Eye Movements as Indicators of Deception: A Machine Learning Approach

Valentin Foucher*
General Psychology
Ulm University
Ulm, Baden-Württemberg, Germany
valentin.foucher@uni-ulm.de

Santiago de Leon-Martinez†*
Kempelen Institute of Intelligent Technologies
Bratislava, Slovakia
Faculty of Information Technology
Brno University of Technology
Brno, Czech Republic
santiago.deleon@kinit.sk

Robert Moro
Kempelen Institute of Intelligent Technologies
Bratislava, Slovakia
robert.moro@kinit.sk

Abstract

Gaze may enhance the robustness of lie detectors but remains under-studied. This study evaluated the efficacy of AI models (using fixations, saccades, blinks, and pupil size) for detecting deception in Concealed Information Tests across two datasets. The first, collected with Eyelink 1000, contains gaze data from a computerized experiment where 87 participants revealed, concealed, or faked the value of a previously selected card. The second, collected with Pupil Neon, involved 36 participants performing a similar task but facing an experimenter. XGBoost achieved accuracies up to 74% in a binary classification task (Revealing vs. Concealing) and 49% in a more challenging three-classification task (Revealing vs. Concealing vs. Faking). Feature analysis identified saccade number, duration, amplitude, and maximum pupil size as the most important for deception prediction. These results demonstrate the feasibility of using gaze and AI to enhance lie detectors and encourage future research that may improve on this.

CCS Concepts

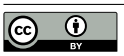
• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Classification and regression trees**; *Feature selection.*

Keywords

Eye Movements, Gaze, Pupil, Deception Detection, Concealed Information Test, Machine Learning, Feature Importance

ACM Reference Format:
Valentin Foucher, Santiago de Leon-Martinez, and Robert Moro. 2025. Eye Movements as Indicators of Deception: A Machine Learning Approach. In *2025 Symposium on Eye Tracking Research and Applications (ETRA '25), May 26–29, 2025, Tokyo, Japan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3715669.3723129

*Both authors contributed equally to the paper
†Also with Kempelen Institute of Intelligent Technologies.



This work is licensed under a Creative Commons Attribution 4.0 International License.

ETRA '25, Tokyo, Japan

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1487-0/25/05
https://doi.org/10.1145/3715669.3723129

1 Introduction

In a world where the lines between truth and lies are increasingly indistinct, the development of reliable deception detection has become paramount. Traditional methods of detecting deception have relied on the measurement of physiological signals, such as heart rate, respiration rate or skin conductance while asking crime-relevant questions [Raskin and Hare 1978]. However, these techniques are not only intrusive but also demonstrate low reliability and can easily be manipulated by experienced deceivers [Ben-Shakhar 2002]. A category of test called Concealed Information Test (CIT), emerged in crime contexts to undermine concealed knowledge [Ben-Shakhar and Elaad 2003]. This test aims to examine participants' reactions to the presentation of crucial information to detect unspoken information. Traditionally focused on response time, this test has been taking advantage of the pervasive accessibility and ease of use of eye-tracking devices to integrate eye movements with the objective of enhancing its resistance to countermeasures.

1.1 Related Works

Concealing a known face has been found to correlate with a decrease in the number of fixations on the concealed face along with an increase in fixation duration [Millen et al. 2017; Schwedes and Wentura 2012]. Similarly, fewer but longer fixations were observed on crime-related details when viewing images related to a mock crime [Peth et al. 2013], and fewer fixations were directed toward an accomplice when deceiving during online interview [Pak and Zhou 2013]. Deception would also influence saccades, with a higher saccade frequency and velocity when lying [Vrij et al. 2015], but smaller saccade amplitudes [Lim et al. 2013]. However, the direction of their effect remains ambiguous, and contradictory outcomes have been reported in two analogous questionnaire studies [Fang et al. 2021]. Research on the impact of deception on blinking behaviours consistently shows a decrease in the number of blinks during interviews or when concealing the recognition of a target stimulus compared to control stimuli [Fukuda 2001; Leal and Vrij 2008, 2010; Peth et al. 2013]. The impact of deception on blink duration remains ambiguous, with some evidence suggesting a decrease in maximum blink duration when lying in response to mock-crime-related questions versus responding truthfully [Marchak 2013]. Pupil size has also been demonstrated to be a potential indicator of deception. Research has shown that greater pupil dilation occurs in response to false statements made in contexts involving benign scenarios

[Dionisio et al. 2001] or when reading statements about a mock crime that participants committed [Cook et al. 2012; Webb et al. 2009]. Greater pupil dilation has also been observed in response to the presentation of mock-crime-related items [Lubow and Fein 1996; Proudfoot et al. 2015; Twyman et al. 2013], familiar faces [Seymour et al. 2013], personal names [Chen et al. 2023], or selected numbers [Foucher and Huckauf 2024] despite concealing them.

Several computational models have demonstrated the predictive contribution of facial and eye parameters in detecting lies from truths ([Prome et al. 2024] for a recent review) as evidenced by their ability to discriminate authentic from deceptive online reviews [Lai and Tan 2019; Ott et al. 2011] or truthful from deceptive statements in real-life trials [Pérez-Rosas et al. 2015]. Indeed, automated deception detection systems exposed the role of non-verbal behaviours, including eye movements, during interviews [O'Shea et al. 2018]. Eye movements appeared to be even more effective at predicting deception than facial micro-movements with an accuracy of 78% [Khan et al. 2021], reporting blinks as the strongest indicator of deception and confirming the results of a previous study that developed a new blink rate metric to differentiate lies from truths [Borza et al. 2018]. Saccades however were not found of importance in their model. Fixations and pupil size could also predict deception in questionnaire surveys with an accuracy of 74% [Fang et al. 2021].

1.2 Present Study

These models sought to differentiate lies from truths. However, to our knowledge, no predictive models have attempted to detect the concealment of information, which would be particularly advantageous in the context of the CIT. Furthermore, concealing or faking information has been shown to induce different eye movement behaviours [Foucher and Huckauf 2024] and may be identified differently in deception recognition systems. The present study aims to fill this gap by evaluating the efficacy of various eye parameters in predicting deception in CIT across two datasets. In both datasets, participants were instructed to select a card value between 1 and 6, and were then presented with all the possible card values on a screen. Participants were instructed to employ only their eyes to either reveal, conceal, or fake the value of their card to their interlocutor. The primary objective of this study was to ascertain the extent to which eye parameters can predict deceptive intentions. The second objective was to determine which eye parameters are the best suited for such prediction. A feature analysis was performed to assess the importance of each feature in the prediction of the deceptive intentions.

2 Experimental Methods

The prediction of deceptive intentions was tested in two different contexts. The Eyelink dataset (collected by the Eyelink 1000) comprises an experiment involving participants deceiving in front of a computer with a chinrest. The Neon dataset (collected by the Pupil Neon) corresponds to an experiment involving participants deceiving an experimenter. The key differences between the two contexts are: 1) the target of the deception (who the participants are lying to) is either a computer or a physical person and 2) the Pupil Neon is a wearable eye tracker that allows much more freedom of movement of the head as compared to a chinrest and was used to improve the ecological validity of the design.

2.1 Design

Eyelink Dataset. Eyelink dataset comprises an experiment in which participants were instructed to play a homemade card game with an interlocutor who could be a partner, an opponent, or a referee, represented on a computer by a partner smiley, an opponent smiley and a referee smiley respectively. In this game, participants began a session with six blue UNO cards, numbered from 1 to 6, placed randomly side by side in front of them. After selecting a number – called the target – participants were informed about the interlocutor with whom they would be interacting. Pairs of numerical cards were then displayed on the screen. The participants' objective was to interact with the interlocutor regarding the target using only their eyes, with the interlocutors described as algorithm-driven systems designed to read the participants' eye movements. There were three types of instructions: (1) *Revealing the truth* is the interaction with the partner: participants have to disclose the target to them. For instance, if they selected the number '2', they want the partner interlocutor to realise that they selected this number. (2) *Faking the truth* is the interaction with the opponent: participants have to mislead them into guessing another card than the target. For instance, if they selected the number '2', they want the opponent interlocutor to guess another number, like the number '5'. The choice of this fake number was up to the participant. (3) *Concealing the truth* is the interaction with the referee: participants have to prevent them from guessing the target. For instance, if they selected the number '2', they do not want the referee interlocutor to realise that they selected this number.

Thirty pairs of numbers were presented 5s each with pseudo-randomized positions and orders. Each number appeared ten times. A fixation cross accompanied by a pair of neutral cards was presented between each pair of numbers for 3s to allow the pupil size to return to a baseline level. Nine sessions comprised interactions with each of the three randomly ordered interlocutors, leading to the presentation of 270 pairs in total. Each session started with an eye-tracker calibration during which participants were instructed to follow a dot's movement from the centre of the screen to six predetermined positions along the borders. The requirement to communicate exclusively through eye movements was implemented to minimise body movements during the experiment, while ensuring participants maintained eye contact with the screen. No guidelines were provided concerning the means to achieve the desired behaviour leaving the participants to develop their own strategies for conveying the intended information to their interlocutor.

Neon Dataset. The Neon dataset constitutes an experiment that is analogous to the one presented in the Eyelink dataset, with the exception that participants engaged in a game against an experimenter seated in front of them, who poses as a mentalist as opposed to a computer. A screen was positioned centrally on the table to present visual stimuli, and a computer was situated adjacent to the participants to collect their responses. At the beginning of each round, participants were presented with six UNO cards numbered from 1 to 6 on the screen and were required to select one as the

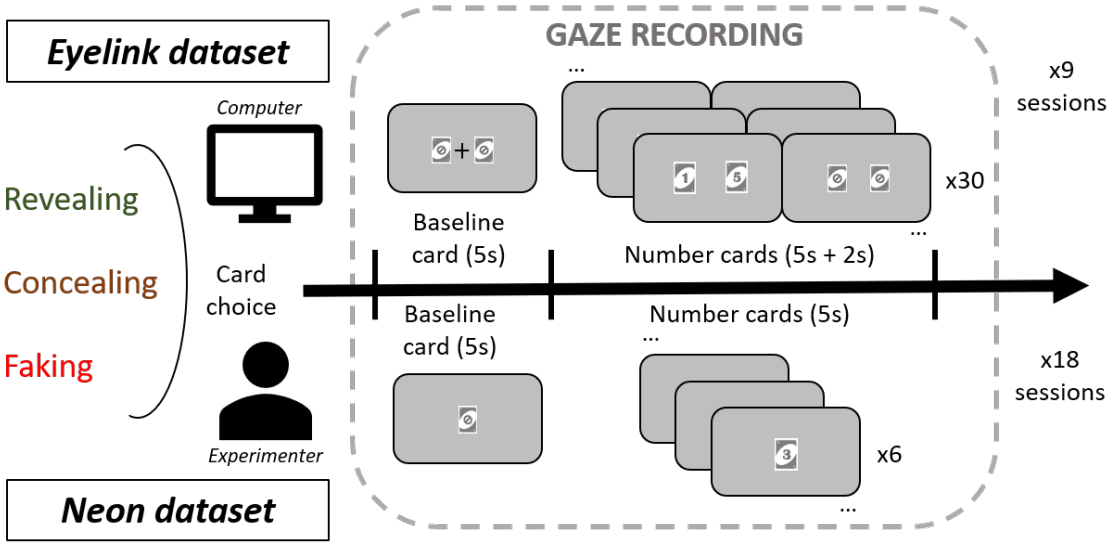


Figure 1: Experiments from both datasets contain the selection of a target, followed by the presentation of a baseline card and the number cards one after another for 5s each. In the Eyelink dataset, the number cards are separated by baseline cards for 2s. Participants are instructed to use their eyes to reveal, conceal, or fake the value of their target to either a computer or an experimenter.

target. Thereafter, a baseline card was displayed for 5 seconds, followed by the presentation of all six number cards in a random order, one after another, at the centre of the screen for 5 seconds each. As in the preceding experiments, the objective for the participants was to interact with the experimenter regarding the value of their target just by using their eyes. The three interactions from the previous experiments could occur, also referred to as "Revealing", "Faking", and "Concealing" the truth, except that the interlocutor was the experimenter who behaved as either a partner, an opponent or a referee. The experimenter was unaware of the value of the target and the interaction participants were performing. Participants were instructed to communicate with the experimenter according to their designated interaction's instruction employing only eye movements, without any requirement to look at any specific location or object. Eighteen sessions comprised the three balanced interaction instructions, leading to the presentation of 108 cards in total. It is noteworthy that no calibration was necessary to record data with the Neon system. The experimental design from both datasets is illustrated in Fig. 1.

2.2 Apparatus & Stimuli

In the Eyelink dataset, eye movements were measured using the EyeLink 1000 from SR Research Ltd with a 1000Hz sample frequency. The setup consisted of one table supporting a screen and the eye-tracker. Participants were seated at a distance of 57cm from a screen in a room with low ambient light, with their heads on a chin rest to prevent noise movements. The instructions were white-coloured and presented against a grey background on a display monitor (54", 1920x1080px, 100Hz). The interaction instructions given to participants regarding their interlocutor's role were conveyed through the utilisation of emoticons (e.g. a green smiley face indicated the partner, an orange angry face indicated the opponent and a yellow neutral face indicated the referee). The use of emoticons in the instructions was not part of the eye-tracking recordings.

In the Neon dataset, eye movements were recorded using the Neon glasses from Pupil Labs GmbH with a 200Hz sample frequency and the "Just act natural" frame. Neon glasses have the specific property of not requiring calibration to record gaze data. The participant and the experimenter faced each other at a distance of 100cm, in a room with ambient lighting. Between them, a screen (31x17.5cm, 1920x1080px, 60Hz) was placed horizontally on the table at a distance of 50cm from both the participant and the experimenter, ensuring that the stimuli material for the experiment remained constantly visible.

In both datasets, UNO cards were used due to their unambiguous design which facilitated immediate identification of the number. The numbers 1 to 6 were used, with the Skip card serving as the baseline. The numeric UNO cards displayed on the screen were created by numeric replicas transposed in grey and white colour with simple pixel discrimination, and displayed on a grey background to minimize the differences in brightness distribution in the stimuli. In the Eyelink dataset, the area of interest associated with each card was defined with an additional degree of visual angle around the borders of the card, defining a rectangle of 7.9 x 6.1 degrees of visual angle. In the Neon dataset, the area of interest associated with each card was defined using Pupil Labs markers at the tablet's corners.

2.3 Participants

The experiments were approved by the Ethics Committee of Ulm University. Prior to participation, all participants were informed of the option to discontinue their involvement at any stage and

provided written consent. For the Eyelink dataset, the study was publicised using the local university study management system and participants received credit hours as compensation for their participation. For the Neon dataset, the study was publicised through an external recruitment agency and participants received a 20€ voucher as compensation for their participation. The prerequisites for participation included in both cases normal or corrected-to-normal vision and the absence of epileptic seizures. The Eyelink dataset contains 87 participants, including 56 females (M=22.1yo, sd=2.1yo). The Neon dataset contains 36 participants, including 21 females (M=34.3 yo, sd=8.8).

2.4 Preparation of the Data

The datasets contain eye movement data that was exported from the Eyelink Data Viewer software package (Eyelink dataset) [SR Research Ltd. 2022] and the Pupil Labs Cloud (Neon dataset), and were preprocessed on R [R Core Team 2023]. Data is organised by participants and trials, with each trial corresponding to the presentation of a card. To capture meaningful changes, only trials containing a target were considered. To remove artefacts in the eye movements' definition, usual thresholds were used: fixations shorter than 60ms were removed, as blinks shorter than 60ms and longer than 700ms, and saccades shorter than 15ms and longer than 400ms [Foucher and Huckauf 2024; Rayner 1998]. Fixations longer than 5s were also removed since cards were presented for 5s each. Following recent pupillometry recommendations [Mathôt and Vilotijević 2022], pupil information was computed from the gaze located on the cards during the first 2.5s after the card presentation, including a cubic interpolation on the Eyelink data to account for missing data due to blinks. Baselines were defined by the 50ms before the presentation of each card, enabling the calculation of a baseline-corrected pupil signal. Outliers with an absolute Z score above 3 were excluded. Data was finally down-sampled from 1000Hz to 20Hz by averaging pupil size in 50ms-bins to capture meaningful size variations.

As the literature does not provide clear evidence that would rule out a specific eye movement, a comprehensive array of fundamental first-order features that characterize all these eye movements was tested. The resulting dataset for each trial and participant comprised the following features: fixation number and duration, saccade number, duration and amplitude, blink number and duration, mean, maximum and minimum pupil size, and the mean pupil size over 2500ms in 50ms-windows separated into fifty bins. Consequently, 60 eye parameters were investigated and could be separated into 7 eye-movement-related features and 53 pupil-related features.

3 Prediction of Deceptive Intentions

Since the experiments are Concealed Information Tests, originally designed to recognize the concealing of information, we perform the following predictive tasks: 1) predict the trials where the participants concealed the value of the target, compared to when they revealed it to the interlocutor, and 2) consider a third type of deceptive intention by predicting the trials where the participants concealed, faked or revealed the value of the target to the interlocutor. Therefore, task 1 is a binary classification between Concealing and Revealing the truth, while task 2 is 3-class classification between Concealing, Faking, and Revealing the truth. The XGBoost classifier was used for all predictive tasks.

3.1 Datasets

Target Distribution. The Eyelink dataset contained the following instances of the following classes: 1996 Concealing, 2065 Faking, and 1996 Revealing. The Neon dataset contained the following instances of the following classes: 160 Concealing , 161 Faking, and 181 Revealing. Due to the reasonable balance of the data, in both the binary classification and 3-class classification no sample or loss weighting/balancing was used for the predictive models.

Feature Datasets. To allow a better understanding of which parameters benefit the performances of the model, and to account for the bigger proportion of the pupil size in the number of parameters, the features were separated into three groups: 1) one including only eye movements (fixations count & duration, blink count & duration, saccade count & duration & amplitude), 2) one including only pupil size (mean, maximum, minimum, fifty time-bins of 50ms), and 3) one including all the eye parameters. Both the Eyelink and Neon datasets were separated into the 3 separate feature datasets for resulting 6 datasets. Each of these were then used for both predictive tasks resulting in 12 conditions, each with 5 replications with different train/test splits. This resulted in 60 trained and tuned final models and a total of 150,600 trained models across all conditions and replications (including cross validation). All conditions and their replications follow the same methodology (splitting, hyperparameter tuning, and evaluation) within and across datasets.

3.2 Model and Training Methodology

Dataset Splitting and Training. Datasets were first split into 5-folds stratified by participant ID (all data of one participant was guaranteed to be in only one fold) of roughly 20% each. These provide 5 different replications (per each of the 12 conditions) each using one of the 5 different test folds (20%) and the remaining data for train and validation (80%). Therefore, 5 different final tuned models were trained and evaluated on 5 different train/validation and test sets for one dataset, where one of these 5 we call a replication.

After a train/validation and test set were fixed, the train/validation dataset was split (stratified by participant ID) into a train and parameter validation set (70%) and early stopping (for # of trees) validation set (10%). Finally, the train and parameter validation set (70%) was split into stratified 5-folds for cross validation of hyperparameters, resulting in a train set (56%) and parameter validation set (14%). Each combination of hyperparameters were trained and evaluated on the 5 different folds and averaged across the 5 to determine the best hyperparameter set. For efficiency, early stopping (for # of trees) was also performed for each fold, but was evaluated on the same parameter validation set and only used to find the best score for that fold for a selection of parameters and later discarded (instead of being added to the best hyperparameter set).

Once the best hyperparameter set was found, a new model was trained on the train and parameter validation set (70%) using those parameters. Then the early stopping validation set (10%) was used to determine the best number of trees. With the best hyperparameter

set and the best number of trees found from early stopping, a final model was trained on the whole train/validation set (80%) and then evaluated on the left out test set (20%).

Evaluation Metrics. Accuracy was used to determine the best hyperparameter set by averaging accuracy scores of the 5 cross validation splits. Logloss (mlogloss in 3-class) was used for finding the best iteration in all cases of early stopping. For test evaluation, we report the average of the 5 replication accuracy scores achieved on the test sets along with standard deviation.

Model: XGBoost Classifier. The model used for all conditions was the XGBoost classifier [Chen and Guestrin 2016]. XGBoost is a gradient boosting algorithm that builds ensembles of decision trees that predict the appropriate class. It is well known inside and outside of the machine learning community for its performance and efficiency on tabular data, compared to other approaches. Like other decision tree models it does not require standardization of features and provides methods for measuring the importance of features in predicting a target. Initially, we also used tabNet [Arik and Pfister 2020], an explainable deep learning model for tabular data, but abandoned it due to poor results when compared to XGBoost.

Hyperparameters. All models were set to have a maximum of 10,000 trees (parameter: n estimators) with 35 early stopping rounds except for the final model, which used the best number of trees from the early stopping validation set and did not have early stopping. For hyperparameter selection, 100 combinations were sampled from the following hyperparameters and their distributions: learning rate (uniform from 0.001-0.191), subsample (uniform from 0.5-1.0), colsample by tree (uniform from 0.5-1.0), min child weight (1, 3, 5, 7), max depth (0, 3-15), alpha (0, 0.01, 1, 2, 5, 7, 10, 50, 100), lambda (0, 0.01, 1, 2, 5, 7, 10, 50, 100).

3.3 Results of Classification and Feature Importance

The results are shown in Table 1. As a baseline, we use a dummy model that predicts the proportion of the majority of classes found in the whole dataset rather than reporting the proportion for each train and test replication that varies only slightly. After evaluation of all 60 final models (5 replications of each 12 conditions), a feature importance analysis for each model was performed. We use a popular game theoretic approach of Shapley (SHAP) values to determine the importance of each feature in explaining the prediction [Shapley 1953]. For each data point in the test set, a SHAP value is calculated for each feature representing the contribution of the feature in prediction. To find the importance of a particular feature in predicting the output, the values are summarized by taking the absolute value and the mean for each feature. The values themselves cannot be compared across models (without normalization or other processing). In order to compare the 60 final models and determine shared important features, we take the top 5 nonzero features for each model and show their occurrence across the 5 replications for each of the 12 conditions.

4 Discussion

As shown in Table 1, the prediction of Concealing and Revealing tasks achieved an accuracy of 74% on the Eyelink dataset from the full features, and 64% on the Neon dataset. These results were considerably above the chance level, thereby proving evidence that the eyes can be used as indicators of deceptive intentions. The enhanced accuracy observed in Eyelink dataset could be attributed to an expanded sample size, the use of a chinrest, and a higher sample frequency of the Eyelink than the Neon which should result in a more accurate gaze estimation. However, the Neon dataset exhibited a satisfactory level of accuracy, despite employing a smaller sample size and using gaze data recorded from mobile eye-trackers, which are supposedly associated with diminished data quality. Interestingly, the Eyelink dataset exhibited higher accuracy with eye-movement features, while the Neon dataset demonstrated a stronger association with pupil features. This discrepancy can be interpreted in the context of the disparity in experimental design between the two datasets. The presentation of two cards simultaneously in Eyelink dataset, as opposed to the presentation of a single card in Neon dataset, could have led to an increased number of eye movements in the former. This may have resulted in less reliable pupil measures compared to the latter, where the eyes may have been more focused on the card's area of interest. Nevertheless, these results also serve to reinforce the reliability of mobile eye-trackers in terms of pupil detection.

The prediction of Concealing, Faking, and Revealing tasks achieved an accuracy of 49% on the Eyelink dataset from the full features and eye features, and 47% on the Neon dataset with just the pupil features. These results follow the same trend, in that only using eye-movement features to predict in the Neon dataset leads to poor results, while sole use of eye features leads to similar or better results than all features in the Eyelink dataset. The lower accuracy can be attributed to the increased complexity of predicting three classes compared to binary classification. However, previous research already showed that Revealing and Faking tasks in a CIT induce similar eye movements [Foucher and Huckauf 2024]. While AI enhances differentiation, achieving a clear distinction between these behaviours remains challenging.

Regarding the importance of the eye movement features, among the 60 investigated, saccade number, duration and amplitude consistently constituted the most important features across the two datasets with maximum pupil size being the only pupil feature shared by all feature models and pupil feature models. While it is plausible that other eye parameters may also have contributed to the model's accuracies, these findings serve to strengthen the role of saccades in the prediction of deception in the context of CIT, contradicting previous findings [Borza et al. 2018]. Additionally, pupil size had a role in deception prediction with a strong influence on the maximum, minimum and mean pupil size, as well as mean pupil size between 800ms and 1500ms after card presentation. This aligns with prior research [Fang et al. 2021] and underscores the potential of pupil size as a reliable indicator of deception.

These results show the potential of ocular indicators of deception in a study environment. In contrast to our study, lying usually serves a purpose. A higher stakes experimental design (with motivation for successful lies and/or consequences for lying) could increase emotional load and physiological parameters. Furthermore, the robustness of a lie detector system could be improved by coupling eye parameters with other physiological modalities such as brain activity [Rosenfeld 2020].

Table 1: Results for the 12 conditions, calculated by averaging the test accuracy of the final models for each of the 5 replications. Additionally, the top 5 nonzero features are shown for each of the 5 replications and their occurrence in parentheses. Note that the maximum occurrence possible is (5) meaning that every final model of all 5 replications found the feature to be important and we only show features that have (2) or more occurrences across the 5 replications.

Revealing / Concealing	Baseline Majority Class	All Features (n=60)		Eye-movement Features (n=7)		Pupil Features (n=53)	
		Avg Acc (std)	Top 5 Features (>1)	Avg Acc (std)	Top 5 features (>1)	Avg Acc (std)	Top 5 Features (>1)
Eyelink dataset	0.500	0.736 (0.015)	Saccade amplitude (5) Saccade number (5) Saccade duration (5) Max pupil size (3) Fixation duration (3)	0.733 (0.026)	Saccade amplitude (5) Saccade number (5) Saccade duration (5) Blink duration (3) Fixation duration (3) Fixation number (2) Blink number (2)	0.585 (0.024)	Min pupil size (5) Pupil window 801 (5) Max pupil size(4) Pupil window 1001 (3) Mean pupil size (2) Pupil window 2301 (2)
Neon dataset	0.531	0.638 (0.069)	Saccade amplitude (5) Max pupil size (5) Saccade duration (3) Pupil window 951 (2) Pupil window 1451 (2)	0.581 (0.033)	Saccade amplitude (5) Saccade duration (5) Fixation duration (4) Saccade number (3) Blink duration (3) Fixation number (2)	0.616 (0.054)	Max pupil size (5) Pupil window 1051 (3) Pupil window 351 (3) Mean pupil size (2) Pupil window 1251 (2)
Revealing / Concealing / Faking							
Eyelink dataset	0.341	0.485 (0.015)	Saccade amplitude (5) Saccade number (5) Saccade duration (5) Fixation duration (3) Pupil window 1 (2)	0.486 (0.020)	Saccade amplitude (5) Saccade number (5) Saccade duration (5) Fixation duration (5) Blink duration (4)	0.377 (0.017)	Min pupil size (5) Mean pupil size (5) Max pupil size (4) Pupil window 1001 (3) Pupil window 2451 (3) Pupil window 801 (2) Pupil window 751 (2)
Neon dataset	0.361	0.456 (0.064)	Saccade amplitude (5) Pupil window 1251 (5) Saccade duration (4) Max pupil size (3) Pupil window 951 (2)	0.413 (0.068)	Saccade amplitude (5) Saccade duration (5) Fixation duration (5) Blink duration (5) Saccade number (3)	0.469 (0.048)	Max pupil size (5) Pupil window 1251 (4) Pupil window 701 (3) Pupil window 901 (2) Pupil window 951 (2) Pupil window 1001 (2) Pupil window 1051 (2)

Ethics and privacy statement. As mentioned in Section 2.3, the experiments were approved by the Ethics Committee of Ulm University and all participants provided their written consent. All data were pseudonymised before being processed. While the findings of this study indicate the potential of incorporating gaze and AI into lie detection systems, it is important to acknowledge the limitations of these methods. They should not be relied upon as the sole basis for determining culpability, but rather be regarded as supplementary sources of information that provide guidance.

5 Conclusion

To conclude, our findings demonstrated that AI can effectively predict deception from gaze. Specifically, the detection of card concealment achieved accuracies ranging from 64% in the Neon dataset to 74% in the Eyelink dataset. The three-classification task proved more complex, with 49% in the Eyelink dataset and 46% in the Neon dataset. The feature selection analysis identified saccade number, duration, and amplitude along with maximum pupil size as the most important predictors of deception. These results imply the potential of integrating gaze analysis with AI to enhance lie detection methods and advocate for further research to refine and advance this approach.

Acknowledgments

This project has received funding from the European Union's Horizon Europe research and innovation funding programme under grant agreement No 101072410. The authors would like to thank Professor Anke Huckauf for participating in elaborating on the initial research idea, and Professor Maria Bielikova for support in the modeling process.

References

Sercan O. Arik and Tomas Pfister. 2020. TabNet: Attentive Interpretable Tabular Learning. arXiv:1908.07442 [cs.LG] https://arxiv.org/abs/1908.07442

Gershon Ben-Shakhar. 2002. A critical review of the Control Questions Test (CQT). *Handbook of polygraph testing* 103 (2002), 38. https://www.basicknowledge101.com/pdf/health/Control%20Questions%20Test.PDF

Gershon Ben-Shakhar and Eitan Elaad. 2003. The validity of psychophysiological detection of information with the Guilty Knowledge Test: A meta-analytic review. *Journal of Applied Psychology* 88, 1 (2003), 131–151. doi:10.1037/0021-9010.88.1.131 Place: US Publisher: American Psychological Association.

Diana Borza, Razvan Itu, and Radu Danescu. 2018. In the Eye of the Deceiver: Analyzing Eye Movements as a Cue to Deception. *Journal of Imaging* 4, 10 (Oct. 2018), 120. doi:10.3390/jimaging4100120 Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.

Ivory Y. Chen, Aytaç Karabay, Sebastiaan Mathôt, Howard Bowman, and Elkan G. Akyürek. 2023. Concealed identity information detection with pupillometry in rapid serial visual presentation. *Psychophysiology* 60, 1 (2023), e14155. doi:10.1111/psyp.14155 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/psyp.14155.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge*

Discovery and Data Mining (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 785–794. doi:10.1145/2939672.2939785

Anne E. Cook, Douglas J. Hacker, Andrea K. Webb, Dahvyn Osher, Sean D. Kristjansson, Dan J. Woltz, and John C. Kircher. 2012. Lyin' eyes: Ocular-motor measures of reading reveal deception. *Journal of Experimental Psychology: Applied* 18, 3 (2012), 301–313. doi:10.1037/a0028307 Place: US Publisher: American Psychological Association.

D. P. Dionisio, E. Granholm, W. A. Hillix, and W. F. Perrine. 2001. Differentiation of deception using pupillary responses as an index of cognitive processing. *Psychophysiology* 38, 2 (March 2001), 205–211.

Xinyue Fang, Yiteng Sun, Xinyi Zheng, Xinrong Wang, Xuemei Deng, and Mei Wang. 2021. Assessing Deception in Questionnaire Surveys With Eye-Tracking. *Frontiers in Psychology* 12 (Nov. 2021), 15. doi:10.3389/fpsyg.2021.774961 Publisher: Frontiers.

Valentin Foucher and Anke Huckauf. 2024. Unveiling Deceptive Intentions: Insights from Eye Movements and Pupil Size. *Proceedings of the ACM on Human-Computer Interaction* 8, ETRA (May 2024), 1–17. doi:10.1145/3655612

K. Fukuda. 2001. Eye blinks: new indices for the detection of deception. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology* 40, 3 (April 2001), 239–245. doi:10.1016/s0167-8760(00)00192-6

Wasiq Khan, Keeley Crockett, James O'Shea, Abir Hussain, and Bilal M. Khan. 2021. Deception in the eyes of deceiver: A computer vision and machine learning based automated deception detection. *Expert Systems with Applications* 169 (May 2021), 114341. doi:10.1016/j.eswa.2020.114341

Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 29–38. doi:10.1145/3287560.3287590

Sharon Leal and Aldert Vrij. 2008. Blinking during and after lying. *Journal of Nonverbal Behavior* 32 (2008), 187–194. https://idp.springer.com/authorize/casa?redirect_uri=https://link.springer.com/article/10.1007/s10919-008-0051-0&casa_token=IxCeB-5zTmsAAAAA:QWAd-Qk5o3djzGvLe6FAH_TOM98Tf2dUotsEqqUr2uFG4uX4ghp9FnAU24up_jLq5Aa9Rd2Zi5ib Publisher: Springer.

Sharon Leal and Aldert Vrij. 2010. The occurrence of eye blinks during a guilty knowledge test. *Psychology, Crime & Law* 16, 4 (May 2010), 349–357. doi:10.1080/10683160902776843 Publisher: Routledge.

Kai Keat Lim, Max Friedrich, Jenni Radun, and Kristiina Jokinen. 2013. Lying through the eyes: detecting lies through eye movements. In *Proceedings of the 6th workshop on Eye gaze in intelligent human machine interaction: gaze in multimodal interaction (GazeIn '13)*. Association for Computing Machinery, New York, NY, USA, 51–56. doi:10.1145/2535948.2535954

R. E. Lubow and Ofer Fein. 1996. Pupillary size in response to a visual guilty knowledge test: New technique for the detection of deception. *Journal of Experimental Psychology: Applied* 2, 2 (1996), 164–177. doi:10.1037/1076-898X.2.2.164 Place: US Publisher: American Psychological Association.

Frank M. Marchak. 2013. Detecting false intent using eye blink measures. *Frontiers in Psychology* 4 (2013), 736. doi:10.3389/fpsyg.2013.00736

Sebastiaan Mathôt and Ana Vilotijević. 2022. Methods in cognitive pupillometry: Design, preprocessing, and statistical analysis. *Behavior Research Methods* 55, 6 (Aug. 2022), 3055–3077. doi:10.3758/s13428-022-01957-7

Ailsa E. Millen, Lorraine Hope, Anne P. Hillstrom, and Aldert Vrij. 2017. Tracking the Truth: The Effect of Face Familiarity on Eye Fixations during Deception - Ailsa E. Millen, Lorraine Hope, Anne P. Hillstrom, Aldert Vrij, 2017. https://journals.sagepub.com/doi/10.1080/17470218.2016.1172093

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. doi:10.48550/arXiv.1107.4557 arXiv:1107.4557 [cs].

James O'Shea, Keeley Crockett, Wasiq Khan, Philippos Kindynis, Athos Antoniades, and Georgios Boultadakis. 2018. Intelligent Deception Detection through Machine Based Interviewing. In *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, Rio de Janeiro, Brazil, 1–8. doi:10.1109/IJCNN.2018.8489392 ISSN: 2161-4407.

Jinie Pak and Lina Zhou. 2013. Eye Gazing Behaviors in Online Deception. In *Proceedings of the Nineteenth Americas Conference on Information Systems*. AMCIS, Chicago, Illinois, US, 10.

Judith Peth, Johann S. C. Kim, and Matthias Gamer. 2013. Fixations and eye-blinks allow for detecting concealed crime related memories. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology* 88, 1 (April 2013), 96–103. doi:10.1016/j.ijpsycho.2013.03.003

Shanjita Akter Prome, Neethiahnanthan Ari Ragavan, Md Rafiqul Islam, David Asirvatham, and Anasuya Jegathevi Jegathesan. 2024. Deception detection using machine learning (ML) and deep learning (DL) techniques: A systematic review. *Natural Language Processing Journal* 6 (March 2024), 100057. doi:10.1016/j.nlp.2024.100057

Jeffrey G. Proudfoot, Jeffrey L. Jenkins, Judee K. Burgoon, and Jay F. Nunamaker. 2015. Deception is in the eye of the communicator: Investigating pupil diameter variations in automated deception detection interviews. In *2015 IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, Baltimore, MD, USA, 97–102. doi:10.1109/ISI.2015.7165946

Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. 2015. Deception Detection using Real-life Trial Data. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)*. Association for Computing Machinery, New York, NY, USA, 59–66. doi:10.1145/2818346.2820758

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

David C. Raskin and Robert D. Hare. 1978. Psychopathy and Detection of Deception In a Prison Population. *Psychophysiology* 15, 2 (March 1978), 126–136. doi:10.1111/j.1469-8986.1978.tb01348.x

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin* 124, 3 (1998), 372. Publisher: American Psychological Association.

J. Peter Rosenfeld. 2020. P300 in detecting concealed information and deception: A review. *Psychophysiology* 57, 7 (July 2020), e13362. doi:10.1111/psyp.13362

Charlotte Schwedes and Dirk Wentura. 2012. The revealing glance: eye gaze behavior to concealed information. *Memory & Cognition* 40, 4 (May 2012), 642–651. doi:10.3758/s13421-011-0173-1

Travis L. Seymour, Christopher A. Baker, and Joshua T. Gaunt. 2013. Combining blink, pupil, and response time measures in a concealed knowledge test. *Frontiers in Psychology* 3 (2013), 614. https://www.frontiersin.org/articles/10.3389/fpsyg.2012.00614/full Publisher: Frontiers Media SA.

Lloyd S Shapley. 1953. A Value for n-Person Games. In *Contributions to the Theory of Games II*, Harold W. Kuhn and Albert W. Tucker (Eds.). Princeton University Press, Princeton, 307–317.

SR Research Ltd. 2022. *EyeLink Data Viewer*.

Nathan Twyman, Ryan M. Schuetzler, Jeffrey Gainer Proudfoot, and Aaron Elkins. 2013. A systems approach to countermeasures in credibility assessment interviews. *Information Systems and Quantitative Analysis Faculty Proceedings & Presentations*. 15 (2013), 20. https://digitalcommons.unomaha.edu/cgi/viewcontent.cgi?article=1016&context=isqafacproc

Aldert Vrij, João Oliveira, Annie Hammond, and Howard Ehrlichman. 2015. Saccadic eye movement rate as a cue to deceit. *Journal of Applied Research in Memory and Cognition* 4, 1 (2015), 15–19. doi:10.1016/j.jarmac.2014.07.005 Place: Netherlands Publisher: Elsevier Science.

Andrea K. Webb, Douglas J. Hacker, Dahvyn Osher, Anne E. Cook, Dan J. Woltz, Sean Kristjansson, and John C. Kircher. 2009. Eye Movements and Pupil Size Reveal Deception in Computer Administered Questionnaires. In *Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience*, Dylan D. Schmorrow, Ivy V. Estabrooke, and Marc Grootjen (Eds.). Springer, Berlin, Heidelberg, 553–562. doi:10.1007/978-3-642-02812-0_64

Funded by the European Union