https://doi.org/10.2352/EI.2023.35.15.COLOR-194

Active learning approaches to analysis of thin-film printed sensors for determining nitrate levels in soil [1]

Xihui Wang, Ali Shakouri, Bruno Ribeiro, George T.C. Chiu, Jan P. Allebach;
Purdue University, West Lafayette, Indiana, United States

Abstract

In order to train a learning-based prediction model, large datasets are typically required. One of the major restrictions of machine learning applications using customized databases is the cost of human labor. In the previous papers [3, 4, 5], it is demonstrated through experiments that the correlation between thin-film nitrate sensor performance and surface texture exists. In the previous papers, several methods for extracting texture features from sensor images are explored, repeated cross-validation and a hyperparameter auto-tuning method are performed, and several machine learning models are built to improve prediction accuracy. In this paper, a new way to achieve the same accuracy with a much smaller dataset of labels by using an active learning structure is presented.

Introduction

The ultimate goal of this research is to develop an image-based machine learning technique that will enable real-time quality evaluation of each sensor in the manufacturing pipeline. Prior research has demonstrated that the texture of the sensor ion-selective membrane (ISM) layer is related to its performance [1, 2]. Our strategy aims to train a learning-based system to predict the performance of any given sensor from a still image of the sensor's active region using the established correlation. This is an alternative to random sample testing which is time and labor-intensive and hence cannot account for all of the individual sensors.

This study presents a follow-up project building on previous research [3, 4, 5]. The most recent improvement made in this research is the introduction of an active learning approach [8, 9] and two additional machine learning models [18, 19] to further optimize our training model while reducing the size of the training dataset. The new prediction system will comb through the unlabeled dataset in search of the data in which it is least confident in processing. Prioritizing human labor to label just the data that the learning model struggles to identify, therefore reduces labor costs and improves training effectiveness.

[1]This material is based upon work supported by the National Science Foundation under Grant No. 2134667.

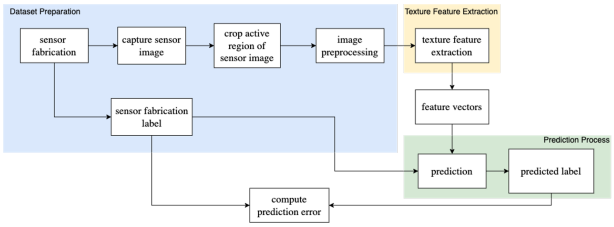


Figure 1. *Overview of the Prediction System.*

The methodology of the study is illustrated in Figure 1. The texture feature vector extracted from the sensor image serves as the feature vector, while the sensor fabrication date is used as the label. The active learning structure with varying base models is employed as the prediction model in this study. The output trained model will be the trained base model in the saturation region.

Dataset Preparation

The data for this project is collected in-house, including the feature vectors and their labels.

For research purposes, the labels will be the different fabrication settings during the sensor manufacture procedure and will be represented by the sensor fabrication date.

Feature Extraction

The feature vector utilized in this research is extracted from sensor images. As hypothesized by the underlying physics model [2], variations in fabrication factors are expected to influence the texture difference exhibited by the sensor image, thereby impacting the overall sensor performance.

In order to obtain the feature vector, a series of steps are necessary, as outlined in this research [3]. To avoid distracting our prediction system, we need to crop the sensor active region, following the protocol depicted in Figure 2. Subsequently, the cropped sensor active region image undergoes the preprocessing procedure, as illustrated in Figure 3, to enhance the texture differences.

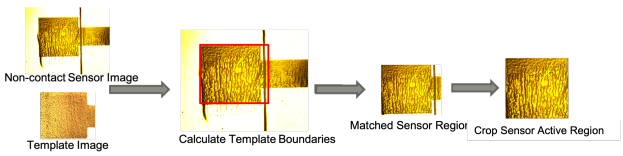


Figure 2. *Crop the sensor ROI Procedure.*

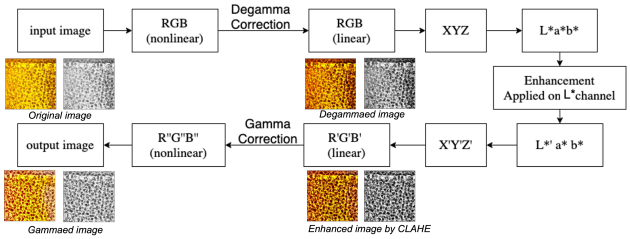


Figure 3. *Sensor Image Preprocessing Procedure.*

After processing the sensor images, the feature vectors are extracted from the preprocessed sensor images. Two methods are employed in this experiment to extract the feature vectors: the uniform local binary pattern (LBP) method [6] and the angularly averaged power spectrum (AAPS) method [7]. These two methods were chosen based on their consistent and precise prediction outcomes in previous experiments [5].

Summary of the Dataset

As illustrated in Figure 4, several samples of the preprocessed sensor region of interest (ROI) are presented for each sensor fabrication date. It should be noted that each sensor fabrication label class comprises both defective and normal sensors. For the purposes of this study, the customized dataset is comprised of eight classes, corresponding to the sensor fabrication labels, and in total includes 968 sensors. Although this dataset is unbalanced, the active learning structure ensures that the unbalanced dataset will not negatively affect the accuracy of the prediction.

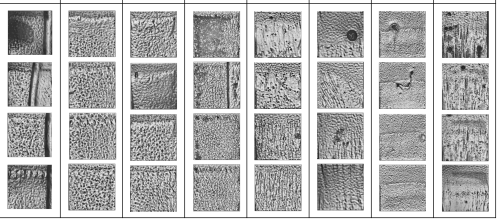
Sensor Fabrication Date	19-08-15	19-09-12	19-09-20	19-12-10	20-01-30	20-03-10	21-02-25	21-05-10
# of sensors	88	126	123	97	150	140	129	115
Examples of the cropped & pre-processed sensor ROI								

Figure 4. *Summery of the Dataset.*

Prediction System

The active learning algorithm, a form of semi-supervised machine learning algorithm, has the potential to achieve superior levels of accuracy while utilizing a smaller quantity of training labels [8, 9]. This outcome may be attainable by enabling the algorithm to prioritize the data that contributes the most to its learning progress.

Overview of Active Learning Algorithm

The active learning algorithm leverages both labeled and unlabeled data to train the model. Initially, the base model of the active learning algorithm is trained using a limited labeled dataset. Following this, a query strategy is selected to determine the uncertainty of the data points in the unlabeled dataset. The algorithm then selects the feature vectors that will be most beneficial to learn from based on the sampling strategy chosen, and we will provide the labels for the queried subset of the unlabeled dataset. The updated labeled dataset is then employed to retrain the base model, as depicted in Figure 5. Consequently, the labeled dataset increases dynamically during the training phase, significantly reducing the quantity of labeled data required to train the model compared to traditional machine learning algorithms, and ultimately decreases cost.

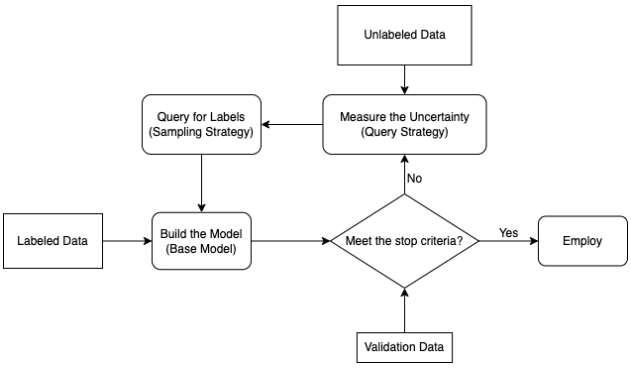


Figure 5. *Overview of Active Learning Strategy.*

In the active learning algorithm, three critical components require consideration, namely the query strategy, the sampling strategy, and the calibrated base model. Moreover, we define the number of query iterations as the number of times the active learning model queries for labels (also referred to as the number of repeating cycles). The initial training process will be denoted as query iteration 0.

Query Strategy

Various query strategies are employed to calculate the uncertainty of the unlabeled data. Typically, the data points with low confidence are considered the most uncertain, such as those that reside near the class boundaries. By selecting these data points, the base model can benefit more and acquire more information during the training process.

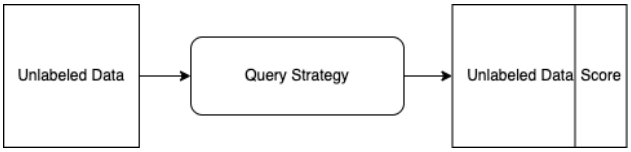


Figure 6. *Query Strategy.*

The least confidence (LC) strategy [10] is a query strategy used to select the data point which has the least likelihood in its most likely label as shown in equation (1). For each feature vector x, the predicted class is denoted as $\hat{y}$. Since the LC strategy tends to choose data points with low confidence, such as those that lie near the class boundaries, it may provide more information for the base model to learn from. In the example experiment illustrated in Figure 7, the accuracy increases significantly in the first few iterations, but then gradually slows down over time.

$$S_{LC} = \arg\min_{x}(P(\hat{y}\,|\,x)) \tag{1}$$

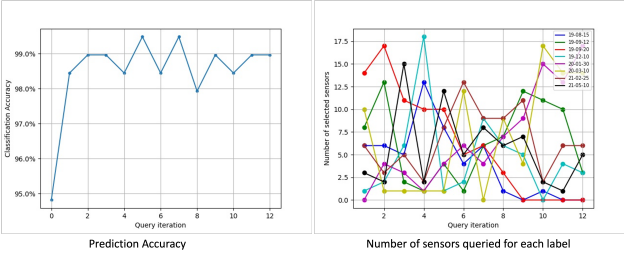


Figure 7. *Example of Least Confidence Strategy.*

The margin sampling (MS) strategy [11] illustrated by equation (2) selects the data point with the smallest difference between the two most probable labels. The most likely predicted class is denoted as $\hat{y}max$ and the second most likely predicted class as $\hat{y}max-1$. As shown in Figure 7 and Figure 8, the MS and LC strategies have different selection criteria that influence the accuracy of the algorithm. In the initial iterations, the MS strategy focuses on deciding between the two most likely labels, while the LC strategy selects the most unsure data points. The accuracy of the algorithm with MS strategy improves rapidly in the later iterations.

$$S_{MS} = \arg\min_{x}(P(\hat{y}_{max}\,|\,x) - P(\hat{y}_{max-1}\,|\,x)) \tag{2}$$

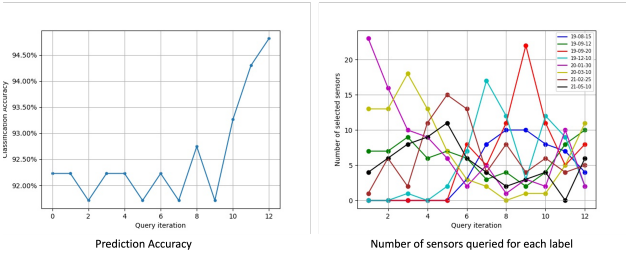


Figure 8. *Example of Margin Sampling Strategy.*

The entropy sampling (ES) strategy [12] illustrated by equation (3) selects the data with the greatest entropy value, where entropy indicates the randomness of the data. The accuracy curve performs between the LC and the MS strategies as shown in Figure 9.

$$S_{ES} = \arg\max_{x}(-\sum_{i} P(\hat{y}_i\,|\,x))\,\log(P(\hat{y}_i\,|\,x)) \tag{3}$$

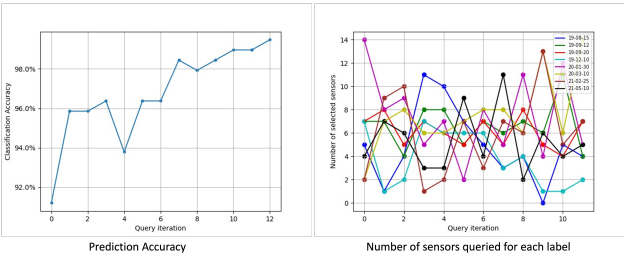


Figure 9. *Example of Entropy Sampling Strategy.*

The linear combination strategy will combine the uncertainty score of the three strategies linearly by assigning the strategies different weights as shown in equation (4). The accuracy curve will depend on the different weights assigned to each strategy.

$$S_{comb} = w_1 S_{LC} + w_2 S_{MS} - w_3 S_{ES} \tag{4}$$

For the experiment outlined in this paper, the least confidence (LC) strategy will be chosen to measure the uncertainty score. This decision is motivated by the fact that the LC strategy boosts the prediction accuracy at early iterations and therefore reduces the size of the labeled training dataset that is needed.

Sampling Strategy

In active learning, there are several sampling strategies that can be applied to select the most informative data points from the unlabeled dataset. These include stream-based selective sampling, pool-based sampling, and membership query synthesis. The selected subset will then be combined with the labeled training dataset and used to train the base model again. This process is illustrated in Figure 10.

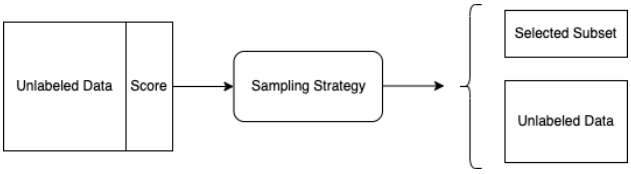


Figure 10. *Sampling Strategy.*

Membership query synthesis [15] involves the model generating its own feature vectors for labeling. This strategy does not apply to our case as we are dealing with real sensors manufactured with different fabrication settings.

In the stream-based strategy [14], the algorithm is presented with a stream of unlabeled data, with each data point being considered individually based on a fixed threshold. As a result, the number of data points in the selected subset may vary each time, and some data points may never be selected for the subset, thus never being used to train the base model.

On the other hand, pool-based sampling [13] involves selecting a fixed number of elements from the entire unlabeled dataset based on the uncertainty score and using that as the selected subset. This strategy has been selected for our experiment as it allows us to use all the data points in the unlabeled dataset to train our model, as long as the active learning algorithm is run for a sufficient number of query iterations.

Base Model Selection

The choice of the base model for an active learning algorithm plays a significant role in its overall performance. Previous studies have mainly used support vector for classification (SVC) [16] and random forest (RF) models [17]. As part of this study, we explore two additional methods, namely the XGBoost and KNN models.

The XGBoost model [18] is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning model. It creates an ensemble model by sequentially combining several weak decision trees to improve the overall accuracy of the system.

On the other hand, the KNN (k nearest neighbors) model [19] is a non-parametric, supervised model based on the k-nearest neighbors' vote. The weight of each neighbor can be uniform or based on the distance between the testing data point and each neighbor point.

Calibration Procedure

Calibrating the base model is crucial for calculating the uncertainty of each data point, as it relies on the true likelihood [21].

A well-calibrated model ensures that the confidence level of an event is accurately reflected in the model's predictions. A model is well-calibrated if, for any probability value p, a prediction of a class with confidence p is correct p percent of the time, as shown in Equation 5.

$$P(\hat{y} = y \,|\, \hat{P} = p) = p \tag{5}$$

To ensure the accuracy of the uncertainty measurements, it is necessary to calibrate the base model. In this study, we use a post-training calibration layer on top of the base model, which maps classifier confidence levels to better probabilities, as illustrated in Figure 11. Specifically, the base model will serve as the classification model and output the probability of each class. The calibration model, on the other hand, will be a regression model to map the true likelihoods to the predicted probabilities. In this experiment, we adopt the Sigmoid method [20] as the calibration model.

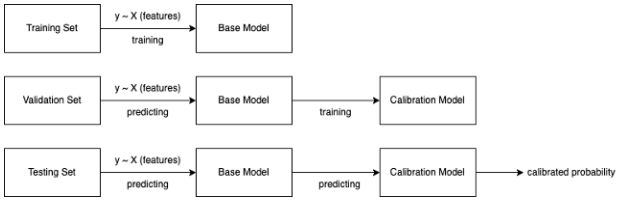


Figure 11. *Calibration Procedure.*

It is important to note that there is no correlation between calibration and accuracy, so accuracy may be improved or reduced by calibrating the model.

Auto-tuning Procedure

To improve the accuracy of the predicted result, we will follow the previous work to auto-tune the base model in order to get the optimized hyperparameter settings for each base model [5].

Experimental Results

The prediction system used in this experiment is illustrated in Figure 12. The dataset is composed of two components, the feature vectors extracted from sensor images and the associated fabrication labels. The customized dataset includes 968 sensors. Initial training data constitutes 20% of the entire dataset, followed by testing data constituting another 20%, and the remaining 60% is considered "unlabeled". During each iteration, the algorithm selects 5% of the data from the "unlabeled" set, queries for the label, and trains the base model with the newly acquired data. As a result, by iteration 12, the "unlabeled" dataset is exhausted and the base model behaves like a traditional machine learning model.

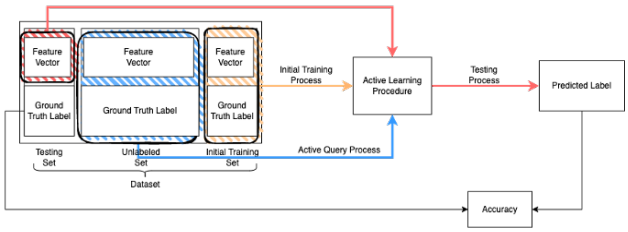


Figure 12. *Overview of the Prediction System.*

In this experiment, we used four calibrated base models: RF, XGBoost, KNN, and SVC models [17, 16, 18, 19]. For research purposes, the stopping criterion for the active learning procedure is to run the algorithm until the "unlabeled" set is empty. We extracted two types of feature vectors: LBP features [6] and AAPS features [7] as discussed in the Dataset Preparation section.

With each base model and type of feature, we run the active learning algorithm 200 times with randomly shuffled training, testing, and "unlabeled" sets.

Table 1 presents the results of the active learning algorithm using different base models with the feature vectors extracted using the LBP method. The SVC model provided the highest average accuracy and the smallest standard deviation at saturation. Additionally, the active learning algorithm with the SVC base model saturated at iteration 4, while the RF base model saturated at iteration 2. That means the active learning structure with the RF model could be well-trained using 30% of the entire dataset, which is about 291 sensors. The SVC model with active learning structure will reach saturation status with 40% of the entire dataset. The accuracy curve of the SVC model for the LBP feature is illustrated in Figure 13.

Table 1: Prediction Results for LBP Feature Vector

Base Model	Average Accuracy	Standard Deviation	Saturated Query Iteration
RF	0.9651	0.01416	**2**
XGBoost	0.9583	0.01466	4
KNN	0.9737	0.01119	3
SVC	**0.9743**	**0.01108**	4

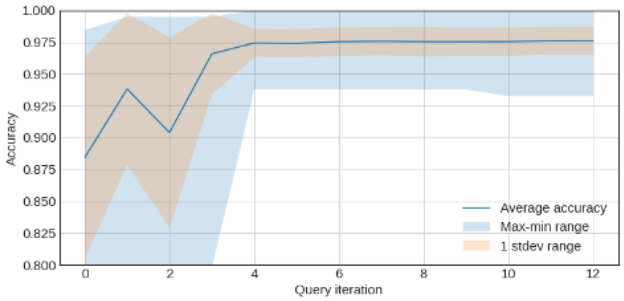


Figure 13. *Active learning results for the SVC model with the LBP feature set.*

Table 2 shows the results of the active learning algorithm using different base models with the feature vectors extracted from the AAPS method. The KNN model gives us the best average

accuracy and smallest standard deviation at saturation, while the active learning algorithm with KNN base model will saturate at iteration 4 with 40% of the entire dataset. Figure 14 shows the accuracy curve of the KNN model for the AAPS feature. It should be note that the highest achievable accuracy with the AAPS feature set is 14.1% lower than that obtained with the LBP feature set.

Table 2: Prediction Results for AAPS Feature Vector

Base Model	Average Accuracy	Standard Deviation	Saturated Query Iteration
RF	0.8208	0.02731	4
XGBoost	0.8221	0.02552	**3**
KNN	**0.8333**	**0.02432**	4
SVC	0.8239	0.02495	**3**

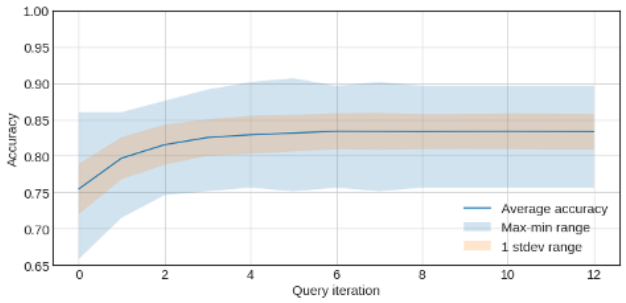


Figure 14. *Active learning results for the KNN model with the AAPS feature set.*

Conclusion

With the newly implemented prediction model with the active learning structure, the same accuracy can be achieved with a smaller number of training labels. The active learning algorithm prioritizes the data that contributes the most to its learning progress, reducing the quantity of labeled data required to train the model, and ultimately decreasing cost.

The customized dataset comprised of eight classes and a total of 968 sensors demonstrates the effectiveness of the proposed approach. We are able to reach about 97.43% accuracy with 40% of the whole dataset as the training dataset.

References

[1] J. Hu, A. Stein, and P. Bühlmann, "Rational design of all-solid-state ion-selective electrodes and reference electrodes," TrAC Trends in Analytical Chemistry, vol. 76, pp. 102–114, 2016.

[2] X. Jin, A. Saha, H. Jiang, M. R. Oduncu, Q. Yang, S. Sedaghat, D. K. Maize, J. P. Allebach, A. Shakouri, N. J. Glassmaker, et al., "Steady-state and transient performance of ion-sensitive electrodes suitable for wearable and implantable electro-chemical sensing," IEEE Transactions on Biomedical Engineering, vol. 69, no. 1, pp. 96-107, 2021.

[3] X. Wang, K. Maize, Y. Mi, A. Shakouri, G. T. Chiu, and J. P. Allebach, "Thin-film nitrate sensor performance prediction based on preprocessed sensor images," Electronic Imaging, 2021.

[4] X. Wang, R. Wu, A. Sara, Y. Mi, A. Shakouri, G. T. Chiu, and J. P. Allebach, "Thin-film nitrate sensor performance prediction based on image analysis and credibility data to enable a certify as built framework," Manufacturing Science and Engineering Conference, 2022.

[5] X. Wang, Y. Mi, A. Shakouri, G. T. Chiu, and J. P. Allebach, "Improvements to color image and machine learning based thin-film nitrate sensor performance prediction: New texture features, repeated cross-validation, and auto-tuning of hyperparameters," Electronic Imaging, vol. 34, pp. 1-6, 2022

[6] T. Ojala, M. Pietikainen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," Pattern Recognition, vol. 29, no. 1, pp. 51-59, 1996.

[7] J. S. Bendat and A. G. Piersol, Engineering Applications of Correlation and Spectral Analysis, John Wiley and Sons, New York, 1980.

[8] A. J. Joshi, F. Porikli and N. Papanikolopoulos, "Multi-class active learning for image classification," 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2372–2379, 2009.

[9] B. Settles, Active learning literature survey, 2009.

[10] M. Li and I.K. Sethi, "Confidence-based active learning," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 8, pp. 1251-1261, 2006.

[11] M. F. Balcan, A. Broder, and T. Zhang, "Margin based active learning," In Learning Theory: 20th Annual Conference on Learning Theory, pp. 35-50, 2007.

[12] L. M. Tiwari, S. Agrawal, S. Kapoor, and A. Chauhan, "Entropy as a measure of uncertainty in queueing system," 2011 National Postgraduate Conference, pp. 1–4, 2011.

[13] A. McCallum and K. Nigam, "Employing EM and Pool-Based Active Learning for Text Classification," In ICML, vol. 98, pp. 350-358, 1998.

[14] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," Machine Learning, vol. 15, no. 2, pp. 201–221, 1994.

[15] D. Angluin, "Queries and concept learning," Machine Learning, vol. 2, pp. 319–342, 1988.

[16] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," Advances in Large Margin Classifiers, vol. 10, no. 3, pp. 61-74, 1999.

[17] G. Biau and E. Scornet, "A random forest guided tour," Test, vol. 25, no. 2, pp. 197–227, 2016.

[18] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016.

[19] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, 1967.

[20] A. Niculescu-Mizil, and R. Caruana, "Predicting good probabilities with supervised learning." In Proceedings of the 22nd International Conference on Machine Learning, pp. 625-632, 2005.

[21] T. Silva-Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull and P. Flach, "Classifier Calibration: How to assess and improve predicted class probabilities: a survey," ARXIV E-prints, 2021.

Author Biography

Xihui Wang *received her B.S. (2016) and M.S. (2019) in Electrical Engineering from Purdue University and is currently a Ph.D. candidate in Purdue ECE. Her research focuses on image processing, computer vision, and machine learning.*

Ali Shakouri *is an electrical and computer engineering professor and director of the Birck Nanotechnology Center at Purdue University. He received his Ph.D. from Caltech. His research focuses on quantum electronics, mutual interaction of heat, light, and electricity in nanomaterials and devices, lock-in imaging, and advanced image processing with applications to nanoscale thermal measurements and roll-to-roll process*

monitoring. He leads a team to manufacture low-cost smart internet of thing (IoT) devices and sensor network for applications in advanced manufacturing and agriculture.

***Bruno Ribeiro** is an Assistant Professor in the Department of Computer Science at Purdue University. After earning his doctorate from the University of Massachusetts Amherst, he went on to become a postdoctoral fellow at Carnegie Mellon University between 2013 and 2015. His research primarily focuses on developing learning algorithms and techniques for complex structured data, such as graphs and hypergraphs, and exploring the link between invariant theory, causality, and machine learning. In 2020, Ribeiro was honored with an NSF CAREER award and multiple best paper awards.*

***George T. C. Chiu** is a Professor with the School of Mechanical Engineering with courtesy appointments in the School of Electrical and Computer Engineering and the Department of Psychological Sciences at Purdue University. He worked on designing printers and multifunction devices with Hewlett-Packard, Palo Alto, CA, USA. From 2011 to 2014, he served as the Program Director for the Control Systems Program with the National Science Foundation while on leave from Purdue University, West Lafayette, IN, USA. Chiu is a Fellow of ASME and a Fellow of the Society for Imaging Science and Technology (IS&T) and an IEEE Senior Member.*

***Jan P. Allebach** is Hewlett-Packard Distinguished Professor of Electrical and Computer Engineering at Purdue University. He is a Fellow of the National Academy of Inventors, IEEE, the Society for Imaging Science and Technology (IS&T), and SPIE. He was named the Electronic Imaging Scientist of the Year by IS&T and SPIE, and was named Honorary Member of IS&T, the highest award that IS&T bestows. He has received the IEEE Daniel E. Noble Award, the IS&T/OSA Edwin Land Medal, the IS&T Johann Gutenberg Prize, and is a member of the National Academy of Engineering.*