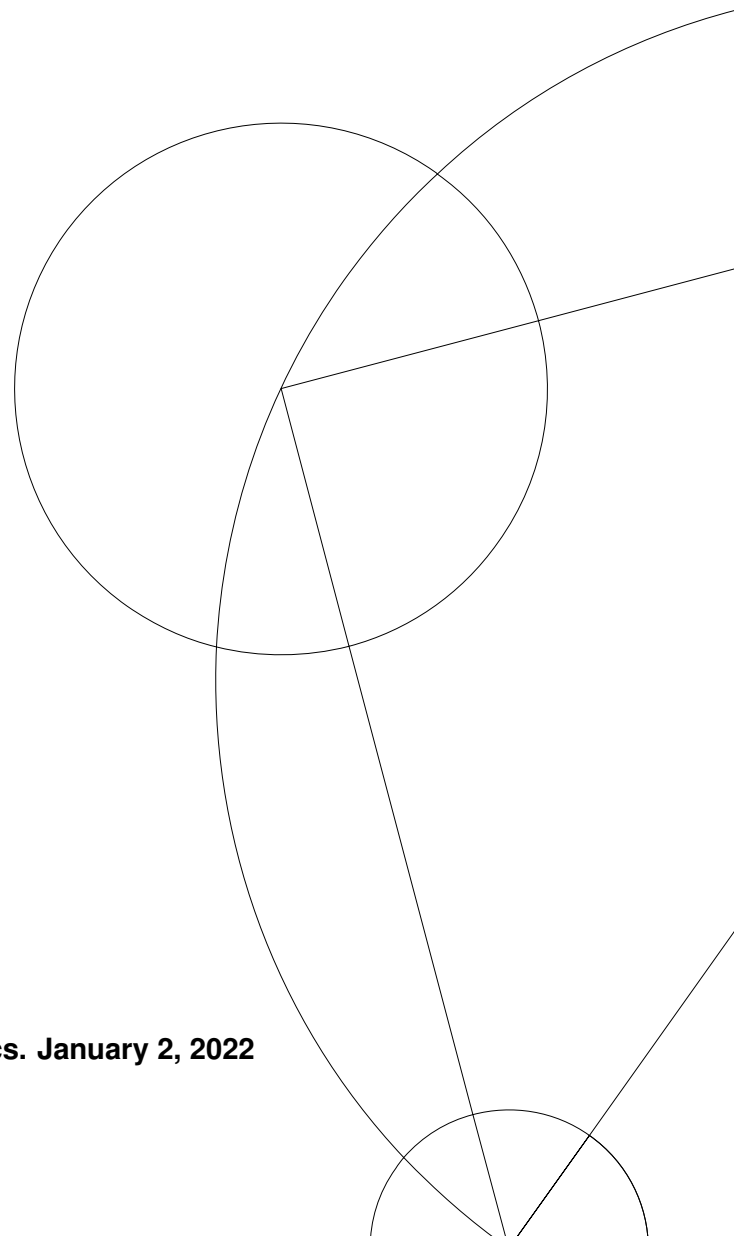
KØBENHAVNS UNIVERSITET

Cognitive Science 3 Exam Project

Task Independent Gender Classification

Classifying gender from IQ tests and passive nature picture viewing



Supervisor: Nora Hollenstein

Department of Scandinavian Studies and Linguistics. January 2, 2022

Group members

Navn	KUid
Christoffer Vang Roed	mth305
Lasse Goul Jensen	kvh190
Niels Krog	lkt259

Abstract

In recent years there have been an increasing interest in biometric identification. Various studies has been conducted aiming to detect individual or unique demographic traits. Results from recent studies have given reason to believe that the ability to predict demographic information using eye tracking is highly dependent on the stimulus. Based on this, this study performs a comparison using two datasets built on different stimulus: IQ tests and passive image viewing. This study found significant differences in the accuracy achieved on the two datasets using a Random Forest (RF) and Long-Short Term Memory network (LSTM) as classifiers. The random forest performed best with an accuracy of 0.85 on the passive image viewing dataset and 0.70 on the IQ dataset. Furthermore, we investigated the effect of using micro movements of the eye as features in the model, which has proven to be a useful feature in recent studies within biometric identification. Lastly, we report the classifiers' ability to perform task independent predictions of gender, using one dataset for training and another for testing. The result from this experiments was not satisfying and close to chance level. However, combining the dataset for training and testing resulted in the random forest having an overall accuracy on the combined dataset of 0.74.

Division of Labour

For this project, all group members have contributed equally to the content of this report. This includes data acquisition, data prepossessing, the development and implementation of models, literature research and the writing of the paper. The paper is a result of great collaboration and everyone vouch for the content originality of this paper. However, below are the main responsible members for each section.

Section	Written by
1.0	Christoffer
2.0	Christoffer & Lasse & Niels
3.0	Niels
4.0	Lasse & Niels
5.0	Lasse
6.0	Christoffer & Lasse & Niels

Keystrokes: 23999

Dependencies

All models were implemented in Python using Scikit-learn and Keras and matplotlib for figure plotting. The code and data sources used in this project is publicly available through Github

Contents

1 Introduction

Eye tracking can be used for many things, including human-computer interaction, human attention and cognitive load. It can also reveal a long series of traits in humans, like personality, personal, age or gender, given the right stimulus. However, is it also possible to predict such traits, e.g. gender no matter what stimulus the user is exposed to?
Current research mainly rely on two types of movements: macro movements and micro movements. The macro movements consists of fixations and saccades. Fixations occur when the eye looks at a region for a period of time around (250ms) with saccades being fast the movements (30-80 ms) from one fixation to the next [4, p. 2]. The second type of movement, micro movements, can be found as very small involuntary movements during the fixation called microsaccades. Microsaccades have a duration of around 20ms [11]. The character of microsaccadesare, described as an event where the eyes shortly drift away from the fixation center. Psychological studies suggest that the mechanisms involved with the eye movements are highly individual and can function as an open window for the explorations of personal characteristics and cognition. The human eye movements are driven by the ocolumotor system, which is built up by six muscles that support movements on all three axes. The movement of the eyes is a highly complex process which consists of both voluntary and involuntary movements [4, p. 1]. These personal characteristics of the oculomotor systems come to light though the macro- and micro movements [13].

Some of the earlier studies related to eye movements started to look in to the connections between cognitive processes and eye movements in reading and picture related tasks [13, p. 2]. Later on, important advances in the understanding of the oculumotor system was achieved, followed by a range of new applications using eye movements such as touch-less applications and applications for human computer interaction [13]. Later followed a range of studies which suggested that eye movements had many unique characteristics related to the individual viewer[13]. Researchers showed that features of the eye movements such as fixations and saccades had unique characteristics related to each unique viewer. Following these findings, an area dealing with biometric recognition emerged.

The first studies in this area dealt with biometric identification using mainly macro movements. The research within biometric identification started around 2001 and has mostly dealt with the identifications of individuals, often related to authentication tasks [13, p. 2]. Authentication tasks cover the research related to improving security systems that uses eye signature or movements as identification [8], similarly to fingerprint readings on smartphones or laptops. Recently, studies have introduced the use of micro movements which has showed to improve biometric identification [4]. The area of biometric identification has furthermore evolved into investigating biometric identification of genders and age etc. This emerged from existing psychological and behavioral studies that had already proven differences between men and women in various cognitive processes [15, p. 44]. Using this knowledge, some studies have been conducted in the area of gender classification with varying success. The success of the gender classification is often related to the visual stimuli and the feature selection [15]. The content of the visual stimuli can in gender classification be divided into two groups, human faces and bodies and other objects/subjects. For other objects/subjects previous studies has showed that males and females use similar macro movements when looking at them, but when viewing human faces or bodies, studies has showed significant differences in scan patterns and fixations between genders [12]. Despite the similar scan patterns between male and females, recent studies found that the micro movements during fixations and saccades did show unique characteristics between genders during passive indoor picture viewing [15]. The data used in these studies has been collected though various eye tracking tasks, where the participants have been told to perform a specific task (reading, looking at images etc.) while the movements of the eye have been tracked.

2 Related work

As mentioned, various studies in the area of biometric identification has been conducted with successful results [13]. Biometric identification of genders using eye tracking is an area that arose from the findings of gender differences in cognitive processes such as visual attention [10]. Gender prediction from eye tracking data has been done before, however, most commonly using specific stimuli, like free-viewing images of faces [12]. The study conducted by Pouretemad et al., demonstrated high accuracy in the classification of gender, as women tend to be better at recognizing faces than men [12]. Their most effective feature for prediction was regions of interest in the face, as men and women tend to look at different areas in faces. However, these features are task dependent and cannot be generalized between types of stimuli [12]. But the gender related differences is not only limited to human face or body viewing, which is demonstrated in Sargezeh et al., Hwang and Lee[15, 3]. Sargezeh et al.examined their participants during passive indoor picture viewing and achieved results well above a chance with and accuracy of 84%, but below those achieved by Pouretemad et al. having a accuracy of 94.7% [15]. Hwang and Lee examined their participants' gaze patterns while online shopping. They succeeded in detecting differences between genders, because male and females tend to focus on different information. For example, males showed increased interest in e.g. product information areas compared to females [3]. One method by Thapak et al.[18] uses pupil width as a feature for predicting gender with a certain success, where they achieved an accuracy of 62%. Though, they mention nowhere in the article, exactly how they managed to do so.

The key differences between biometric identification and demographic prediction is, that the demographic data is applied to more people. Two subjects may have different microsaccadic amplitude, but is that affected by their gender or is it simply unique to them? Demographic data is more general, as it covers more people than a single subject. There is to the best of our knowledge done no previous studies in the area of task independent identification of genders. Therefore, this relies on more general studies on task independent classification, which have dealt with the prediction of individuals across varying eye tracking tasks, as opposed to gender [8, 16, 1]. Task independent biometric identification can be divided into semi independent classification which uses data generated from different tasks but in the same context (e.g. using two different images) and completely independent where training data comes two completely independent sources (e.g. from images and texts) [1, 8]. A study conducted by Schröder et al. [16], investigates the differences in training and testing on data extracted from the same stimulus versus training and testing on data extracted from different stimulus using a random forest and a radial basis function network, as proposed by the winner of BioEye2015 [13]. Their results when training and testing on data derived from the same stimuli varied in accuracy from 0.817 to 0.941, where as their results when training and testing on data derived from different stimulus varied from 0.046 to 0.235 (multiclass), which clearly states the challenges in task independent classification [16]. The study [8] conducted by Schröder et al. investigates task independent biometric identification of individuals, to improve biometric authentication. Schröder et al. [8] used gaze patterns achieved from video viewing as input features and used a machine learning model which has previously been used in modern text-independent speaker recognition. [8]. Their predictions achieved results significantly above chance and showed improvements based on the amount of input training data [8].

3 Datasets

In order to perform proper task independent generalization, two datasets were used for this paper. The datasets had to fulfill some criteria: availabe participant gender and age, comparable features or raw data included and the participants had to be presented with different stimuli. The following two datasets were selected; the TüEyeQ IQ test dataset[7], and the Doves free image viewing set[9]. The final target value for prediction was settled to be gender, as it was more comparable between the two datasets than age.

3.1 TüEyeQ

The TüEyeQ dataset[7] contains data generated by 315 participants (94 males, 217 females and 4 unknown (which were removed)), with ages ranging from 18 to 30 and a mean age of 23.27 years.

The TüEyeQ dataset further more includes a wide range of demographic information which was not included in this study, because the information was not comparable to other datasets. The demographic information extracted was gender, which was later used in the experiments. The participants in the TüEyeQ dataset performed IQ tests (specifically the CFT 20-R test) while their eye movements were recorded. The features used from this dataset are the following:

- Fixation duration
- Fixation mean coordinates
- Mean pupil diameter
- Microsaccade count
- Microsaccade peak amplitude
- Microsaccade peak velocity
- Saccade duration
- Saccade start coordinates
- Savvade end coordinates

3.2 Doves

The Doves dataset[9] contains data generated by 29 participants (18 males and 11 females), with ages ranging from 17 to 45 and a mean age of 26.58 years. From the Doves dataset, the age was available and extracted in same way as for TüEyeQ to be used in later experiments. The stimuli used in the Doves dataset consists of 101 images. The images were selected from the Natural stimuli collection created by Hans Van Hateren[2] and display varying grey scale images of nature. They have varying sizes, and were therefore modified by cropping to a standard size using only the central 1024x768 pixels. Each of the participants viewed all 101 of the images as a free viewing task. An example of such image is seen in Figure 1. The dataset had sparse features available, so the same features as TüEyeQ were extracted from available raw data which is seen in 3.1. This process is describes in subsection 4.2.

Figure 1: An example of an image the participants were presented.

4 Methodology

4.1 Training Methodology

To generalize the classification between the two datasets, we first tune our models by both training and testing on a single dataset, using 5-fold cross validation. This way the models are optimized within a more controlled environment, than if it were applied to two datasets. When the models are fit to the data, we perform a series of tests on them. All combinations are listed below:

1. Train on DOVES, test on DOVES subset
2. Train on TüEyeQ, test on TüEyeQ subset
3. Train on DOVES, test on TüEyeQ
4. Train on TüEyeQ, test on DOVES

5. Train and test on a concatenated dataset consisting of DOVES and TüEyeQ

Performing all these trainings will gather insight in, how (and if) data is generalized across stimuli. The two first versions are simply for control, to validate the datasets, models and features, while three, four and five will reveal how well the models generalize on different data. We hypothesize, that five perform better than three and four, as the test data here will be more foreign from the training data.

4.2 Feature extraction

The Doves dataset used in this report comes with a very limited set of extracted features, and lacking documentation of the methods applied. For these reasons, we decided to perform our own extraction on the raw data. The main task was to identify when saccades and fixations occurred, as well as when microsaccades occurred within each of these fixations. After this, several higher level-features where extracted from this data to be used in the classification task.

4.2.1 Saccades and Fixations

The first part of feature extraction on eye-tracking data, is separating and labeling fixations and saccades. Since there is no definitive criteria for when fixations start and end, there exists a host of different approaches and definitions, which produce a wide variety of results when applied [14, 6]. Since the data in the Doves dataset was relatively simple, with relatively few statistics recorded (x and y coordinates for one eye, as well as pupil dilation), we decided on a relatively simple detection algorithm, the velocity-based Velocity-Threshold Identification or I-VT [14, p. 73]. I-VT uses a simple velocity threshold to determine whether to label a point as part of a fixation or as a saccade. The fixation groups (grouping of eye coordinates that constitute a fixation) are then returned in the format of $<x,y,t,d>$ with x and y being the coordinates of the centroid of the fixation group, t as the time of the first point, and d as the duration of the fixation. An approximate velocity threshold may need to be inferred when only point to point velocities are known, such as in the case of the Doves data [14, p. 73]. Furthermore, we collapsed any groups shorter than 10ms in duration, to avoid noise in the form of outliers. This left us with clearly defined groups of saccades and fixations.

4.2.2 Microsaccades

As mentioned in section 1, microsaccades are very important to demographic identification, and so we needed to extract this data as well. We used a thresholding definition similar to the I-VT algortihm used to identify fixations, only with a lower threshold of $3°/s$ and an upper threshold of $100°/s$ as defined by Otero-Millan et. al [11], as well as a duration around 20ms [11].

4.2.3 Higher-level Features

Once the basic categorisation of fixations, saccades and microsaccades was complete, higher-level features could be extracted. We focused mainly on descriptive statistics of the main feature groups, such as mean, minimum, maximum and standard deviation of fixation, saccade, microsaccade and pupil data, as well as counts of occurrences of these. We also implemented a meta-feature of total fixation duration to total saccade duration reported by Sargezeh et. al. to help in gender classification [15].

4.3 Models

We compared two models for gender classification: random forest and Long-Short Term Memory (LSTM). Both models were compared to a dummy classifier baseline to ensure, that they were performing better than chance.

4.3.1 Random Forest

The random forest (RF) is an ensembe classifier consisting of a series of decision trees. It supports non-linearity and is often a good first choice for both classification and regression issues, as it has short training time and accurate results. Other models were compared, results can be seen in appendix A.

4.3.2 LSTM

The LSTM is a type of Recurrent Neural Network and works well on sequenced data like language and weather readings. As the eye tracking data is sequential, we theorized, that an LSTM would be able to infer from the relation between time steps, which the random forest cannot. The model consists of six layers seen in Figure 2. These hyperparameters were fine-tuned with a random search.

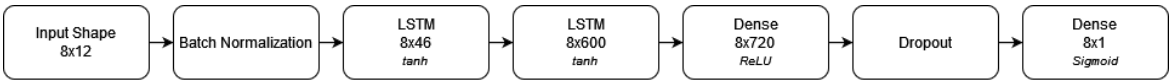


Figure 2: LSTM architecture

The LSTM model uses binary cross entropy as loss and binary accuracy for metrics. Binary accuracy is justified, as neither datasets are highly unbalanced in terms of distribution of targets and false positives and negatives are not of importance. The improving training loss and accuracy during training is seen in Figure 3. The model used early stopping with a patience of 10 epochs to prevent overfitting and the validation data makes up 10% of the training data.

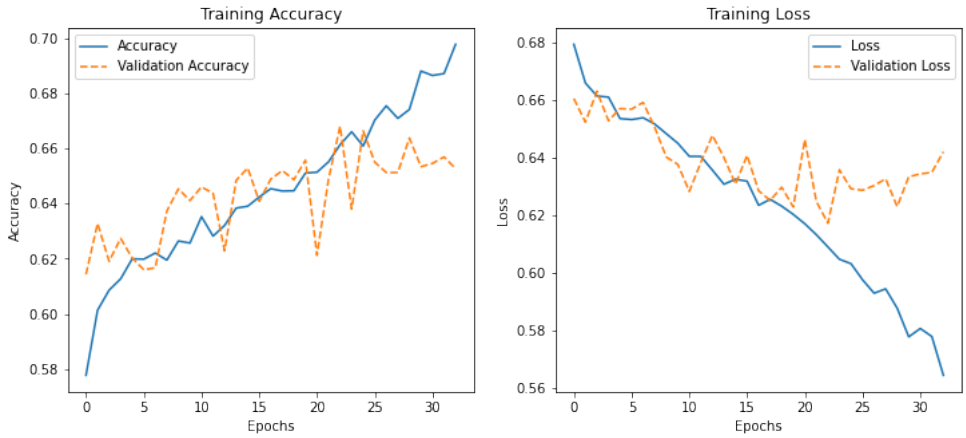


Figure 3: Accuracy and loss for LSTM training on TüEyeQ.

5 Evaluation

The final architectures for the LSTM and RF were fit to all combinations of training and testing data. The metrics used for evaluationg them are Area Under Curve (AUC) of the Receiver operating characteristic (ROC) curve and accuracy. The performance metrics are presented in Table 1 and ROC curves are shown in Figure 4.

Train Data	Test Data	LSTM Accuracy	LSTM AUC	RF Accuracy	RF AUC	Baseline Accuracy	Baseline AUC
Doves	**Doves**	0.76	0.68	**0.85**	**0.94**	0.55	0.50
TüEyeQ	**TüEyeQ**	0.63	**0.70**	**0.70**	**0.70**	0.51	0.50
Doves	**TüEyeQ**	**0.53**	**0.63**	0.50	0.61	0.51	0.50
TüEyeQ	**Doves**	0.47	**0.57**	**0.62**	**0.57**	0.54	0.50
Both	**Both**	0.70	0.68	**0.74**	**0.74**	0.52	0.50

Table 1: The performance metrics on both RF, LSTM and Baseline on all dataset combination. Best performing highest metric on each dataset combination is highlighted in bold.

The random forest has similar of best performing results on all combinations in all metrics, except when training on Doves and testing on TüEyeQ, where the LSTM had a slightly better performance. However, both models were better than the baseline in the majority of trials.

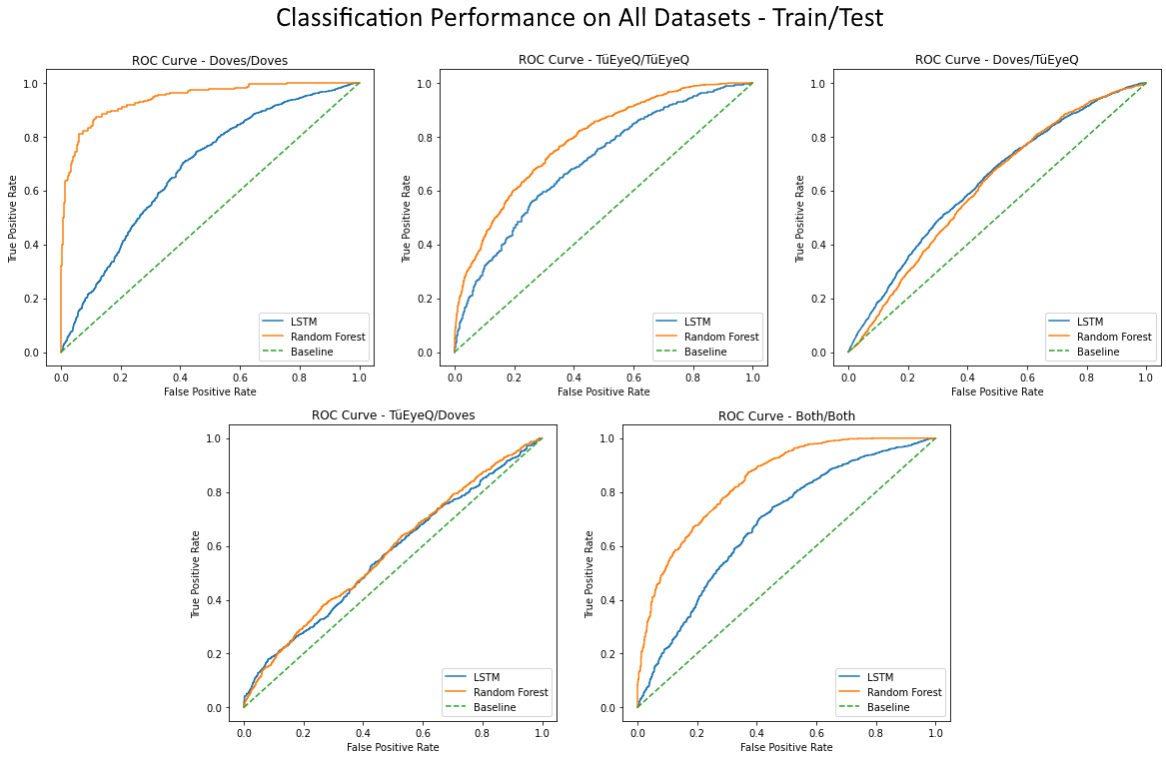


Figure 4: ROC curves for all combinations and models.

5.1 Feature Evaluation

To evaluate on the features we employed in our models, we employed Scikit-Learn's feature importance method, which returns a ranking of features used in a model, based on how much they contribute to the classification task. To get an overview of what the top performers were, we evaluated on our full extracted feature set of the TüEyeQ dataset, which amounted to 75 different features. 9 of the 15 top performers were related to either microsaccades or pupil dilation, indicating that these areas contribute greatly to gender classification. Furthermore, we tested several features ad hoc one by one, including the full list of x and y eye coordinates, which proved to contribute greatly to the classification task. However, as exact coordinates on the screen do not generalize well to other setups and stimuli, they were dropped from the suite of features. These evaluations were found using the random forest, as the LSTM used all available features.

6 Discussion

Our results show that it was possible to distinguish between the two genders using data gathered in different settings, i.e. during IQ testing and free viewing of nature images.

We found that specific features had a greater effect on the accuracy and therefore infer that these features varies more between genders and are task independent. Specifically the microsaccadic movements of the eye and pupil dilation affected the model positively, as implied by previous research in the field [18, 4]. The total fixation duration to total saccade duration ratio used by Sargezeh et. al. [15] did not make much of a difference in classification.

The classifiers did not achieve equally good results on both datasets. Differences in accuracy related to the task has been seen in previous work [12, 15, 1], but it has not yet been measured on two different datasets using the same classifiers. The results show significantly better results on Doves using both RF and LSTM classifiers, with RF and LSTM respectively having an accuracy of *0.85* and *0.76* on Doves and an accuracy of *0.70* and *0.63* on TüEyeQ when both training and testing on the same dataset. This finding supports that differences seen in male and female visual attention patterns vary more between image viewing than during IQ tests. Studies made by Pouretemad et

al. and Sargezeh et al. [12, 15] already supported a difference between viewing human faces and passive viewing of indoor images. The difference in the results between Doves and TüEyeQ might be explained as a passive versus active viewing task, as the IQ task is categorized as an active task. Existing research suggests that active viewing tasks generally should increase the differences in fixations between men and women compared to a passive viewing task [12, 5]. If that were to be true, it would be expected that the classifiers performed better on TüEyeQ, which is not the case. Therefore, it is suspected that the differences is due to the content that is seen. This finding could support further research in establishing to what extent the difference in visual attention between genders is related to the stimulus.

The robustness of the models were also tested having them perform independent classification, where the models were trained and tested on different datasets. As expected, the results from these experiments showed the classifiers had limited ability to train on data from one dataset and then test on the other. The existing results on individual task independent classification achieved by Kinnunen et al. [8] and Schröder et al. [16], showed the same patterns with accuracies slightly above chance but well below the results seen when the models were tested an trained on the same data. The findings of our study show accuracies are slightly above chance with random forest having an acurracy of *0.62* as the best result. As seen in the study by Schöder et al. [16], the transfer is asymmetrical, (i.e. training on one dataset and testing on another will not necessarily the give the same results if the datasets are swapped) between the two dataset. Random forest and LSTM trained on TüEyeQ and tested on Doves did not achieve the same AUC as when trained on Doves and tested on TüEyeQ. Our approach faced the same issues as previous attempts aiming to predict individuals across different tasks [8, 16]. Also of note is the fact that some of these tests resulted in extremely low recall or precision such as the random forest trained on Doves and tested on TüeyeQ (recall 0.03, precision 0.48). This means that in some cases the model most likely just learned to guess a single class most of the time. In this case, a respectable accuracy is achieved, but in practice the model is useless.

Combining the dataset improved the classification significantly with random forest having an accuracy of *0.74*. The accuracy achieved when combining the two dataset was slightly better than the accuracy achieved when training and testing on TüEyeQ alone. The reason perhaps being, that Doves contribute more to the combination. Further research could be relevant to clarify if the performance when combining dataset would continue to improve, when adding more unique dataset or at least not decrease.

6.0.1 Ethical Considerations

With eye tracking technology becoming better and webcam-based trackers emerging, the ability to predict demographic information about a user is worth considering. Cameras facing the user are in all modern devices, including laptops, smartphones and virtual reality (VR) headsets. Massive quantities of data can quickly be gathered when tracking people while they use technology in their everyday life. This would result in a database consisting of eye tracking from a multitude of different tasks, similarly to our last experiment, combining the two datasets. And as we conclude, combining this data makes the generalization more robust, allowing for a potentially privacy-invasive tool.

The solution to this issue lies in the hand of manufacturers and is already present is some devices. Most laptop webcams had a light turning on, indicating when the webcam is being used and apps on smartphones need permission to access cameras. The solution may also lie within the eye tracking software, where the readings will be altered to prevent user identification[17]. However, with Facebook being a large manufacturer of VR with eye tracking, while also being a significant advertising firm, it is worth considering that eye movement can reveal more about someone than commonly believed.

7 Conclusion

This study investigated different approaches in the area of gender classification using eye movements. One of the surprising findings came from a comparison of training and testing on two dataset of participants experiencing different stimuli: IQ tests and free viewing of images. We found significant differences in the accuracy achieved on the two datasets using a random forest and an LSTM as classifiers. The random forest performed best with accuracies of *0.85* on passive nature viewing dataset and *0.70* on the IQ dataset. Furthermore, we investigated the effect of using micro movements in the model, which expectedly improved our classifier performances. Lastly, we tested the classifiers'

ability to perform task independent predictions, using one dataset for training and another for testing. This approach resulted in prediction accuracies only slightly above chance, using both random forest and LSTM. However combining the dataset for training and testing provided much better results with the random forest having an overall accuracy on the combined dataset of *0.74*.

References

[1] A. Darwish and M. Pasquier. Biometric identification using the dynamic features of the eyes. In *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–6, 2013.

[2] J. H. v. Hateren and A. v. d. Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings: Biological Sciences*, 265(1394):359–366, Mar 1998.

[3] Y. M. Hwang and K. C. Lee. Using an eye-tracking approach to explore gender differences in visual attention and shopping attitudes in an online shopping environment. *International Journal of Human–Computer Interaction*, 34(1):15–24, 2018.

[4] L. Jaeger, I. Makowski, P. Prasse, S. Liehr, M. Seidler, and T. Scheffer. Deep eyedentification: Biometric identification using micro-movements of the eye. *CoRR*, abs/1906.11889, 2019.

[5] H. JJ, P. MM, and S. DI. Females scan more than males: a potential mechanism for sex differences in recognition memory. *Psychol Sci.*, 7:1157–63, 2013.

[6] R. Karsh and F. W. Breitenbach. Looking at looking: The amorphous fixation measure. pages 53–64, 1983.

[7] E. Kasneci, G. Kasneci, T. Appel, J. Haug, F. Wortha, M. Tibus, U. Trautwein, and P. Gerjets. Tüeyeq, a rich iq test performance data set with eye movement, educational and sociodemographic information. *Sci Data*, 8(154), 2021.

[8] T. Kinnunen, F. Sedlak, and R. Bednarik. Towards task-independent person authentication using eye movement signals. 2010.

[9] I. V. D. Linde, U. Rajashekar, A. C. Bovik, and L. K. Cormack. Doves: a database of visual eye movements. *Spatial vision*, 22(2):161–177, 2009.

[10] P. Merritt, E. Hirshman, W. Wharton, B. Stangl, J. Devlin, and A. Lenz. Evidence for gender differences in visual selective attention. *Personality and individual differences*, 43(3):597–609, 2007.

[11] J. Otero-Millan, J. L. A. Castro, S. L. Macknik, and S. Martinez-Conde. Unsupervised Clustering Method to Detect Microsaccades. *Journal of Vision*, 14(2):18–18, 02 2014.

[12] H. Pouretemad, C. Eslahchi, and A. Salahirad. Gender classification based on eye movements: A processing effect during passive face viewing. *Advances in Cognitive Psychology*, (13):232–240, September 2017.

[13] I. Rigas and O. V. Komogortsev. Current research in eye movement biometrics: An analysis based on bioeye 2015 competition. *Image and Vision Computing*, 58:129–141, 2017.

[14] D. D. Salvucci and J. H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. *ETRA '00: Proceedings of the 2000 symposium on Eye tracking research applications*, pages 71–78, November 2000.

[15] B. A. Sargezeh, N. Tavakoli, and M. R. Daliri. Gender-based eye movement differences in passive indoor picture viewing: An eye-tracking study. *Physiology Behavior*, 206:43–50, 2019.

[16] C. Schröder, A. Zaidawi, S. M. Klim, M. H. Prinzler, S. Maneth, and G. Zachmann. Robustness of eye movement biometrics against varying stimuli and varying trajectory length. pages 1–7, 2020.

[17] J. Steil, I. Hagestedt, M. X. Huang, and A. Bulling. Privacy-aware eye tracking using differential privacy. *In 2019 Sym- posium on Eye Tracking Research and Applications (ETRA '19)*, 27:1–9, 2019.

[18] P. Thapak, A. K. Shrivastava, T. Bhopal, and T. Bhopal. A high sensitive approach for gender prediction by using pupil dilation. *Journal of Global Research in Computer Sciences*, (9):23–27, September 2012.

A Random Forest Compared to Other Models

Random Forest is often a good choice when choosing a model. However, to make sure of the choice, the RF was compared to a support vector machine (SVM) and logistic regression. Below are the ROC curves for different dataset combinations.

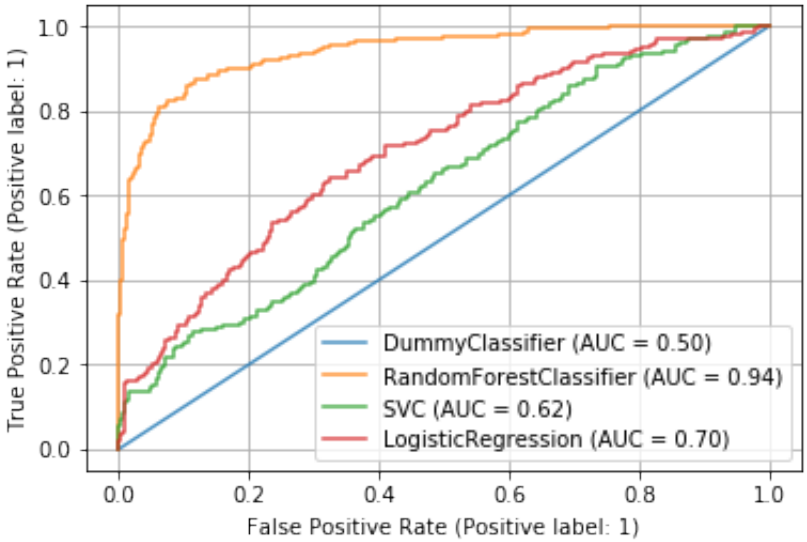


Figure 5: ROC curve for gender classification on the Doves dataset

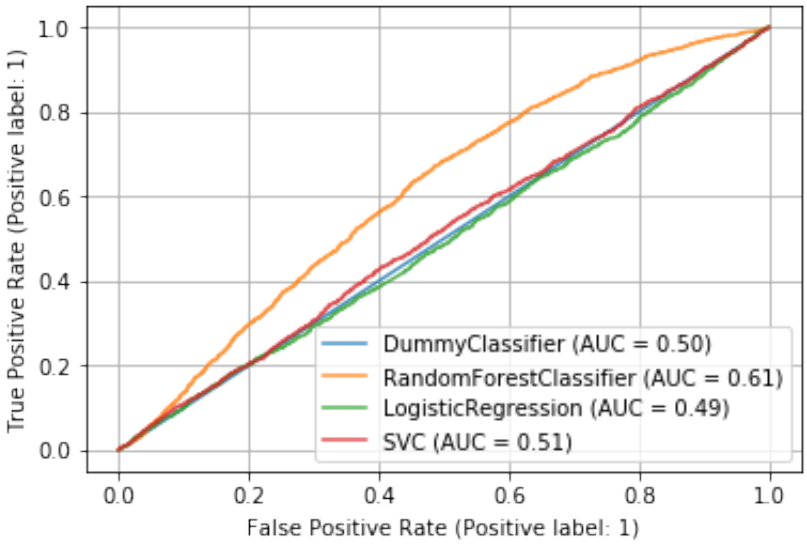


Figure 6: ROC curve for training on Doves and testing on TüEyeQ

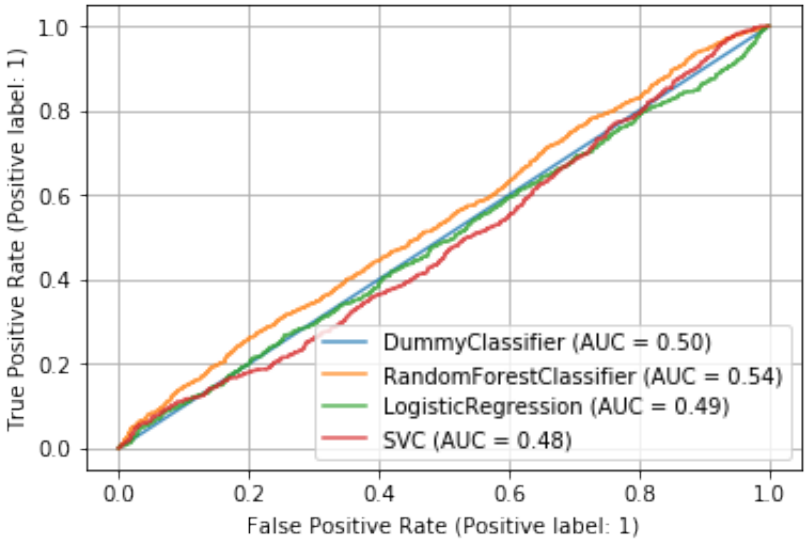


Figure 7: ROC curve for training on TüEYeQ and testing on Doves.