Impacts of Behavioral Biases on Active Learning Strategies

Deepesh Agarwal
Department of Electrical and Computer Engineering
Kansas State University
Manhattan, Kansas, USA
deepesh@ksu.edu

Obdulia Covarrubias-Zambrano
Department of Cancer Biology
The University of Kansas Medical Center
Kansas City, Kansas, USA
ocovarrubias@kumc.edu

Stefan Bossmann
Department of Cancer Biology
The University of Kansas Medical Center
Kansas City, Kansas, USA
sbossmann@kumc.edu

Balasubramaniam Natarajan
Department of Electrical and Computer Engineering
Kansas State University
Manhattan, Kansas, USA
bala@ksu.edu

***Abstract*—Cyber-Physical-Human Systems (CPHS) interconnect humans, physical plants and cyber infrastructure across space and time. Industrial processes, electromechanical systems operations and medical diagnosis are some examples where one can see the intersection of humans, physical and cyber components. Emergence of Artificial Intelligence (AI) based computational models, controllers and decision support engines have improved the efficiency and cost effectiveness of such systems and processes. These CPHS typically involve a collaborative decision environment, comprising of AI-based models and human experts. Active Learning (AL) is a category of AI algorithms which aims to learn an efficient decision model by combining domain expertise of the human expert and computational capabilities of the AI model. Given the indispensable role of humans and lack of understanding about human behavior in collaborative decision environments, modeling and prediction of behavioral biases is a critical need. This paper, for the first time, introduces different behavioral biases within an AL context and investigates their impacts on the performance of AL strategies. The modelling of behavioral biases is demonstrated using experiments conducted on a real-world pancreatic cancer dataset. It is observed that classification accuracy of the decision model reduces by at least 20% in case of all the behavioral biases.**

***Index Terms*—Active Learning, Behavioral Biases, Cyber-Physical-Human Systems, collaborative decision environment, human behavior modelling**

I. Introduction

Active Learning (AL) is a form of semi-supervised machine learning (ML) approach where the learning algorithm leverages information from external sources in order to predict labels for the unlabeled instances in the dataset. The primary motive is to accomplish a higher prediction accuracy with fewer labelled instances as compared to traditional supervised ML methodologies. It has proved to be advantageous in modern ML frameworks involving expensive or wearisome labelling procedures [1]. The learning algorithm in AL settings is referred to as *Active Learner* and the external information source is termed as the *Oracle*. The AL framework can be represented as a collaborative decision environment comprising of Artificial Intelligence (AI) engine, in the form of ML-based classification/regression models; and human experts, in the form of Oracle (i.e., a domain expert). Typically, in such environments, the aim is to learn an efficient decision model by combining domain expertise of the human expert and computational capabilities of the AI model.

Although there is a plethora of literature published on handling practical AL challenges, like cold-start problem, oracle uncertainty, variable labelling costs and performance evaluation in the absence of ground truth, the collaboration of human and AI engine in a decision environment is neither straightforward nor well understood. There are anomalies and biases associated with both human and AI components of the decision environment. Algorithmic biases (like, selection bias, sampling bias, correlation fallacy, etc.) arises due to inability of algorithms to appropriately adjust to differences in data from different population subgroups [2]. On the other hand, behavioral biases (like, overconfidence, cognitive bias, hot hand fallacy, regret aversion bias, etc.) creep in due to uncertainties associated with human decisions [3]. This paper, first simulates different behavioral biases in an AL context. Then, the impact of these behavioral biases on the performance of AL strategies is quantified by comparing against an ideal case, where behavioral biases are absent.

A. Related Work

Human experts are crucial components of AI-enabled services in cyber-physical-human systems (CPHS). They form a collaborative decision environment with the support of AI-based computational models. This is pertinent in a wide variety of domains, including fault diagnosis, predictive maintenance, optimal control, process and manufacturing industry operations and medical diagnosis. Given the compelling role of humans in such decision environments, it is an important research challenge to model, predict and use the limits of

978-1-6654-5818-4/22/$31.00 ©2022 IEEE

human behavior (e.g., behavioral bias and cognitive fatigue) in CPHS design [4]. Modeling human behavior in a decision environment is not straightforward. Humans use cognitive mechanisms and decision heuristics to process information and make decisions under uncertainty [5], [6].

Behavioral biases have been studied in numerous fields, including investment and finance [7], radiology [8], medical diagnosis [9], and human-in-the-loop systems [10]. The existence of behavioral biases for investment decisions is studied, supported by evidence from the Indian stock market [11]. Protte et al. [3] have presented the impacts of overconfidence bias and hot-hand fallacy with the help of an experimental framework involving surveillance drone piloting. Cognitive bias and carelessness are parameterized, and their impact on users' reliability is evaluated for personal context recognition [12]. Furthermore, several recent studies have proposed methodologies to address algorithmic biases using effective sampling approaches [13], [14], [15], [16] and adversarial learning [17], [18], [19]. Among all these studies presented in the literature, a generic framework for modelling and prediction of behavioral biases in the context of a collaborative decision environment has not been proposed so far. An interactive framework with the flexibility to simulate and predict different behavioral biases would be highly beneficial to study, analyze and use the limits of human decision under uncertainty in human-AI collaborative decision-making tasks.

B. Contributions

The article demonstrates the simulation of different behavioral biases in an AL context. AL is represented as a collaborative decision environment consisting of AI engines (ML-based models) and human experts (Oracle). The user inputs are designed to be taken in two steps: agreement or disagreement with the AI model, followed by class labels based on human judgement (in case of disagreement). The behavioral biases are simulated by providing pre-engineered human decisions during the input steps. All the bias models are validated by performing experiments on a real-world pancreatic cancer dataset. Further, the impacts of all the simulated biases on the performance of AL strategies are examined by comparing classification accuracy against an ideal case, which does not subsume any type of behavioral bias. It is observed that the accuracy score of the decision model is reduced by at least 20% (in cases of hot-hand fallacy and representative bias) to around 85% (in case of gambler's fallacy). Such a collaborative decision framework, with the flexibility to study multiple behavioral biases, has not been proposed in the literature and is a novel contribution of this work.

The remainder of this article is organized as follows: background on AL is presented in Section II. Section III elaborates upon the modelling of behavioral biases within AL frameworks, followed by experimental setup in Section IV and results in Section V. The article ends with concluding remarks in Section VI.

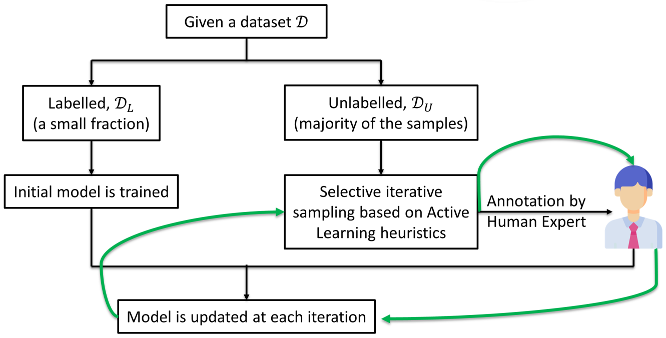
Given a dataset $\mathcal{D}$
Labelled, $\mathcal{D}_L$ (a small fraction)
Unlabelled, $\mathcal{D}_U$ (majority of the samples)
Initial model is trained
Selective iterative sampling based on Active Learning heuristics
Annotation by Human Expert
Model is updated at each iteration

Fig. 1. General Approach of AL frameworks.

II. Background on Active Learning

The general approach of AL frameworks is presented in Figure 1. Given a dataset $\mathcal{D}$, a small fraction of the samples ($\mathcal{D}_L$) are labelled and majority of them ($\mathcal{D}_U$) are unlabeled. The primary aim is to accurately predict labels for all instances in $\mathcal{D}_U$ with much fewer labelled instances for training, as compared to the conventional supervised ML frameworks. This is executed by allowing the Active Learner to choose the data it wants to learn from – by posing *queries* to Oracle in the form of unlabeled instances and requesting for the corresponding labels. An initial ML model is trained using $\mathcal{D}_L$. This is followed by the iterative selection of queries by optimizing appropriate AL heuristics, like entropy of class probabilities, margin uncertainty or classifier uncertainty. The model is updated at each query step by including the query and associated label within $\mathcal{D}_L$. This interactive modelling procedure between AI engine (ML model) and human (Oracle, i.e., domain expert) can be well represented as a collaborative decision environment.

The inherent assumptions in AL frameworks give rise to several challenges during implementation in practical scenarios. There is a wealth of literature on methodologies for handling each of these practical challenges: cold-start problem [20], [21], [22], oracle uncertainty [23], [24], [25], hybrid query strategies [22] and performance evaluation in the absence of ground truth [25]. However, the representation of AL in the form of a collaborative decision environment is not well examined in the literature. The human and AI components in this collaboration engender behavioral and algorithmic biases respectively. In this work, different types of behavioral biases are simulated within an AL context and their impacts is studied by comparing the accuracy score of associated AL strategy with an ideal case which does not incorporate any sort of behavioral bias.

III. Behavioral Bias Models

The irrational behaviors of humans which abstractly hinder the logical decision process are known as behavioral biases. The human decision or judgement methodically deviates from rationale, under the influence of these biases. This can lead to serious consequences, especially in domains like human health and medicine, where the stakes associated with decision-making are high. In this work, we consider the following behavioral biases: herding bias, cognitive bias, hot-hand fallacy, representative bias, anchoring bias, gambler's fallacy and regret-aversion bias.

In order to simulate the behavioral biases, the AL framework has been designed to query human experts in two steps:

(I) Firstly, the instance selected by the Active Learner is labelled as per the AI model trained at the current step. This instance is then presented to the human expert along with the predicted label, who is asked to specify whether he/she agrees or disagrees with the decision of the AI model.

(II) If the human expert agrees with the decision of AI model, the predicted label is considered to update the model, otherwise the human expert is prompted to provide a label as per his/her judgement.

In this work, we simulate each of the behavioral biases by supplying pre-engineered human decisions during both the input steps, based on the foundational understanding of respective biases. On the other hand, the human inputs corresponding to "Ideal Case" are formulated based on the ground-truth labels in the dataset, which justifies the absence of behavioral biases.

Herding bias is the tendency of humans to take a specific decision just because it is being supported by many other people, rather than relying on their own judgement. This is simulated in our AL environment by making the human expert to indiscriminately agree with the AI model during step (I) of each query. Cognitive bias arises from the generation of a strong, falsified preconceived notion in human minds. Henceforth, there is a tendency to form mental shortcuts to process the information quickly, rather than making rational decisions. We simulate this by making the human expert to blindly disagree with the AI model during step (I) of the input process. Further, their decisions are simulated by supplying uniformly distributed random numbers as shown in (1) during step (II) of each query. Here, C is the number of classes and d_j is decision of the human expert at step (II) for jth query.

$$d_j \sim U(1, C) \tag{1}$$

Hot-hand fallacy causes humans to overconfidently believe that their decision will be correct based on sequences of immediate correct decisions in the past. This is a "fallacy" because a future outcome is independent of the past performance. This is simulated by considering ground-truth labels during an initial set of queries, similar to that in Ideal Case. After an initial set of queries, the inputs are formulated so as to make the human expert to always disagree with AI model in step (I) and generating uniformly distributed random numbers in step (II) to mimic the overconfidence effect in hot-hand fallacy. Representative bias leads to decisions being taken based on an erroneous prototype already existing in the human minds. This "prototype" is typically the most relevant example of a particular object or event. It results in overestimation of similarity between two things that are being compared by the humans. In the AL environment, ground truth labels are considered during initial fraction of queries. Once the representative bias sets in, the inputs in step (I) are designed to have the human expert randomly agree/disagree with the AI model, followed by uniformly distributed random numbers in step (II), as indicated in (1).

Anchoring bias induces the human decisions to over-rely on first piece of information about a particular event or object. This skews the human judgement and prevent them from making rational decisions. This is simulated in our AL environment by having inputs so as to make the human expert to always agree with the AI model after an initial set of queries. This emulates the decision of human experts to be anchored based on the information learn during initial queries. Gambler's fallacy causes humans to erroneously predict the probability of a random event based on the outcomes corresponding to sequences of immediate events in the past. Although the human expert would have made a series of incorrect decisions, he/she would still go ahead for another wrong decision overconfidently, in the hope of making a correct one. We simulate this by having the human experts to forcibly make wrong decisions, i.e., shuffling the ground-truth class labels for a fraction of instances in the query set.

Regret-aversion bias occurs when human experts make decisions, so as to avoid regretting alternate decisions in the future. Under the influence of this bias, the expert prefers to select the option that would carry the least regret, even if it is not the most appropriate choice. We simulate this in our AL environment by modifying the ground-truth labels to replace them with the ones corresponding to a pessimistic choice (for example, replacing the label corresponding to lower grade of a disease with the one corresponding to higher grade of the same) for a fraction of instances in the set of queries.

IV. Experimental setup

In this work, we demonstrate simulation of all the behavioral biases discussed in Section III in an AL context, on a pancreatic cancer dataset adapted from [26]. It comprises of data from 159 participants, classified into 4 classes (healthy, pancreatitis, localized and metastatic) on the basis of an enzymatic signature consisting of arginase, matrix metalloproteinase-1, -3, and - 9, cathepsin-B and -E, urokinase plasminogen activator, and neutrophil elastase [26]. 10% of the total instances in the dataset are selected randomly to create an initial labeled dataset, which is used to train an initial ML-based classification model. Further, 50% of the total instances are used for querying iteratively (one query per iteration), and the classification model is updated after each query step. k-Nearest Neighbors (kNN) is chosen as the base classification method because it is versatile, simple and easy to implement and a non-parametric classification algorithm. Moreover, it does not make any inherent assumptions about the distribution of input data. Uncertainty Sampling (US) query strategy is implemented in Python 3.8 to select instances for annotation by the human experts. US selects instances for querying from the pool of unlabeled samples which minimizes the classifier

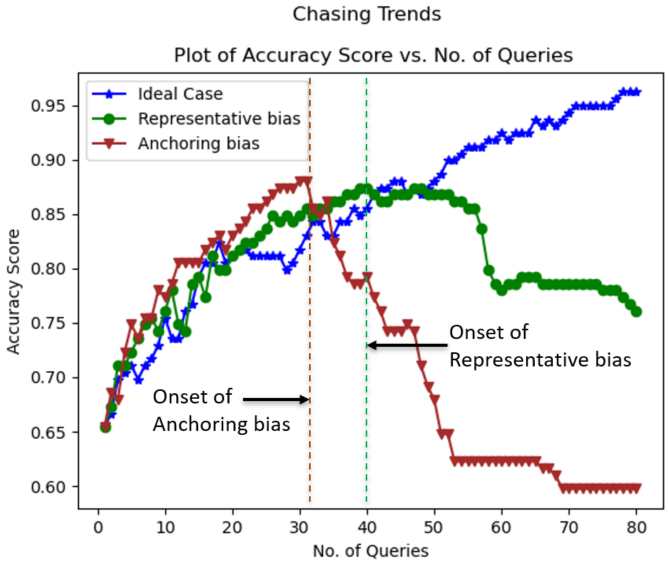

Fig. 2. Plot of Accuracy Score vs. No. of Queries: Chasing Trends.

uncertainty, as described mathematically in (2). Here, $\hat{y}$ is the predicted label for the instance x under the model θ. In the US query strategy, $\hat{y}$ is the prediction with the highest posterior probability under the model θ (indicated in eq. (3)) and x^* is the instance chosen for annotation by the human expert.

$$x^* = \arg\min_x P_\theta(\hat{y}|x) = \arg\max_x 1 - P_\theta(\hat{y}|x) \tag{2}$$

$$\hat{y} = \arg\max_y P_\theta(y|x) \tag{3}$$

V. Results

For the sake of convenience, the behavioral biases discussed in Section III are categorized into 4 categories: Representative bias and Anchoring bias are classified as *Chasing Trends*; Hot-hand and Gambler's fallacies are *Overconfidence* biases; Herding bias and Cognitive bias fall under *Limited Attention Span*; and *Regret-aversion* bias is treated as a separate category. In order to study the impacts of all these behavioral biases in AL setting, the performance (i.e., classification accuracy score) of the model is recorded after each query step for all the cases. The plots of accuracy scores for all categories of biases are presented in Figures 2 - 5.

It can be seen that the accuracy score increases with increase in no. of queries for the Ideal Case. The model trained with initial labelled dataset classifies around 65% of the instances correctly. This score gradually increases to around 96% after 80 queries are made to the human annotator and model being updated after each query step. The corresponding confusion matrix is shown in Table I. However, this trend is not observed in case of any of the behavioral biases. For Representative bias (Figure 2), the accuracy score increases upto 40% of the queries. The inputs are provided so as to set in the Representative bias at this point. Once its sets in, the accuracy score reduces with increasing no. of queries. This is because the human decisions are biased due to an erroneous prototype already existing in their minds. Similarly, for the cases of Anchoring bias (Figure 2) and Hot-hand fallacy (Figure 3), the accuracy score start decreasing after the corresponding biases set in at 50% and 60% query steps respectively. Further, it can be seen that in the case of Cognitive bias (Figure 4), the accuracy score consistently reduces with increase in no. of queries. This is because the human experts make biased decisions due to a strong, falsified preconceived notions. They tend to form mental shortcuts for quick information processing, rather than making rational decisions. Similar trends can be observed for Gambler's fallacy (Figure 3), Herding bias (Figure 4) and Regret-aversion bias (Figure 5). In each of these cases, the human experts make decisions biased on several

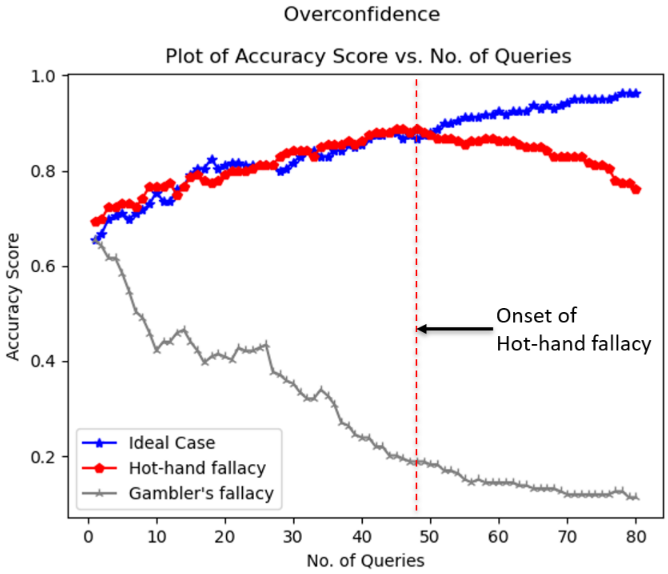

Fig. 3. Plot of Accuracy Score vs. No. of Queries: Overconfidence.

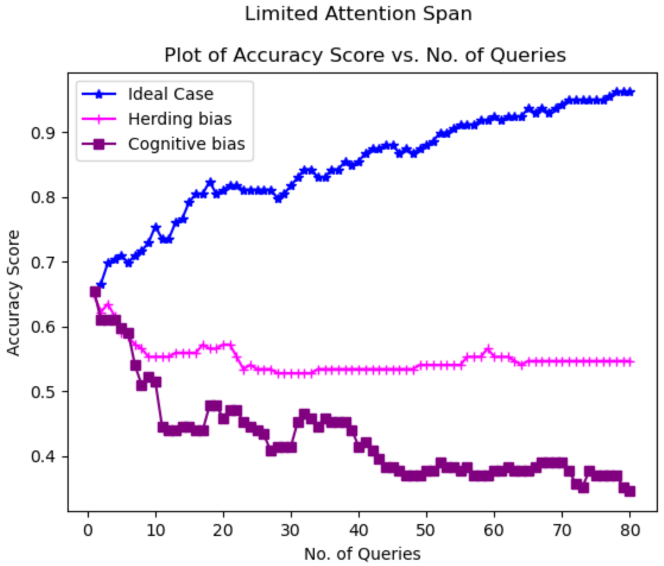

Fig. 4. Plot of Accuracy Score vs. No. of Queries: Limited Attention Span.

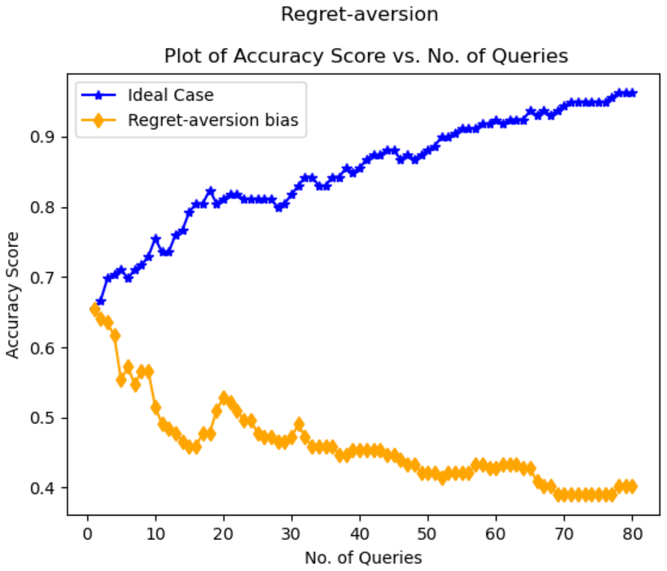

Fig. 5. Plot of Accuracy Score vs. No. of Queries: Regret-aversion.

TABLE I
CONFUSION MATRIX FOR IDEAL CASE AFTER 50% QUERIES

		Predicted Class			
True Class		Healthy	Pancreatitis	Localized	Metastatic
	Healthy	50	0	0	0
	Pancreatitis	2	23	0	1
	Localized	0	0	32	1
	Metastatic	0	2	0	48

factors as discussed in Section III, rather than relying on their own logical judgement.

VI. CONCLUSION

In this paper, the impact of seven behavioral biases, namely, herding bias, cognitive bias, hot-hand fallacy, representative bias, anchoring bias, gambler's fallacy and regret-aversion bias is illustrated using experiments conducted on a real-world pancreatic cancer dataset. Firstly, AL is represented in the form of a collaborative decision environment of AI engines and human experts, and the annotation by human experts is formulated as a two-step process. Secondly, the behavioral biases are simulated by dispensing pre-manipulated user inputs based on the foundational understanding of respective biases during the iterative query steps. Finally, the impacts of these biases on the performance of AL strategies are assessed by comparing classification accuracy score of the decision model against a reference case, which does not assimilate any sort of behavioral bias. It is observed that the performance deteriorates significantly when the human decisions are influenced by each of the behavioral biases. Future extensions of this work include ways to detect behavioral biases within a collaborative decision setting, incorporate algorithmic biases and implement the corresponding framework across datasets from different domains.

ACKNOWLEDGEMENT

This material is based upon work supported by National Science Foundation under award number 2129617.

REFERENCES

[1] S. Hao, P. Hu, P. Zhao, S. C. Hoi, and C. Miao, "Online active learning with expert advice," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 12, no. 5, pp. 1–22, 2018.

[2] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[3] M. Protte, R. Fahr, and D. E. Quevedo, "Behavioral economics for human-in-the-loop control systems design: Overconfidence and the hot hand fallacy," *IEEE Control Systems Magazine*, vol. 40, no. 6, pp. 57–76, 2020.

[4] Y. Yildiz, "Cyberphysical human systems: An introduction to the special issue," *IEEE Control Systems Magazine*, vol. 40, no. 6, pp. 26–28, 2020.

[5] D. Kahneman, "Maps of bounded rationality: Psychology for behavioral economics," *American economic review*, vol. 93, no. 5, pp. 1449–1475, 2003.

[6] D. Ariely and S. Jones, *Predictably irrational*. Harper Audio New York, NY, 2008.

[7] S. A. Zahera and R. Bansal, "Do investors exhibit behavioral biases in investment decision making? a systematic review," *Qualitative Research in Financial Markets*, 2018.

[8] L. P. Busby, J. L. Courtier, and C. M. Glastonbury, "Bias in radiology: The how and why of misses and misinterpretations," *Radiographics*, vol. 38, no. 1, pp. 236–247, 2018.

[9] E. D O'Sullivan and S. Schofield, "Cognitive bias in clinical medicine," *Journal of the Royal College of Physicians of Edinburgh*, vol. 48, no. 3, pp. 225–231, 2018.

[10] E. Wall, L. M. Blaha, L. Franklin, and A. Endert, "Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics," in *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, IEEE, 2017, pp. 104–115.

[11] S. Mehta and J. Chaudhari, "The existence of behavioural factors among individual investors for investment decision in stock market: Evidence from indian stock market," *Global Journal of Research in Management*, vol. 6, no. 1, p. 57, 2016.

[12] F. Giunchiglia, M. Zeni, and E. Big, "Personal context recognition via reliable human-machine collaboration," in *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, IEEE, 2018, pp. 379–384.

[13] A. Amini, A. P. Soleimany, W. Schwarting, S. N. Bhatia, and D. Rus, "Uncovering and mitigating algorithmic bias through learned latent structure," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 289–295.

[14] V. Iosifidis and E. Ntoutsi, "Dealing with bias via data augmentation in supervised learning scenarios," *Jo Bates Paul D. Clough Robert Jäschke*, vol. 24, 2018.

[15] M. Ngxande, J.-R. Tapamo, and M. Burke, "Bias remediation in driver drowsiness detection systems using generative adversarial networks," *IEEE Access*, vol. 8, pp. 55 592–55 601, 2020.

[16] F. P. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 3995–4004.

[17] M. Alvi, A. Zisserman, and C. Nellåker, "Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[18] Z. Wang, K. Qinami, I. C. Karakozis, *et al.*, "Towards fairness in visual recognition: Effective strategies for bias mitigation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8919–8928.

[19] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 335–340.

[20] A. Primpeli, C. Bizer, and M. Keuper, "Unsupervised bootstrapping of active learning for entity resolution," in *European Semantic Web Conference*, Springer, 2020, pp. 215–231.

[21] J. Shao, Q. Wang, and F. Liu, "Learning to sample: An active learning framework," in *2019 IEEE International Conference on Data Mining (ICDM)*, IEEE, 2019, pp. 538–547.

[22] D. Agarwal, P. Srivastava, S. Martin-del-Campo, B. Natarajan, and B. Srinivasan, "Addressing practical challenges in active learning via a hybrid query strategy," *arXiv preprint arXiv:2110.03785*, 2021.

[23] M.-R. Bouguelia, S. Nowaczyk, K. Santosh, and A. Verikas, "Agreeing to disagree: Active learning with noisy labels without crowdsourcing," *International Journal of Machine Learning and Cybernetics*, vol. 9, no. 8, pp. 1307–1319, 2018.

[24] R. Saeedi, K. Sasani, and A. H. Gebremedhin, "Collaborative multi-expert active learning for mobile health monitoring: Architecture, algorithms, and evaluation," *Sensors*, vol. 20, no. 7, p. 1932, 2020.

[25] D. Agarwal, P. Srivastava, S. Martin-del-Campo, B. Natarajan, and B. Srinivasan, "Addressing uncertainties within active learning for industrial iot," in *2021 IEE World Forum on Internet of Things (WF-IoT)*, IEEE, in press.

[26] M. Kalubowilage, O. Covarrubias-Zambrano, A. P. Malalasekera, *et al.*, "Early detection of pancreatic cancers in liquid biopsies by ultrasensitive fluorescence nanobiosensors," *Nanomedicine: Nanotechnology, Biology and Medicine*, vol. 14, no. 6, pp. 1823–1832, 2018.