Calibration of Cognitive Classification Systems for Radar Networks for Increased Reliability

Svenningsson, Peter; Kruse, Nicolas; Fioranelli, Francesco ; Yarovoy, Alexander

DOI
10.1109/RadarConf2248738.2022.9764292

Publication date
2022

Document Version
Final published version

Published in
2022 IEEE Radar Conference (RadarConf22) Proceedings

Citation (APA)
Svenningsson, P., Kruse, N., Fioranelli, F., & Yarovoy, A. (2022). Calibration of Cognitive Classification Systems for Radar Networks for Increased Reliability. In *2022 IEEE Radar Conference (RadarConf22) Proceedings* Article 9764292 (Proceedings of the IEEE Radar Conference). IEEE. https://doi.org/10.1109/RadarConf2248738.2022.9764292

Calibration of Cognitive Classification Systems for Radar Networks for Increased Reliability

Peter Svenningsson, Nicolas Kruse, Francesco Fioranelli and Alexander Yarovoy
Microwave Sensing, Signals and Systems (MS3)
Delft University of Technology
Delft, The Netherlands
{p.o.svenningsson, n.c.kruse, f.fioranelli, a.yarovoy}@tudelft.nl

***Abstract*—Cognitive radar frameworks rely on the ability to quantify and reason on future uncertainty, which allows for the selection of an optimal decision policy. These methods require that the uncertainty estimates provided by the underlying statistical model are well-calibrated, i.e. consistent with true uncertainty. In this work, the utilization of probability calibration techniques for target classification is explored. It is shown from simulations and experimental data that the proposed techniques can be used to correct errors in uncertainty estimates caused by incorrect modeling assumptions, such as the independence of sensors and the independence of classification covariates. This correction improves classification performance and the reliability of cognitive systems so that resources are utilized in accordance with user-defined cost functions.**

***Index Terms*—Cognitive radar, probability calibration, resource management, target classification**

I. Introduction

Networks of radar sensors can be utilized as an extension to single-radar observation and classification tasks, as demonstrated in [1] and [2]. This improved performance is attributable to e.g. complementary target information that spatially separated sensors can acquire. To leverage this potential, it is necessary to ensure consistency between the confidence of the classifier and the true posterior probability for a given prediction [3]. This confidence calibration is not only required for the selection of an appropriate number of measurements in a budget-limited radar network, but also for scenarios where class probabilities are not used in isolation. An example of the latter is sensor fusion or the evaluation of a Bayes estimator.

Radar resource management is a topical issue. In [4], an adaptive model is developed which estimates the predicted measurement uncertainty of a single radar sensor in order to make a decision on when to take additional measurements. This model is extended in [5] to include a decision on the type of waveform to be used, prioritizing either tracking or classification accuracy. A multi-target tracking solution is proposed in [6], where resources are to be allocated in the form of sensor time per target. Radar Resource management for classification is covered more extensively in e.g. chapter 5 of [7]. The resource allocation problem extends to networks of multiple sensors.

In this work, the calibration of a cognitive radar system is explored by the definition of a minimal radar resource problem. A simulation of such a cognitive system will be employed to investigate two methods of calibration: isotonic regression and logistic regression. Furthermore, potential improvements in classification accuracy when dealing with correlated sensors under a budget constraint will be demonstrated. Using an experimentally acquired dataset of human motion, the improved reliability of a classifier after calibration will be shown.

The main contributions in this work are as follows:

- Introduction of the calibration of probabilities for a full radar network performing a classification task.
- Systematic exploration of calibration through a simulated model, and validation with an experimental dataset from five radar nodes.
- Demonstration of the need for probability calibration in order to utilize radar resources in accordance with mission objectives.

In Section II, the cognitive framework itself will be defined. Section III contains a description of the case study which is conducted to investigate calibration of the cognitive system, followed by the results of this study in Section IV. For reproducibility, code has been made available[1].

II. Method

In this section, the constituent parts of the cognitive framework are described. First, the procedure of class estimation is described, followed by a definition of the resource management problem. Finally, the two types of calibration employed in this work are outlined. The notation is described in Table I.

A. Class Estimation

The class of a target y is estimated from an observation of features x which are modeled as independent and normally distributed. The posterior is estimated as,

$$p(y \mid x) \propto p(x \mid y)\, p(y), \tag{1}$$

which follows from Bayes' theorem for a prior distribution $p(y)$. A point estimate $\hat{y}$ of the posterior is selected as the maximum a posteriori estimate,

$$\hat{y} = \mathrm{MAP}\left(p(y \mid x)\right). \tag{2}$$

[1] https://github.com/petersvenningsson/probability-calibration

2022 IEEE Radar Conference (RadarConf22) | 978-1-7281-5368-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/RADARCONF2248738.2022.9764292

978-1-7281-5368-1/22/$31.00 ©2022 IEEE

TABLE I: Table of notation

$\mathrm{H}(x)$	Entropy operator	Entropy of a random variable x.
$\mathbb{E}_{p(x,y)}[\mathrm{H}(y \mid x)]$	Expected entropy	The expected entropy of $y \mid x$ under the joint distribution $p(y,x)$.
$\mathrm{MAP}(p(y \mid x))$	Maximum a posteriori estimate	A point estimate of $y \mid x$.
$Q(y \mid x)$	Prediction confidence	The probability of the MAP estimate of $p(y \mid x)$ being true.
$\mathrm{PCC}(x,y)$	Pearson correlation coefficient	The degree of linear correlation between random variables x and y.
$\mathrm{PCC}(x^{(i,l)}, x^{(i',l)})$	Feature correlation	The degree of correlation between two features measured by sensor l.
$\mathrm{PCC}(x^{(i,l)}, x^{(i,l')})$	Sensor correlation	The degree of correlation for a feature when measured by sensor l and l'.
$\langle\cdot\rangle$	Ensemble	Mean value of quantity $\cdot$ over a large number of samples.

B. Resource Management Problem

The resource management problem under consideration consists of the selection of a sufficient amount of sensor measurements to reach a goal posterior confidence q_τ. Let y indicate a class variable in a binary classification problem and x_k a measurement taken from k sensors. Following the methodology of [4], a constraint is imposed that a goal entropy H_τ should be reached. An optimization problem can then be defined as:

$$\begin{aligned} \min_k \quad & k \\ \text{s.t.} \quad & \mathbb{E}_{p(x,y)}\left[\mathrm{H}(y \mid x_k)\right] > H_\tau . \end{aligned} \tag{3}$$

Here, k denotes the number utilized sensors and H the entropy operator. To aid interpretability, the goal entropy H_τ is mapped to a goal confidence q_τ. For the binary entropy function,

$$\begin{aligned} \mathrm{H}(y \mid x_k) = & -p(y \mid x_k)\log p(y \mid x_k) - \dots \\ & (1-p(y \mid x_k))\log(1-p(y \mid x_k)), \end{aligned} \tag{4}$$

let the confidence be defined as

$$\mathrm{Q}(y \mid x) = \mathrm{H}^{-1}\mathrm{H}(y \mid x), \tag{5}$$

where H^{-1} is restricted to the range $[0.5, 1]$. The optimization problem can then be expressed as

$$\begin{aligned} \min_k \quad & k \\ \text{s.t.} \quad & \mathbb{E}_{p(x,y)}\left[\mathrm{Q}(y \mid x_k)\right] > q_\tau , \end{aligned} \tag{6}$$

where

$$\mathbb{E}_{p(x,y)}\left[\mathrm{Q}(y \mid x_k)\right] \approx \langle \mathrm{Q}(y \mid x_k)\rangle , \tag{7}$$

is estimated over N samples drawn from the recorded dataset.

A prediction $\hat{y}$ has a probability $p(y=\hat{y} \mid x)$ of being true. It follows from the symmetry of the binary entropy $\mathrm{H}(y \mid x)$ that many predictions drawn from $p(y \mid x_k)$ will reach $\mathrm{Q}(y \mid x_k)$ accuracy and that q_τ can be interpreted as a goal accuracy.

C. Calibration

Predicted probabilities estimated by a classification model can generally not be equated to the probability of the prediction being true. As investigated in [8], various classification models tend to produce class probabilities that are incorrect estimates of true posterior probability due to their methods of estimation. In this section, two calibration techniques are described that are aimed to improve these estimates.

Calibration through logistic regression relies on the assumption that class predictions are related to true posterior probabilities through a sigmoidal transformation. The method, proposed by Platt[9], employs a function with two parameters to calibrate the class predictions $\hat{y}_i$,

$$\mathrm{m}\left(\hat{y}_i\right) = \frac{1}{1+\exp\left(A\hat{y}_i + B\right)} \tag{8}$$

where A and B are the function parameters. The parameters are fit to a training set by minimizing the negative log likelihood of the set,

$$\begin{aligned} \mathrm{argmin}_{A,B}\Big\{ -\sum_i & y_i \log(\mathrm{m}(\hat{y}_i)) + \dots \\ & (1-y_i)\log(1-\mathrm{m}(\hat{y}_i))\Big\} \end{aligned} \tag{9}$$

with y_i indicating the true class labels.

A generalization of the logistic calibration described above is calibration by isotonic regression, where the restriction of sigmoid shape is released. Following [10], the isotonic calibration function,

$$\hat{\mathrm{m}} = \mathrm{argmin}_{\mathrm{m}}\left\{\sum_{i\in\mathcal{V}}\left(y_i - \mathrm{m}(\hat{y}_i)\right)^2\right\}, \tag{10}$$

is only restricted to be piecewise-constant and non-decreasing. The regression can be fit using the pair-adjacent violators algorithm [11].

III. Case Study

A case study in two parts is conducted to investigate probability calibration of a cognitive system. The first part is a simulation of such a system, followed by experimental validation.

A. Simulation

To demonstrate the effects of probability calibration for the resource management problem defined in (6), a cognitive system is simulated. The classification model from Section II-A is employed to perform a binary classification task whilst the cognitive system can select the amount of required measurements from a set amount of sensors. Each sensor draws a measurement from a bivariate normal distribution which comprise two classification features, $x^{(1,l)}$ and $x^{(2,l)}$ drawn from sensor l. The distribution has unit variance and means shown in Table II.

TABLE II: Parameters of Gaussian distribution $x \mid y$.

		Mean		Variance	
Class	$p(y)$	$x^{(1)}$	$x^{(2)}$	$x^{(1)}$	$x^{(2)}$
Class 1	0.5	0.5	-0.5	1.0	1.0
Class 2	0.5	1.0	0.0	1.0	1.0

Correlations in measurements are induced in two distinct ways. Pearson correlation coefficient is defined as,

$$\mathrm{PCC}(A, B) = \frac{\mathrm{cov}(A, B)}{\sigma_A \sigma_B}$$

for random variables A and B and their respective standard deviations σ_A and σ_B. Sensor correlation is defined as the correlation coefficient of feature i,

$$\text{Sensor correlation} = \mathrm{PCC}(x^{(i,l)}, x^{(i,l')}),$$

for sensors l and l'. Feature correlation is defined as the correlation coefficient of the two features which comprise the bivariate feature distribution measured by the same sensor,

$$\text{Feature correlation} = \mathrm{PCC}(x^{(i,l)}, x^{(i',l)}),$$

for feature i and i'.

The goal accuracy is set according to operational goals. The effectiveness of probability calibration is evaluated under varying degrees of sensor correlation. Feature correlation is fixed at 0.0 unless otherwise indicated. The conditional distributions and the calibration parameters are fit from 10^6 samples and are evaluated on 10^5 samples. The total number of available sensors is 50 and the predictions are estimated from a uniform prior.

It is expected that as the correlation between classification increases, the classification model described in Section II-A should provide estimates of $p(y \mid x)$ which diverges from the true posterior. This is a consequence of that the model assumes that there is no correlation between the features. The posterior estimate of $p(y \mid x)$ is then corrected by the calibration methods described in Section II-C.

B. Experiment

To validate the results obtained from the simulated system, the same methodology is applied to experimental data recorded from a human target which is either walking or stationary. Full details of the experimental setup can be found in [12]. Measurements are taken by a set of five PulsON P410 Ultra Wideband radars. The monostatic, omnidirectional radars are arranged in a semicircle with a diameter of $6.38\,\mathrm{m}$ and have a center frequency and bandwidth of $4.3\,\mathrm{GHz}$ and $2.2\,\mathrm{GHz}$ respectively. The acquired dataset can be found in [13].

Radar measurements are processed to yield velocity-time representations, or spectrograms, of the complete sequences. The spectrograms are divided into segments of $1\,\mathrm{s}$, and 20 features are extracted from each segment through Principal Component Analysis. Class estimation is subsequently performed as described in Section II-A.

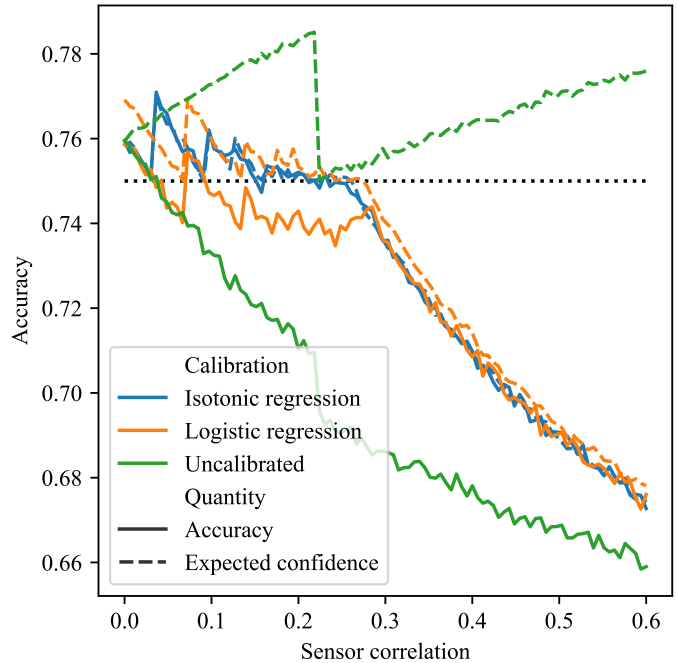

Fig. 1: Accuracy versus sensor correlation for independent features. Goal confidence is indicated by a dotted black line. Accuracy decreases immediately with increasing sensor correlation in the case of an uncalibrated model. The calibrated models are able to reach and maintain the goal accuracy until the point where all sensors have been utilized.

IV. Results

In this section, the results for the case study described in Section III are presented.

A. Simulation Results

Figure 1 shows the expected confidence and the accuracy of the classification model with and without calibration. The horizontal axis represents the amount of correlation between the sensors and the goal confidence is set to $75\,\%$. At 0.0 sensor correlation both the calibrated and uncalibrated models can achieve the required goal accuracy, as the expected confidence is correctly estimated. For increasing values of sensor correlation, the uncalibrated accuracy diverges from the goal accuracy as the uncalibrated model incorrectly estimates the expected posterior distribution, whilst the calibrated models increase the amount of sensor measurements to maintain the required accuracy.

The increased sensor utilization is displayed in Figure 2. The rapid growth in sensor usage for greater sensor correlation is explained by the diminishing contributions of additional sensors. For sensor correlation greater than 0.3, the calibrated systems are no longer able to utilize additional sensors, which is seen in Figure 1 as the inability to reach the goal accuracy. Important to note is that, whilst the calibrated models can no longer reach the goal accuracy, their expected confidence and accuracy are decreasing in equal measures. This contrasts with

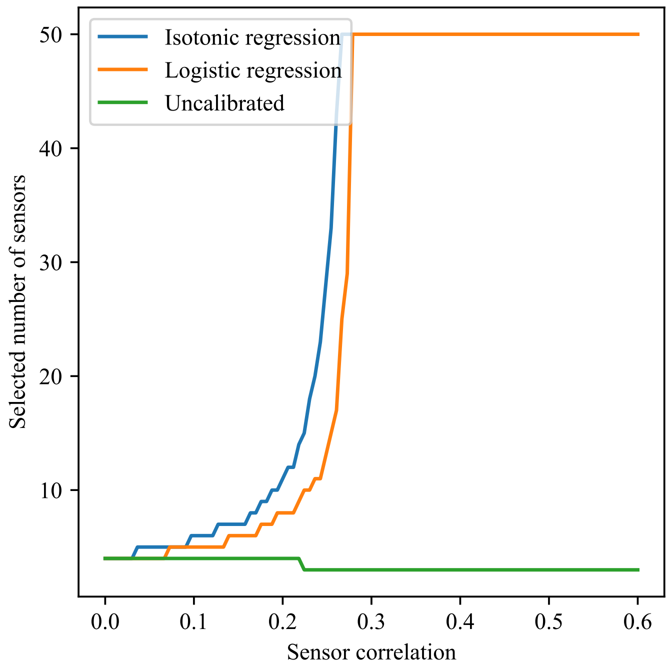

Fig. 2: Sensor utilization versus sensor correlation for simulated data. Apparent features of the graph are the rapid increase in required sensor utilization for the calibrated models and the contrasting lack of sensor utilization for the uncalibrated model.

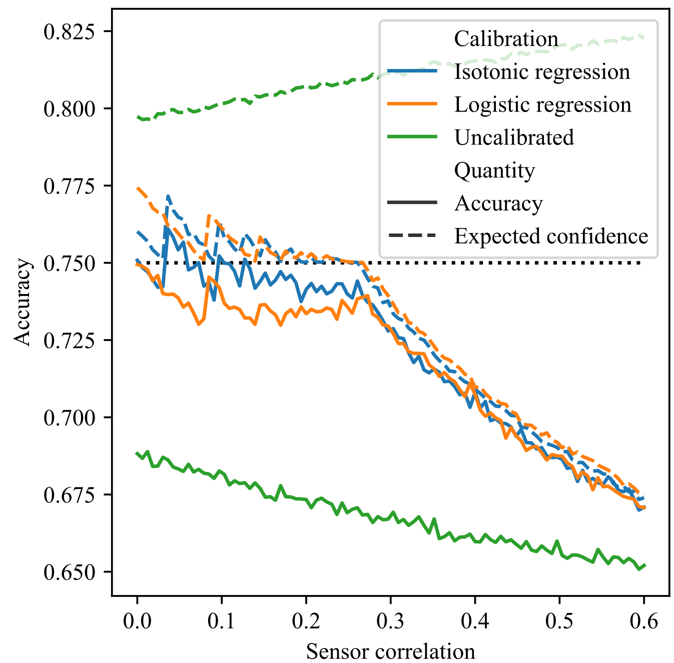

Fig. 3: Accuracy versus sensor correlation for correlated features. Goal confidence is indicated by a dotted black line. Of note is the inability of the uncalibrated model to utilize enough sensors at 0.0 sensor correlation.

the uncalibrated model and indicates that a good estimate can be made on the ability to reach a goal accuracy with a given amount of sensors.

In Figure 3, both the features and the sensors are correlated with feature correlation fixed at 0.3. It can be seen in the figure that even for completely independent sensors, the uncalibrated classification model is overconfident and the cognitive system does not take enough measurements to reach the goal accuracy. In the case of human activity classification, it is realistic that features are correlated to some degree, as they are different characterizations of the same measurement. This result thus shows the improved performance of calibrated classification models.

To further demonstrate the improved sensor utilization of the calibrated models, a plot of accuracy versus goal accuracy is shown in Figure 4 for a fixed sensor correlation of 0.3. Under ideal circumstances, a cognitive system will make a perfect estimate of the true posterior probability and will thus tend to the diagonal in this figure. This implies that proximity to the diagonal is indicative of a well-calibrated model. For lower goal accuracies, all three models achieve an equal accuracy of approximately 0.64. This is due to the minimum requirement of a single measurement resulting in a minimum accuracy, regardless of calibration. Moving towards higher goal accuracies, it can be seen that the calibrated models remain closer to the optimal diagonal, as the improved expected confidence provides a good indication of when to utilize an additional sensor. This trend continues until the maximally attainable accuracy of approximately 0.78 is reached, after which the expected confidence for the calibrated models remains constant.

Finally, Figure 5 displays reliability curves for all models, to indicate the relations between predicted confidence and true posterior probability. As mentioned in Section II, given enough samples, the predicted confidence of a well-calibrated model should approach the true posterior probability. Thus, the diagonal in the figure again represents ideal calibration. The inverse-sigmoidal shape of the uncalibrated model is typical for statistical classifiers which assume feature independence [8]. Due to this shape and the symmetry of the uncalibrated reliability curve around $(0.5, 0.5)$, the assumptions for logistic regression to produce a well-calibrated model are largely fulfilled. Calibration through isotonic regression provides the most significant improvement over the uncalibrated model.

B. *Experimental Results*

A plot of accuracy versus goal accuracy is shown in Figure 6. The results are similar to those acquired from simulation, with two notable differences. First, the uncalibrated model is overconfident to a greater extent than in the simulated results. This may be explained by high degrees of correlation between the features used for activity classification. Secondly, fewer steps are present in all plots due to the amount of sensors being limited to five. It is apparent that the calibrated models almost exclusively have an accuracy above the goal accuracy, up until the point where there are no additional sensor measurements to

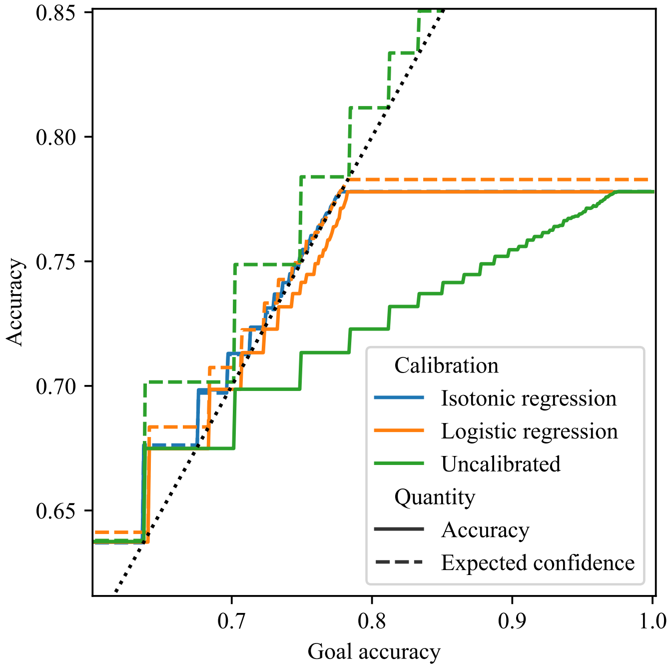

Fig. 4: Accuracy versus goal accuracy for simulated data for a maximum of 50 sensors. Dotted black line indicates equality between accuracy and goal accuracy. Every vertical step in the plots indicates the utilization of an additional sensor.

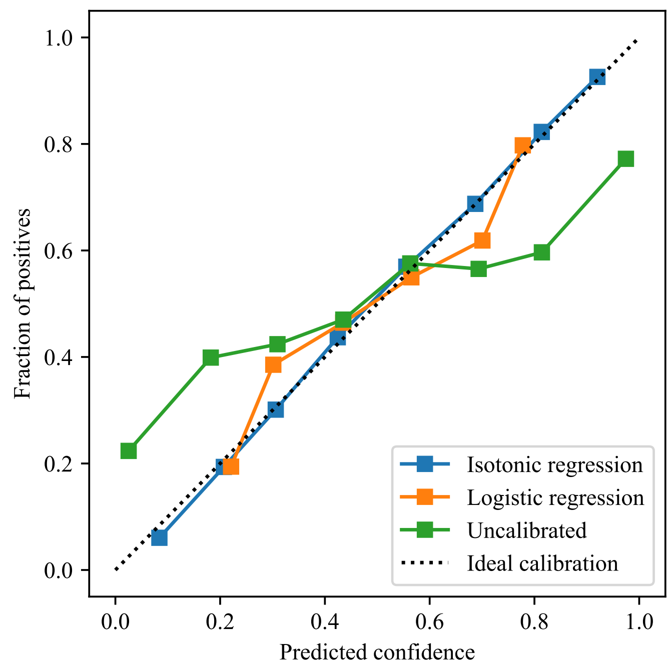

Fig. 5: Reliability curves for calibrated and uncalibrated models for simulated data. Ideal calibration is indicated with a dotted black line. Overconfidence of the uncalibrated model is indicated by the inequality of the predicted confidence and the fraction of positives.

utilize. Furthermore, the expected confidence of the calibrated models bears far better correspondence to the accuracy than in the case of the uncalibrated model.

In Figure 7, the reliability curves for all three models are shown. It can be seen that, when compared to simulation results, logistic regression provides less improvement in calibration effectiveness. This is primarily due to the uncalibrated model not fulfilling the assumption of sigmoidal shape that is necessary for the effective application of logistic calibration. The calibration can however be seen to be effective for the interval $[0.8, 1.0]$, which explains the improvement over the uncalibrated model in Figure 6. Isotonic regression shows minimal deviation from ideal calibration when compared to the uncalibrated model and can be considered to be an effective method in this scenario.

V. Conclusions

In this work, calibration of class estimation is explored in the context of a cognitive system resource management problem. Improvements in reliability are demonstrated for a simulated system through the usage of logistic and isotonic probability calibration. The improved reliability results in better estimates of the required number of measurements to achieve a goal accuracy in a classification task. The simulation results are experimentally verified using radar data of human activities. It is shown that in both cases, isotonic regression is a more effective method of calibration.

The task of class estimation in this work has been performed through statistical modeling. In future works, applicability to approaches based on non-statistical models such as e.g. support vector machines or other machine learning algorithms may be investigated. Additionally, enhancements in simulation fidelity may lead to novel insights. This may for example be achieved through the simulation of nonlinear correlations in data, or by mimicking realistic sensor correlations for a certain radar network geometry. Finally, the method proposed in this work may be extended for a dynamical model, to develop improved real-time decision processes for cognitive systems.

Acknowledgment

Research was sponsored by the Army Research Laboratory and was accomplished under Grant Number W911NF-2020-185. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

The authors are grateful to R. G. Guendel and all the participants in the data collection.

References

[1] H. Li, A. Mehul, J. Le Kernec, S. Z. Gürbüz, and F. Fioranelli, "Sequential Human Gait Classification with Distributed Radar Sensor Fusion," *IEEE Sensors Journal*, vol. 21, no. 6, pp. 7590–7603, 2020.

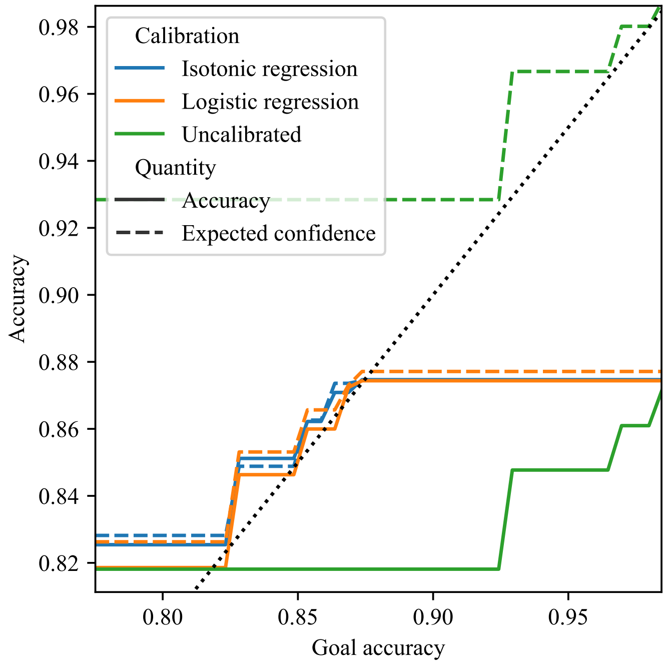

Fig. 6: Achieved accuracy versus goal accuracy for experimental data collected with five radar sensors. Dotted black line indicates equality between accuracy and goal accuracy. Every vertical step in the plots indicates the utilization of an additional sensor.

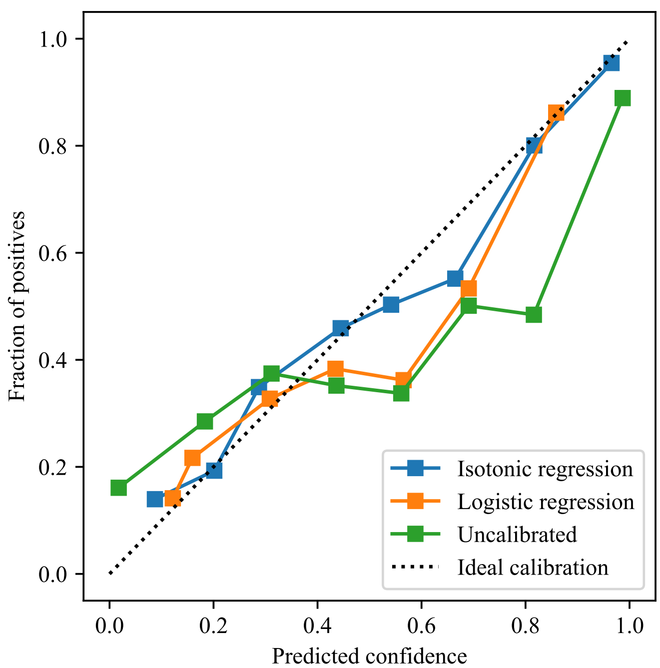

Fig. 7: Reliability curves for calibrated and uncalibrated models for experimental data. Ideal calibration is indicated with a dotted black line. Overconfidence is indicated by the inequality of predicted confidence and fraction of positives.

[2] N. Kern, M. Steiner, R. Lorenzin, and C. Waldschmidt, "Robust Doppler-Based Gesture Recognition with Incoherent Automotive Radar Sensor Networks," *IEEE Sensors Letters*, vol. 4, no. 11, pp. 1–4, 2020.

[3] S. Särkkä, *Bayesian filtering and smoothing*. Cambridge University Press, 2013, no. 3.

[4] K. L. Bell, G. E. Smith, A. E. Mitchell, and M. Rangaswamy, "Fully adaptive radar for target classification," in *2019 IEEE Radar Conference (RadarConf)*, 2019.

[5] K. Bell, C. Kreucher, and M. Rangaswamy, "An evaluation of task and information driven approaches for radar resource allocation," in *2021 IEEE Radar Conference (RadarConf21)*, 2021, pp. 1–6.

[6] M. I. Schöpe, H. Driessen, and A. Yarovoy, "Multi-task sensor resource balancing using lagrangian relaxation and policy rollout," in *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, 2020, pp. 1–8.

[7] M. I. Schöpe, "Approximately optimal resource management for multi-function radar algorithmic solutions using a generic framework," Ph.D. dissertation, Delft University of Technology, 2021.

[8] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proceedings of the 22nd International Conference on Machine Learning (ICML 2015)*, 01 2005, pp. 625–632.

[9] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*. MIT Press, 1999, pp. 61–74.

[10] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 08 2002.

[11] M. Ayer, H. D. Brunk, G. Ewing, W. T. Reid, and E. Silverman, "An empirical distribution function for sampling with incomplete information," *Annals of Mathematical Statistics*, vol. 26, pp. 641–647, 1955.

[12] R. G. Guendel, M. Unterhorst, E. Gambi, F. Fioranelli, and A. Yarovoy, "Continuous human activity recognition for arbitrary directions with distributed radars," in *2021 IEEE Radar Conference (RadarConf21)*, 2021, pp. 1–6.

[13] R. G. Guendel, M. Unterhorst, F. Fioranelli, and A. Yarovoy, "Dataset of continuous human activities performed in arbitrary directions collected with a distributed radar network of five nodes," Nov 2021. [Online]. Available: https://data.4tu.nl/articles/dataset/Dataset_of_continuous_human_activities_performed_in_arbitrary_directions_collected_with_a_distributed_radar_network_of_five_nodes/16691500/2