Master's thesis
Digital Humanities
Cognitive Science

Can a deep neural network predict the political affiliation from facial images of Finnish left and right-wing politicians?

Anton Berg

November 4, 2022

Supervisor(s): Anna-Mari Rusanen, Matti Nelimarkka

University of Helsinki
Faculty of Humanities

Pietari Kalmin katu 5, 00014 University of Helsinki

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Koulutusohjelma — Utbildningsprogram — Degree programme
Faculty of Humanities	Digital Humanities Cognitive Science

Tekijä — Författare — Author

Anton Berg

Työn nimi — Arbetets titel — Title

Can a deep neural network predict the political affiliation from facial images of Finnish left and right-wing politicians?

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidantal — Number of pages
Master's thesis	November 4, 2022	35

Tiivistelmä — Referat — Abstract

This master's thesis seeks to conceptually replicate psychologist Michael Kosinski's study, published in 2021 in Nature Scientific Reports, in which he trained a cross-validated logistic regression model to predict political orientations from facial images. Kosinski reported that his model achieved an accuracy of 72%, which is significantly higher than the 55% accuracy measured in humans for the same task. Kosinski's research attracted a huge amount of attention and also accusations of pseudoscience.

Where Kosinski trained his model with facial features containing information for example about head position and emotions, in this thesis I use a deep learning convolutional neural network for the same task. Also, I train my model with Finnish data, consisting of photos of the faces of Finnish left- and right-wing candidates gathered from the 2021 municipal elections.

I research whether a convolutional neural network can learn to predict from candidates' faces whether a member of a Finnish party belongs to either the right-wing Coalition Party (Coalition) or the left-wing Left Alliance (Left Alliance) with better than 55% accuracy, and what is the possible role of color information on the classification accuracy of the model. On this basis, I also consider the wider ethical issues surrounding these types of models and the technological advances they bring.

There has been a recent ethical debate on the widespread use of facial recognition technology in relation to issues such as human autonomy, privacy, and civil liberties. In the context of previous scientific findings, there has also been debate about the potential ability of facial recognition technologies to reveal information about our most personal traits, such as sexual orientation, personality, and emotional states. Thus, facial recognition technologies are also closely related to privacy issues.

In his original article, Michael Kosinski did not underestimate the many problematic ethical issues that the use of facial recognition technology can raise. He did, however, underline the role of science in trying to determine the function, capability, and accuracy of these technologies. Only through research can we gain insights into these technologies, which can then potentially be used to inform societal decision-making. This research approach is also the aim of this Master's thesis.

Avainsanat — Nyckelord — Keywords

deep learning, neural networks, face recognition, ethics of artificial intelligence

Säilytyspaikka — Förvaringsställe — Where deposited

Helsinki University Library

Muita tietoja — Övriga uppgifter — Additional information

Contents

1. Introduction

In 2021, psychologist Michael Kosinski published an article in Nature Scientific Reports titled: *Facial recognition technology can expose political orientation from naturalistic facial images* [19]. He trained a facial recognition algorithm to predict the political orientation of liberal and conservative face pairs. His dataset was collected from three countries (the U.S., the U.K., and Canada) and consisted of 1,085,795 images in total. Kosinski reported that his model gained 72% classification accuracy. He argued that if one compares the result to chance (50%) or to previously tested human accuracy (55%), it is evident that questions concerning the protection of our most personal attributes and privacy and civil liberties, in general, should be raised to the table.

Kosinski used a cross-validated logistic regression model, which he gave face descriptors for learning. The face descriptors consisted of vectors that held information not just about age, gender, and ethnicity but also about transient facial features such as the values of combined head pose and orientation (roll, yaw, and pitch) and emotional expressions (probabilities of expressing anger, disgust, sadness, surprise, and fear). Although Kosinski stated that facial recognition algorithms are not directly interpretable, he hypothesized that head orientation and emotional expression probably gave the model most of its predictive powers. The article immediately raised many critical voices and accusations of conducting racist pseudoscience, previously known as *physiognomy*. Kosinski defended his study by stating that although physiognomists were wrong when they claimed that they could accurately judge a person's whole character based on facial appearance (for example moral traits), it is still plausible to think that faces contain at least some information about them.

In this master's thesis, I choose to *conceptually replicate* Kosinski´s study by using a deep neural network architecture trained on Nordic, Finnish data consisting of facial images of left- and right-wing political candidates. I decided to use a neural network implementation because Kosinski also stated that although he used logistic regression for his model, *[...] "alternative methods produced identical results"*, such as with a deep neural network classifier. Conceptual replication is not an *exact* or *direct* replication of the initial study. In a direct replication, one seeks to determine if the particular operationalization of a set of dependent and independent variables produces

the same effect. In conceptual replication one is instead testing for the same effect with an alteration of the operationalization of the variables [17]. For my study, this means for example, that instead of giving the model face descriptions as vector data, I am letting the model learn these by itself.

Conceptual replication is also supported by the fact that I am aiming to replicate Kosinski's study in a country with a very different political climate and party structure from the countries he used for data. Finland has a multiparty system and the opinions are not as polarized - when compared, for example, to the current situation in the U.S. In order to bring my data at some level in line with Kosinski´s original study, I decided to only use images from the far ends of the Finnish political spectrum, the right-wing party *Kokoomus*, and the left-wing party *Vasemmistoliitto*. Given this "material mismatch", I believe that my research still has relevance for different disciplines. Although many theoretical studies have been conducted in the political and social sciences on the differences between the North American and Nordic political systems, empirical data-driven studies that utilize computational social science methods are still very sparse. When it comes to the demarcation of political ideologies and scientific interests and areas such as image- and face recognition, cognitive science, and the ethics of artificial intelligence, the number of studies approaches zero.

The reason for choosing this topic and methodology is grounded also in similar ethical concerns that Kosinski already raised in his article. I think the debate around Kosinski's article focused too much on whether the classification of the algorithm is based on biological and genetic (face morphology, emotions, etc.) or cultural traits (color, posing, beard styles, etc.). In the end, and from a purely pragmatic ethical standpoint, I do not think it is what matters the most. What matters is whether the classification works. If it works, this has direct ethical consequences, for example, for human autonomy and democratically established societies build on the foundational idea of freedom and its many derivatives. This is why I also chose a simplistic model without fine-tuning or added emotion detection. The rationale is this: if the most basic model could gain high accuracy in this given task, it increases the probability space of potential implementors and benefittors, ranging from individual tech hobbyists to big enterprises and even nation-level actors.

The thesis begins with a theoretical background that first goes through the research questions. I will then take some time to go through some of the previous research and the history of deep learning, image- and face recognition. This historical recursion is important, especially for readers that are not familiar with the different stages of the development of machine- and deep learning. After this, I will tie this history into Kosinski´s article and its experimental background. In the methodology section, I will introduce my process for the empirical study: explaining the different stages of data

collection, cleaning, and model building. I will also explain the notion of *accuracy* in machine learning, and how this relates to the evaluation of the quality of these kinds of models. Then I will continue with the results, where I reveal how the classification accuracy of my trained model behaved with the different test datasets and what is its significance. Last, I will conclude with a discussion and conclusion section, where I also examine and discuss the results in an ethical context.

2. Background

2.1 Research questions

My main research questions are: 1) Can a convolutional neural network learn to predict from candidates' faces whether a member of a Finnish party belongs to either the right-wing Coalition (Kokoomus) or the left-wing Left Alliance (Vasemmistoliitto) with better than 55% accuracy, as humans have been reported to do in a similar task, and 2) What is the role of color information in the classification accuracy of the model? Following on from these, I also consider a broader ethical question in the final part of this Master's thesis: why does this matter?

The main idea is to experimentally test the hypothesis put forward by Michael Kosinski that a neural network could achieve similar accuracy to the cross-validated logistic regression model he used in his original study [19]. Kosinski trained a face recognition algorithm to predict the political orientation of liberal and conservative face pairs using data collected from three countries (USA, UK, and Canada), consisting of a total of 1 085 795 images. Kosinski used a more complex model in which facial descriptors consisted of vectors containing information on age, gender, and ethnicity, as well as transient facial features such as values for head posture and orientation (roll, yaw, and pitch) and emotional expressions (probabilities of expressing anger, disgust, sadness, surprise, and fear). If my model succeeds in gaining similar or higher classification accuracy, this would not only mean conceptually replicating Kosinski´s initial results with Finnish data but also achieving it with a simpler model.

2.2 Building blocks for image recognition

2.2.1 Small history of machine learning and neural networks

The beginning of modern machine- and deep learning technology, and in many ways also image recognition, can be traced to begin from a paper called *A Logical Calculus of Ideas Immanent in Nervous Activity* which was published in 1943 by researchers Jerry Lettvin, Walter Pitts, and Warren McCulloch, who hypothesized that a *neural network*

consisting of appropriately wired neurons could be reduced to propositional logic [23]. Although they later discovered that the complexity of the brain's neural circuitry was more significant than they anticipated, they never the less created the starting point of what was later to become new kinds of ways to model cognitive functioning and architecture. In addition to this, in 1957, Franck Rosenblatt developed a concept of the *perceptron* at the Cornell Aeronautical Laboratory. The biological principles of neurons also inspired him [32]. The model developed by Rosenblatt is binary. The integration of the inputs is implemented by adding weighted inputs with fixed weights obtained in the training phase. This is similar to the concept of the so-called *Hebbian learning*, a neuroscientific idea introduced by a third innovator, Donald Hebb in his 1949 book The Organization of Behaviour [16]. The basic concept of Hebb´s theory is that when our brain learns something new, neurons are activated and connect with other neurons to form a neural network. In cognitive science, this theoretical invention is known as *connectionism*, and it came to oppose the prevailing *symbolistic* model of learning, where there always needs to be a one-to-one correspondence between symbols and the concepts they represent [9].

After a long "winter" in the study of neural networks, caused by a fierce critic in the form of an influential book against the perceptrons written by the philosopher Marvin Minsky [24], in the 1990s, the development and research of machine learning and neural networks recovered and starter to gain popularity again. The models also progressed from simple perceptrons to multi-layer *deep learning networks* whose ability to recognize and classify, for example, handwritten numbers were eagerly received in the postal and banking services of their time (see for example [21]). As the 21st century approached, the hypothetical and theoretical potential of machine learning began to materialize. However, the computational speed of computers and the amount of data available for teaching formed a bottleneck that led many researchers to reject new fields utilizing artificial intelligence such as machine vision or computational linguistics. However, this began to change with the simultaneous development of processor (CPU) and graphics processor (GPU) technology, growing Internet use, increased image circulation, and cloud computing services.

The current field of machine learning divides into three major areas: supervised-, unsupervised- and reinforcement learning. In supervised learning, the system learns from categorized and labeled data, while in unsupervised learning, it needs to make sense of the data. In reinforcement learning, the system learns by interacting with a specific environment and a set of rewards and penalties [7]. Deep learning models are deeply rooted in statistics and, when taught with enough data, can be efficiently used to find hidden patterns or correlations from the data. A basic neural network model has an input layer, some amount of hidden layers, and an output layer. The input

layer receives the observation values; for example, in the case of an image, these would be pixel values. The hidden layer consists of nodes that can be used for computing and transforming the input values by multiplying them with different weights, a set of predetermined numbers. The power of the neural network lies in the optimization of these weights. Weighted connections are then put forward to either a new hidden layer or the output layer, which returns an output value. This kind of model can learn by association, allowing a system to generalize from images and use this information to make novel predictions and classifications of images it has not encountered before.

2.2.2 Images as data

In recent years, the amount of information available on social media sites such as Instagram, Facebook, and Twitter has exploded, and documents from public institutions such as hospitals and public administrations have started to be digitized. This new digital revolution has also created new research opportunities. However, due to the scale of the material, visual content analysis with traditional qualitative approaches is somewhat limited. Luckily, the rapid development of machine- and deep learning has enabled new data collection and analysis methods, which are also scalable as the data size continues to grow.

As our social processes in digital environments evolve, they generate more non-textual data. Images, memes, and videos have become a new way of communicating and building digital identities. Images and visual material have also become increasingly important data sources for researchers interested in these new media environments. They allow researchers to test existing theories and also push them to develop new ones that can shed light on the impact of these media formats on our social life [37].

In 2007, Professor Fei-Fei Li of Princeton University used Amazon Mechanical Turkey to label three million images into a single dataset that became *ImageNet*. Li's work gave birth to the current renaissance of machine vision, and image recognition, which has led to the construction of countless image recognition systems and services capable of viewing images and identifying different objects and categories [7]. We have also come a long way from simplistic configurations and architectures. A modern neural network architecture used for image recognition, such as the CIFAR-10 built by Alex Krizhevsky, winner of the ImageNet competition and later named "AlexNet," contains a staggering amount of 650,000 individual neurons in 8 layers connected to 60 million adjustable and optimizable weights. Using these models makes it possible to analyze vast amounts of images' pixel by pixel', then use this extracted information for different classification purposes. For example, a model can be taught to identify different stages of melanoma [2], or in potential social science applications, image recognition can be

used, for example, to predict the size of a crowd or to detect possible image alterations and modifications [37].

2.3 Face recognition as a cognitive phenomenon

2.3.1 Theories and models of perception related to face recognition

Facial recognition is one of the most special abilities of the human brain and our perception system. Perception has not only been a central topic for Cognitive Psychology but is more specific to Cognitive Science already from its beginning in the 1970s [31, 14]. A cognitive component of face recognition is pattern recognition, which also heavily links to other cognitive processes such as memory and thinking. Human pattern recognition is a mixture of knowledge distinguishing, acquisition, storage, and retrieval, where an input signal from the world is being compared to existing models of categorizations. In cognitive psychology, there are existing theoretical models for describing these processes such as *the Theory of Template*, *the Theory of Prototype*, or *the Theory of Feature* [31].

The Theory of Templates is based on the idea that people store various mini-copies of external patterns formed in the past in long-term memory. These copies (templates) correspond to external stimulus patterns one by one. When a simulation affects people's sensory organs, the simulated information is first encoded, compared, and matched with the patterns stored in the brain. Then one particular pattern in the brain is identified that best matches these. Although template theory explains some human pattern recognition, it has many limitations. Template theory for example assumes that people must first store a matching template before they can recognize patterns. As a result, a huge number of these templates are needed. This in turn not only overloads memory processes but also leads to inflexible and rigid pattern recognition. It also fails to fully explain the overall human pattern recognition process [31].

Prototype theory assumes that memory stores, not models that match one-by-one with external models, but prototypes. A prototype is an internal copy of a given model - a kind of abstraction that encapsulates the characteristics of all individuals of a given type or category. For example, people know different faces, but a specific face with two eyes, a nose, and two ears may work as a prototype of a face. According to prototype theory, in the process of pattern recognition, external stimuli need only be compared with the prototype, and the meaning of objects emerges from the relations between the objects - the correspondence between the input information and the prototype. When

this external stimulus information best matches a particular prototype in the brain, the information can be classified into the category of that prototype. and identified [31]. However, prototype theory has its own problems. For example, it only includes top-down processing but not bottom-up processing, which is sometimes more important for matching prototypes in the human perception process. Cognitive Neuroscientist Irving Biederman has proposed a complementary theory he calls *Recognition-By-Components* as a solution [3]. The core assumption of this theory is that an object is composed of some basic shapes or components - geometries. These include a block, cylinder, sphere, arc, and wedge. Although the number of components does not seem sufficient to identify all objects, these geometries can be used to effectively describe them, since the different spatial relationships of all geometries can form a myriad of configurations. Biederman´s theory has been also influential to the Swedish cognitive scientist Peter Gärdenfors, who has introduced his own *Theory of Conceptual Spaces* [14].

The third influential theory has been the Feature theory, which also tries to explain the perception of patterns and shapes. According to it, people tend to match the features of a pattern to the features stored in memory, rather than matching the whole pattern to a model or prototype. The feature theory model is currently highly applied in computer vision and pattern recognition. However, its problems are in some ways the reverse of those of prototype theory. Feature theory is a bottom-up processing model that lacks the implementation of top-down processing. Therefore, it also fails to fully explain the human perceptual system [31, 14].

In psychological models and computational theorizing of the face, the concept of face space is used. In the face-space model of recognition, faces can be viewed as points in a multidimensional space. The axes of the space represent the "features" by which the face is encoded. Thus, faces can be represented by coordinates on axes that define the value of the face at the following values in each feature dimension. At the psychological level, it is often not necessary to define the nature of the features that constitute the axes of space. It is usually sufficient to know that a face can be described using a set of feature dimensions (e.g. face width, the distance between eyes, eye color, etc.) and that different faces differ in terms of combinations of features. At the computational level, physical face spaces form the core of most automated face images. In face recognition algorithms, feature axes represent physical features of the face, often extracted statistically through for example using principal component or independent component analysis [28].

2.3.2 Modern tools

Machine pattern recognition developed rapidly in the early 1960s and immediately began to be successfully applied in areas such as weather prediction, satellite flight map interpretation, character perception and recognition, voice and fingerprint recognition, and medical image analysis. Today, machine-based pattern recognition is mainly divided into two basic methods: 1) statistical pattern recognition and 2) structural (syntactic) pattern recognition. Structural pattern recognition is based on the structural properties of the image and performs recognition using the information of the layered structure of the pattern. Statistical pattern recognition, which has a wider range of applications, is based on probability theory and is integrated with, for example, Bayesian decision-making [31].

Modern Computer Vision research has taken the advantage of pattern recognition and implemented it into its project of building intelligent systems that utilize a diverse set of signals such as micro-expressions, electroencephalography (EEG), gestures, tone of voice, facial expressions, and emotions [4]. Combined with advances in machine learning and the classification accuracy of deep learning, these technologies have combined to enable many of the conveniences of the modern world, such as various forms of digital login and authentication.

Modern face detection is based on teaching computers to find desired regions of interest (ROIs) from image data. For many systems, this is still a very hard problem to solve, especially when operating "in the wild", where many kinds of noise can affect the system's ability to perform face recognition tasks such as movement, lighting, poses, or distance [39]. For handling this noise, developers have come up with many sorts of fixes, such as illumination change normalization or a plethora of different noise filters, among others.

One of the methods for detecting faces, and the one I also used in my methodology, is a Multi-task Cascade Convolution Neural Network (MTCNN). A MTCNN has a layered Convolutional Neural Network (CNN) structure, that goes through three stages before outputting the result [10]. In the first stage, it obtains candidate windows and regression vectors that constitute the bounding box that separates the face from the rest of the image. The second stage is for another CNN, which is being fed those candidates for false candidate detection. The third stage is quite similar to the second, but this time an output that contains five facial landmark positions is derived in the end.

A very important phase of manually building an actual classifier that uses faces as input data is choosing the relevant features (facial features). The quality of these is key for good accuracy. In Computer Vision, several methods have been developed for

feature extraction, such as *Local Binary Patterns*, *Active Appearance Model*, *Optical Flow*, and *Action Units* [4]. In his original study, Michael Kosinski used VGGFace2, a dataset for recognizing faces across pose and age, to convert his facial image data into face descriptors, which were 2,048-value-long vectors containing the core features [6]. To identify the political orientation of individuals, he also compared the facial descriptions with the average facial descriptions of liberals and conservatives. These descriptors were then fed into a cross-validated logistic regression model with self-reported political orientation (conservative vs. liberal) as the objective. Interestingly, he also made the claim that almost identical results were obtained using alternative methods: such as with a deep neural network classifier.

A CNN is a type of neural network that can also be used for image classification problems. This is because CNNs are able to detect and identify underlying patterns that may often be too complex for the human eye. The input image is run through several hidden layers of the CNN, which decode it into features. These features are then used for classification by a function (e.g. Softmax) that retrieves the highest probability from the probability distribution of classes as the predicted class [4]. A CNN model can be enhanced in many ways to overcome a classical problem of overfitting, where the model might perform well with a test set from the original data, but generalizes poorly. Approaches for enhancing the model are, for example, increasing the models' complexity (adding more layers), tuning various parameters during model training (epochs, batch size, learning rate, etc.), or simply increasing the training data [9, 5, 37].

A rudimentary CNN image recognition algorithm is based on first turning pixel-level information about color intensities into predictions of different types of image labels or categories. The images are three-dimensional matrix representations for the model that it uses as input values. Given the three-dimensional nature of images (image has over 150,000 features that range from 0 to 255), a fully connected neural network would need a very long time to train. Convolutional layers tackle this problem. The image's red, green, and blue values (RGB) can be processed separately as matrices of pixel-level intensities of these values. The filter of the convolutional layer *convolves* (slides) over these input volumes and creates a new output volume for the model. Finally, there will be a sum across all three filter dimensions [37]. This can then work as a new input for yet another convolutional layer. The modern state-of-art convolutional neural networks also have *pooling* layers that downsample the dimensionality of output volumes. This is needed because often, the number of parameters to learn will increase the risk of overfitting the model.

2.3.3 Comparing the abilities of humans and machines

It has been suggested that the ability of humans to recognize faces can be used as a benchmark for evaluating the performance of automatic face recognition algorithms. However, it is often unclear which factors influence the accuracy of human face recognition. For example, O'Toole et al. have proposed that these factors can be classified into constraints on facial structure and gaze parameters [28]. Facial structure constraints include factors such as facial typicality, gender, and ethnicity. Gaze parameters include factors such as changes in illumination and viewpoint and perceptual complications that arise when we see faces and people in motion. O'Toole et al. have further argued that human experience and familiarity with faces could solve many, if not all, of the challenges of face recognition [28]. They argue that computer algorithms should therefore seek to mimic the ways in which humans learn to recognize faces. However, applying these principles to the design of algorithms that could address the pressing challenges of face recognition in naturalistic imaging environments has proven challenging.

According to O´Toole et al. facial typicality is one of the best-known and most robust predictors of human face recognition performance (see e.g. Light et al. (1979)). But what should be considered "typical" in this context? Prototype theory, for example, assumes that one could find some kind of "average" face. There is instead a kind of paradox in automatic face recognition, that individual faces are easier to recognize. O´Toole et al. infer, that presumably this happens: *"because typical faces are plentiful and so there are more faces that could be falsely mistaken for any given typical face than for any given distinctive face"* [28].

Since 2005, there have been face recognition competitions where human and computer vision performance has been systematically compared. In their 2014 review article, Phillips and O´Toole made a comparison of human and computer performance across these different face recognition experiments and competitions [30]. Their analysis showed, that when it came to face recognition tasks where the data comprised of frontal faces in still images, algorithms were systematically and consistently superior to humans. They wrote that: *"machines were able to represent a person´s identity primarily by encoding information extracted from the face, whereas information from the body, hair, or head was ignored"* [30]. Based on their results they also concluded that humans were better with video and difficult still-face pairs because humans could take advantage of all available identity cues for recognition.

Michael Kosinski also stated in his article [19], that humans are relatively poor at doing face-based judgments. He made reference to a 2013 study made by Tskhay Rule [35], where people were asked to distinguish between two faces that belonged either to conservative or liberal people. In this study, humans were correct about

55% of the time, which was only slightly above chance (50 %) Kosinski also stated, that the low human accuracy in these kinds of tasks does not represent the limits of what algorithms can achieve. He makes an adept mark on the fact, that algorithms already excel in many pattern recognition tasks with such large datasets, that are already almost incomprehensible to humans. Algorithms are already reported to be able to for example outperform us in face-based judgments of intimate attributes, such as personality [18, 29] or sexual orientation [36, 22].

3. Methodology

3.1 Data collection

The data for my model was collected during the Finnish municipal elections held on 13.6.2021. A web-scraping program was coded in the Python programming language that could crawl parties' websites and download all candidates' pictures from each municipality. The program utilized Selenium, a free software tool for automating web applications for testing purposes [33]. I also scraped images from the parties' youth organizations for extended testing purposes. A crucial ethical note: I did not hack into any databases, and overall there was no ethical obstacle to collecting the data since each candidate was running for a socially important public office. An example of the data can be seen in this picture:

Figure 3.1: Example of a candidate image after initial collection.

3.2 Preparing the data

As a precautionary first step for preparing the data, I manually reviewed the whole dataset to check for duplicate images that could end up biasing the model's training. As a second step, I looked at the symmetry of the dataset. I found no alarming irregularities or other anomalies. The gender balance and cultural factors such as spectacles, mustache, and beard styles, were balanced. After this stage I was left with 1799 images in total which is illustrated in this table:

	Kokoomus	Vasemmistoliitto
Female	385	296
Male	520	498
Eyeglasses (male)	209	201
Eyeglasses (female)	149	194
Beard or moustache	186	250
Total	905	894

Table 3.1: Data distribution after cleaning.

Because I wanted to train the model with only facial images, I needed to detect and extract only the faces from the images. For this, I created a custom face-classifier that utilized a Multi-Task Cascaded Convolutional Neural Network (MTCNN) [39]. This was easy to implement in Python programming language because of the already existing *mtcnn* library. I also resized the images to 128 x 128 pixels at this stage. Then, to test the model's classification basis, I made versions of the images with all color information removed. As a result, I wanted to have two separate models: one trained with color images and one trained without color information. To take the color information out from the images, I used a free source command-line tool *ImageMagick* [34], which can do many kinds of image transformations and alterations. The final data looked like this:

(a) with colors. **(b)** without colors.

Figure 3.2: Final datasets.

After all these steps, the data was ready, and I could start building and training the models. The initial modeling plan is illustrated in this schematization:

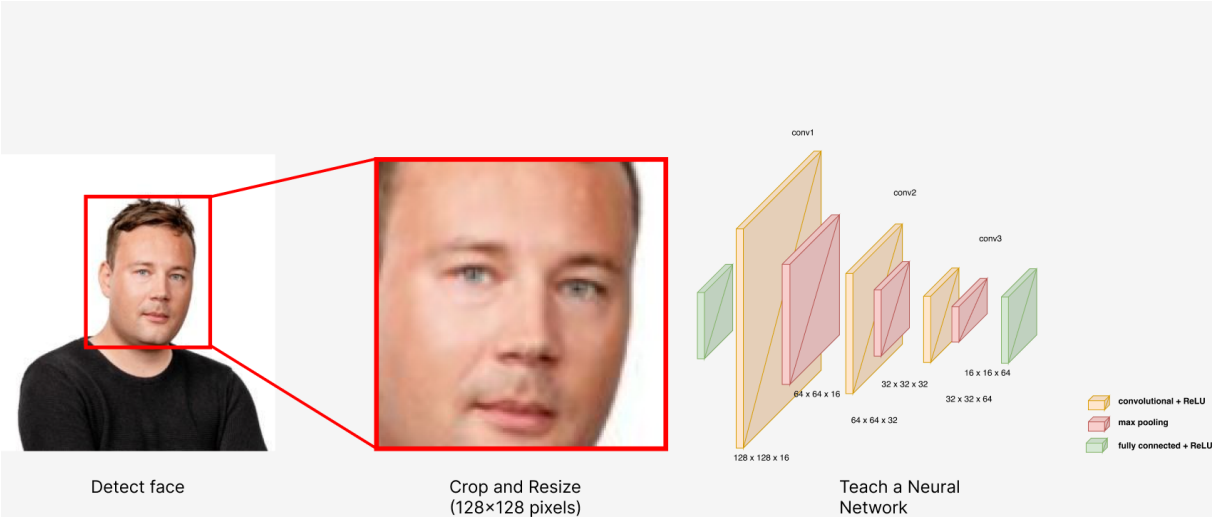

Figure 3.3: The initial modeling plan.

3.3 Building and training the models

For model building, I used TensorFlow, an open source library for developing and training machine learning models [1]. I created the models and documented the whole process in Google Colab notebooks (Direct links to notebooks can be found in the Appendix). This was to promote transparency and reproducibility. With TensorFlow Keras utils, splitting the data into training (80%) and validation (20%) and setting some initial parameters was easy. I used a batch size of 32 and an image height and width of 128 – the original pixel size of every image in the dataset.

I used the Sequential model that consisted of three convolution blocks with an integrated max pooling function layer. As I already stated when explaining the architecture of a basic CNN, using a pooling layer after convolutional layers is a common practice in ordering the structure. The convolutional layer summarizes the presence of features in an input image, and pooling is required to downsample the detection of features in feature maps. To use maximum pooling means calculating the maximum value for each patch of the feature map. Finally, there is a fully connected layer consisting of 128 units on top, and a rectified linear unit activation function (ReLU) is used to transform the summed weighted inputs from the nodes either as positive values or, if negative, will output zero. This can be illustrated with a simple code function example:

```
if input > 0:
    return input
else:
    return 0
```

Or described mathematically as:

$$g(z) = max\{\theta, z\}$$

To conclude, the overall model summary looked like this:

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 rescaling_1 (Rescaling)     (None, 128, 128, 3)       0

 conv2d (Conv2D)             (None, 128, 128, 16)      448

 max_pooling2d (MaxPooling2D  (None, 64, 64, 16)       0
 )

 conv2d_1 (Conv2D)           (None, 64, 64, 32)        4640

 max_pooling2d_1 (MaxPooling  (None, 32, 32, 32)       0
 2D)

 conv2d_2 (Conv2D)           (None, 32, 32, 64)        18496

 max_pooling2d_2 (MaxPooling  (None, 16, 16, 64)       0
 2D)

 flatten (Flatten)           (None, 16384)             0

 dense (Dense)               (None, 128)               2097280

 dense_1 (Dense)             (None, 2)                 258

=================================================================
Total params: 2,121,122
Trainable params: 2,121,122
Non-trainable params: 0
_________________________________________________________________
```

Figure 3.4: Model summary.

3.4 Accuracy in machine learning models

The performance of a machine learning model can be evaluated by using and interpreting different statistical performance metrics such as *accuracy*, *precision*, or *recall*. Here I will cover accuracy because that is what I used for model evaluation.

Accuracy can be understood as the calculated ratio between the models' number of correct predictions to the total number of predictions. Accuracy is fundamentally describing how a model is performing across all the classes it is trying to predict and

it can also be given this statistical formulation:

$$Accuracy = \frac{True_p + True_n}{True_p + True_n + False_p + False_n}$$

A known problem in machine learning model evaluation is, that most precision metrics are built only for two-class binary problems [20]. However, there are some ways to also deal with multi-class problems, such as Cohen's kappa metric. The difference between any of the used methods lies in the scoring of the actual classification rates. Cohen's kappa for example scores successes independently for each category and combines them [13]. Because the intention of my model is to be predicting between two classes (right-wing vs. left-wing), the problem of multi-classes doesn't need to be addressed here.

A bigger challenge, and one that I need to address, is that the performance of a machine learning model may not exactly match its performance after deployment, because the data on which the model was trained and evaluated may be very different from the cases in which the model will eventually be used in the real world. For example, my own data consists of images of political candidates posing for elections. The setting and posing are rigid and quite different from, say, images that would be posted on social media. It is therefore important to give weighted thought to how much trust my model should in the end be given if applied to different contexts. For example, Yin et al. (2019) showed that people's trust in a machine learning model is affected by the stated accuracy and also its observed accuracy and that the first can change depending on the former [38]. If a model with high accuracy would be used outside its original training examples, societal problems, such as misunderstandings, and false beliefs could generate because of this misplaced trust.

4. Results

When trained with color images of candidates' faces, the model reached an 83% validation accuracy. This is a stronger evaluation score than in Kosinski's initial study and indicates that the model could actually accomplish its task of predicting if a face belongs to either a left-wing or right-wing member quite well. The results are illustrated in Figure 4.1:

Figure 4.1: Training and Validation accuracy/loss when trained with color images.

Figure 4.1 also shows that there was no significant overfitting; instead, the model

performed very well on the test images. I also found out that the model performed very well even with data not belonging to the initial training data. This result was achieved by testing the model with data containing faces from a different age cohort (the youth organizations).

When all the color information had been taken away from the images, and the model was built the same, it surprisingly still reached a validation accuracy of 78%. This means that color information only increased the model's predictive power by 5%. The model could also again predict well even when tested with images not belonging to the initial dataset. The results for non-color images are illustrated in Figure 4.2:

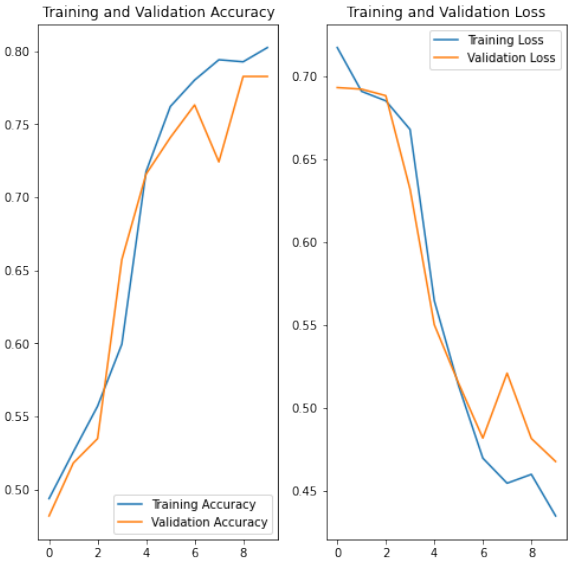

Figure 4.2: Training and Validation accuracy/loss when trained without color information.

Based on these results I can conclude that I have achieved a conceptual replication of the Kosinski article. A basic deep convolutional neural network model could successfully predict the political affiliation of Finnish left- and right-wing politicians from just a single facial image and with a relatively small dataset for training. The color space did not seem to affect the classification accuracy much, which is somewhat

surprising. Because of the symmetry of the dataset for cultural factors, such as eyeglasses and amount of bearded men, it seems that the model is pulling out information for the basis of its prediction from the face itself. Because the model is a *black box*, it is hard to know the exact factors used for classification. Although the previous observation is not the subject of this paper, I will make a few educated guesses as to what the model's capabilities might be based on: 1) there actually might be minimal differences in the morphological features of the face and head pose and orientation that the model recognizes, also 2) the degree of obesity could be an explanatory factor? The differences in obesity are visible in the face, carrying information about genetic and cultural differences. I would not consider emotion to be a good explanatory factor because of the specific nature of the images (candid pose).

5. Discussion

5.0.1 Ethical dimensions and reflections

Based on my results, I think it can be argued that the use of this kind of facial recognition technology raises a number of concerns, especially in liberal democracies. As I have shown, these kinds of models are fairly easy to build, which means that they can be used quite easily, for example for various lobbying purposes. This in turn has direct ethical implications for example to privacy and human autonomy. Of course, there is nothing very new in these concerns. Already since the early days of machine learning in the 60s and 70s, ethical problems related to its development have been considered, speculated, and evaluated. Artificial intelligence researcher Brian Christian has aptly stated that the history of artificial intelligence often appears to us as "cycles of hope and gloom" [7]. Christian has also defined what he calls as *The Alignment Problem* [7]. An alignment problem occurs when there is a contradiction between the original design goals of technology and the way it is being implemented. The alignment problem situates under the ethics of artificial intelligence, which shares the fundamental basic questions of a wider field of philosophy and ethics of technology: what kind of technology would best serve a good human life, and what kind of social solutions would make this line of development a reality? [26]. When taking this understanding of the alignment problem, it is hard for me to believe that the original developers of image recognition technology would have wanted it to be used in this way, for predicting people's political ideologies.

During the process of writing this Master's thesis, I constantly asked myself, is this the right and responsible way to use this kind of technology? I also found myself reflecting quite a bit on the instrumental notion of technology as something ontologically neutral, which some academics have stated as being the status quo for example in the social sciences [11]. It would seem, for example, that certain technological developments are not contingent and do not give rise to different forms of social organization, but instead, that certain technologies push societies in a more totalitarian, controlled direction. This is true in digital dictatorships like China as well as in Western democracies, even if their political ideology and cultural traits are quite different. It follows

that in the future it will also make sense to consider the agency, nature, and role of technologies such as facial recognition.

I myself came to the conclusion that facial recognition technology should not be used to identify personal characteristics, such as political convictions, in liberal democracies. The risk of abuse of human freedoms and autonomy is too high. The mere fact that something is technologically possible should not imply that it should be adopted for general use. This kind of thinking easily leads to technological determinism. We are currently living in a world where the industrial revolution brought about by artificial intelligence is challenging our existing ethical and moral concepts and their limits. Simultaneously, research has revealed serious flaws in many AI technologies. These developments also require us to reflect on the objectivity and autonomy of science. These questions are huge, and it is not possible to focus on them in any great depth here. However, I will reflect a little on my own thoughts in this regard. While it is true that science should strive to objectively study and measurements of different phenomena, it is clear that science cannot be a zone free of human dignity and values. Also, the relationship of science to the construction of reality is not epistemologically neutral, but science as an institution and scientists as individuals are also involved in power and thus to some extent also constructing social realities - either consciously or unconsciously. The pursuit of some form of hard ontological realism should not, in my view, be used as any kind of metaphysical "bypass lane" to go round the responsibilities of science as a builder of human good. In liberal democracies, this *good* is determined by commonly constructed goals, norms, and values. Science in Finland is publicly funded so that it can provide information to support decision-making, in addition to basic research. This decision-making should aim to benefit society. I feel that my research contributes to this value.

5.0.2 Impact on the scientific community

Many in the scientific community have been recognizing the concerns I reflected on above. It has recently become somewhat familiar to talk about biased software, which refers to a broader problem of algorithm biases and digital discrimination. We have many known examples of how modern artificial intelligence technology can be biased at every stage of the developmental process and end up discriminating against people [27, 15, 8]. Some researchers have pointed out that digital discrimination is on the rise [27, 12]. Many have demanded responsibility and remedial actions from the technology community, and it has also heavily impacted the scientific community. Over the past few years, research on themes and concepts of artificial intelligence ethics, like fairness, transparency, and other dimensions of safety, now forms a substantial portion

of work presented at the most prominent artificial intelligence and machine learning conferences. Some see them as the most dynamic and fastest-growing areas in all of science [7]. Unfortunately, biases and digital discrimination can be very hard to detect. The underlying problem is that machine and deep learning systems are incredibly good at finding hidden patterns and correlations in the data. They cannot be made 100% blindfolded. This means that discrimination can happen to a degree that is almost imperceptible to us. Google's famous Word2Vec algorithm can, for example, find very subtle differences in the dictation or syntax manners between job applications written by men and women. These can be slight grammatical preferences of prepositions or synonyms. When someone is applying for a position historically known to be male-dominant, for example, for software engineering positions, the algorithm can create a bias towards preferring applications written by men, even when sex is a protected variable in the classification and ranking algorithm. The deep neural network architecture that I used in this master´s thesis, is also somewhat of a black box when it comes to transparency.

Philosopher Vincent Müller highlights most of the challenges and research areas in the ethics of modern AI and robotics in his recent contribution to The Stanford Encyclopedia of Philosophy (2020), *Ethics of Artificial Intelligence and Robotics* [25]. Among other topics, the volume discusses AI and robotics as part of the development of humanity in the near future, the threats, risks, and their control associated with the development of AI and AI systems, and ethical issues related to individual freedom and privacy in the digital age. According to Müller, new emerging technologies have generated both fear and excitement throughout history. However, he argues that the most important thing is to focus on identifying the actual threats while at the same time nullifying unjustified fears. According to Müller, a major general challenge related to the ethics of AI is the retroactive occurrence of potential harms. He says this has been repeated in history with many other technological innovations such as nuclear power, cars, and plastics. Müller also criticises the current public discourse on AI. In this discourse, plans for AI are often described as straightforward and requiring only minor technological tweaking. However, the reality is quite different.

According to Müller, one of the significant ethical issues is the freedom of the individual, and in particular, privacy and ownership of the data collected about a person. Will we have the right to be unidentifiable in the future, or are we doomed to constant digital manipulation and behavioral modification? This is an essential remark from the perspective of this master's thesis. We already know that, based on the data collected about us, we are bombarded with personalized advertisements on various social media platforms created by artificial intelligence algorithms that are constantly learning and becoming better in their predictions. Our freedom to make independent

decisions and choices could therefore be under threat. A facial recognition system that could predict our political ideology from our Instagram or Facebook pictures could pose serious problems and even be erosive to our autonomy and, in the bigger picture, our democratic institutions. If a person were to be subjected to subliminal influence aimed at changing his or her political thinking, this would not serve the fundamental values of democracy, according to which the voter must have the freedom to choose for himself or herself and his or her will must not be unduly influenced. Political opinion is also a characteristic that is covered by the right to privacy. Violation of this protection is punishable in liberal democracies.

6. Conclusions

In this master's thesis, I conceptually replicated Michael Kosinski's finding that a convolutional neural network could also predict political affiliation from facial images. This is also the case when the model is trained on Nordic, Finnish data. The neural network implementation trained with color images achieved an 82% accuracy, which is very high for a dataset with only a couple of thousand images. Even when all the color information was extracted away, the accuracy remained as high as 78%. This has significant consequences for our understanding of what these technologies can do, especially if misused to affect our decision-making without our consent or under the radar of our conscious attention.

The widespread use of facial recognition technology in predicting political ideologies poses dramatic risks to all human autonomy, privacy, and civil liberties. If these technologies are being used this way, they become prime examples of the alignment problem and are likely not what the initial designers of these systems intended. Previous scientific discoveries have also shown that widely used facial recognition techniques could reveal information on our most personal traits, such as sexual orientation, personality, and emotional states. Therefore, the privacy threats posed by facial recognition technology are unprecedented in many ways. Moreover, there seems to be no limit to the development of computer vision and artificial intelligence, or at least no imminent slowdown. Moreover, even small predictions can have a considerable impact when applied to large populations in electoral situations. Unfortunately, this can also increase the effectiveness of manipulation.

My hope is shared with Kosinski - that scientists, policymakers, engineers, and citizens will start to pay more and more attention and that the ethical dimension of the debate will take center stage.

Bibliography

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] A. Adegun and S. Viriri. Deep learning techniques for skin lesion analysis and melanoma cancer detection: a survey of state-of-the-art. *Artificial Intelligence Review*, 54(2):811–841, 2021.

[3] I. Biederman. *Visual object recognition*, volume 2. MIT press Cambridge, MA, USA, 1995.

[4] D. Canedo and A. J. Neves. Facial expression recognition using computer vision: A systematic review. *Applied Sciences*, 9(21):4678, 2019.

[5] D. Canedo and A. J. R. Neves. Facial Expression Recognition Using Computer Vision: A Systematic Review. *Applied Sciences*, 9(21):4678, Nov. 2019.

[6] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[7] B. Christian. *The Alignment Problem - Machine Learning and Human Values.* Norton & Company, New York, 2020.

[8] K. Crawford. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence.* Yale University Press, 2021.

[9] P. Domingos. *The master algorithm: How the quest for the ultimate learning machine will remake our world.* Basic Books, 2015.

[10] S. S. Farfade, M. J. Saberian, and L.-J. Li. Multi-view face detection using deep convolutional neural networks. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 643–650, 2015.

[11] A. Feenberg. *Transforming technology: A critical theory revisited*. Oxford University Press, 2002.

[12] S. Feldstein. *The Rise of Digital Repression: How Technology is Reshaping Power, Politics, and Resistance*. Oxford University Press, 2021.

[13] S. García, A. Fernández, J. Luengo, and F. Herrera. A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability. *Soft Computing*, 13(10):959–977, 2009.

[14] P. Gardenfors. *Conceptual spaces: The geometry of thought*. MIT press, 2004.

[15] T. Gebru. Race and Gender. In *The Oxford handbook of ethics of ai*, pages 251–269. 2020.

[16] D. O. Hebb. *The organization of behavior: A neuropsychological theory*. Psychology Press, 2005.

[17] R. Hudson. Explicating exact versus conceptual replication. *Erkenntnis*, pages 1–22, 2021.

[18] A. Kachur, E. Osin, D. Davydov, K. Shutilov, and A. Novokshonov. Assessing the big five personality traits using real-life static facial images. *Scientific Reports*, 10(1):1–11, 2020.

[19] M. Kosinski. Facial recognition technology can expose political orientation from naturalistic facial images. *Scientific Reports*, 11(1):100, Dec. 2021.

[20] T. C. Landgrebe and R. P. Duin. Efficient multiclass roc approximation by decomposition via confusion matrix perturbation analysis. *IEEE transactions on pattern analysis and machine intelligence*, 30(5):810–822, 2008.

[21] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[22] J. Leuner. A replication study: Machine learning models are capable of predicting sexual orientation from facial images. *arXiv preprint arXiv:1902.10739*, 2019.

[23] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[24] M. L. Minsky and S. A. Papert. Perceptrons: expanded edition, 1988.

[25] V. C. Müller. Ethics of artificial intelligence and robotics. 2020.

[26] I. Niiniluoto. Tekniikan filosofia. 2020.

[27] C. O'Neill. Weapons of math destruction. *How Big Data Increases Inequality and Threatens Democracy*, 10:3002861, 2016.

[28] A. J. O'Toole, F. Jiang, D. Roark, and H. Abdi. Predicting human performance for face recognition. *Face processing: Advanced models and methods*, 2006:293–320, 2006.

[29] I. S. Penton-Voak, N. Pound, A. C. Little, and D. I. Perrett. Personality judgments from natural and composite facial images: More evidence for a "kernel of truth" in social perception. *Social cognition*, 24(5):607–640, 2006.

[30] P. J. Phillips and A. J. O'toole. Comparison of human and computer performance across face recognition experiments. *Image and Vision Computing*, 32(1):74–85, 2014.

[31] Y. Pi, W. Liao, M. Liu, and J. Lu. Theory of cognitive pattern recognition. *Pattern recognition techniques, technology and applications*, pages 433–463, 2008.

[32] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[33] Selenium-Project. Selenium webdriver @ONLINE. https://github.com/seleniumhq/selenium, June 2022.

[34] A. Thyssen. Imagemagick @ONLINE. https://legacy.imagemagick.org/Usage/, Nov. 2012.

[35] K. O. Tskhay and N. O. Rule. Accuracy in categorizing perceptually ambiguous groups: A review and meta-analysis. *Personality and social psychology review*, 17(1):72–86, 2013.

[36] Y. Wang and M. Kosinski. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of personality and social psychology*, 114(2):246, 2018.

[37] N. Webb Williams, A. Casas, and J. D. Wilkerson. *Images as Data for Social Science Research: An Introduction to Convolutional Neural Nets for Image Classification*. Cambridge University Press, 1 edition, Aug. 2020.

[38] M. Yin, J. Wortman Vaughan, and H. Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–12, 2019.

[39] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, oct 2016.

Appendix A. Links to the code

Google Colab:

Code with color images: https://colab.research.google.com/drive/1OP7a1m4yXph1UelP-EGBJXThiHUXlfjc?usp=sharing

Code without color information: https://colab.research.google.com/drive/1JllTVX1vnS-_YG7TxQe_3dH35INZljJr?usp=sharing