Reliable Driver Gaze Classification based on Conformal Prediction

Simone Dari[1,2], Eyke Hüllermeier[2]

[1]Safety Department, BMW Group
Knorrstraße 147, 80788 Munich

[2]Heinz Nixdorf Institute and Department for Computer Science,
Paderborn University
Pohlweg 51, 3098 Paderborn

E-Mail: simone.dari@bmw.de, eyke@upb.de

1 Introduction

Machine learning is increasingly used in practical applications that can be categorized as safety-critical, such as AI-assisted driving. In this context, we recently considered the problem of driver monitoring, which plays an essential part in avoiding accidents by warning the driver in time and shifting the driver's attention to the traffic scenery in critical situations [1]. This may apply for the different levels of automated driving, for take-over requests as well as for driving in manual mode. More specifically, we tackled the problem of predicting the driver's gazing direction. Distinguishing eight different regions, this problem can be formalized as a classification task, in which each region corresponds to a class (cf. Figure 1). We proposed a deep learning approach to predict gaze regions, which is based on informative features such as eye landmarks and head pose angles of the driver. Moreover, we introduced different post-processing techniques that improve the accuracy by exploiting temporal information from videos and the availability of other vehicle signals. Our main interest is to leverage accurate gaze prediction for improved human-computer-interaction. In this regard, it is arguably important to guarantee a certain level

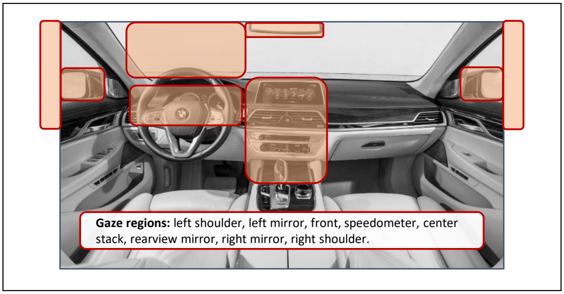

Gaze regions: left shoulder, left mirror, front, speedometer, center stack, rearview mirror, right mirror, right shoulder.

Figure 1: Spatial zones distinguished in the driver gaze classification task.

of awareness of the computer (AI system) of its own certainty or uncertainty in a prediction [3].

In this work, we therefore leverage so-called *conformal prediction* (CP) to increase the reliability of such predictions [10, 12]. Instead of predicting a single class, CP produces a *set-valued* prediction, i.e., a subset of all candidate classes that comprises the true class with high probability. This way, the system is able to express ambiguities (several regions appear plausible, because the driver's gaze cannot be determined precisely) as well as a partial or complete lack of knowledge — for example, the driver may look in a completely different direction, which does not correspond to any of the eight pre-specified regions (thereby producing so-called out-of-distribution data).

Conformal prediction can be seen as a meta-learning technique, which can be put on top of any base learner, i.e., any standard classifier producing "point predictions." It merely requires a measure of (non-)conformity of a (hypothetical) data point, i.e., a measure of how well a combination of feature values and gaze directions fits with the training data seen so far. While the required level of confidence — the predicted set contains the true class with a pre-specified probability (such as 95%) — is guaranteed regardless of the conformity measure, the latter has a strong influence on the precision of predictions, i.e., the (average) size of the predicted sets.

In this work, we evaluate different types of conformity scores to construct conformal predictors for driver gaze classification, including scores derived from kernel density estimation as proposed in [2, 5], and compare them with regard to the quality of set-valued predictions as well as time efficiency. Moreover, we elaborate on a specific characteristic of our problem, namely the fact that our output space exhibits a natural (topological) structure induced by the spatial relationship between the classes — unlike standard classification problems, where the classes constitute a simple set with no relationships between its elements. As a consequence, there are more meaningful (*viz.* topologically connected) and less meaningful (unconnected) set-valued predictions. To assure semantically meaningful set-valued predictions, we propose an extension of standard CP.

The work is structured as follows. In Section 2, we shortly review the gaze classification system with its results. Section 3 explains the conformal prediction method more closely. In Section 4, we apply the method to the gaze dataset. Section 5 discusses the results while Section 6 concludes with some final remarks.

2 Gaze Classification

The problem of driver monitoring was recently studied in [1]. This section briefly summarizes the method used and the key results obtained. For detailed information, the interested reader is referred to [1].

2.1 Problem Statement and Dataset

The goal of the gaze classifier is to reliably classify the region the driver is looking at, based on an image of the driver. Certain regions are of special interest and are displayed in Figure 1. The underlying dataset was extracted from a naturalistic driving study in which participants were driving a car for several months while being recorded with an RGB camera installed at the A-pillar. Sample images are provided in Figure 2. The examined dataset consists of 75 video snippets from 20 subjects (5 female, 15 males). Driver videos were recorded in size 980 $\times$ 540 at 15 frames per second. The important regions of interest (also *classes*, *labels*) with the number of images available are given in table 1.

Table 1: Number of classes

Class	Abbr.	Number
left shoulder	ls	98
left mirror	lm	967
speedometer	sp	228
front	f	2,296
inner mirror	inm	713
center console	cc	332
right shoulder	rs	72
right mirror	rm	356

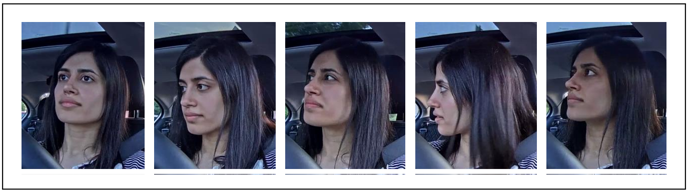


Figure 2: Example images from the dataset.

2.2 Method

The pipeline of the gaze classification system is depicted in Figure 3. In a pre-processing step, meaningful features on the driver's head pose and the eyes are generated from existing image-based methods and then fed into a fully connected neural net. For an inserted image, the driver's face is detected [4] and the three head pose angles are computed [8]. The angles describe the orientation of the head, where the rotation around the x-axis is called *pitch* (i.e. from up to down), around the y-axis *yaw* (i.e. from left to right), and around the z-axis *roll* (i.e. from left to right shoulder). For the eyes, the eye landmark detector by Park et al. [7] is employed. We make use of 15 landmarks per eye that describe the eyelid and the iris. The generated features are fed as input to a neural network architecture. The output of this network is scaled by the softmax-activation function, which produces a vector with probabilities for each class.

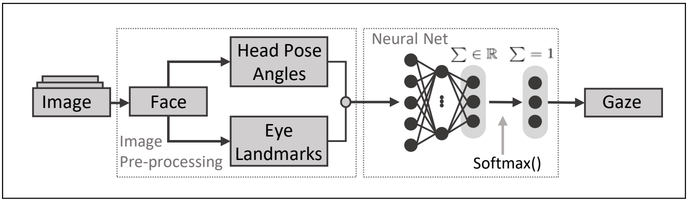
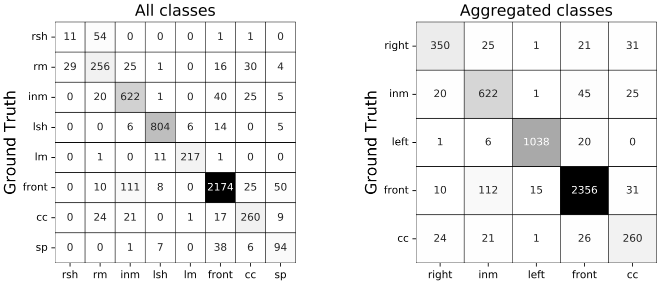


Figure 3: Pipeline of the Gaze classification system.


All classes

Ground Truth	rsh	rm	inm	lsh	lm	front	cc	sp
rsh	11	54	0	0	0	1	1	0
rm	29	256	25	1	0	16	30	4
inm	0	20	622	1	0	40	25	5
lsh	0	0	6	804	6	14	0	5
lm	0	1	0	11	217	1	0	0
front	0	10	111	8	0	2174	25	50
cc	0	24	21	0	1	17	260	9
sp	0	0	1	7	0	38	6	94

Aggregated classes

Ground Truth	right	inm	left	front	cc
right	350	25	1	21	31
inm	20	622	1	45	25
left	1	6	1038	20	0
front	10	112	15	2356	31
cc	24	21	1	26	260

Figure 4: Results after Cross-validation [1].

2.3 Results

As there might be driver-dependent characteristics in the features, we propose training with a leave one-driver out cross-validation, training on 19 drivers, while testing on the remaining driver. The results from the test set after every iteration are aggregated into the confusion matrix given in Figure 4. In total, the model achieves an accuracy of 87.1%. After aggregating the classes from the left and the right side, as well as the *speedometer* with the *front* class, accuracy increases to 91.4%. Misclassification occurs for the classes *front* and *inner mirror*, as well as for *inner mirror*, *front* and the *right side*.

2.4 Discussion

In general, the error rate of 12.9% can be narrowed down to three types of misclassifications: (i) misclassifications between similar classes that can be aggregated together without a higher loss of information (e.g., *right shoulder* and *right mirror*) (4.3%), (ii) misclassifications between classes far apart (e.g., *left mirror* and *right mirror*) which make up 1.4% and (iii) misclassifications among classes close to one another. Indeed, there is a high number of misclassifications for the classes *front*, *inner mirror*, *right mirror* and *center console*. One can possibly assume that classes away from the camera are harder to perceive. If the driver is looking through the front windshield and directly beneath the inner mirror, e.g., while focusing on a vehicle far ahead on the right side, it becomes difficult from the camera point of view to correctly annotate this situation, for both, the human annotator and apparently also the system.

3 Conformal Prediction

In cases of uncertainty, set-valued predictions are supposed to produce reliable predictions, i.e., subsets of classes comprising the true one with high probability, very much like confidence intervals as known from classical statistics. One way of obtaining such sets is through *conformal prediction* [10, 12], which is based on the idea of reducing prediction to hypothesis testing: Given a query instance, a class label is included as a candidate in the set-valued prediction unless the hypothesis that this label corresponds to the ground truth can be rejected at a pre-specified level of confidence. The test itself relies on assigning each instance/label combination a measure of *non-conformity*, reflecting how “strange” this combination appears in light of the data seen so far. Its counterpart is the *conformity* measure, reflecting the similarity to the data seen so far.

The original idea of CP was introduced for the setting of online learning [12]. Here, we present a version adapted to the standard setting of supervised learning, called *Inductive Conformal Prediction* (ICP) [6]. We only focus on the case of classification, for which we have seen n examples in the training data and seek to predict the label of a new query instance.

Formally, for previous observations $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$ with $(x_i,y_i) \in \mathscr{Z} = \mathscr{X} \times \mathscr{Y}$, a set-valued predictor $\Gamma^{\varepsilon} : \mathscr{X} \rightarrow 2^{\mathscr{Y}}$ is constructed on the basis of a permutation-invariant *non-conformity measure* $\phi : \mathscr{Z} \times \mathscr{Z}^n \rightarrow \mathbb{R}$ that indicates how strange a hypothetical example $z = (x,y) \in \mathscr{Z}$ is compared to previous examples in a set $A \in \mathscr{Z}^n$. The outputs of the non-conformity measure are called non-conformity scores and are formally described as

$$\begin{aligned} \phi_i &= \phi(z_i, A_i) = \phi((x_i, y_i), \{z_1, \dots, z_{n+1}\} \setminus \{z_i\}) && (1) \\ &:= \phi(y_i, \hat{f}_{A_i}(x_i)), && (2) \end{aligned}$$

where $\hat{f}_{A_i} : \mathscr{X} \to \mathscr{Y}$ is the point prediction rule learned on A_i. Then, for a new instance x_{n+1}, all possible candidate values $y \in \mathscr{Y}$ are considered for y_{n+1} by testing the hypothesis $H_0 : y_{n+1} = y$ (against $H_1 : y_{n+1} \neq y$) and computing the p-values

$$p_y = \frac{\sum \mathbf{1}\{\phi_i \geq \phi_{n+1}\}}{n+1}. \tag{3}$$

The set-valued prediction is then given by

$$\Gamma^{\varepsilon}(x_{n+1}) = \{y : p_y > \varepsilon\}. \tag{4}$$

Under some technical assumptions[1], it can be shown that this prediction fulfills

$$\mathbb{P}(y_{n+1} \in \Gamma^{\varepsilon}(x_{n+1})) \geq 1 - \varepsilon. \tag{5}$$

In ICP, the dataset is split into three parts: the training set, the *calibration* set and the testing set. The training set is used to train the point predictor $\hat{f}$. The calibration set $\{z_1, \dots, z_n\}$ is used to compute the non-conformity scores $\{\phi_1, \dots, \phi_n\}$ only once. For a new instance z_{n+1} from the test dataset, the non-conformity score ϕ_{n+1} is computed as usual:

$$\begin{aligned} \phi_i &= \phi(z_i, A \setminus \{z_i, z_{n+1}\}) \quad \forall i \in \{1, \dots, n\} && (6) \\ \phi_{n+1} &= \phi(z_{n+1}, A \setminus z_{n+1}). && (7) \end{aligned}$$

Then, the non-conformity score ϕ_{n+1} is compared to the scores from the calibration set to eventually compute its p-value according to (3).

The guarantee (5) holds “on average”, that is, when assuming new samples (x_{n+1}, y_{n+1}) to be drawn according to the underlying probability measure on

[1] A key assumption is the condition of exchangeability [12].

Table 2: Non-Conformity Measures

A)	Kernel Density (KDE)	$\phi((x,\tilde{y}),A) = (1+\hat{p}_A(x\|y=\tilde{y})^{-1}$
B)	Distance to mean (DTM)	$\phi((x,\tilde{y}),A) = \|\bar{x}_{A,\tilde{y}} - x\|$
C)	1 Nearest Neighbour (1NN)	$\phi((x,\tilde{y}),A) = \frac{\min\{\|x_{A,y}-x\|,y=\tilde{y}\}}{\min\{\|x_{A,y}-x\|,y\neq\tilde{y}\}}$

$\mathscr{X}$. It does not hold, however, *conditional* to a specific class $\tilde{y} \in \mathscr{Y}$, i.e., under the condition that $y_{n+1} = \tilde{y}$. In other words, predictions might be more valid for some (ground truth) classes and less for others. Therefore, in cases of strong class imbalance, where the set sizes vary too strongly among the classes for the same choice of confidence level $1-\varepsilon$, it appears meaningful to choose $\varepsilon_{\tilde{y}}$ for each class $\tilde{y}$ separately. This is also known as *Mondrian Conformal Prediction* [11].

In the following, we consider several measures for $y_{n+1} = \tilde{y}$ which are given in table 2. There, $\bar{x}_{A,\tilde{y}}$ is the mean over all x-vectors in A labeled with class $\tilde{y}$, and $x_{A,y}$ the instance in A with smallest (Euclidean) distance to x among those with label y. Moreover, $\hat{p}_A$ denotes the class-conditional density, estimated on the set A by means of kernel density estimation with Gaussian kernel learned.

4 Results

In this section, the results for the different non-conformity measures are reported for the gaze dataset introduced earlier. For every new instance in the test set, the p-value p_y for each class $y \in \mathscr{Y}$ is returned. The latter can be used in two different ways: (i) The class with the highest p-value is chosen as a point prediction (the p-value itself is then called the *credibility* of the prediction). (ii) For a given confidence level $1-\varepsilon$, the set of labels Γ^{ε} is returned as a set-valued prediction. For (i) the error of this predictor is reported, while for (ii), the average size of the predicted sets (at different confidence levels) is of specific interest.

4.1 The Gaze Dataset

Similar to [6, 2], we apply the method of conformal prediction by extracting the output of the neural network before applying the softmax function and use it as the input feature for ICP (cf. Figure 3). In this way, we circumvent the disadvantages of the softmax transformation [2]. For calibration, 300 instances per class (100 instances for the classes *left shoulder* and *right shoulder*) are sampled. The test set is a newly annotated dataset that consists of 5 videos with 3,138 frames.

We report the results for the three conformity measures KDE, DTM, and 1NN. A stacked bar plot is employed to visualize the set sizes for each conformity measure. It is provided in Figure 5. The set sizes at confidence level $1-\varepsilon$ are displayed in different colors. The black graph corresponds to the accuracy of the single predictions while the red graph represents the accuracy of all predicted sets (including non-empty sets as well). More information is provided in Table 3 with the average credibility of the class with the highest *p*-value. Furthermore, the table contains information on the average size of the non-empty sets and the accuracies for all sets at different confidence levels $1-\varepsilon \in \{0.85, 0.90, 0.95, 0.98\}$.

From Table 3, it can be observed that the error of the point predictor is lowest at 10.6% for the KDE measure. The other measures produce error rates between 16.1% and 18.3%, while the error rate for the baseline gaze classifier with the softmax-generated output is at 15.2%. The favorable, i.e., the highest average credibility is reached by KDE and 1NN at 32.19% and 34.14%. From the plots in Figure 5, it can be noticed that there are only slight differences in the number of empty set predictions (colored in green) and single set predictions (in purple). The number of sets with more than one label is highest for all confidence levels for the measure 1NN. Table 3 shows that the average size of non-empty sets is always lowest for KDE. Also, the average set size for non-empty sets is for all three measures similar at lower confidence levels, e.g., $1-\varepsilon = 0.85$. With increasing confidence levels, the sizes vary more strongly, e.g., 4.81 for DTM and 1.69 for KDE at confidence level $1-\varepsilon = 0.98$. The (statistical) guarantee (5) is met for both, 1NN and KDE. DTM misses

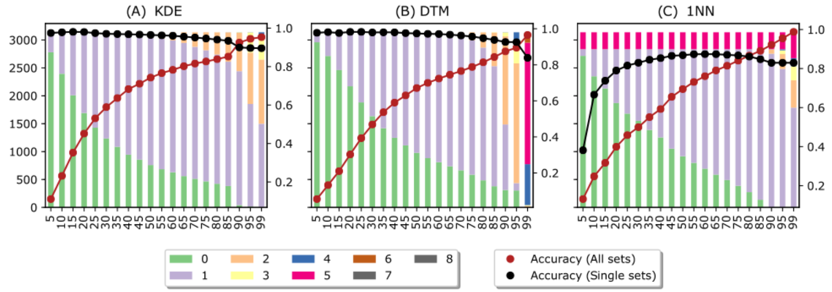


Figure 5: Stacked bar graph that visualizes the set sizes.

the required confidence level by a few percentage points at higher confidence levels.

The average computational time per method, i.e., computing the conformity scores for the calibration set and the testing set, is shortest for DTM with 64 seconds. For the KDE method and 1NN, 379 resp. 723 seconds are needed on average.

4.2 Structure of Prediction Sets

As the label space $\mathscr{Y}$ consists of eight regions in our application, there are $2^8 = 256$ possible prediction sets that might be produced by CP. Even if all these sets are valid in a statistical sense, not all of them appear to be semantically meaningful. In fact, $\mathscr{Y}$ is not just a set of distinct classes. Instead, the classes are spatially related to each other. Intuitively, one would therefore expect that prediction sets correspond to spatially neighbored regions. Or, stated differently, prediction sets that include certain regions while omitting regions “in-between” may appear less meaningful. For example, if *right mirror* and *inner mirror* are included, one would expect *front* to be included, too.

To check for the semantic meaningfulness of CP predictions, we examine the test data further. For the confidence level $1-\varepsilon = 0.96$, the predicted sets and their number are displayed in Table 4. The sets $\{\textit{right mirror}, \textit{inner mirror}\}$ and $\{\textit{inner mirror}, \textit{center console}\}$ are examples of arguably less meaningful

Table 3: Results Gaze Dataset

Conf-Meas.		KDE	DTM	1NN
Error*		0.106	0.161	0.17
Time (s)		379	64	723
Avg. Cred.		32.19	39.68	34.14
85	Size	1.2	1.28	1.4
	Acc.	0.853	0.849	0.895
90	Size	1.34	1.73	1.44
	Acc.	0.932	0.878	0.931
95	Size	1.56	2.38	1.59
	Acc.	0.947	0.897	0.962
98	Size	1.69	4.81	2.06
	Acc.	0.953	0.965	0.988

*Error of the Baseline model: 0.152.

predictions, as they omit the in-between class *front*. These sets are marked with (*). Only once, the predicted set contains classes which are evidently not meaningful $\{$*right mirror*, *left mirror*,*front*$\}$ marked with (**). This combination was produced by 1NN. In total, 48 of the 256 theoretically possible sets are predicted.

5 Discussion

While the statistical guarantee of correctness holds with an increasing number of instances, regardless of the non-conformity measure chosen, this measure has an important influence on the *efficiency* of CP, that is, the size of prediction sets: The more suitably the non-conformity measure is chosen, the smaller these sets will be. In our case, non-conformity scores derived from KDE and 1NN provide significantly smaller sets than DTM, which is in line with the theory presented in [9]. Indeed, one should note that the distance to the mean is a rather crude measure, which ignores a lot of information about the class distributions.

Table 4: Sets at confidence level $1 - \varepsilon = 0.96$

Set Size	1NN	DTM	KDE	Sets
0	0	0	8	{}
1	2307	26	1718	{rm}, {lsh}, {sp}, {inm}, {lm}, {cc}, {fr}
2	344	1007	1009	{rm, cc}, {rm, inm}*, {rm, fr}*, {lsh, fr}*, {fr, cc}, {lm, lsh}, {lm, fr}, {cc, sp}, {inm, sp}*, {inm, cc}*, {fr, sp}, {inm, fr}
3	120	591	397	{lm, cc, sp}, {rsh, rm, cc}, {rsh, inm, fr}*, {rm, lm, fr}**, {rm, inm, fr}*, {rm, inm, cc}*, {inm, fr, cc}, {inm, cc, sp}*, {inm, fr, sp}, {lm, fr, sp}, {rm, fr, sp}*, {inm, lm, fr}, {lm, lsh, fr}, {fr, cc, sp}
4	53	1445	6	{inm, fr, cc, sp}, {rm, lm, fr, sp}*, {lm, lsh, fr, sp}, {lm, fr, cc, sp}, {rsh, rm, fr, cc}, {rsh, rm, inm, cc}, {rm, inm, fr, sp}*, {rm, inm, fr, cc}
5	312	60	0	{inm, lm, fr, cc, sp}, {rm, inm, fr, cc, sp}, {rsh, rm, inm, fr, cc}, {rsh, rm, inm, lm, fr}*
6	2	9	0	{rm, inm, lm, fr, cc, sp}, {rsh, rm, inm, fr, cc, sp}

Number of predicted sets: 48.

As for the semantic meaningfulness of the CP predictions, we observed that CP seems to capture the spatial structure of the classes quite well, with only a few exceptions. In total, 48 of the 256 possible sets were returned as predictions, only 13 of which displayed minor gaps and a single one severe "gaps" in the associated spatial region, i.e., the union of the regions associated with the classes in the set.

CP can also be used as a point predictor. In that sense, regarding solely the error made when choosing the class with the highest p-value, CP with the KDE measure as non-conformity score even outperforms the original gaze classification system, while also providing additional information. This indicates that the relations among the neural net's outputs cannot solely be disclosed with the softmax function, and that the highest value in the output layer does not always correspond to the best prediction.

6 Conclusion

In safety-relevant applications of machine learning, such as AI-assisted driving, a predictive model should produce reliable predictions and be aware of its own uncertainty. In this paper, we considered the problem of predicting the driver's gazing direction and elaborate on the use of conformal prediction to represent uncertainty. Instead of guessing a single class label (gazing direction), even in cases of uncertainty, conformal prediction yields set-valued predictions that are guaranteed to cover the true class with high probability. Our first experimental results with different variants of conformal prediction are rather promising. In particular, we have seen that the extension of our original gaze classification system by means of CP can indeed decrease the error rate of the model while still providing important information on the confidence of the estimates. Especially promising is the non-conformity measure based on kernel density estimation, as it yields the smallest set sizes at high confidence levels.

A possible application of this method, which we seek to investigate in future work, is the handling of out-of-distribution data for classes not covered by the gaze classification system (e.g. blinks). Moreover, we plan to elaborate on

Mondrian Conformal Prediction with class-specific confidence levels, where the confidence levels are determined by the Pareto optimum between different criteria (e.g. average set size and accuracy of single predictions).

References

[1] S. Dari, N. Kadrileev, and E. Hüllermeier. “A Neural-Network Based Driver Gaze Classification System with Vehicle Signals”. In: *IEEE International Joint Conference on Neural Networks 2020.* In press.

[2] Y. Hechtlinger, B. Póczos, and L. A. Wasserman. “Cautious deep learning”. *CoRR*, abs/1805.09460, 2018.

[3] E. Hüllermeier and W. Waegeman. “Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods.” *CoRR*, abs/1910.09457, 2019.

[4] Z. Liu, P. Luo, X. Wang, and X. Tang. “Deeplearning face attributes in the wild”. In: *2015 IEEE InternationalConference on Computer Vision, ICCV 2015,* pages 3730–3738. IEEE Computer Society. 2015.

[5] S. Messoudi, S. Rousseau, and S. Destercke. “Deep conformal prediction for robust models.” In: *18th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems* , IPMU 2020, pages 528–540. Springer. 2020.

[6] H. Papadopoulos. “Inductive conformal prediction: Theory and application to neural networks.” In: *Tools in Artificial Intelligence,* chapter 18. IntechOpen, Rijeka, 2008.

[7] S. Park, X Zhang, A. Bulling, and O. Hilliges. “Learning to find eye region landmarks for remote gaze estimation in unconstrained settings.” In: *Proceedings of the 2018 ACM Symposiumon Eye Tracking Research and Applications - ETRA ’18,* pages 1–10. ACM Press. 2018.

[8] N. Ruiz, E. Chong, and J. M. Rehg. “Fine-grained headpose estimation without keypoints.” *CoRR, abs/1710.00925,* 2017.

[9] M. Sadinle, J. L., and L. A. Wasserman. "Least ambiguous set-valued classifiers with bounded error levels." *CoRR,abs/1609.00451,* 2016.

[10] G. Shafer and V. Vovk. "A tutorial on conformal prediction". *J. Mach. Learn.* vol. 9, pages 371–421, 2008.

[11] P. Toccaceli and A. Gammerman. "Combination of inductive mondrian conformal predictors". *Mach. Learn.*, vol. 108(3), pages 489–510, 2019.

[12] V. Vovk, A. Gammerman, and G. Shafer. "Algorithmic Learning in a Random World." Springer-Verlag, New York, 2005.