Journal of Research High School

Vol. 2023 (2) 226 – 234
ISSN 2688-3651 [online]

Predicting Election Outcomes from Facial Images of Candidates Using an Unbiased Machine Learning Model

Raymond (Deming) Lin,[1] * Jadelyn Tran[1]

[1]Webber Academy, Calgary, AB, Canada
*Corresponding Author: raymond.lin.336@gmail.com

Advisor: Tyler Giallanza, tylerg@princeton.edu

Received January 5, 2023; Revised July 4, 2023; Accepted, July 24, 2023

Abstract

In the past, studies on psychology have shown that humans can create instantaneous judgments of a stranger's personality and characteristics, just based on a picture of their face. The most successful study reached a 72.4% accuracy in predicting election outcomes. There had been machine learning studies that tried to replicate this success, but some kind of human input and therefore bias were often present. This project aimed to create a bias-free and independent machine learning model that only uses the image of political candidates to predict their success. With no other information than a candidate's face, the model achieved a 70.43% accuracy predicting election results. Not only did the different approaches in this experiment give a quantitative way to compare different types of human thinking, but it can also be used as a benchmark for future research that further investigates the relationship between facial traits, human judgments, and machine learning.

Keywords: Election prediction, Machine learning, Classifier, Transfer learning, Facial images, Physiognomy

1. Introduction

Once every few years, the topic of elections floods the media. Typically, voters always try to judge candidates for who they are and how their platform would benefit them. However, no matter how hard individuals try to combat their varying levels of prejudice, electoral decisions are still heavily influenced by the appearance of the candidates.

Studies had shown that individuals can infer various traits about a stranger just from their face (Olivola & Todorov, 2010; Olivola & Todorov, 2009; Todorov, 2018), and a lot of that subconscious interpretation of information heavily affects electoral decision making (Ballew & Todorov, 2007). There had been numerous studies (Olivola & Todorov, 2010; Olivola & Todorov, 2009; Todorov, 2018) that tried to test how well personality and other traits can be detected by humans by just looking at faces. In addition to simplistic traits like attractiveness and emotions, short exposures to pictures of strangers can lead to predictions of more complex features such as level of social class (Todorov, 2018) and political competence (Ballew & Todorov, 2007).

For example, a study of facial judgements and election outcomes was conducted by Alexander Todorov and Charles C. Ballew (2007). They presented sixty-four test subjects with pairs of senators who were running against each other and asked the subjects to predict who was more likely to win the election. Test subjects were only allowed to judge on the candidates that they had no familiarity with, and they successfully predicted 72.4% of the Senate races in 2006. Interestingly, when Todorov and Ballew (2007) asked participants to "deliberate and make a good judgment", their predictive accuracy decreased. Participants were also shown to perform better when constrained to 250 milliseconds and thus forced to make a rapid, unreflective judgment. Not only was this study a demonstration of how much physical appearance influences electoral results, it also suggests that the more accurate judgments are often subconscious, as the more the participants consciously tried to analyze, the worse their results were.

Inspired by the positive results of the human experiments, researchers tried to replicate these results by building machine learning models to predict election outcomes from images of faces (Joo et al., 2015; Todorov et al., 2005; Ventura et al., n.d.). Many of these models had near-human accuracy. For example, an extensive study (Joo et al., 2015) was able to reach a 67.9% accuracy in predicting the US governor races by combining classic computer vision methods with a Support Vector Machine (SVM). Specifically, the researchers trained the SVM on two types of data: personality/demographic characteristics, and physical attributes. To determine personality/demographic characteristics, the researchers asked human test subjects to evaluate the comparative characteristics of two candidates at a time (e.g., which candidate looks wealthier, which candidate looks older), and compiled those results into a series of perceived characteristics. To determine the physical attributes, they also split each of the training images into regions and extracted physical attributes, such as if they were smiling or if they were wearing glasses.

Despite the impressive accuracy of this method, some degree of further work was needed to fully explore this problem. For example, the researchers themselves named the categories of attributes that they want to extract, which could have led to certain physical features left ignored or under-represented. They also did not mention the logic behind determining the categories of perceived characteristics that they want to evaluate, and the logic of why those attributes were the most significant. Furthermore, a significant portion of the input data used for the training came from human participants to begin with; the model likely would not have achieved such a high accuracy without human contribution. In summary, previous experiments (Joo et al., 2015) likely relied on assumptions that could have introduced biases into how the algorithm could have made its prediction.

This project aimed to create a series of machine learning models that would provide a deeper understanding of both how artificial intelligence can be used to emulate subconscious human analysis during elections, and how AI can be used as an additional tool for election prediction and as a new perspective into subconscious human psychology. The following is the approach:

1. To create a new machine learning model that can generate a prediction of the relative electoral success of a politician by only using an image of their face. Unlike previous studies, this new model would use no human test subjects to provide the input data, and to not manually label certain features of the face for the model to focus on. By creating a machine learning model that does not rely on any form of human analysis, this could eliminate the bias that previous studies in this field could have introduced.
2. To create two different approaches to solve this problem, with both machine learning models having the same lack of reliance on human test subjects. The first approach purely uses the pixels of the images of candidates as the input to the machine learning model. The other approach simplifies the image down to facial characteristics perceived by a machine learning model, and to use that to generate a prediction. These two approaches could be interpreted as representations of two different ways of human thinking when asked to predict stranger's traits from their faces. These could act as a quantifiable comparison between human's conscious and subconscious judgments. Testing two different approaches would also provide an opportunity to try different methodologies to search for a higher accuracy.
3. To see how much facial appearance influenced election outcomes compared to other electoral factors that are commonly known to influence election results, such as the incumbency status and the amount of the campaign budget for the candidate. Models that predict electoral success of politicians based on these other factors will also be developed. By comparing those models to the model that analyzes facial appearance, it could lead to a better understanding of how much of a role does facial appearance of a politician plays compared to these other more well-established factors.

2. Materials and Methods

2.1 Problem Formulation

The aim of this project was to train a machine learning model to predict the outcome of an election from a facial image of a candidate. Most previous studies on this subject all paired politicians during their experimental or data collection process, training the model to predict which of the two candidates will win the election. This pairing was often either derived from the two candidates running against each other (Ballew & Todorov, 2007), or the two candidates with a similar perceived age (Joo et al., 2015). This logic might have seemed sound, as one would think that to evaluate the competence of a politician, there had to be a relative comparison. However, this new experiment was designed to challenge that preconception by not using any pairing in the process. All the data that were used were individually selected from an array and individually evaluated, with “winner” or “loser” as the only label. This dramatically reduce the amount of data needed to train the model due to the quadratic nature of pairwise predictions.

2.2 Data Collection

Political Candidates Dataset

Specifically for this experiment, a brand-new dataset for the facial images of politicians was created. We found public datasets with names of the winning and losing politicians from the US House of Representatives elections from 2000 to 2020 (total of 9 elections) and Canadian House of Commons elections from 2000 to 2021 (total of 8 elections). In total, there were around 18,000 candidates from the Canadian Parliament and 6,880 candidates from the US House of Representatives. This data formed a list of names for the politicians, and the information about what elections they won and lost also was used for labeling the dataset.

After the names of the candidates were found, BeautifulSoup was used to extract the HTML information from either google searches, specifically the knowledge panel, or from Wikipedia. 7068 facial images were found, one per candidate, per election.

Out of those images, an automatic filter via Deepface (Serengil, 2022) was ran to detect the face of the person in the images. This is important as Deepface is used later again to crop out parts of the image and only leave the face of the candidate. The images that did not have a detectable face were discarded, and 5,511 images remained. A round of manual filtering was also conducted, eliminating images that, for example, had unusual lightning or multiple faces; 4,501 images remained. For the candidates in this dataset, their images were standardized as much as possible by using an official campaign photo, where most candidates are smiling, facing the camera directly, and are in good lightning conditions. For many of the politicians that were removed, a different, usable photo were often manually found, increasing this dataset to a total of 4,859 images. Among these images, there were candidates that lost/won multiple times, and only one of those images were kept in the dataset. Candidates that have both won and lost had their images eliminated entirely.

The number of “winners” and “losers” had to be the same while training the program. During training, 2,018 unique images of candidates (1 image per candidate) were used, 1,009 for each category, with some “winner” images left unused (randomly selected during each train/test split). Using Deepface, these images were converted to grayscale and cropped to 48 by 48 pixels centered in their face, with the background and clothing cropped out.

Incumbency and Campaign Financing Data

Although the primary focus of this study is on facial images, this project also sought to explore how much incumbency and campaign financing information influences the predictions made by the models. BeautifulSoup was used to collect data from Ballotpedia, and using the information found by web scraping the HTML code, the information of 609 US candidates were recorded. The candidates that had this information were trained again (later fully explained under Models - Method 3), this time with incumbency and financial spending as additional input variables.

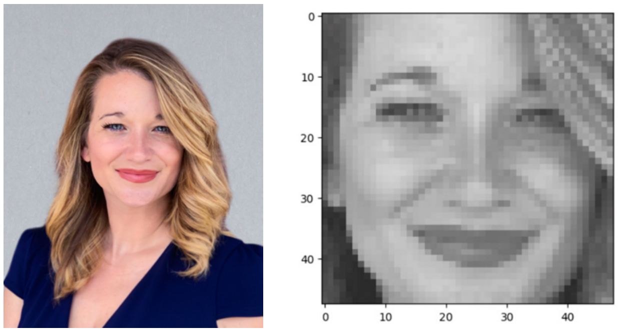

Figure 1a & 1b. This is an example of a candidate in the “Loser” category (Aliscia Andrews, Virginia 10th District). These are the original image that was collected and the processed image used for training, respectively.

Examples

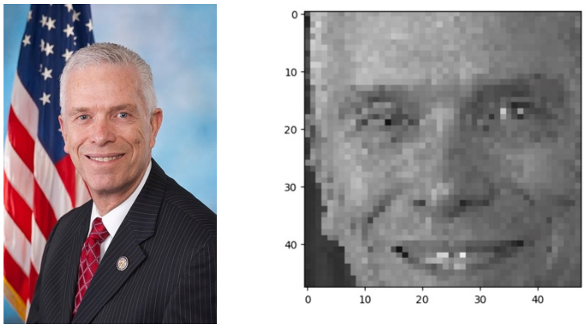

Figure 2a & 2b. This is an example of a candidate in the “Winner” category (Bill Johnson, Ohio 6th District). These are the original image that was collected and the processed image used for training, respectively.

2.3. Models

To fully understand this problem, four methods were designed. The method of input and the training process were the manipulated variables in order to obtain a more comprehensive suite of results. The four methods utilized a classifier for images, a classifier for images with an autoencoder, a logistic regression with Deepface.analyze (Serengil, 2022), and a classifier with Deepface.represent respectively. All coding was developed with Python and Google Colab (“Google”, 2017). Specifically, PyTorch (“PyTorch”, n.d.) was used for all of the programming with the machine learning models. Deepface (Serengil, 2022), a python library with extensive functions for facial recognition and image analysis, was used as well.

Method 1 – Pixel-Based Classifier

Using PyTorch (“PyTorch”, n.d.), a neural network was developed. It had linear layers and it classified the given inputs into either the “winner” or the “loser” category. This classifier was trained only using normalized arrays of the 48x48 pixel, gray scale images, with everything but the politician’s face cropped out of the photo. The model was trained with BCELoss as the loss function and a learning rate of 0.0001, until it reached a point of diminishing returns (around 750-1000 epochs).

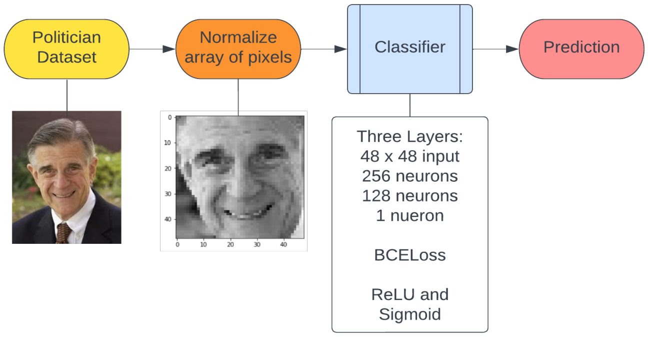

Figure 3. This is the general structure of Method 1. The classifier uses pixels of facial images to determine a result.

In the classifier, there are three linear layers in total. The first layer accepts 2,304 inputs (48 by 48 pixels), and has 256 outputs. The second layer decreases that to 128 outputs, and the third layer decreases it to 1 (a binary unit indicating if the candidate is a winner or a loser). ReLU was the activation function of the first two layers, and Sigmoid for the last layer.

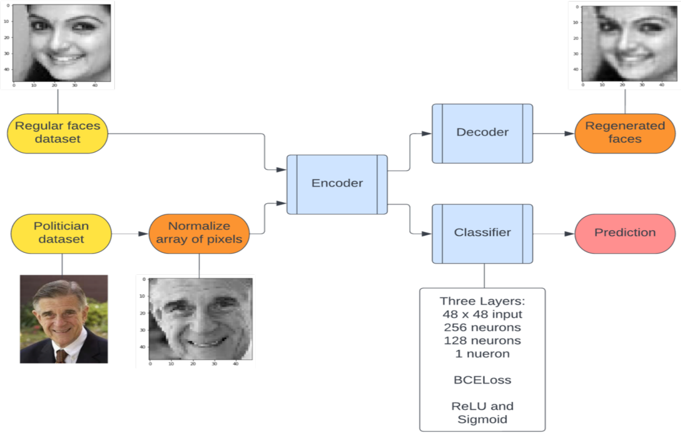

Figure 4. This is the general structure of Method 2a. The difference between Method 1 and 2a was that Method 2a used an autoencoder to implement transfer learning, and its encoder was used with the classifier to improve its feature extraction abilities.

Method 2a – Pixel-Based Classifier with Autoencoder

Due to the limited size of the training dataset, Method 2a sought to improve the performance of the model by implementing transfer learning to the classifier with the use of an autoencoder. First, an autoencoder was created to be trained on an independent dataset (Arora, 2020) of regular faces, with the autoencoder trained to compress and then reconstruct the facial images with as much detail as possible. By doing so, the autoencoder learned what part of a human face is essential for reconstructing the facial image, while discarding other irreverent details. Then, the autoencoder was used with the classifier in Method 1, to hopefully increase the efficiency of the model. This was because the autoencoder could assist the classifier to only focus on parts of the face that are of significance, narrowing down the facial information that influences election results.

This autoencoder is a Convolutional Neural Network that has three Conv2d layers for encoding, and three ConvTranpose2d layers for decoding. The activation function was three ReLU functions for encoding, and two ReLU and one Sigmoid function for decoding. The weights of the autoencoder were saved and then applied to the classifier from Method 1.

Method 2b – Pixel-Based Classifier with Deepface Representations

To implement another method of transfer learning, a pre-trained model was used to extract features from the candidates' images. The returned values of Deepface.represent (Serengil, 2022) are the input variable for this method. Deepface.represent was meant to be used for facial recognition by representing faces with vector embeddings. These embeddings were then used as the input for a classifier. This classifier is similar to the classifier used in Method 1 and 2a; the difference is that since the input variables were changed, the expected input of the first layer was changed as well.

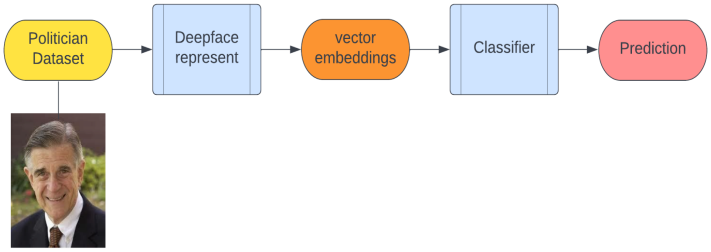

Figure 5. This is the general structure of Method 2b. Just like Method 2a, this was another implementation of transfer learning using Deepface represent.

Method 3 – Demographic-Informed Classifier

As a comparison to the prior 2 methods, which predicted election outcomes from facial images directly, for this method a model was trained on demographic/facial expression information. The Python library Deepface, and specifically the Deepface.analyze function was used, to generate these features. The function received the images of the politicians, and it returned a series of perceived values of the politicians, such as their age, composition of race, likelihood of certain emotions, gender, etc. These values were used as input variables for a Logistical Regression model from SkLearn ("Learn: Machine", n.d.). This model was also retrained with the incumbency status and the amount spent in campaign financing as additional input variables.

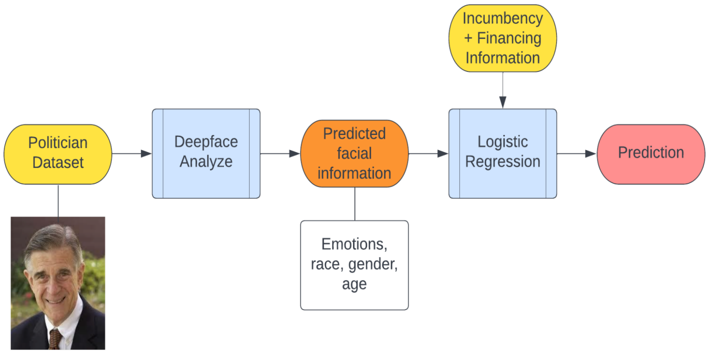
Incumbency + Financing Information
Politician Dataset
Deepface Analyze
Predicted facial information
Logistic Regression
Prediction
Emotions, race, gender, age

Figure 6: This is the general structure of Method 3. The perceived traits from Deepface.analyze were used as input for a logistic regression model that predicted a result

Deepface contained pre-trained models that can analyze attributes of facial images. Deepface.analyze is a function that can return multiple attributes from a facial image. All of the returned information were utilized, which included likelihood of certain emotions (angry, neutral, happy, etc.), likelihood of race (White, Latino, Asian…), age and gender.

3. Results

Evaluations of the accuracy of the model with both the training and testing datasets were recorded. The Receiver Operating Characteristics curves (ROC curves) was plotted and the Area Under the Curve (AUC) was calculated. The following are the results.

Table 1: Accuracy and AUC value of respective methods

	Method 1	Method 2a	Method 2b	Method 3
Training Accuracy (%)	94.20	95.36	61.16	60.45
Testing Accuracy (%)	70.43	66.96	16.57	59.95
AUC	0.77	0.74	0.37	0.61

3.1 Results for Method 1, 2a, 2b

ROC curves are used here to display the performance of the models as it is a common way to evaluate a binary machine learning classifier. True positive rate (y-axis) is the proportion of positive cases that are correctly evaluated, while the false positive rate (FPR) is the proportion of positive cases that are incorrectly evaluated. The ROC curve is used for its ability to display the effectiveness of the model at a range of thresholds, or cutoff values that determines if the output of the machine learning model should be classified as positive or negative (in this case, Winner or Loser). The AUC value, the area under the blue line, is a simple numerical representation of the information of the ROC curve. It is a more wholistic evaluation compared to an accuracy percentage as it represents the effectiveness of the model in a variety of thresholds.

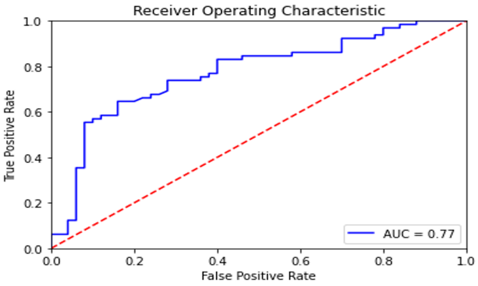

Figure 7: ROC graph for Method 1

The worst performance for a model would be the dotted red line, with an AUC value of 0.5, which would indicate that the model would have the same proportion of correct and incorrect predictions, or around just a 50% accuracy. The most ideal ROC curve would have the blue line closest to the top left of the graph, or an AUC value close to 1, indicating that at whatever threshold, the model would be able to maintain a high true positive rate, or a high rate of a correct positive evaluation.

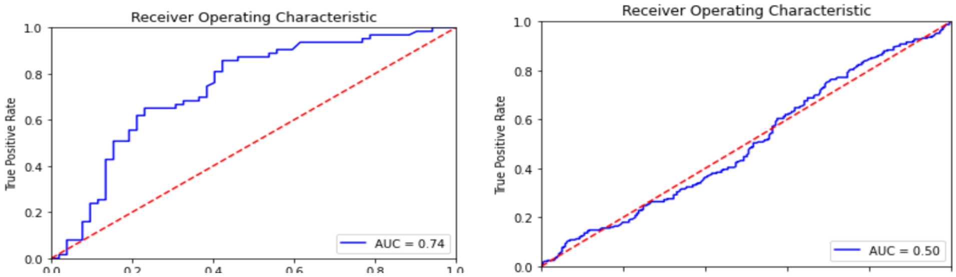

Figure 8a & 8b: ROC graph for Method 2a and Method 2b respectively

Table 2: Confusion Matrix for Method 1

	Actually Winners	Actually Losers
Predicted Winners	44	6
Predicted Losers	28	37

Table 3: Confusion Matrix for Method 2a

	Actually Winners	Actually Losers
Predicted Winners	36	16
Predicted Losers	22	41

From these results, it's clear that Method 1, with its 70.43% accuracy was the best approach. Method 2a was extremely close, with a 3.47% worse testing accuracy. It correctly predicted 4 candidates less than Method 1. However, Method 2a provided a better accuracy for predicting Losers. Method 2b, however, was very unsuccessful. It only achieved a 16% testing accuracy, which was worse than simply guessing with a 50% chance.

3.2 Results for Method 3

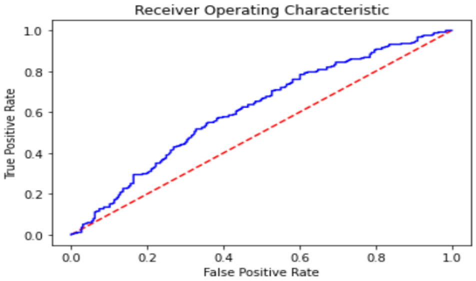

Figure 9: ROC curves for Method 3

Method 3 was relatively successful, achieving a 59.95% testing accuracy, with an AUC value of 0.61. It was not close to the success with Method 1, but its accuracy was significantly higher than 50%, the accuracy of making predictions with random guesses.

As stated before, Method 3 was retested by adding incumbency and amount spent on the campaign as additional parameters for the input of the linear regression model.

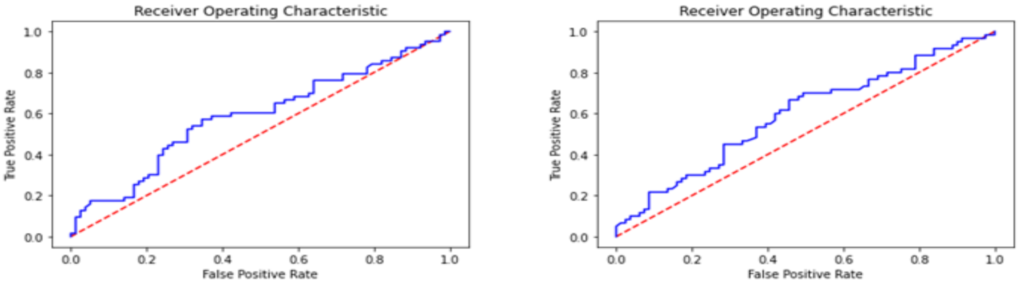

Figure 10A & 10B: ROC curves for variations of Method 3. These were results with the inclusion / exclusion of incumbency information. (330 politicians)

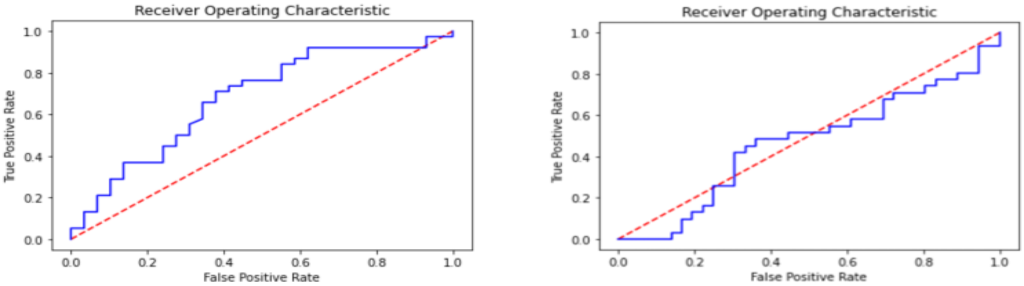

Figure 10C & 10D: ROC curves for variations of Method 3. These were results with the inclusion / exclusion of campaign financing information. (133 politicians)

The incumbency and financing information were only available for some candidates (330 for incumbency and 133 for financing). Figure 10B and 10D were baseline testing that had no information besides a candidate's face, while figure 10A and 10C were the results after incumbency / financial information is added to each group. Both with incumbency and financing, it demonstrated that the addition of this information greatly helped with the results in the respective groups.

The training with financing information as the only additional information resulted in the best ROC curve, with the highest AUC value out of the four (0.677). The training with incumbency information was close behind in those terms, but has a higher accuracy (highest of the four) of 58.79%.

4. Discussion

The model from Method 1, while only using the pixels of the images of a candidate's face, was able to reach 70.43% accuracy. It had no information on the political affiliation of the candidate, and no information of the political inclination of the region they are running in. This model was able to accomplish this without any other information, completely disregarding the competency, incumbency, or campaign spending of the candidate. In contrast to prior work (Joo et al., 2015), these new models had no human contribution as it only relied on an image of the politician.

These results are quite significant, as it suggests that artificial intelligence on its own, with no human influence, can emulate the appearance-based bias that effects election outcomes. Since this model had no other information to rely on, it is also a demonstration on how the facial appearance of a candidate is a heavy contributing factor in the minds of a voter. It is true that in real life, voters could be influenced by a variety of superficial aspects beyond appearance. Voters might also watch a video of a candidate or see them in person, which results in a lot more information on the candidate's appearance than just an image. However, by just using a picture, the machine learning model was still able to achieve a significantly greater-than-chance accuracy, and future improvements of this project that can analyze video samples, for example, might even further improve the results.

To make up for the relatively small size of the dataset, this project also designed two methods of transfer learning (Method 2a & 2b). They used generic data of faces train facial feature extraction, which then can be applied onto the election prediction model. Neither approach was more successful than Method 1 and Method 3. This suggests that the factors that are important for judging faces for elections are somewhat unique, and not the exact same as facial information that that are preserved in an autoencoder. A conclusion that could be reached from this is that on a dataset specifically targeted for a purpose, a large amount of data may not be necessary to achieve a significant result.

In previous human psychological studies on this topic, test subjects often only had a second or less to make a prediction. For most humans, that is not enough time to consciously analyze the candidate's faces and infer traits. Studies have shown that the shorter the time taken, the better the prediction (Ballew & Todorov, 2007). The explanation given was that humans' subconscious judgments or their "intuitions" are incredibly good at reflecting their internal judgments and biases, such as the preference when looking at an electoral candidate. In these cases, humans' subconscious made even better predictions than conscious analysis.

There is still debate about exactly how humans can do this, and the true nature of these judgments are yet to be fully understood. Are these judgments values that can be easily represented by numbers and categories of perceived emotions or traits, or are they reflections of nebulous, underlying factors that cannot be simply understood with a few words? The results from this project could provide a quantifiable way to compare these two possibilities.

Method 1 acts as a representation of human judgments that include underlying factors, while Method 3 only uses quantifiable and simplified traits to form a prediction. Method 3 reached 59.41% accuracy, showing that to a certain extent, simple and quantifiable traits such as emotions, race and age can still be incredibly useful information to analyze how voters would view a political candidate. However, Method 1 achieving the highest accuracy of 70.43% shows that there are factors unaccounted for in Method 3, and some useful traits cannot be successfully simplified or represented with understandable and basic categories.

In this project, Method 1 and Method 3 could be seen as representations of two ways of human thinking. Method 1 (by using all the pixels of a facial image) represents intuition: the subconscious, unreflective judgments that cannot be represented with a few words or values. Method 3 (by simplifying facial information into emotions, age, gender) represents consciousness: when humans intentionally try to categorize information into specific and familiar groups. This machine learning model may be a quantifiable assessment between the two methods of human judgment. With further integration into existing systems of political forecasting, the model's ability to assess voter bias can be crucial to improving election predictions. Furthermore, this project can be useful for any researchers that want to build upon these findings, use it as a benchmark, and further explore the nuances of human judgments and machine learning. Future studies based on different theories and principles can also use this model to compare with to determine which features are truly important for predicting election outcomes.

5. Conclusion

This study established a machine learning model that reached a 70.43% accuracy and an AUC value of 0.77 while predicting election outcomes by only analyzing the facial images of candidates. This was accomplished by gathering an unprecedented dataset with over two thousand facial images of political candidates, and creating four different methods that unlike previous studies, did not rely on the involvement of human test subjects. Not only did this project demonstrate how a politician's facial appearance has a consistent correlation to his or her political success, it also displayed how artificial intelligence can replicate voter's subconscious bias during elections, and therefore possibly other instantaneous and unreflective human judgments.

Acknowledgment

Majors thanks to Tyler Giallanza for his guidance and mentorship.

References

Arora, N. (2021). Age, Gender and Ethnicity (Face Data) csv. Kaggle.

Ballew, C. C., & Todorov, A. (2007, November 13). Predicting political elections from rapid and unreflective face judgments. Retrieved July 3, 2023, from https://www.pnas.org/doi/10.1073/pnas.0705435104

Google Colaboratory. (2017). computer software. Retrieved January 2, 2023.

Joo, J., Steen, F. F., & Zhu, S.-C. (2015, December). Automated Facial Trait Judgment and Election Outcome Prediction: Social Dimensions of Face. International Conference on Computer Vision.

Olivola, C. Y., & Todorov, A. (2010, January 23). Elected in 100 milliseconds: Appearance-Based Trait Inferences and Voting.

Olivola, C., & Todorov, A. (2009, May 5). The Look of a Winner. Retrieved January 2, 2023, from https://www.scientificamerican.com/article/the-look-of-a-winner/

PyTorch. (n.d.). Version (1.13.1). *From Research To Production*. Retrieved January 2, 2023, from https://pytorch.org/.

scikit-learn. (n.d.). Version (1.2.0). *scikit-learn Machine Learning in Python*. Retrieved January 2, 2023, from https://scikit-learn.org/stable/.

Serengil, S. I. (2022, May 10). DeepfaceVersion (0.0.75). *deepface 0.0.75*. PyPI. Retrieved January 2, 2023, from https://pypi.org/project/deepface/.

Todorov, A. (2018, May 14). *Can we read a person's character from facial images?* Scientific American Blog Network. Retrieved January 2, 2023, from https://blogs.scientificamerican.com/observations/can-we-read-a-persons-character-from-facial-images/

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005, June 10). Inferences of Competence from Faces Predict Election Outcomes. Washington D.C.; American Association for the Advancement of Science.

Ventura, C., Masip, D., & Lapedriza, A. (n.d.). Interpreting CNN Models for Apparent Personality Trait Regression. Institute of Electrical and Electronics Engineers.