Model agnostic calibration of image classifiers

Paolo Giudici[1,*,†], Giulia Vilone[2,*,†]

[1] *University of Pavia, Pavia, Italy*
[2] *Tobii AC, Galway, Ireland*

Abstract

Predictions obtained from deep neural networks can be extremely accurate but not very robust, leading to uncertainty in their predictions. This crucial problem is getting growing attention from the Machine Learning (ML) community. A pragmatic solution that is increasingly applied is computing confidence bounds of the predictions of ML models. Most confidence bounds in the literature are theoretically sound but unfeasible from a practical perspective. This paper contributes to the literature by proposing probabilistic confidence bounds based on conditional probabilities. It demonstrates their operational validity using a real-world application: predicting car drivers' sleeping states.

Keywords

Confidence measure, Confidence calibration, Driver's drowsiness

1. Introduction

Modelling the uncertainty of Deep Neural Network (DNN)'s predictions is a critical problem receiving growing attention from the Machine Learning (ML) community. Deep learning architectures can achieve outstanding results in various domains. However, their application in high-risk fields, such as autonomous driving, demands that DNNs be accurate and indicate when their predictions are likely incorrect due to out-of-distribution data. A DNN should provide a calibrated measure of its confidence that its predictions are correct, meaning they correspond to the ground truth [1]. Ideally, such an indication should be the outcome of a transparent process that helps an end-user interpret the output and the related level of confidence of a DNN to take informed countermeasures in case of erroneous predictions. Modelling the uncertainty of DNN predictions is crucial for improving the trust, safety, fairness, reliability and interpretation of AI systems, as well as for enhancing human-AI collaboration and allowing for safer decision-making [2, 3]. Uncertainty modelling can also guide the development of more robust architectures and training procedures, help identify weaknesses and limitations in existing models, and inform decisions about which models to use.

This paper focuses on a novel calibration method to compute probabilistic confidence measures for DNNs applied to image classification problems. It presents a real-world application of the proposed confidence calibration method to a DNN trained for detecting whether a car driver

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta

[*] Corresponding author.

[†] These authors contributed equally.

✉ paolo.giudici@unipv.it (P. Giudici); giulia.vilone@tobii.com (G. Vilone)

iD 0000-0002-4198-0127 (P. Giudici); 0000-0002-4401-5664 (G. Vilone)

is alert or has a microsleep. The proposed calibration method's advantage is that confidence bounds can be calculated for individual predictions without requiring heavy computations, and it is based on a transparent and interpretable process.

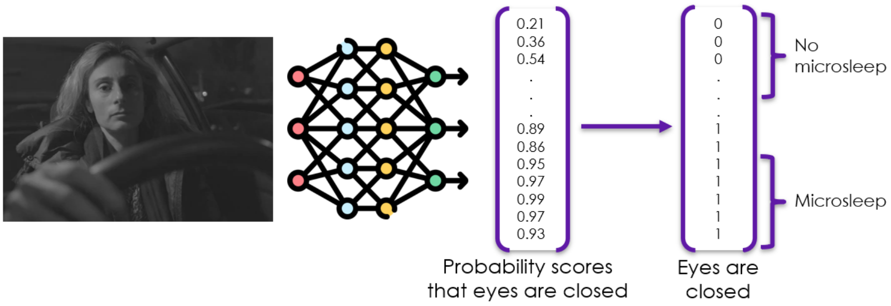
2. Literature review

Confidence calibration aims to estimate uncertainty via matching the confidence level of a set of samples with their prediction accuracy [1, 4]. For instance, a model should correctly classify 90 out of 100 samples if its confidence level on such predictions is 0.9. More formally, given the input $X \in \mathcal{X}$ and label $Y \in \mathcal{Y} = \{1, \cdots, K\}$ both random variables following a ground truth joint distribution $\pi(X, Y) = \pi(Y|X)\pi(X)$, a DNN f with $f(X) = (\hat{Y}, \hat{P})$ where $\hat{Y} \in \{1, \cdots, K\}$ is a predicted class and $\hat{P}$ is its associated confidence level, *perfect calibration* can be defined as [1] $P(\hat{Y} = Y|\hat{P} = p) = p, \quad \forall p \in [0, 1]$.

The most recent DNNs are poorly calibrated [1]. Depth, width, weight decay, and batch normalisation influence calibration. Consequently, achieving perfect calibration in practical, real-world settings is impossible. Scholars have proposed different solutions to improve calibration that can be clustered into "scaling-based", "binning-based", "similarity-based", and "Bayesian-based" methods. *Scaling-based methods* adjust the probability returned by a model that an input belongs to an output class by learning one or more scalar parameters so that this probability accurately represents the likelihood of that particular class. Standard methods for confidence calibration in the classification domain are Platt scaling [5], Beta calibration [6] and temperature scaling methods [1]. *Binning-based methods* divide samples into multiple bins based on samples' confidence and calibrate each bin. Popular binning-based methods include Bayesian Binning into Quantiles (BBQ) [7], histogram binning [8] and Ensemble of Near Isotonic Regression [9]. However, existing binning-based calibration methods fail to see the proximity bias issue, which is the tendency of models to be overly confident in low-proximity samples (samples lying in sparse density regions of the input space) rather than high-proximity ones. Thus, models suffer from inconsistent miscalibration, limiting the capabilities of calibration methods to deliver reliable and interpretable uncertainty estimates [10]. *Similarity-based methods* estimate the confidence level based on the distance of the instances in the input dataset that are closer (or more similar) to the test sample and their output class. For instance, [11] proposed to estimate confidence levels using a non-conformity measure, calculated as the average k-neighbour proximity for all the samples in the same predicted class for a given sample, to indicate how 'atypical' this sample is relative to the other samples. *Bayesian-based methods* quantify the uncertainty related to inputs and parameters' calibration via a posterior distribution of the model's parameters, which balances the prior probability of the parameters with the likelihood function learned from the available data [12, 13]. However, exact Bayesian inference is not tractable in DNNs due to its sophisticated implementation and high computational cost [14]. Furthermore, these methods are often harder to scale and can suffer from sub-optimal performance [15]. Scholars have proposed many techniques to approximate the intractable posterior distributions derived by Bayesian inference for DNNs [16, 17], a popular one being Markov Chain Monte Carlo [18].

Finally, Conformal Prediction (CP) is a framework for assessing the uncertainties of AI systems. Given a sample, CP returns a prediction interval in regression problems and a set

Figure 1: Structure of the AI system that predicts a driver's alert/microsleep states based on a DNN that returns the probability of whether the driver's eyes are closed. If this probability is higher than 80%, eyes are considered closed. A microsleep consists of 15 consecutive eyes-closed frames.



of classes in classification problems guaranteed to cover the true value with high probability. However, CP is computationally inefficient as it requires retraining a model over a calibration set containing $n + 1$ samples w.r.t. the previous iteration [19]. Some real-world applications, like autonomous cars, need lightweight DNNs as the computational resources are required to process the signals received from various devices, such as sensors and cameras. Furthermore, most of these methods are complex and thus hard to understand, explain, and debug.

3. The proposed calibration method

The proposed calibration method was inspired by the request of a car manufacturer to show an indicator of a driver's state (either alert or microsleep) augmented by confidence levels in the predictions made by a DNN. The methodology to calculate this indicator must a) calculate and display in real-time; hence, its computation cannot be resource greedy; b) be understandable, especially in incremental innovations concerning existing practices, to ensure meeting safety and industry quality standards; and 3) return two confidence levels, depending on whether or not the driver is in a microsleep state. These conditions led us to develop a methodology that, while mathematically sound, is also simple to understand and implement.

The DNN, based on a Shufflenet backbone, returns a probability score on whether the eyes of the driver are open or closed. The eyes are considered open if the probability score is lower than 80%; above this threshold, eyes are considered closed. This threshold was determined by maximising the prediction accuracy of the network on a Tobii proprietary dataset containing several videos of people driving at a simulator at different times of the day and night. When the eyes are predicted as open, the driver is in a no-microsleep, or alert, state. A microsleep starts after 15 consecutive frames (the car manufacturer required this) are labelled as "eyes closed" and ends when there occurs a frame labelled as "open eyes" (see Figure 1).

The proposed calibration method considers the confusion matrices to calculate the DNN's confidence levels. The false positive and negative rates reported in the confusion matrices provide insight into how often the network confuses the two output classes. The positive cases correspond to eyes open and no microsleep, and an example of the two confusion matrices calculated over the network's predictions is reported in Table 1. The number of instances can

Table 1
Confusion matrices of the DNN's predictions on whether (a) a driver's eyes are open or closed or (b) is alert or having a microsleep.

		Predicted state	
		Open eyes	Closed eyes
Actual	Open eyes	19,501 (97%)	499 (2%)
state	Closed eyes	580 (3%)	19,240 (98%)

(a)

		Predicted state	
		Alert	Microsleep
Actual	Alert	17,863 (93%)	2,137 (10%)
state	Microsleep	1,255 (7%)	18,745 (90%)

(b)

be transformed into probabilities of correctly or wrongly classifying frames by dividing each value by the sum of its column. For example, the values in the first row of the open/closed eyes confusion matrix are divided by $19,501 + 580 = 20,081$. The resulting probabilities are reported in brackets in the same table.

When predicting a microsleep state, the DNN's confidence levels are based on the conditional probabilities of open or closed eyes. The prior conditions are the true and false positive/negative rates derived from the two confusion matrices and the probability scores returned by the DNN. When the DNN's prediction is "eyes open", its confidence level $P(O)$ is the sum of the probability $P(\hat{C})$ that the eyes are closed returned by the DNN multiplied by the true and false positive rates, respectively (see Eq. 1).

$$P(O) = P(\hat{O})P(O|\hat{O}) + P(\hat{C})P(O|\hat{C}) \tag{1}$$

where $P(\hat{O})$ is the DNN's probability score that the eyes are open and is computed as $P(\hat{O}) = 1 - P(\hat{C})$. $P(O|\hat{O})$ and $P(O|\hat{C})$ are the true and false positive rates of the eyes-open/close confusion matrix, respectively. Similarly, the confidence level that the eyes are truly closed $P(C)$ can be calculated per Eq. 2.

$$P(C) = P(\hat{C})P(C|\hat{C}) + P(\hat{O})P(C|\hat{O}) \tag{2}$$

where $P(C|\hat{C})$ and $P(C|\hat{O})$ are the true and false negative rates of the eyes-open/close confusion matrix, respectively.

The confidence level that the driver is truly alert ($P(A)$) follows the same logic (see Eq. 3. The probability scores correspond to the probability that the eyes are open calculated per Eq. 1.

$$P(A) = P(O)P(A|\hat{A}) + (1 - P(O))P(A|\hat{M}) \tag{3}$$

where $P(A|\hat{A})$ and $P(A|\hat{M})$ are the true and false positive rates as per the microsleep confusion matrix, respectively.

The confidence level of microsleeps differs from the previous cases because the microsleep probability scores correspond to the eyes-closed probability $P(C)$ raised to the power of the number of frames that are missing to reach the microsleep state (see Eq. 4 and Eq. 5). For instance, if the DNN assigned the label "eyes closed" to three consecutive frames, $P(C)$ must be raised to the power of 12. This corresponds to the probability of independently randomly sampling 12 frames predicted as "eyes closed". $P(C)$ of the last frame is the best estimate of the probability that the following frames will be predicted as "eyes closed" because it is impossible to know what probability scores will be returned by the DNN for future frames.

$$P(M) = P(\hat{M})P(M|\hat{M}) + (1 - P(\hat{M}))P(M|\hat{A}) \tag{4}$$

where $P(M|\hat{M})$ and $P(M|\hat{A})$ are the true and false negative rates as per the microsleep confusion matrix, respectively.

$$P(\hat{M}) = P(C)^{(15-F_{CE})} \tag{5}$$

where $F_{CE} = \sum_{0 \leq n \leq 14} \mathbb{1}_{ClosedEyes}$ represents the number of consecutive frames (up to 15) labelled as "eyes closed" by the DNN.

4. The experiment

The proposed method was applied to the public dataset Night-Time Yawning-Microsleep-Eyeblink-driver Distraction (NITYMED)[1] [20]. NITYMED contains 21 videos with 25 frames per second, each lasting approximately 2 minutes, of 11 male and eight female drivers in real cars under nighttime conditions. The drivers talk, look around, and have microsleeps. The NITYMED videos are not labelled. The ground truth labels were created by applying a key point detector to extract the facial landmarks of each frame and calculate the Eye Aspect Ratio (EAR). It was decided that the driver's eyes are closed when EAR is below 20%. The frames classified as "eyes closed" were visually inspected to ensure this threshold was not too high, thus labelling as "eyes closed" frames with open eyes. We remark that the described labelling procedure provides an "expert-based ground truth," which is not objective. This is the case in many other machine learning applications, where the model is assessed not against an "objective" truth but a subjective one. This does not alter the generality of the proposed method.

5. Results

The model reaches a prediction accuracy of 88.1% in classifying the frames of the NITYMED videos as eyes open/closed, and 95.2% prediction accuracy of the alert or microsleep states with the EAR considered as ground truth (see table 2). Noticeably, the true negative rates of the DNN are 78% and 74% in predicting eyes open/close and the alert/microsleep stats, which are quite low and are expected to significantly impact the resulting confidence levels on the alert or microsleep states. The DNN predicted slightly less than 50% of the microsleeps detected with the EAR (87 out of 185). This is due to the high EAR threshold (20%) used to determine when the driver's eyes are closed. This gap can be easily closed by reducing this threshold. A further inspection of the frames classified as "eyes closed" with EAR highlighted a few where the eyes are still partially open, but it is possible to see most of the eyelids. This threshold allowed testing of the proposed calibration method under suboptimal conditions where the network's accuracy is not high. This is the typical situation where a DNN should not always

[1] https://datasets.esdalab.ece.uop.gr/

Table 2
Confusion matrix of the network's predictions on whether a NITYMED driver has (a) open or closed eyes or (b) is in the alert or microsleep state.

		Predicted state	
		Open eyes	Closed eyes
Actual state	Open eyes	46,682 (89%)	1,289 (22%)
	Closed eyes	5,674 (11%)	4,683 (78%)

(a)

		Predicted state	
		Alert	Microsleep
Actual state	Alert	54,498 (96%)	367 (26%)
	Microsleep	2,433 (4%)	1,030 (74%)

(b)

be trusted, and confidence levels can support and improve decision-making. A video[2] shows examples of microsleeps detected by the model and its alert/microsleep levels of confidence.

Figure 2: Alert/microsleep confidence levels computed on a microsleep event in a NITYMED video.

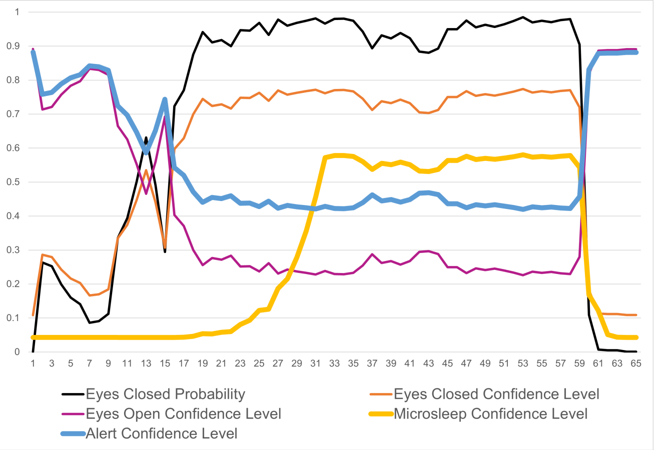


Figure 2 shows an example of the model's confidence levels of the alert and microsleep statuses of a NITYMED driver. The microsleep confidence level remains constantly low until the DNN returns a frame with closed eyes. Then, it quickly increases as the number of consecutive eyes-closed frames increases. However, this confidence level never goes above 60% even when the number of consecutive eyes-closed frames is far higher than 15, and the DNN assigns high eye-closed probability scores to these frames, meaning that the chances that the driver is truly having a microsleep are pretty high. This is, as expected, due to the combined effect of the low eyes-open/close and microsleep true negative rates. One desired requirement for a confidence level assessment method like the proposed one is to compute calibrated uncertainty estimates. However, this method was not expected to meet this requirement as the confidence levels are computed using confusion matrices that consider the DNN's errors throughout the entire dataset. This assumption was tested by binning the frames of the NITYMED videos according to their microsleep confidence levels and checking whether these levels match the prediction accuracy. The results are reported in Table 3 confirm this assumption. When the confidence level for the microsleep state is below 50%, only one frame out of 57,000 was correctly labelled as microsleep by the model. Conversely, the prediction accuracy is higher than the confidence levels when they are in the range 50-58%. This issue could be easily overcome by extracting other confusion matrices for the frames with mid-range confidence levels (the frames where the eyes are not fully open or closed) and calculating the confidence levels with these matrices. Confidence levels and prediction accuracy

[2] https://vimeo.com/870309458

match the two extreme tails of the data distribution. The confidence levels cannot be higher than 59%, and correspondingly 61% of the frames are correctly labelled as showing a microsleep event.

6. Conclusions

Table 3: Number of NITYMED frames grouped by their microsleep confidence level and model's prediction accuracy.

Confidence level	# frames	Prediction accuracy
$< 50\%$	56,927	0%
50%	24	63%
51%	35	77%
52%	42	76%
53%	63	79%
54%	96	79%
55%	107	83%
56%	140	84%
57%	184	81%
58%	385	72%
59%	320	61%

This paper presents a probabilistic method to calibrate predictions arising from ML models. It demonstrates its operational validity using a real-world application concerning the forecast of car drivers' alert and sleeping states. The proposed calibration method has returned a reliable estimate of the confidence level of the predictions made by a DNN that considers the true and false positive/negative rate to assess the network's confidence. This method brings the following advancements compared to other calibration methods: it is 1) simple to implement, 2) comprehensible, and 3) not resource-greedy and does not require high computation power. We remark that our analysis is conditional on the available data. If data allow, further analysis can be entertained. For example, in the paper, we assumed that the cost of type I and type II errors is the same. If a cost function were known, we could incorporate it into our model. Similarly, if more data on the drivers were known, we could assess other aspects, such as the gender and ethnicity fairness of the proposed algorithm. Other future research directions include testing this method on datasets containing data other than images from different application domains than autonomous driving cars. Theoretically, the proposed calibration method is model-agnostic and should apply to other learning algorithms, such as support vector machines. From a methodological viewpoint, it would be important to examine how the proposed confidence bounds change when the two prediction errors in the confusion matrix are assigned different costs.

References

[1] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: International conference on machine learning, PMLR, 2017, pp. 1321–1330.

[2] I. Cortés-Ciriano, A. Bender, Deep confidence: a computationally efficient framework for calculating reliable prediction errors for deep neural networks, Journal of chemical information and modeling 59 (2018) 1269–1281.

[3] B. Ji, H. Jung, J. Yoon, K. Kim, et al., Bin-wise temperature scaling (bts): Improvement in confidence calibration performance through simple scaling techniques, in: International Conference on Computer Vision Workshop, IEEE/CVF, 2019, pp. 4190–4196.

[4] M. Minderer, J. Djolonga, R. Romijnders, F. Hubis, X. Zhai, N. Houlsby, D. Tran, M. Lucic,

Revisiting the calibration of modern neural networks, Advances in Neural Information Processing Systems 34 (2021) 15682–15694.

[5] J. Platt, et al., Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, Advances in large margin classifiers 10 (1999) 61–74.

[6] M. Kull, T. Silva Filho, P. Flach, Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers, in: Artificial Intelligence and Statistics, PMLR, 2017, pp. 623–631.

[7] M. P. Naeini, G. Cooper, M. Hauskrecht, Obtaining well calibrated probabilities using bayesian binning, in: Proceedings of the AAAI conference on artificial intelligence, volume 29(1), 2015.

[8] B. Zadrozny, C. Elkan, Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers, in: Icml, volume 1, 2001, pp. 609–616.

[9] M. P. Naeini, G. F. Cooper, Binary classifier calibration using an ensemble of near isotonic regression models, in: 16th International Conference on Data Mining, IEEE, 2016, pp. 360–369.

[10] M. Xiong, A. Deng, P. W. W. Koh, J. Wu, S. Li, J. Xu, B. Hooi, Proximity-informed calibration for deep neural networks, Advances in Neural Information Processing Systems 36 (2024) 68511–68538.

[11] S. Bhattacharyya, Confidence in predictions from random tree ensembles, in: 2011 IEEE 11th International Conference on Data Mining, IEEE, 2011, pp. 71–80.

[12] Y. Gal, R. Islam, Z. Ghahramani, Deep bayesian active learning with image data, in: International conference on machine learning, PMLR, 2017, pp. 1183–1192.

[13] A. Kendall, Y. Gal, What uncertainties do we need in bayesian deep learning for computer vision?, Advances in neural information processing systems 30 (2017).

[14] S. Seo, P. H. Seo, B. Han, Learning for single-shot confidence calibration in deep neural networks through stochastic inferences, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 9030–9038.

[15] L. Blier, Y. Ollivier, The description length of deep learning models, Advances in Neural Information Processing Systems 31 (2018).

[16] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, Advances in neural information processing systems 30 (2017).

[17] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1492–1500.

[18] R. M. Neal, Bayesian learning for neural networks, volume 118, Springer Science & Business Media, 2012.

[19] H. Papadopoulos, V. Vovk, A. Gammerman, Conformal prediction with neural networks, in: 19th International Conference on Tools with Artificial Intelligence, volume 2, IEEE, 2007, pp. 388–395.

[20] N. Petrellis, S. Zogas, P. Christakos, P. Mousouliotis, G. Keramidas, N. Voros, C. Antonopoulos, Software acceleration of the deformable shape tracking application: How to eliminate the eigen library overhead, in: Proceedings of the European Symposium on Software Engineering, 2021, pp. 51–57.