates (Neutral)	1,865	0.611	0.585	0.636	0.0114	57%
Point Coordinates (Neutral) + Apyest	1,865	0.610	0.584	0.635	0.0114	58%
Mesh Coordinates	18,443	0.620	0.609	0.632	0.0049	58%
Mesh Coordinates + Apyest	18,443	0.622	0.611	0.633	0.0050	59%
Mesh Coordinates (Happy)	15,095	0.624	0.615	0.633	0.0040	59%
Mesh Coordinates (Happy) + Apyest	15,095	0.625	0.617	0.634	0.0038	59%
Mesh Coordinates (Neutral)	1,870	0.597	0.562	0.633	0.0157	57%
Mesh Coordinates (Neutral) + Apyest	1,870	0.594	0.558	0.630	0.0161	57%

Table 14

Model Metrics – White Females – Gun – Reduced Images

	White Females Gun - Reduced					
	n (of smaller group)	Area Under Curve (AUC)	CI Lower	CI Upper	σM	Accuracy
Age	8,935	0.520	0.507	0.532	0.0054	51%
Pitch/Yaw	8,935	0.502	0.495	0.510	0.0034	50%
Emotions	8,935	0.548	0.533	0.563	0.0065	53%
Sex Typicality	8,935	0.509	0.498	0.521	0.0051	51%
Combined Apyest Model	8,935	0.545	0.535	0.555	0.0043	53%
Features - Whole Image	8,935	0.648	0.636	0.659	0.0050	60%
Features (WI) + Apyest	8,935	0.648	0.636	0.660	0.0052	61%
Features - Masked Image	8,935	0.658	0.645	0.671	0.0057	61%
Features (MI) + Apyest	8,935	0.658	0.645	0.671	0.0057	61%
Point Coordinates	8,935	0.628	0.610	0.646	0.0079	59%
Point Coordinates + Apyest	8,935	0.631	0.612	0.650	0.0085	60%
Point Coordinates - No Mouth	8,935	0.627	0.610	0.645	0.0077	59%
Point Coordinates - No Mouth + Apyest	8,935	0.630	0.611	0.649	0.0084	60%
Point Coordinates (Happy)	6,573	0.631	0.619	0.634	0.0054	60%
Point Coordinates (Happy) + Apyest	6,573	0.633	0.622	0.645	0.0049	60%
Point Coordinates (Neutral)	828	0.580	0.556	0.603	0.0100	56%
Point Coordinates (Neutral) + Apyest	828	0.584	0.554	0.613	0.0130	56%
Mesh Coordinates	8,955	0.612	0.599	0.626	0.0062	58%
Mesh Coordinates + Apyest	8,955	0.614	0.599	0.628	0.0064	58%
Mesh Coordinates (Happy)	7,233	0.614	0.600	0.628	0.0061	58%
Mesh Coordinates (Happy) + Apyest	7,233	0.615	0.601	0.629	0.0062	58%
Mesh Coordinates (Neutral)	830	0.561	0.514	0.609	0.0210	55%
Mesh Coordinates (Neutral) + Apyest	830	0.575	0.532	0.617	0.0186	56%

Table 15

Model Metrics – White Males – Immigration – All Images

	White Males Immigration					
	n (of smaller group)	Area Under Curve (AUC)	CI Lower	CI Upper	σM	Accuracy
Age	8,742	0.647	0.630	0.664	0.0076	61%
Pitch/Yaw	8,742	0.554	0.547	0.560	0.0029	54%
Emotions	8,742	0.547	0.535	0.558	0.0051	53%
Sex Typicality	8,742	0.534	0.521	0.547	0.0057	54%
Combined Apyest Model	8,742	0.661	0.644	0.679	0.0077	62%
Features - Whole Image	8,742	0.762	0.745	0.780	0.0076	70%
Features (WI) + Apyest	8,742	0.769	0.752	0.786	0.0075	70%
Features - Masked Image	8,742	0.758	0.745	0.772	0.0060	70%
Features (MI) + Apyest	8,742	0.763	0.749	0.776	0.0060	70%
Point Coordinates	8,742	0.712	0.693	0.732	0.0085	66%
Point Coordinates + Apyest	8,742	0.724	0.699	0.750	0.0111	66%
Point Coordinates - No Mouth	8,742	0.657	0.669	0.705	0.0081	64%
Point Coordinates - No Mouth + Apyest	8,742	0.711	0.687	0.736	0.0107	66%
Point Coordinates (Happy)	5,158	0.717	0.703	0.731	0.0061	66%
Point Coordinates (Happy) + Apyest	5,158	0.727	0.708	0.746	0.0083	67%
Point Coordinates (Neutral)	1,535	0.680	0.660	0.701	0.0092	63%
Point Coordinates (Neutral) + Apyest	1,535	0.702	0.682	0.721	0.0085	65%
Mesh Coordinates	8,702	0.663	0.654	0.672	0.0040	62%
Mesh Coordinates + Apyest	8,702	0.703	0.691	0.715	0.0052	65%
Mesh Coordinates (Happy)	5,595	0.654	0.639	0.669	0.0067	61%
Mesh Coordinates (Happy) + Apyest	5,595	0.698	0.679	0.718	0.0087	64%
Mesh Coordinates (Neutral)	1,517	0.662	0.641	0.682	0.0091	62%
Mesh Coordinates (Neutral) + Apyest	1,517	0.690	0.671	0.710	0.0086	64%

Table 16

Model Metrics – White Males – Immigration – Reduced Images

	White Males Immigration - Reduced					
	n (of smaller group)	Area Under Curve (AUC)	CI Lower	CI Upper	σM	Accuracy
Age	5,224	0.649	0.625	0.673	0.0105	61%
Pitch/Yaw	5,224	0.530	0.516	0.544	0.0063	52%
Emotions	5,224	0.555	0.539	0.572	0.0073	54%
Sex Typicality	5,224	0.514	0.495	0.534	0.0087	54%
Combined Apyest Model	5,224	0.664	0.642	0.686	0.0097	62%
Features - Whole Image	5,224	0.755	0.732	0.778	0.0103	69%
Features (WI) + Apyest	5,224	0.761	0.738	0.783	0.0101	69%
Features - Masked Image	5,224	0.756	0.740	0.773	0.0073	69%
Features (MI) + Apyest	5,224	0.760	0.744	0.777	0.0071	70%
Point Coordinates	5,224	0.707	0.680	0.734	0.0120	65%
Point Coordinates + Apyest	5,224	0.719	0.687	0.751	0.0141	66%
Point Coordinates - No Mouth	5,224	0.680	0.654	0.706	0.0115	63%
Point Coordinates - No Mouth + Apyest	5,224	0.705	0.673	0.737	0.0140	65%
Point Coordinates (Happy)	3,595	0.718	0.700	0.735	0.0077	66%
Point Coordinates (Happy) + Apyest	3,595	0.729	0.714	0.745	0.0070	67%
Point Coordinates (Neutral)	883	0.663	0.627	0.699	0.0159	62%
Point Coordinates (Neutral) + Apyest	883	0.687	0.655	0.719	0.0142	64%
Mesh Coordinates	5,206	0.663	0.650	0.676	0.0058	62%
Mesh Coordinates + Apyest	5,206	0.705	0.687	0.723	0.0080	65%
Mesh Coordinates (Happy)	3,622	0.663	0.643	0.683	0.0088	61%
Mesh Coordinates (Happy) + Apyest	3,622	0.703	0.680	0.726	0.0100	64%
Mesh Coordinates (Neutral)	877	0.656	0.601	0.670	0.0152	60%
Mesh Coordinates (Neutral) + Apyest	877	0.665	0.640	0.690	0.0111	62%

Table 17

Model Metrics – White Females – Immigration – All Images

	White Females Immigration					
	n (of smaller group)	Area Under Curve (AUC)	CI Lower	CI Upper	σM	Accuracy
Age	7,782	0.656	0.643	0.670	0.0059	62%
Pitch/Yaw	7,782	0.513	0.505	0.522	0.0038	51%
Emotions	7,782	0.533	0.523	0.544	0.0048	53%
Sex Typicality	7,782	0.505	0.491	0.519	0.0063	51%
Combined Apyest Model	7,782	0.660	0.647	0.674	0.0060	62%
Features - Whole Image	7,782	0.760	0.745	0.776	0.0067	70%
Features (WI) + Apyest	7,782	0.769	0.753	0.786	0.0073	70%
Features - Masked Image	7,782	0.761	0.743	0.779	0.0080	70%
Features (MI) + Apyest	7,782	0.769	0.750	0.788	0.0084	70%
Point Coordinates	7,782	0.721	0.700	0.742	0.0093	66%
Point Coordinates + Apyest	7,782	0.736	0.712	0.761	0.0109	68%
Point Coordinates - No Mouth	7,782	0.703	0.681	0.725	0.0096	65%
Point Coordinates - No Mouth + Apyest	7,782	0.727	0.701	0.752	0.0113	67%
Point Coordinates (Happy)	5,799	0.726	0.709	0.743	0.0076	67%
Point Coordinates (Happy) + Apyest	5,799	0.744	0.725	0.764	0.0086	69%
Point Coordinates (Neutral)	877	0.684	0.641	0.728	0.0192	64%
Point Coordinates (Neutral) + Apyest	877	0.700	0.658	0.742	0.0185	65%
Mesh Coordinates	7,809	0.643	0.631	0.655	0.0054	61%
Mesh Coordinates + Apyest	7,809	0.700	0.682	0.717	0.0079	65%
Mesh Coordinates (Happy)	5,826	0.653	0.638	0.668	0.0067	61%
Mesh Coordinates (Happy) + Apyest	5,826	0.706	0.687	0.724	0.0082	65%
Mesh Coordinates (Neutral)	914	0.620	0.575	0.666	0.0202	58%
Mesh Coordinates (Neutral) + Apyest	914	0.644	0.602	0.687	0.0188	61%

Table 18

Model Metrics – White Females – Immigration – Reduced Images

	White Females Immigration - Reduced					
	n (of smaller group)	Area Under Curve (AUC)	CI Lower	CI Upper	σM	Accuracy
Age	3,808	0.684	0.666	0.701	0.0079	64%
Pitch/Yaw	3,808	0.509	0.497	0.520	0.0051	51%
Emotions	3,808	0.535	0.512	0.558	0.0102	53%
Sex Typicality	3,808	0.506	0.485	0.527	0.0093	50%
Combined Apyest Model	3,808	0.684	0.668	0.700	0.0071	64%
Features - Whole Image	3,808	0.763	0.746	0.780	0.0075	70%
Features (WI) + Apyest	3,808	0.772	0.753	0.792	0.0086	71%
Features - Masked Image	3,808	0.762	0.748	0.775	0.0061	70%
Features (MI) + Apyest	3,808	0.769	0.755	0.783	0.0063	71%
Point Coordinates	3,808	0.727	0.702	0.752	0.0110	67%
Point Coordinates + Apyest	3,808	0.750	0.721	0.778	0.0125	69%
Point Coordinates - No Mouth	3,808	0.712	0.684	0.739	0.0122	66%
Point Coordinates - No Mouth + Apyest	3,808	0.740	0.711	0.769	0.0128	68%
Point Coordinates (Happy)	2,949	0.732	0.710	0.754	0.0097	68%
Point Coordinates (Happy) + Apyest	2,949	0.757	0.733	0.781	0.0106	70%
Point Coordinates (Neutral)	383	0.669	0.633	0.704	0.0157	62%
Point Coordinates (Neutral) + Apyest	383	0.691	0.660	0.721	0.0135	64%
Mesh Coordinates	3,826	0.647	0.631	0.664	0.0072	61%
Mesh Coordinates + Apyest	3,826	0.713	0.693	0.733	0.0088	66%
Mesh Coordinates (Happy)	2,968	0.656	0.632	0.679	0.0104	62%
Mesh Coordinates (Happy) + Apyest	2,968	0.715	0.692	0.734	0.0103	66%
Mesh Coordinates (Neutral)	399	0.612	0.567	0.657	0.0198	60%
Mesh Coordinates (Neutral) + Apyest	399	0.629	0.584	0.675	0.0200	59%

Table 19

Model Metrics – Asian Males – Gun – All Images

	Asian Males Gun					
	n (of smaller group)	Area Under Curve (AUC)	CI Lower	CI Upper	σM	Accuracy
Age	3,010	0.524	0.501	0.547	0.0102	52%
Pitch/Yaw	3,010	0.555	0.542	0.567	0.0054	54%
Emotions	3,010	0.579	0.568	0.591	0.0052	56%
Sex Typicality	3,010	0.645	0.625	0.664	0.0088	59%
Combined Apyest Model	3,010	0.637	0.618	0.657	0.0084	60%
Features - Whole Image	3,010	0.707	0.688	0.723	0.0083	65%
Features (WI) + Apyest	3,010	0.709	0.688	0.730	0.0093	65%
Features - Masked Image	3,010	0.688	0.678	0.697	0.0042	64%
Features (MI) + Apyest	3,010	0.694	0.684	0.704	0.0044	64%
Point Coordinates	3,010	0.687	0.665	0.709	0.0097	64%
Point Coordinates + Apyest	3,010	0.697	0.674	0.720	0.0102	65%
Point Coordinates - No Mouth	3,010	0.679	0.656	0.702	0.0100	64%
Point Coordinates - No Mouth + Apyest	3,010	0.690	0.664	0.716	0.0113	65%
Point Coordinates (Happy)	1,626	0.694	0.662	0.725	0.0140	65%
Point Coordinates (Happy) + Apyest	1,626	0.706	0.670	0.742	0.0169	66%
Point Coordinates (Neutral)	528	0.631	0.591	0.670	0.0175	59%
Point Coordinates (Neutral) + Apyest	528	0.635	0.599	0.671	0.0159	60%
Mesh Coordinates	3,003	0.685	0.669	0.701	0.0070	63%
Mesh Coordinates + Apyest	3,003	0.695	0.678	0.712	0.0076	65%
Mesh Coordinates (Happy)	1,947	0.681	0.667	0.695	0.0062	63%
Mesh Coordinates (Happy) + Apyest	1,947	0.691	0.675	0.707	0.0071	63%
Mesh Coordinates (Neutral)	529	0.626	0.581	0.671	0.0198	59%
Mesh Coordinates (Neutral) + Apyest	529	0.640	0.597	0.682	0.0188	60%

Table 20

Model Metrics – Asian Males – Gun – Reduced Images

	Asian Males Gun - Reduced					
	n (of smaller group)	Area Under Curve (AUC)	CI Lower	CI Upper	σM	Accuracy
Age	1,950	0.532	0.518	0.547	0.0065	52%
Pitch/Yaw	1,950	0.535	0.511	0.558	0.0103	53%
Emotions	1,950	0.569	0.545	0.589	0.0090	55%
Sex Typicality	1,950	0.639	0.610	0.668	0.0128	58%
Combined Apyest Model	1,950	0.629	0.603	0.656	0.0117	60%
Features - Whole Image	1,950	0.696	0.681	0.711	0.0067	65%
Features (WI) + Apyest	1,950	0.698	0.684	0.712	0.0062	66%
Features - Masked Image	1,950	0.670	0.646	0.694	0.0106	62%
Features (MI) + Apyest	1,950	0.676	0.654	0.698	0.0099	63%
Point Coordinates	1,950	0.680	0.661	0.689	0.0082	63%
Point Coordinates + Apyest	1,950	0.693	0.672	0.715	0.0094	64%
Point Coordinates - No Mouth	1,950	0.672	0.651	0.693	0.0091	62%
Point Coordinates - No Mouth + Apyest	1,950	0.689	0.666	0.713	0.0104	64%
Point Coordinates (Happy)	1,201	0.689	0.673	0.705	0.0071	64%
Point Coordinates (Happy) + Apyest	1,201	0.704	0.683	0.724	0.0090	65%
Point Coordinates (Neutral)	323	0.625	0.593	0.657	0.0142	60%
Point Coordinates (Neutral) + Apyest	323	0.639	0.600	0.679	0.0174	62%
Mesh Coordinates	1,952	0.684	0.668	0.700	0.0071	64%
Mesh Coordinates + Apyest	1,952	0.698	0.680	0.717	0.0080	65%
Mesh Coordinates (Happy)	1,317	0.685	0.668	0.702	0.0076	64%
Mesh Coordinates (Happy) + Apyest	1,317	0.700	0.680	0.719	0.0087	65%
Mesh Coordinates (Neutral)	325	0.601	0.562	0.641	0.0176	58%
Mesh Coordinates (Neutral) + Apyest	325	0.613	0.578	0.648	0.0154	59%

Table 21

Model Metrics – Hispanic Males – Immigration – All Images

	Hispanic Males Immigration					
	n (of smaller group)	Area Under Curve (AUC)	CI Lower	CI Upper	σM	Accuracy
Age	2,991	0.644	0.607	0.681	0.0163	60%
Pitch/Yaw	2,991	0.535	0.516	0.555	0.0086	52%
Emotions	2,991	0.554	0.539	0.568	0.0065	55%
Sex Typicality	2,991	0.549	0.526	0.572	0.0100	54%
Combined Apyest Model	2,991	0.671	0.640	0.701	0.0134	62%
Features - Whole Image	2,991	0.711	0.684	0.738	0.0119	65%
Features (WI) + Apyest	2,991	0.718	0.690	0.747	0.0125	67%
Features - Masked Image	2,991	0.716	0.698	0.735	0.0082	66%
Features (MI) + Apyest	2,991	0.722	0.703	0.741	0.0083	67%
Point Coordinates	2,991	0.696	0.676	0.715	0.0087	64%
Point Coordinates + Apyest	2,991	0.717	0.695	0.738	0.0095	66%
Point Coordinates - No Mouth	2,991	0.673	0.652	0.694	0.0091	63%
Point Coordinates - No Mouth + Apyest	2,991	0.702	0.678	0.726	0.0106	64%
Point Coordinates (Happy)	1,696	0.724	0.705	0.743	0.0085	66%
Point Coordinates (Happy) + Apyest	1,696	0.749	0.729	0.769	0.0089	69%
Point Coordinates (Neutral)	664	0.613	0.567	0.658	0.0200	58%
Point Coordinates (Neutral) + Apyest	664	0.637	0.608	0.667	0.0130	61%
Mesh Coordinates	2,995	0.650	0.635	0.666	0.0068	61%
Mesh Coordinates + Apyest	2,995	0.695	0.677	0.713	0.0079	64%
Mesh Coordinates (Happy)	1,701	0.658	0.642	0.675	0.0075	62%
Mesh Coordinates (Happy) + Apyest	1,701	0.716	0.693	0.734	0.0101	66%
Mesh Coordinates (Neutral)	695	0.594	0.562	0.625	0.0140	57%
Mesh Coordinates (Neutral) + Apyest	695	0.619	0.528	0.657	0.0164	59%

Table 22

Model Metrics – Hispanic Males – Immigration – Reduced Images

	Hispanic Males Immigration - Reduced					
	n (of smaller group)	Area Under Curve (AUC)	CI Lower	CI Upper	σM	Accuracy
Age	2,004	0.643	0.617	0.670	0.0116	60%
Pitch/Yaw	2,004	0.509	0.489	0.529	0.0088	50%
Emotions	2,004	0.543	0.526	0.560	0.0075	53%
Sex Typicality	2,004	0.545	0.528	0.563	0.0077	54%
Combined Apyest Model	2,004	0.675	0.648	0.703	0.0121	63%
Features - Whole Image	2,004	0.707	0.685	0.728	0.0095	65%
Features (WI) + Apyest	2,004	0.714	0.692	0.736	0.0098	66%
Features - Masked Image	2,004	0.706	0.687	0.726	0.0087	65%
Features (MI) + Apyest	2,004	0.712	0.692	0.733	0.0089	66%
Point Coordinates	2,004	0.695	0.663	0.728	0.0144	64%
Point Coordinates + Apyest	2,004	0.716	0.684	0.749	0.0143	66%
Point Coordinates - No Mouth	2,004	0.670	0.643	0.697	0.0120	63%
Point Coordinates - No Mouth + Apyest	2,004	0.701	0.669	0.732	0.0138	65%
Point Coordinates (Happy)	1,225	0.714	0.690	0.739	0.0109	66%
Point Coordinates (Happy) + Apyest	1,225	0.743	0.715	0.772	0.0127	68%
Point Coordinates (Neutral)	437	0.594	0.553	0.636	0.0183	57%
Point Coordinates (Neutral) + Apyest	437	0.616	0.586	0.647	0.0135	59%
Mesh Coordinates	2,009	0.652	0.626	0.678	0.0114	61%
Mesh Coordinates + Apyest	2,009	0.698	0.669	0.726	0.0126	64%
Mesh Coordinates (Happy)	1,230	0.660	0.636	0.685	0.0107	61%
Mesh Coordinates (Happy) + Apyest	1,230	0.709	0.686	0.733	0.0104	65%
Mesh Coordinates (Neutral)	437	0.598	0.553	0.644	0.0202	58%
Mesh Coordinates (Neutral) + Apyest	437	0.614	0.557	0.671	0.0250	58%

Table 23

Model Metrics – Hispanic Males – Gun – All Images

	Hispanic Males Gun					
	n (of smaller group)	Area Under Curve (AUC)	CI Lower	CI Upper	σM	Accuracy
Age	2,921	0.502	0.486	0.518	0.0070	50%
Pitch/Yaw	2,921	0.568	0.552	0.584	0.0072	55%
Emotions	2,921	0.581	0.561	0.601	0.0089	55%
Sex Typicality	2,921	0.619	0.600	0.638	0.0083	58%
Combined Apyest Model	2,921	0.637	0.621	0.653	0.0071	59%
Features - Whole Image	2,921	0.680	0.659	0.700	0.0091	63%
Features (WI) + Apyest	2,921	0.686	0.666	0.706	0.0089	64%
Features - Masked Image	2,921	0.684	0.670	0.698	0.0061	64%
Features (MI) + Apyest	2,921	0.689	0.677	0.702	0.0055	64%
Point Coordinates	2,921	0.681	0.656	0.706	0.0111	63%
Point Coordinates + Apyest	2,921	0.691	0.668	0.713	0.0010	64%
Point Coordinates - No Mouth	2,921	0.675	0.650	0.701	0.0112	63%
Point Coordinates - No Mouth + Apyest	2,921	0.687	0.663	0.710	0.0102	64%
Point Coordinates (Happy)	1,728	0.681	0.659	0.703	0.0097	64%
Point Coordinates (Happy) + Apyest	1,728	0.689	0.665	0.713	0.0106	64%
Point Coordinates (Neutral)	523	0.670	0.645	0.696	0.0114	61%
Point Coordinates (Neutral) + Apyest	523	0.671	0.642	0.701	0.0131	64%
Mesh Coordinates	2,924	0.673	0.656	0.690	0.0075	62%
Mesh Coordinates + Apyest	2,924	0.682	0.665	0.700	0.0078	63%
Mesh Coordinates (Happy)	2,022	0.670	0.646	0.694	0.0107	62%
Mesh Coordinates (Happy) + Apyest	2,022	0.685	0.664	0.706	0.0094	64%
Mesh Coordinates (Neutral)	523	0.657	0.618	0.697	0.0174	61%
Mesh Coordinates (Neutral) + Apyest	523	0.666	0.624	0.708	0.0186	62%

Table 24

Model Metrics – Hispanic Males – Gun – Reduced Images

	Hispanic Males Gun - Reduced					
	n (of smaller group)	Area Under Curve (AUC)	CI Lower	CI Upper	σM	Accuracy
Age	1,901	0.501	0.472	0.530	0.0127	50%
Pitch/Yaw	1,901	0.555	0.545	0.566	0.0048	54%
Emotions	1,901	0.591	0.565	0.617	0.0116	56%
Sex Typicality	1,901	0.612	0.585	0.639	0.0121	58%
Combined Apyest Model	1,901	0.645	0.627	0.664	0.0082	60%
Features - Whole Image	1,901	0.667	0.642	0.692	0.0111	63%
Features (WI) + Apyest	1,901	0.671	0.646	0.696	0.0111	63%
Features - Masked Image	1,901	0.648	0.617	0.679	0.0138	62%
Features (MI) + Apyest	1,901	0.654	0.623	0.686	0.0139	62%
Point Coordinates	1,901	0.668	0.640	0.697	0.0125	63%
Point Coordinates + Apyest	1,901	0.685	0.660	0.710	0.0110	63%
Point Coordinates - No Mouth	1,901	0.658	0.633	0.684	0.0112	62%
Point Coordinates - No Mouth + Apyest	1,901	0.679	0.653	0.704	0.0112	63%
Point Coordinates (Happy)	1,249	0.676	0.657	0.695	0.0085	63%
Point Coordinates (Happy) + Apyest	1,249	0.685	0.665	0.705	0.0088	63%
Point Coordinates (Neutral)	318	0.645	0.596	0.701	0.0233	60%
Point Coordinates (Neutral) + Apyest	318	0.657	0.608	0.706	0.0217	60%
Mesh Coordinates	1,904	0.671	0.647	0.696	0.0108	62%
Mesh Coordinates + Apyest	1,904	0.685	0.662	0.709	0.0105	63%
Mesh Coordinates (Happy)	1,393	0.662	0.635	0.689	0.0120	62%
Mesh Coordinates (Happy) + Apyest	1,393	0.675	0.648	0.702	0.0119	63%
Mesh Coordinates (Neutral)	318	0.637	0.575	0.699	0.0274	61%
Mesh Coordinates (Neutral) + Apyest	318	0.639	0.581	0.697	0.0256	58%

Appendix O

Figure 70
Facial Quartile Points Plot – White Males – Gun

Figure 70a - All	Figure 70b - No Mouth

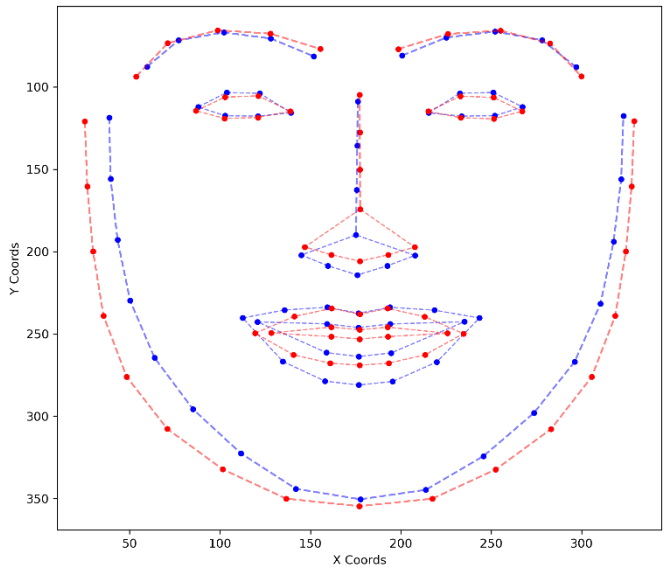
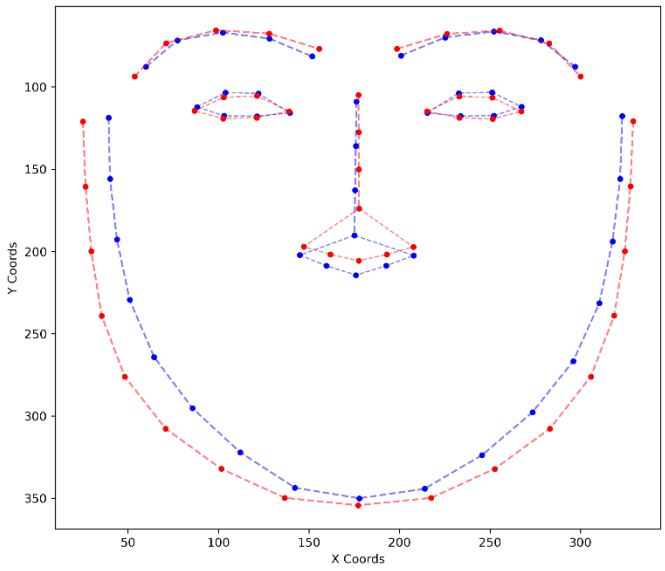
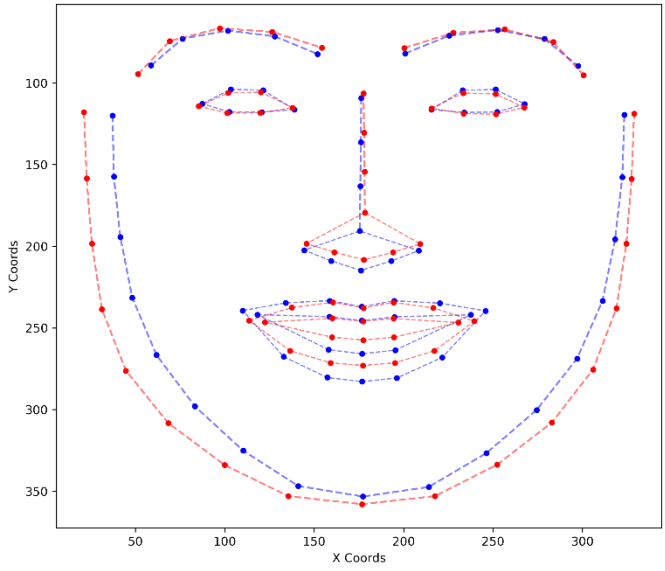
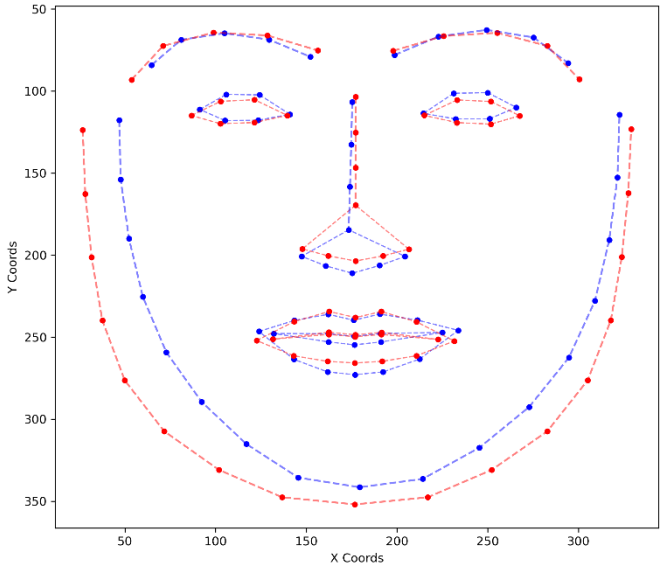

	

Figure 70c - Happy	**Figure 70d - Neutral**
	

Return to relevant section

Figure 71

Facial Quartile Points Plot – White Females – Gun

Figure 71a - All	Figure 71b - No Mouth

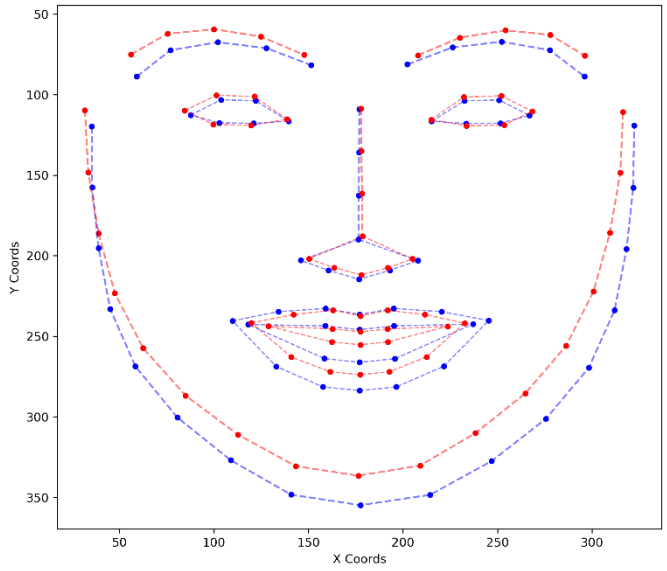
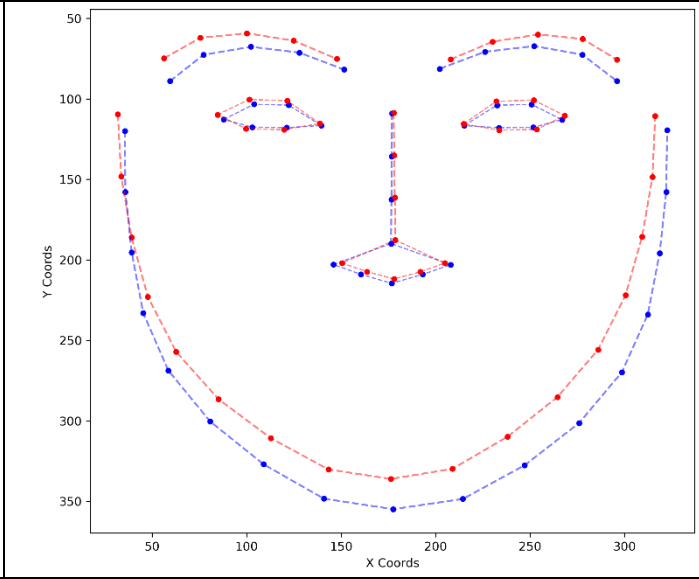

	

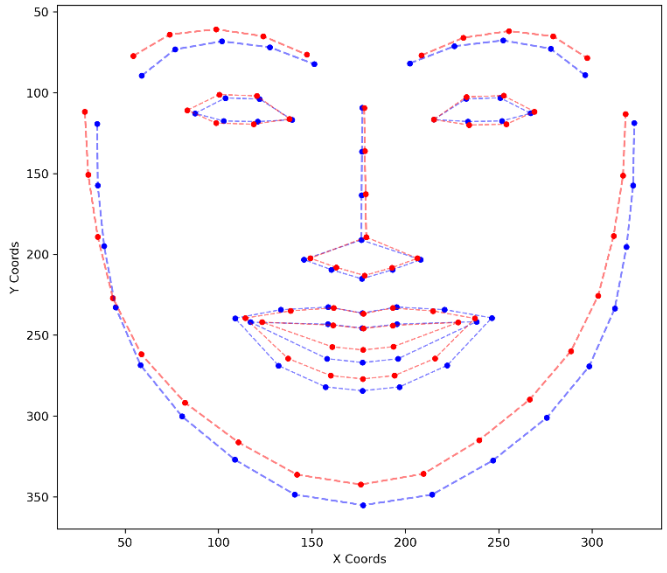
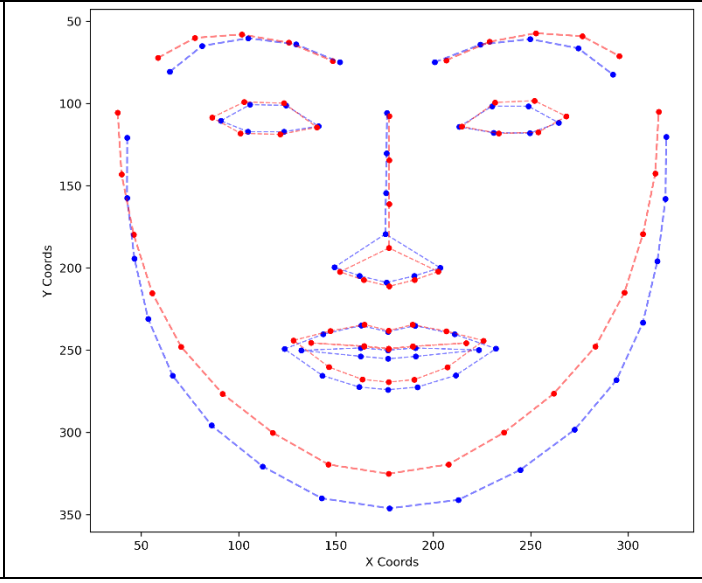
Figure 71c - Happy	**Figure 71d - Neutral**
	

Figure 72

Facial Quartile Points Plot – White Males – Immigration

Figure 72a - All	Figure 72b - No Mouth

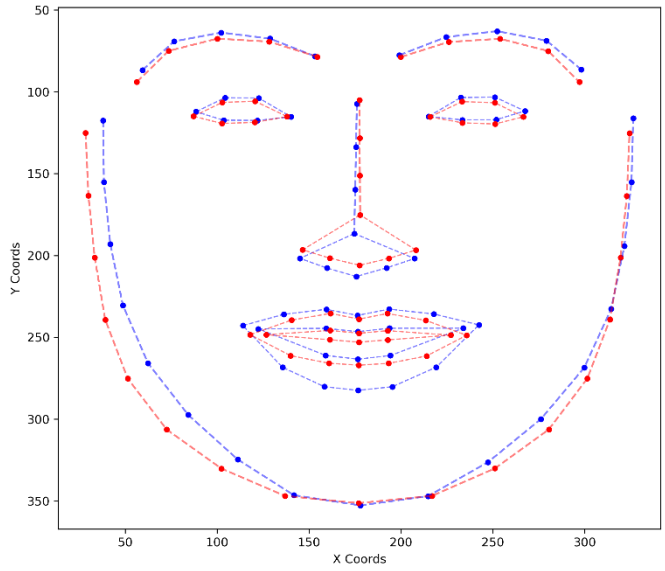
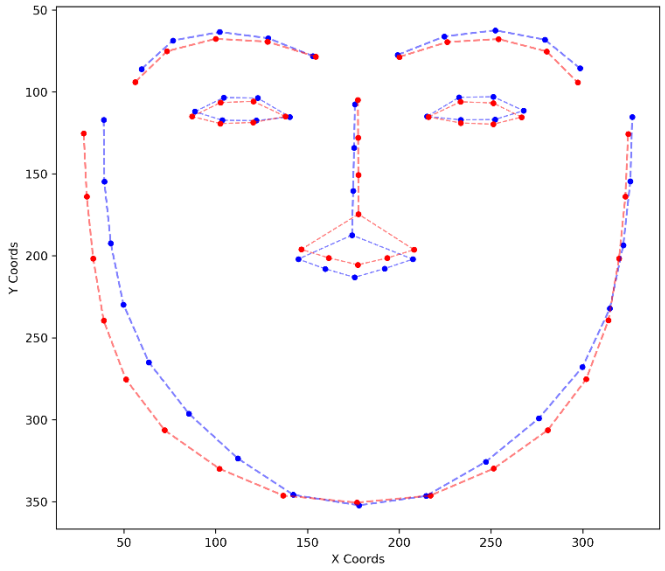

	

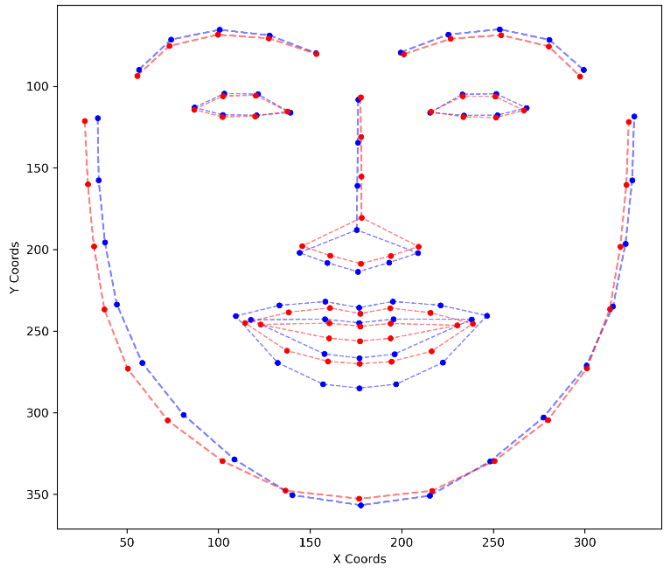
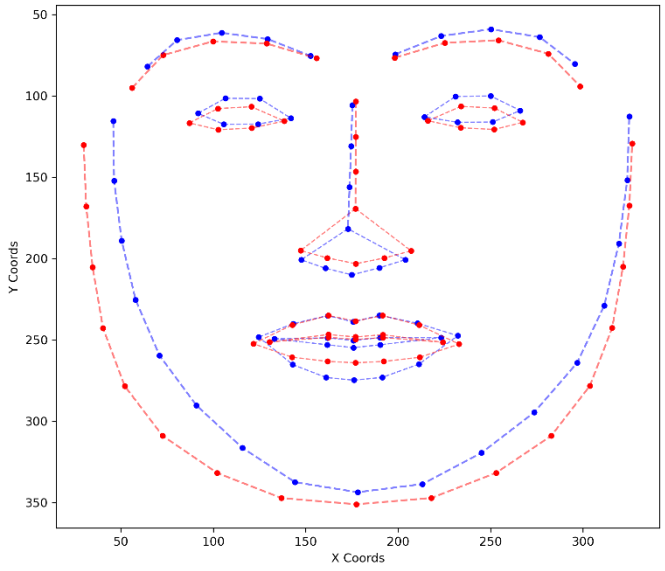
Figure 72c - Happy	**Figure 72d - Neutral**
	

Figure 73

Facial Quartile Points Plot – White Females – Immigration

Figure 73a - All	Figure 73b - No Mouth

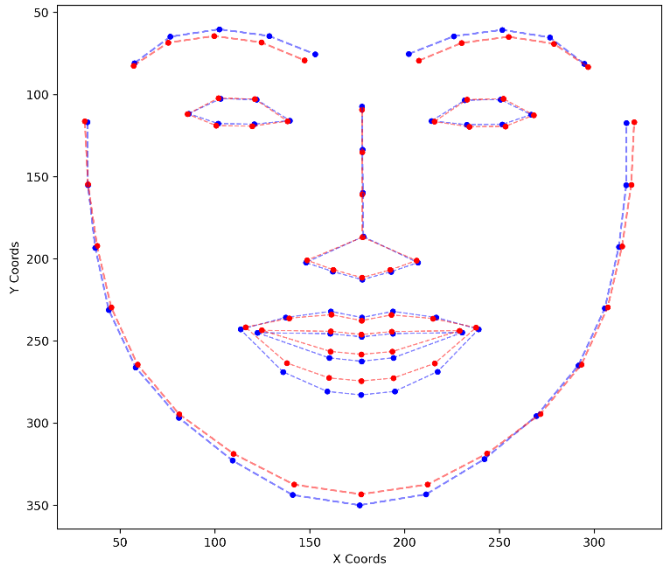
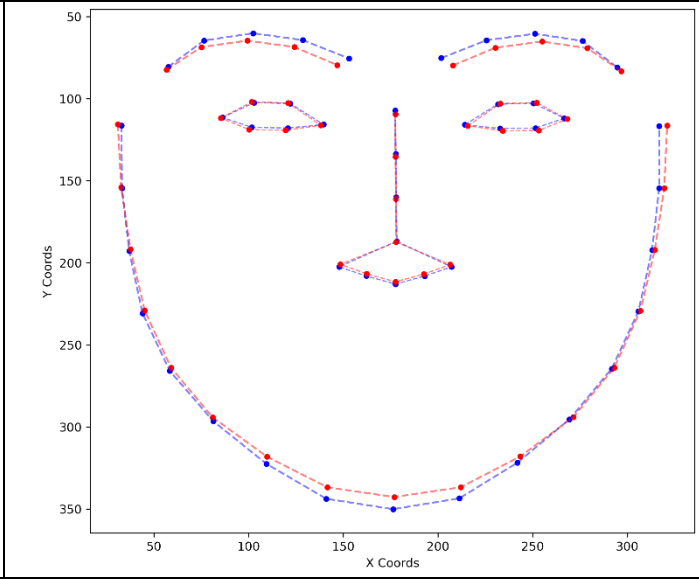

	

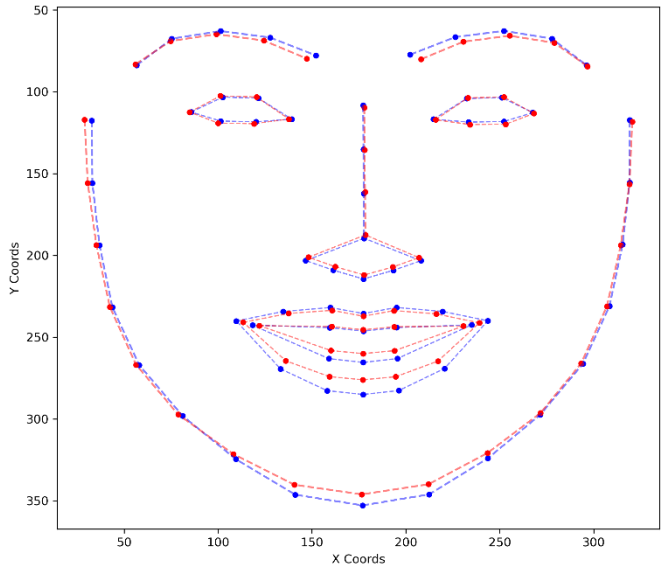
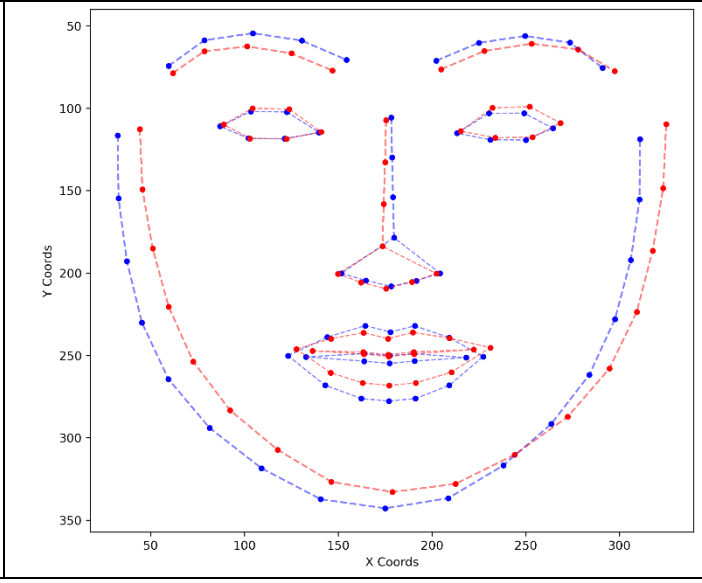
Figure 73c - Happy	**Figure 73d - Neutral**
	

Figure 74

Facial Quartile Points Plot – Asian Males – Gun

Figure 74a - All	Figure 74b - No Mouth

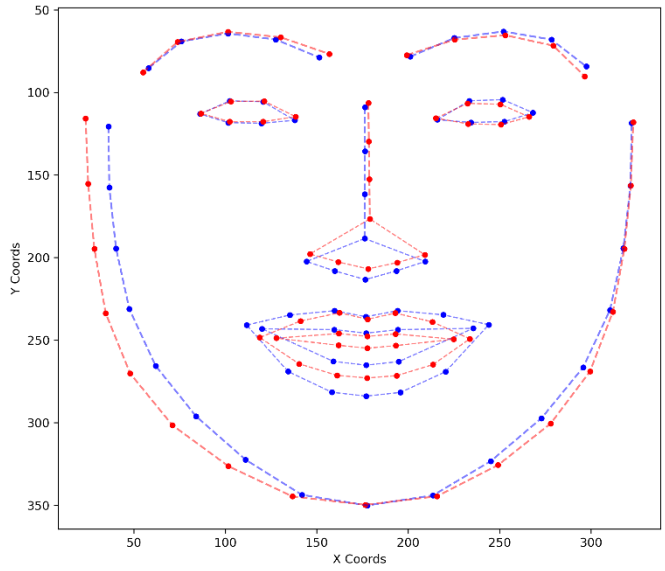
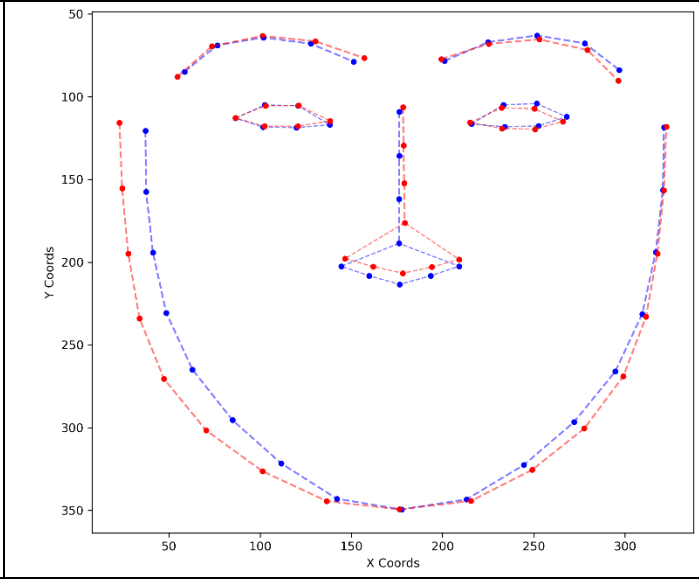

	

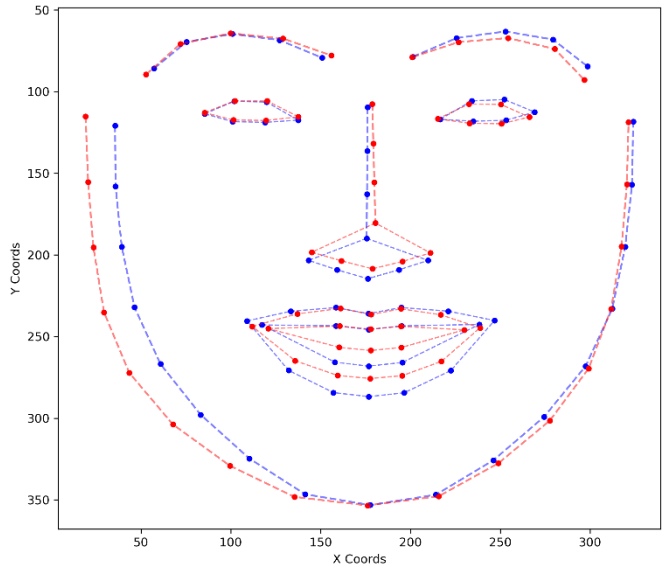
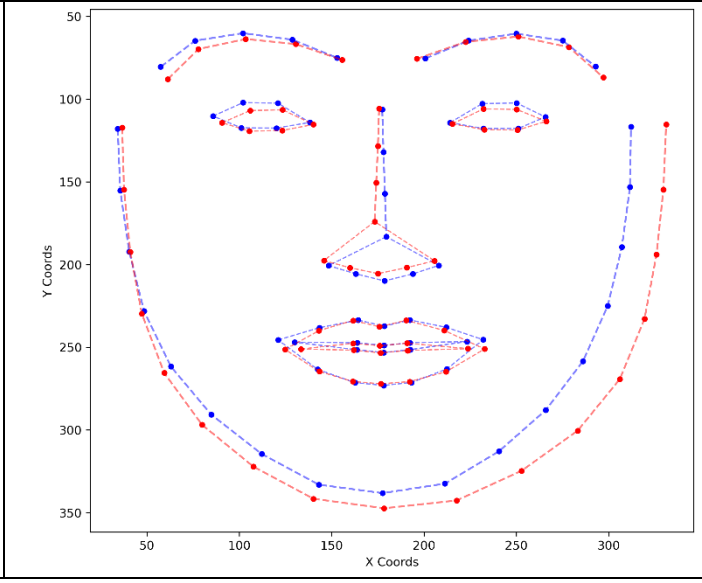
Figure 74c - Happy	**Figure 74d - Neutral**
	

Figure 75

Facial Quartile Points Plot – Hispanic Males – Immigration

Figure 75a - All	Figure 75b - No Mouth

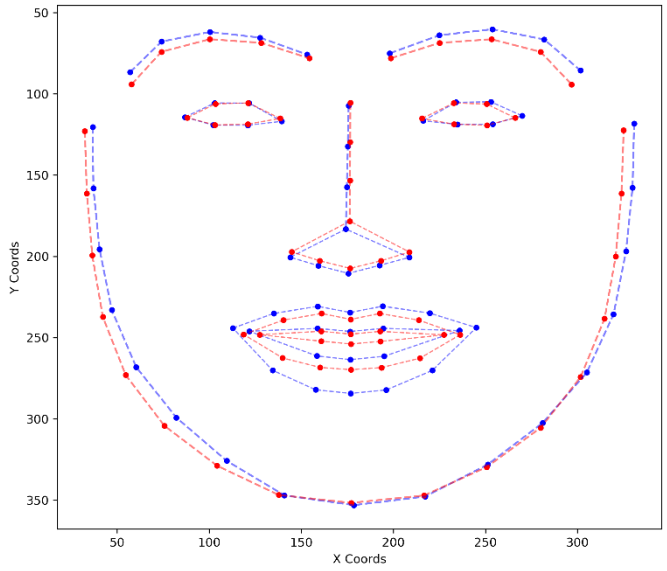
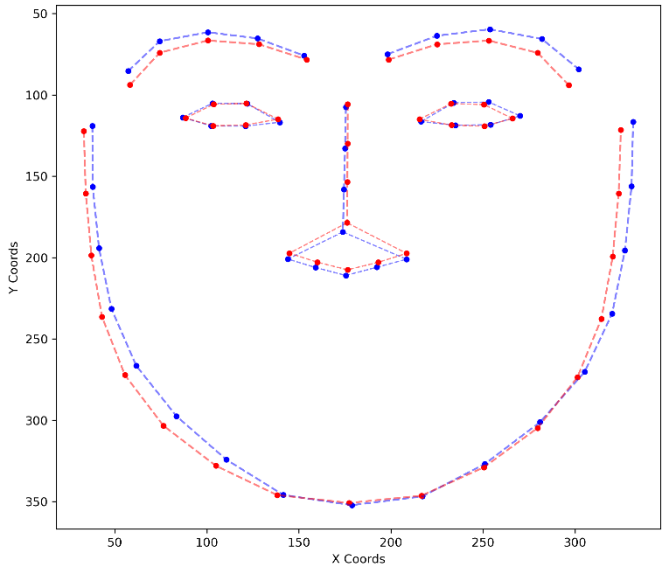

	

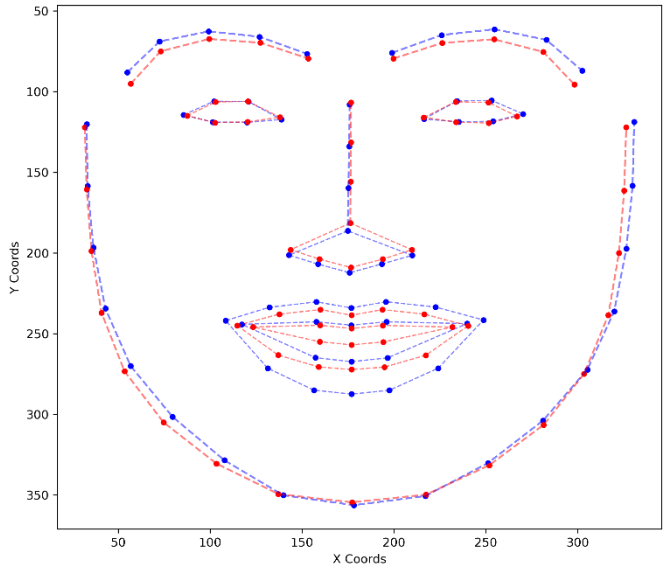
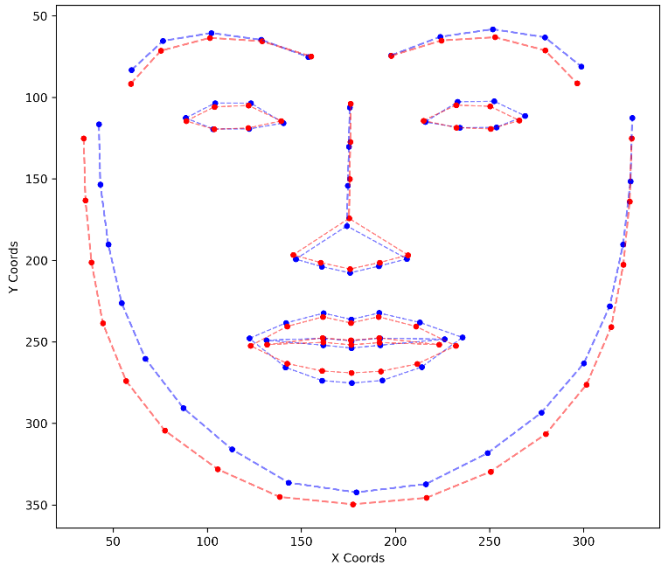
Figure 75c - Happy	**Figure 75d - Neutral**
	

Figure 76

Facial Quartile Points Plot – Hispanic Males – Gun

Figure 76a - All	Figure 76b - No Mouth

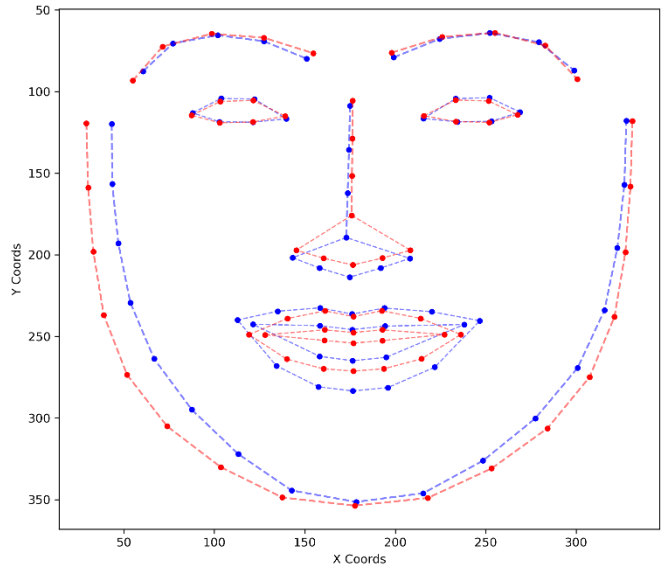
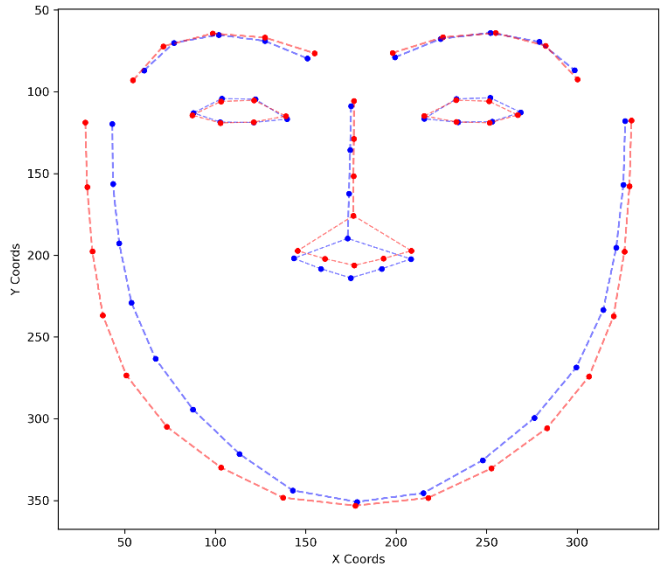

	

Figure 76c - Happy	**Figure 76d - Neutral**