# Simple and Effective Approaches for Uncertainty Prediction in Facial Action Unit Intensity Regression

Torsten Wörtwein and Louis-Philippe Morency
Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA

***Abstract*— Knowing how much to trust a prediction is important for many critical applications. We describe two simple approaches to estimate uncertainty in regression prediction tasks and compare their performance and complexity against popular approaches. We operationalize uncertainty in regression as the absolute error between a model's prediction and the ground truth. Our two proposed approaches use a secondary model to predict the uncertainty of a primary predictive model. Our first approach leverages the assumption that similar observations are likely to have similar uncertainty and predicts uncertainty with a non-parametric method. Our second approach trains a secondary model to directly predict the uncertainty of the primary predictive model. Both approaches outperform other established uncertainty estimation approaches on the MNIST, DISFA, and BP4D+ datasets. Furthermore, we observe that approaches that directly predict the uncertainty generally perform better than approaches that indirectly estimate uncertainty.**

## I. Introduction

Facial action unit (AU) intensity estimation is central to many critical technologies, including assistive technologies in health care, driver fitness evaluations in automotive applications, and screenings in hiring agencies. For many of these applications, trust is also important. We need models that not only predict a primary output, but also a secondary quantity describing the uncertainty of its prediction.

Uncertainty in machine learning models primarily originates from two sources: 1) aleatoric uncertainty, in which observations can be noisy, and 2) epistemic uncertainty, in which the model might not be well-estimated or might have an improper structure [9]. For real-world applications, we need an uncertainty mesaure which combines both types of uncertainty. For regression tasks, we can operationalize this uncertainty as the absolute error between a model's prediction and the ground truth. While prediction uncertainty has been studied in different fields [8], uncertainty in facial AU intensity prediction has not been studied.

Most existing approaches for uncertainty prediction rely on only epistemic uncertainty. In this paper, we study approaches which capture both epistemic and aleatoric uncertainties by predicting the absolute error. We describe two such approaches and compare them to a variety of established approaches[1]. Both approaches have a secondary model that predicts the absolute error of the primary model. Our first approach assumes that uncertainty is a weighted combination of known absolute error from similar reference observations. This assumption has been previously demonstrated to work well [7], [3] when using a $k$-nearest neighbor approach. Our second approach uses a multi-layer perceptron (MLP) to predict the uncertainty. Such an approach has proven to work well in the past where a single perceptron predicts the uncertainty [13], [14]. With these two approaches, we can capture the prediction uncertainty, whether it is caused by epistemic uncertainty, by aleatoric uncertainty, or by a combination of both.

[1]The code is available at `https://github.com/twoertwein/UncertaintyRegression`

## II. Simple and Effective Uncertainty Prediction

### A. DWAR: Similarity-based Error Prediction

Our first approach adopts the non-parametric deep weighted averaging classifier [4] for regression (DWAR) as our secondary model. The DWAR model learns a low-dimensional embedding ($\mathbf{h}$) in which we use an RBF kernel to define similarity between a new observation and the reference data (training data). The predicted uncertainty of a new observation ($\hat{\epsilon}$) is the similarity-weighted average over the reference uncertainty ($\epsilon$)[2].

$$\hat{\epsilon} = \frac{\sum_{i=1}^{\text{ref}} \epsilon_i \; w(\mathbf{h}, \mathbf{h_i})}{\sum_{i=1}^{\text{ref}} w(\mathbf{h}, \mathbf{h_i})} \tag{1}$$

$$w(\mathbf{h}, \mathbf{h_i}) = \exp\left(-\|\mathbf{h} - \mathbf{h_i}\|^2\right) \tag{2}$$

This model is trained end-to-end. During training, only the current batch is used as reference data, leading to a time complexity of $\mathcal{O}(n^2)$ per batch of size $n$. At test time, the entire training dataset (size $N$) is used, resulting in $\mathcal{O}(N)$ time complexity for a single prediction.

### B. U-MLP: Direct Error Prediction

Our second approach predicts uncertainty using an MLP (U-MLP). Empirically, we observe that the U-MLP performs better when provided with the embedding of the last layer of the primary model concatenated with the primary model's prediction, instead of providing it with the original input representation. For a fair comparison, DWAR uses the same input representation as the U-MLP.

## III. Baselines

As aleatoric uncertainty is difficult to measure without the influence of epistemic uncertainty, we name approaches which directly predict uncertainty "supervised approaches".

[2]Using the validation data should result in less biased errors, but we use the validation set for the prediction interval evaluation, and therefore cannot use the validation set to estimate the uncertainty.

### A. Epistemic Baselines

**Ensemble:** The variance of ensembles is an established approach to the quantification of prediction uncertainty [6]. Ensembles often consist of multiple models of the same type trained on bootstrapped data. While this approach does not make any assumption about the error distribution and can be used with any type of model, it can be computationally expensive to train $k$ models instead of one model. We use an ensemble consisting of ten MLP models.

**Dropout:** By using dropout at inference time in a neural network, Bayesian inference can be approximated without the high computational costs associated with training Bayesian models [5]. To derive an estimation of uncertainty, we keep dropout at the second-to-last layer of the primary model activated, and consider the variance over 1,000 inference runs for each data point to approximate Bayesian inference [5] (requiring 999 additional matrix multiplications). Since no additional computations are required during training, this approach is more practical than the ensemble approach.

**Gaussian Process (GP):** Similar to DWAR, we use an MLP to learn a projection into a low-dimensional space, where a sparse GP [16] then uses an RBF kernel to determine the similarity between two data points. All parameters of this model—MLP parameters, scale parameter of the RBF kernel, inducing points of the GP, and the GP's observation noise parameter, which is shared between all inducing points—are trained end-to-end, optimizing the sparse GP's marginal likelihood. The time complexity during training of the sparse GP is $\mathcal{O}(NM^2)$, where $M$ is the number of inducing points ($M = 2,000$ in all our experiments).

### B. Supervised Baselines

**Multi-task MLP:** Training a multi-task MLP for two tasks, one for the AU intensity estimation ($\hat{y}$) and one for the estimated absolute error ($\hat{\epsilon}$), could improve prediction of the absolute error at the cost of worsened AU intensity prediction. We optimize the following combined loss for the two tasks.

$$(y - \hat{y})^2 + 0.5(|y - \hat{y}| - \hat{\epsilon})^2 \tag{3}$$

**Attenuation:** A loss-function-agnostic approach deriving uncertainty is the attenuation the original loss function by allowing the model to estimate its prediction variance ($\sigma^2$) for a prediction ($\hat{y}$) [9].

$$\frac{\text{loss}(y, \hat{y})}{\sigma^2} + \log \sigma^2 \tag{4}$$

The greater the uncertainty estimation of a model ($\sigma^2$), the less confidence it has in its prediction.

## IV. Experimental Setup

### A. Datasets

We focus on two facial AU datasets, and for comparison, we run the same experiments on a (subset of) MNIST to evaluate whether the approaches simply exploit skewed labels[3].

[3]Facial AU datasets are known to be highly skewed towards no/little activation. An uncertainty estimation approach might simply learn to associate a high activation estimation with a high uncertainty estimation.

**BP4D+:** This dataset [17] (version 0.2) consists of videos of 140 subjects that have been annotated for facial AU intensities during emotion-eliciting tasks (AU 6, 10, 12, 14, and 17). We use subject-stratified hold-out sets for training (containing 60% of subjects), validation (20% of subjects), and testing (20% of subjects). Stratification is used to ensure a similar average AU intensity for each set.

**DISFA:** This dataset contains AU-annotated videos of 27 subjects viewing an emotion-eliciting video [11] (AU 1, 2, 4, 5, 6, 9, 12, 15, 17, 20, 25, and 26). We use the same method as for BP4D+ to determine the dataset split. The input representation of each image for these two datasets is the same as for OpenFace 2.0's AU intensity estimation [1]: face-aligned HOG features and the rigid and non-rigid shape parameters.

**MNIST:** We use MNIST [10] to validate the uncertainty approaches on a different domain. To make MNIST similar to BP4D+ and DISFA, we reframe MNIST as a regression task instead of a classification task and only consider numbers between 0 and 5 (the same range as facial AU intensities). To study the impact of skewed datasets on uncertainty, we use the 0-5 MNIST and also generate a skewed 0-5 MNIST that reflects the same skew as AU 6 (cheek raiser) has on the BP4D+ dataset.

To evaluate whether the uncertainty estimation generalizes to different datasets we train a model on BP4D+ and evaluate its performance on the DISFA test set (without any adaptation, denoted as BP4D+ $\rightarrow$ DISFA+).

### B. Implementation Details

All MLP models have the following hyper-parameters validated using the concordance correlation coefficient (CCC) on the validation set: the number of layers, the learning rate of Adam (the learning rate optimizer we use), and the dropout rate. Early stopping is based on the CCC on the validation set for the last 50 epochs. The maximum number of epochs is 500. Aside from the GP, all models are optimized for the mean squared error.

DWAR, U-MLP, and Dropout share the same primary model.

### C. Metrics

We report the performance of the AU intensity predictions using the subject-averaged intra-class coefficient (ICC). For the predicted uncertainty, we use the following two metrics.

**Spearman Correlation coefficient ($\rho$):** We use the Spearman correlation coefficitent to measure the monotonic relationship between the estimated uncertainty and the absolute error. This metric has the advantage that it does not penalize approaches than do not predict the intensities well, but are able to estimate their own error well.

**Prediction Interval Width ($|\mathbf{PI}|$):** We construct prediction intervals using the normalized inductive conformal prediction [12], where the normalization coefficient ($\hat{\sigma}$) is the predicted uncertainty.

$$\alpha = \hat{\sigma}\text{Perc}_{95}\left(\left\{\frac{|y_i - \hat{y}_i|}{\hat{\sigma}_i} | i \in \text{calibration set}\right\}\right) \tag{5}$$

| | DISFA | BP4D+ | BP4D+ $\rightarrow$ DISFA[1] | DISFA[1] |
|---|---|---|---|---|
| DWAR/U-MLP/Dropout | 0.502 | 0.653 | 0.520 | 0.545 |
| Ensemble | 0.341 | 0.664 | 0.495 | 0.535 |
| GP | 0.460 | 0.662 | 0.491 | 0.467 |
| Multi-Task | 0.477 | 0.643 | 0.450 | 0.549 |
| Attenuation | 0.477 | 0.646 | 0.479 | 0.556 |

TABLE I

ICC (HIGHER IS BETTER) OF THE PRIMARY MODELS AVERAGED OVER ALL AUS. [1] AVERAGED OVER THE COMMON AUS BETWEEN DISFA AND BP4D+ (AU 6, 12, AND 17).

$$P(y \in [\hat{y} - \alpha, \hat{y} + \alpha]) \geq 0.95 \ , \tag{6}$$

where $\text{Perc}_{95}$ is the 95th-percentile. The constructed intervals have an asymptotic coverage rate (the probability of containing the true intensity) of 95%, with the assumption that the calibration set (validation set) and the test set are both independent and identically distributed. We report the median interval width as a measure of the efficiency of the prediction intervals [15]. Smaller intervals at the same coverage rate are potentially more useful. This metric is affected by the accuracy of the AU intensity prediction.

Theoretically, a higher correlation ($\rho$) should result in a smaller interval width, and vice-versa when the same primary model is used. In practice, this is not always the case as many outliers, i.e., more than 5% in the uncertainty prediction, can have a negative impact on the intervals.

We test for statistical differences at the subject level between our two described approaches and against all baseline approaches. These tests are conducted with subject-clustered percentile bootstrapping[4]. We do not conduct these tests for DISFA because we have only five subjects in the test set. For 0-5 (Skewed) MNIST, we use bootstrapping without clustering.

## V. RESULTS AND DISCUSSION

Our initial experiment evaluates whether estimating uncertainty degrades the performance of AU intensity estimation, which would influence the prediction interval width (|PI|). Table I shows that almost all models (with exception of the ensemble) predict AUs with comparable performance. The main experimental results are shown in Table II. Table III provides AU-specific results for BP4D+, including the statistical test outcomes, and Table IV demonstrates that the constructed prediction intervals reach their targeted coverage rate.

**Uncertainty under Skew:** To study the effects of label skew, we report results on 0-5 MNIST and 0-5 Skewed MNIST in Table II. Since uncertainty predictions are better on 0-5 Skewed MNIST for almost all approaches, this indicates that these approaches at least partly exploit the label skew. The skew may also explain the different prediction interval widths between DISFA and BP4D+: DISFA is much more skewed and has much smaller prediction interval widths.

**Cross-dataset evaluation:** Testing the BP4D+ models on DISFA provides two particularly intersting results. The first result is a high correlation between the absolute error and the estimated uncerainty for the BP4D+ $\rightarrow$ DISFA evaluation (shown in Table II). The second result is that the coverage rates for the prediction intervals, as reported in Table IV, are closely centered around the targeted 95%, even though the BP4D+ validation set is used for the calibration set. This could indicate that the evaluated uncertainty approaches generalize to data from slightly different conditions.

**Epistemic vs. Supervised Approaches:** The best performing approaches for each dataset and metric are supervised approaches. We hypothesize that supervised approaches perform better because they can use both the aleatoric and epistemic uncertainty to estimate the prediction uncertainty, whereas epistemic approaches can only capture the epistemic uncertainty.

**DWAR:** This non-parametric approach achieves high correlations for almost all evaluations, but tends to perform less well for the prediction interval width: it performs significantly worse on it than other approaches for severely-skewed AUs, e.g., AU 14 and AU 17. Similar to the weighted average, DWAR is confined to the previously-observed range of errors in the training set. This may artificially truncate its correlation and result in large prediction intervals. This approach seems to work very well even across datasets. Compared to the U-MLP, it requires more computational efforts, but also provides transparency. A user can inspect the nearest neighbors, which influence the prediction the most.

**U-MLP:** U-MLP works very well across all datasets and metrics and never performs significantly worse than any other approach (see Table III). It produces remarkably efficient prediction intervals across all datasets, e.g., +/- 0.6 on average for AU intensities on BP4D+, whereas other approaches need around +/- 0.9. In a few cases it is outperformed by DWAR and dropout, but is otherwise always the best performing approach across both families.

**Epistemic Baselines:** The MLP ensemble and dropout are the best performing epistemic baselines. The sparse GP poorly estimates the variance of some AUs (and the 0-5 MNIST). We hypothesize that this occurs because the marginal likelihood of this specific sparse GP is known to have many local minima [2]. Despite this drawback, this specific sparse GP has been shown to better estimate the variance than other sparse GPs [2].

Compared to the MLP ensemble, dropout variance has no overhead during training and is still computationally feasible at test time: there are only $n-1$ additional matrix multiplications for the last layer. It is also already in use in many situations, which makes it a convenient approach that can be implemented easily without re-training or training an additional model to derive uncertainty.

**Supervised Baselines:** The motivation behind a multi-task MLP model and the loss attenuation was to attain good error estimation despite a decrease in AU intensity prediction

[4]We calculate the metric of interest for each cluster (each subject), and then bootstrap the difference between the approaches (5000 re-samplings and a 95%-confidence interval).

| | 0-5 MNIST | | 0-5 Skewed MNIST | | DISFA | | BP4D+ | | BP4D+ → DISFA | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\rho \uparrow$ | $\lvert PI \rvert \downarrow$ | $\rho \uparrow$ | $\lvert PI \rvert \downarrow$ | $\rho \uparrow$ | $\lvert PI \rvert \downarrow$ | $\rho \uparrow$ | $\lvert PI \rvert \downarrow$ | $\rho \uparrow$ | $\lvert PI \rvert \downarrow$ |
| DWAR | $0.923^{U}$ | $0.085^{U}$ | $0.900^{U}$ | $\mathbf{0.001}^{U}$ | 0.646 | **0.303** | 0.681 | 1.819 | **0.754** | 1.486 |
| U-MLP | **0.932** | **0.042** | **0.970** | **0.001** | **0.835** | 0.502 | **0.690** | **1.211** | **0.754** | **0.902** |
| Ensemble | $0.710^{UD}$ | $0.360^{UD}$ | $0.891^{U}$ | $\mathbf{0.001}^{UD}$ | 0.506 | 0.907 | 0.591 | 1.839 | 0.572 | 1.621 |
| Dropout | $0.537^{UD}$ | $0.119^{UD}$ | $0.887^{U}$ | $\mathbf{0.001}^{UD}$ | 0.795 | 0.351 | 0.576 | 1.923 | 0.614 | 1.543 |
| GP | $0.023^{UD}$ | $0.365^{UD}$ | $0.766^{UD}$ | $0.235^{UD}$ | 0.369 | 0.739 | 0.213 | 2.132 | 0.365 | 1.580 |
| Multi-Task | $0.851^{UD}$ | $0.121^{UD}$ | $0.785^{UD}$ | $0.303^{UD}$ | 0.654 | 0.881 | 0.620 | 2.065 | 0.689 | 1.800 |
| Attenuation | $0.617^{UD}$ | $0.301^{UD}$ | $0.834^{UD}$ | $0.411^{UD}$ | 0.576 | 1.162 | 0.632 | 2.091 | 0.735 | 1.672 |

TABLE II

AVERAGED UNCERTAINTY METRICS OVER AUS. MNIST UNCERTAINTY METRICS ARE NOT AVERAGED. FOR MNIST, MARKED RESULTS INDICATE A SIGNIFICANTLY WORSE PERFORMANCE COMPARED TO U-MLP (U) / DWAR (D).

| | AU 6 | | AU 10 | | AU 12 | | AU 14 | | AU 17 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\rho \uparrow$ | $\lvert PI \rvert \downarrow$ | $\rho \uparrow$ | $\lvert PI \rvert \downarrow$ | $\rho \uparrow$ | $\lvert PI \rvert \downarrow$ | $\rho \uparrow$ | $\lvert PI \rvert \downarrow$ | $\rho \uparrow$ | $\lvert PI \rvert \downarrow$ |
| DWAR | **0.723** | $2.693^{U}$ | **0.543** | $2.551^{U}$ | **0.653** | $2.569^{U}$ | $0.749^{U}$ | 1.202 | $0.738^{U}$ | $0.080^{U}$ |
| U-MLP | $0.653^{D}$ | **1.541** | 0.520 | **1.596** | $0.631^{D}$ | **1.485** | 0.783 | 1.423 | **0.862** | **0.012** |
| Ensemble | $0.635^{D}$ | $2.750^{U}$ | $0.391^{D}$ | $2.571^{U}$ | $0.509^{UD}$ | $2.607^{U}$ | 0.761 | $1.189_{D}$ | $0.657^{UD}$ | $0.077^{U}_{D}$ |
| Dropout | $0.507^{UD}$ | $2.505^{U}_{D}$ | $0.418^{UD}$ | $3.356^{UD}$ | $0.389^{UD}$ | $2.617^{UD}$ | $0.702^{UD}$ | $\mathbf{1.107}^{U}_{D}$ | $\mathbf{0.862}_{D}$ | $0.029^{U}_{D}$ |
| GP | $0.339^{UD}$ | $2.524^{U}$ | $-0.035^{UD}$ | $3.130^{UD}$ | $-0.132^{UD}$ | $3.231^{UD}$ | $0.297^{UD}$ | $1.366^{U}$ | $0.598^{UD}$ | $0.407^{UD}$ |
| Multi-Task | $0.557^{UD}$ | $2.930^{U}$ | $0.392^{UD}$ | $2.888^{UD}$ | $0.576^{D}$ | $2.698^{UD}$ | $\mathbf{0.799}_{D}$ | 1.285 | $0.774^{U}_{D}$ | $0.523^{UD}$ |
| Attenuation | $0.632^{D}$ | $2.479^{U}$ | $0.464^{D}$ | $3.039^{UD}$ | $0.602^{D}$ | $2.701^{UD}$ | 0.771 | 1.258 | $0.692^{UD}$ | $0.976^{UD}$ |

TABLE III

STATISTICAL TESTS ON BP4D+. RESULTS MARKED IN SUPERSCRIPT/SUBSCRIPT INDICATE A SIGNIFICANTLY WORSE/BETTER PERFORMANCE COMPARED TO U-MLP (U) / DWAR (D).

| | Mean Coverage Rate | | |
|---|---|---|---|
| | DISFA | BP4D+ | BP4D+ → DISFA |
| DWAR | 0.953 | 0.934 | 0.964 |
| U-MLP | 0.961 | 0.935 | 0.955 |
| Ensemble | 0.962 | 0.938 | 0.937 |
| Dropout | 0.956 | 0.935 | 0.960 |
| GP | 0.944 | 0.941 | 0.953 |
| Multi-Task | 0.951 | 0.944 | 0.951 |
| Attenuation | 0.957 | 0.934 | 0.953 |

TABLE IV

OBSERVED COVERAGE RATE (RATIO OF THE TRUE VALUE BEING IN THE INTERVAL) FOR THE PREDICTION INTERVALS AVERAGED OVER AUS.

performance. The results suggest that estimating the error separately (as done in DWAR and U-MLP) outperforms these two baselines in both regards. However, it is important to note that both baselines have less computational overhead during training and testing than the U-MLP, only requiring back-propagation for an additional variable and one additional dot product at test time.

## VI. CONCLUSION

We evaluated the performance of two supervised approaches to the estimation of uncertaity, and compared their performance to several established approaches. Some of these approaches require the use of slightly different architectures (GP, multi-task MLP, and loss attenuation), some require secondary models (U-MLP, DWAR, and ensemble), and some generally do not require any changes for existing users (dropout). The results suggest that epistemic approaches achieve a worse perform than supervised approaches, perhaps because they do not capture aleatoric uncertainty. The clearly best-performing and most simple approach is the prediction of absolute error with a secondary MLP model (U-MLP). However, a notable results is that dropout provides decent uncertainty estimation while requiring the fewest changes during training.

A future avenue of work is the evaluation of uncertainty prediction for emotion recognition, as well as the developtment of more transparent approaches to ucnertainty (similar to DWAR), to break the cycle of needing an uncertainty estimation for the uncertainty.

## VII. ACKNOWLEDGMENTS

This material is based upon work partially supported by the U.S. National Science Foundation (Awards 1750439, 1722822, 1734868) and U.S. National Institutes of Health. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of U.S. National Science Foundation or U.S. National Institutes of Health, and no official endorsement should be inferred.

## REFERENCES

[1] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International*

*Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE, 2018.

[2] M. Bauer, M. van der Wilk, and C. E. Rasmussen. Understanding probabilistic sparse gaussian process approximations. In *Advances in neural information processing systems*, pages 1533–1541, 2016.

[3] H. Boström, H. Linusson, T. Löfström, and U. Johansson. Accelerating difficulty estimation for conformal regression forests. *Annals of Mathematics and Artificial Intelligence*, 81(1-2):125–144, 2017.

[4] D. Card, M. Zhang, and N. A. Smith. Deep weighted averaging classifiers. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 369–378. ACM, 2019.

[5] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016.

[6] T. Heskes. Practical confidence and prediction intervals. In *Advances in neural information processing systems*, pages 176–182, 1997.

[7] U. Johansson, H. Boström, T. Löfström, and H. Linusson. Regression conformal prediction with random forests. *Machine Learning*, 97(1-2):155–176, 2014.

[8] K. Kasiviswanathan and K. Sudheer. Methods used for quantifying the prediction uncertainty of artificial neural network based hydrologic models. *Stochastic environmental research and risk assessment*, 31(7):1659–1670, 2017.

[9] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017.

[10] Y. LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[11] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[12] H. Papadopoulos, A. Gammerman, and V. Vovk. Normalized nonconformity measures for regression conformal prediction. In *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2008)*, pages 64–69, 2008.

[13] H. Papadopoulos and H. Haralambous. Neural networks regression inductive conformal predictor and its application to total electron content prediction. In *International Conference on Artificial Neural Networks*, pages 32–41. Springer, 2010.

[14] H. Papadopoulos and H. Haralambous. Reliable prediction intervals with regression neural networks. *Neural Networks*, 24(8):842–851, 2011.

[15] H. Papadopoulos, V. Vovk, and A. Gammerman. Regression conformal prediction with nearest neighbours. *Journal of Artificial Intelligence Research*, 40:815–840, 2011.

[16] M. Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Artificial Intelligence and Statistics*, pages 567–574, 2009.

[17] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3438–3446, 2016.