A Crowdsourcing-driven AI Model Design Framework to Public Health Policy-Adherence Assessment

Yang Zhang, Ruohan Zong, Lanyu Shang, Dong Wang
School of Information Sciences
University of Illinois Urbana-Champaign
Champaign, IL, USA
{yzhangnd, rzong2, lshang3, dwang24}@illinois.edu

***Abstract*—This paper focuses on a *public health policy-adherence assessment (PHPA)* application that aims to automatically assess people's public health policy adherence during emergent global health crisis events (e.g., COVID-19, MonkeyPox) by leveraging massive public health policy adherence imagery data from the social media. In particular, we study an *optimal AI model design* problem in the PHPA application, where the goal is to leverage the crowdsourced human intelligence to accurately identify the optimal AI model design (i.e., network architecture and hyperparameter configuration combination) without the need of AI experts. However, two critical challenges exist in our problem: 1) it is challenging to effectively optimize the AI model design given the interdependence between network architecture and hyperparameter configuration; 2) it is non-trivial to leverage the human intelligence queried from ordinary crowd workers to identify the optimal AI model design in the PHPA application. To address these challenges, we develop *CrowdDesign*, a subjective logic-driven human-AI collaborative learning framework that explores the complementary strength of AI and human intelligence to jointly identify the optimal network architecture and hyperparameter configuration of an AI model in the PHPA application. The experimental results from two real-world PHPA applications demonstrate that CrowdDesign consistently outperforms the state-of-the-art baseline methods by achieving the best PHPA performance.**

I. Introduction

EMERGENT global health crises like COVID-19 have revealed the need for rapid, effective assessments of public health policy adherence [1]. It is critical for government authorities to obtain accurate and timely information about the public health policy adherence so that necessary policy adjustments and precautions can be made to protect vulnerable populations and ensure people's health and well-being [2]. Meanwhile, social media platforms emerge as a pervasive paradigm to acquire an unprecedented amount of timely observations of the public health policy adherence by exploring the imagery data contributed by the common citizens [3]. This paper develops a human-AI collaborative framework that leverages collective intelligence to optimize AI model design for public health policy-adherence assessment (PHPA).

Recent progress in AI and deep learning have developed advanced deep neural network models (e.g., ResNet, VGG, Transformer) that can be applied to automatically analyze the massive amount of PHPA data on social media (e.g., images related to public smoking bans and urban environmental hygiene posted by social media users) [2]. Those deep neural network models often include a complex set of network architectures (e.g., network layers, convolutional blocks, activation functions) and hyperparameter configurations (e.g., learning rate, weight decay, training epochs) to capture complex and diverse visual features in the studied images [4]. The combination of a specific network architecture and a particular hyperparameter configuration in an AI model is often referred to as an *AI model design* instance [5]. We note that selecting the appropriate AI model design instance plays a decisive role in the performance of the AI model [6]. Figure 1 shows a PHPA application that assesses whether people follow the mask wearing policy or not. This task is challenging due to multiple reasons. Motivated by the above observations, this paper studies an *optimal AI model design (OAMD)* problem where the goal is to accurately identify the optimal AI model design (i.e., optimal network architecture and hyperparameter configuration combination) that can achieve the desirable PHPA performance.

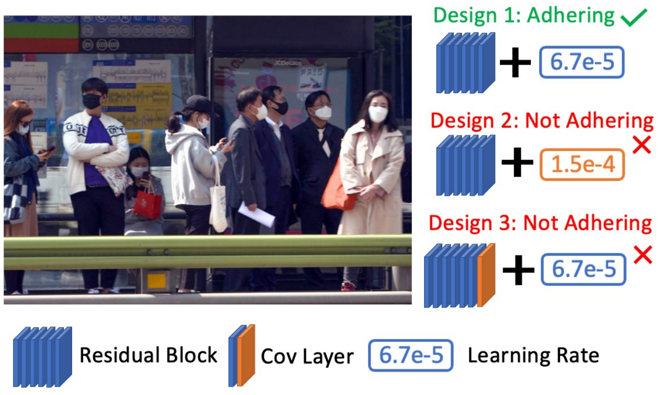


Figure 1. Mask Wearing Policy Adherence Assessment of three Different AI Model Designs. Three AI model design instances have different network architectures and network configurations. A small change in AI model design leads to an incorrect assessment.

Our objective is to develop a novel AI model design framework that can effectively harness both AI and crowdsourced human intelligence without requiring the involvement of AI experts. This approach aims to optimize the network architecture and hyperparameter configurations that are critical to the performance of AI models in public health applications. Our framework is inspired by the observation that humans often can provide more consistent and reliable estimations on the public health policy adherence practice than AI models. For example, human perception can accurately identify that all people follow the mask-wearing policy in Figure 1 while AI models with several design options fail to do that. Such human intelligence can help infer the likelihood of an AI model design in providing the correct PHPA assessment and identify the optimal AI model design instance. We observe that crowdsourcing offers a scalable, cost-effective alternative to enlisting public health specialists in identifying optimal AI model design [7]. Therefore, our framework leverages the timely and scalable human intelligence from crowdsourcing platforms (e.g., Amazon MTurk) that provide pervasive and economic access to a massive amount of freelance crowd workers. However, two technical challenges exist in designing our framework, which we elaborate below.

The first challenge is how to concurrently identify the optimal network architecture and hyperparameter configuration of an AI model given the interdependence between them. There exists a “chicken-and-egg” issue in the OAMD problem. In particular, current neural architecture search (NAS) methods are designed to identify the optimal neural network architecture with a pre-defined hyperparameter configuration [8]. Meanwhile, the recent hyperparameter optimization (HPO) frameworks focus on searching for the optimal hyperparameter configuration of a given neural network architecture [9], [10]. A straightforward solution to address such a problem is to perform the two tasks successively. However, such an approach largely ignores the interdependence between NAS and HPO, which often leads to a suboptimal application performance [6]. There also exist efforts that jointly optimize the network architecture and hyperparameter configuration of an AI model [6], [11]. However, those solutions often require a large amount of training data with high-quality labels (e.g., 1M+ images from ImageNet dataset) to train their AI models to explore the massive NAS + HPO search space [12]. However, there is often a very limited amount of high-quality training data available in PHPA applications due to the emergent nature of the events (e.g., unfolding public health crisis) [13], and the learned AI model design could face an overfitting issue on the validation data set and lead to a non-trivial performance degradation on the unseen data samples in the testing set. Thus, it is challenging to effectively optimize the AI model design given the interdependence between NAS and HPO.

The second challenge is how to effectively transform the complex OAMD problem to a simplified problem that does not require extensive AI expertise and can be solved by ordinary crowd workers. Unlike the AI experts who can directly provide insights on the AI model design (e.g,. configuring network architectures, selecting hyperparameters), crowd workers are usually only capable of providing annotations of assigned labeling tasks [8]. In addition, crowd workers are often unable to diagnose and troubleshoot the AI model when it fails. A key question in designing our crowdsourcing-driven AI model design framework is how to transfer the crowdsourced human intelligence to an effective strategy in identifying the optimal AI model design in a PHPA application. We observe that recent efforts on crowd-AI collaboration systems mainly focus on leveraging crowdsourced human intelligence to retrain the AI models to boost the overall application performance [14], [15]. However, those existing solutions are not developed to address the OAMD problem and the AI models they employ are manually selected and configured by AI experts. Such a manual AI model design process is known to be error-prone and time-consuming [16]. Some initial attempts have been made to utilize crowdsourced human intelligence for optimizing network architecture [8] or hyperparameter configuration [10] individually. However, these solutions are unable to solve the OAMD problem due to the intricate interdependence between neural architecture and hyperparameter configuration, which could lead to suboptimal performance in the studied applications. Therefore, it remains a challenging task to integrate the AI and human intelligence under a principled framework to optimize the overall performance of the studied application.

This paper develops CrowdDesign, a crowdsourcing-driven AI model design framework to solve the OAMD problem in PHPA applications. Utilizing subjective logic, CrowdDesign integrates the strengths of AI and human insights to identify the most effective model designs for assessing public health policy adherence from social media image. In particular, CrowdDesign first designs a joint network architecture and hyperparameter space reduction scheme that effectively reduces the AI model design search space while maintaining a high likelihood of including the optimal AI model design instance in the reduced search space. The CrowdDesign then develops a probabilistic reasoning-based AI model design scheme that leverages the crowdsourced human intelligence to learn the probability of each AI model design instance from the reduced search space in providing correct PHPA labels through subjective logic-based probabilistic reasoning. The learned probability is then used to identify the optimal AI model design instance for the studied PHPA application. To the best of our knowledge, CrowdDesign is the first crowdsourcing-driven AI model design framework to address the OAMD problem in PHPA applications. This paper focuses on the PHPA application as an illustrative example and use case to demonstrate the effectiveness of the CrowdDesign. We envision that CrowdDesign can also be applied to address the OAMD problem in a much broader set of real-world applications (e.g., intelligent transportation, smart health, online recommender systems) that also rely on advanced AI models (e.g., ResNet, VGG, and Transformer) where the application performance is highly sensitive to the AI model design. We evaluate CrowdDesign via two real-world PHPA applications (i.e., mask wearing policy adherence and social distancing policy adherence). The results demonstrate CrowdDesign achieves the best PHPA performance compared to a rich set of state-of-the-art deep neural networks, human-AI models, and AI model optimization baselines. The significance of this research lies in its potential to revolutionize how public health compliance is monitored and evaluated, offering a new tool to quickly facilitate the adjustment of strategies in response to ongoing or emerging health crises.

CrowdDesign is designed to address the challenge of limited high-quality data in emergent public health crises, where agile and adaptive models are critical. For instance, during a public health emergency, such as a disease outbreak, CrowdDesign can be effectively adapted to perform the corresponding PHPA tasks by leveraging labeled social media images, crowdsourced from health professionals, for AI-driven PHPA model optimization. This crowdsourced data helps identify the optimized model design for assessing public health policy adherence by jointly modeling the subjective logic of AI systems and crowd workers. Note that on platforms like Amazon MTurk, we can specifically target and recruit health professionals by setting tasks with required qualifications, such as certifications or related experience. This adaptability is critical for maintaining the relevance and effectiveness of PHPA applications when dealing with rapidly changing health norms and practices.

Beyond PHPA, the CrowdDesign framework's adaptability and its strategy for harnessing crowdsourced data make it a potent tool for a variety of other domains requiring real-time data interpretation and model optimization. For example, in intelligent transportation systems, CrowdDesign could enhance AI models that predict traffic patterns and manage flows by incorporating real-time crowd-sourced data from mobile devices and sensors, adjusting to unexpected conditions such as accidents or road closures instantly. Similarly, in smart health applications, the CrowdDesign could utilize inputs from healthcare workers and patients to continuously improve diagnostic algorithms, particularly in response to emerging medical conditions or in the management of chronic diseases where patient feedback can provide critical insights for treatment adjustments. These application scenarios demonstrate the CrowdDesign framework's capability to not only address specific data scarcity and quality issues inherent in PHPA application but also to leverage the power of crowd intelligence to enhance AI development and applications in dynamic environments across sectors. This unique integration of AI with crowdsourced data input stands out as a significant innovation, positioning CrowdDesign as a versatile, scalable solution for the broader field of AI-driven applications, where real-time adaptability and data enhancement are crucial.

The main contributions of this paper are summarized as follows:

- We develop CrowdDesign, which explores the complementary strengths of AI and crowdsourced human intelligence to solve the OAMD problem.
- We address two important technical challenges, namely the *complex interdependence between NAS and HPO*, and the *crowd-driven AI model design*, in our solution by developing a principle-based subjective logic-driven human-AI collaborative learning framework.
- We perform extensive experiments to evaluate our solution through two real-world PHPA applications. Evaluation results demonstrate significant performance gains of our CrowdDesign scheme compared to state-of-the-art baselines.
- CrowdDesign is the first crowdsourcing-driven AI model design framework to address the OAMD problem in PHPA applications, which can be potentially applied to other AI-based real-world applications where the application performance is highly sensitive to the design of the AI model.

II. Related Work

A. Social Media Sensing in Public Health

The use of social media for sensing in public health has rapidly expanded over the last decade, providing unprecedented opportunities for real-time health monitoring and crisis response [3]. However, the application of this data source in PHPA presents unique challenges and questions that merit thorough discussion [17]. Several studies have highlighted the potential of social media data to improve public health surveillance. For instance, research by Xue *et al.* demonstrated that Twitter data could be used to effectively track disease outbreaks and public sentiment towards health policies [18]. Similarly, Bonnevie *et al.* utilized social media analysis to monitor public reactions to flu vaccinations, illustrating how social platforms can offer real-time insights into public health compliance and attitudes [19]. These methodologies generally employ natural language processing and image recognition technologies to analyze text and visual content from social platforms, identifying trends and patterns related to health behaviors and policy adherence. For example, AI models like those developed by Negri *et al.* can classify images and posts according to their adherence to health guidelines, such as mask-wearing and social distancing during the COVID-19 pandemic [2]. While promising, the use of social media data for PHPA is not without significant limitations. One major challenge is the reliability and accuracy of the data. Social media platforms often contain biased self-reported data and misinformation, which can skew perceptions of public health compliance. This challenge was explored by Afful et al., who found significant discrepancies between actual virus transmission rates and those inferred from social media trends [20]. Another challenge is the ethical concern related to privacy and consent in the use of social media data. Social media users may not anticipate that their posts will be used for public health surveillance, raising ethical questions about informed consent and data anonymization. Studies by Bender *et al.* have called for stringent guidelines to ensure that social media data used in public health research respects user privacy and data security [21]. In light of these challenges, our CrowdDesign framework is positioned to enhance the reliability and ethical use of social media data for PHPA. By leveraging crowdsourced human intelligence, CrowdDesign addresses the accuracy issue by enabling the verification and enhancement of AI-generated insights, ensuring that data used for policy adherence assessment is both accurate and representative. Moreover, CrowdDesign incorporates rigorous data anonymization processes and adheres to ethical guidelines that respect user privacy, thus addressing the ethical concerns prevalent in the current literature. The literature suggests a need for more robust AI models and ethical frameworks to better utilize social media for public health monitoring. Future research should focus on developing advanced machine learning algorithms that can more effectively filter misinformation and biased data. Additionally, establishing global standards for the ethical use of social media data in health research is crucial to address privacy concerns and enhance public trust in health assessment technologies.

B. AI Model Design

AI model design is a critical task in developing robust and effective AI models to address complex real-world problems [5]. Recent advances in AI model design have shown substantial improvements in automatically selecting network architectures and hyperparameter configurations for AI models [22]–[26]. For example, Awad *et al.* designed an AI model design framework that leverages a non-stochastic infinite-armed bandit-based scheme to ensure an effective search of the AI model design space [22]. Chen designed a meta-learning hyperparameter optimization strategy that leverages deep uncertainty estimation to identify the optimal AI model design instance from a large-scale AI model search space [24]. Hirose *et al.* developed an AI model design search framework that jointly searches the optimal AI model design and prunes the search space for efficient AI model optimization [25]. Tan *et al.* introduced a lightweight AI model design approach that includes a factorized hierarchical AI model design searching mechanism to ensure a fast AI model search process via a multi-objective reinforcement learning scheme [26]. However, those approaches often need a large number of high-quality annotated samples to train their AI models to explore the huge AI model design search space, which is not always available in real-world applications (e.g., unfolding public health crises, emerging disaster events, early misinformation detection) [10]. In contrast, our CrowdDesign explicitly leverages the complementary nature of AI and human intelligence to effectively identify the optimal AI model design instance without requiring a significant amount of training data in PHPA applications.

C. Crowd-AI Hybrid Learning

Our work is also relevant to crowd-AI hybrid learning, where complementary strengths of crowdsourced human intelligence and AI are jointly explored to tackle challenging computational problems [14]–[16], [27]–[30]. For example, Hong *et al.* developed an interactive crowdsourcing interface to effectively incorporate real-time crowd annotations to build AI-infused object detection models [27]. Guo *et al.* designed a crowd-AI camera-based sensing scheme that collects diverse questions and answers from crowd workers to boost the performance of automatic question and answering on images captured by surveillance cameras [28]. Lin *et al.* introduced a hybrid crowd-AI system that leverages the crowd-annotated medical imagery samples to troubleshoot the deep image classification model for tropical disease diagnosis [29]. Yoo *et al.* introduced a task-agnostic collaborative learning scheme that utilizes a crowdsourcing-based deep estimation error inference design to identify and fix the AI failure cases in nature scene classification [30]. Shukla *et al.* proposed a crowd-AI collaborative method that utilizes a crowdsourcing uncertainty-aware deep estimation model to identify and fix AI failure cases in image classification. However, the AI models in the current crowd-AI hybrid learning approaches are often manually configured by AI experts, which is known to be suboptimal [16]. More recently, a few initial efforts have been made to explore the crowdsourced human intelligence to enhance the performance of AI systems by optimizing network architecture [8] and hyperparameter configuration [10], respectively. However, these solutions overlook the intricate interdependence between neural architecture search and hyperparameter configuration, which could result in nontrivial performance degradation in the studied applications [11]. In contrast, our CrowdDesign explicitly utilizes the crowdsourced human intelligence to solve the OAMD problem in PHPA applications without the inputs from AI experts.

D. Few-shot and Zero-shot Learning

In recent years, advancements in AI model design, particularly in computer vision, have introduced robust techniques such as few-shot and zero-shot learning, which are incredibly valuable in scenarios marked by data limitations [31]. Few-shot and zero-shot learning enable models to understand new tasks or recognize new objects with very few or zero training examples, effectively mitigating the challenge of data scarcity often encountered in emergent public health scenarios [32]. For example, Hu *et al.* introduce a transformer-based few-shot learning pipeline that achieves accurate and robust classification by incorporating pre-training on external data, meta-training with few-shot tasks, and task-specific fine-tuning [33]. Schlarmann *et al.* develops an unsupervised adversarial fine-tuning approach to enhance the robustness of the CLIP vision encoder, which yields robustness on various zero-shot vision classification tasks [34]. Chen *et al.* design a novel approach to zero-shot learning that employs meta-learning techniques to adapt to new tasks without additional data, aiming to generalize well across diverse unseen classes [35]. Zhao *et al.* integrate few-shot learning strategies with domain adaptation, enabling models to quickly adjust to new data with minimal examples, enhancing performance in dynamic image classification [36]. While powerful for generalizing to new classes without a significant amount of examples, the effectiveness of few-shot and zero-shot learning is often constrained by the precision of the descriptions and the model's capacity to interpret these attributes correctly [32]. In contrast, CrowdDesign focuses on leveraging crowd input to systematically identify and refine the optimal AI model design. This approach not only addresses the inherent limitations of few-shot and zero-shot learning by incorporating a wide range of human insights into the training process but also enhances model adaptability and accuracy through continuous real-world feedback. This dynamic integration allows CrowdDesign to surpass the static nature of predefined data models, offering a more robust solution that evolves with emerging data and real-time scenarios.

III. Problem Description

We first present a few key definitions to define the optimal AI model design problem in PHPA applications.

Definition 1: **PHPA Image ($\boldsymbol{D}$)**: We define $\boldsymbol{D} = \{D_1, D_2, ..., D_I\}$ as a set of public health-crisis related images collected from social media that are used to assess the public health policy-adherence (e.g., the mask wearing policy adherence in Figure 1). In particular, D_i represents the i^{th} collected image. I is the number of images in the studied PHPA application.

Definition 2: **PHPA Label ($\boldsymbol{L}$)**: Our paper focuses on the classification-based PHPA application, where the status of the public health policy adherence can be classified into K different categories. For instance, in a PHPA application that focuses on estimating whether people follow the mask wearing policy or not, the PHPA labels of this application are binary: *adhering mask wearing policy* and *not adhering mask wearing policy*. In particular, we define $L = \{L_1, L_2, ..., L_I\}$ as the *ground-truth* PHPA labels of all studied PHPA images, where L_i represents the PHPA label for D_i.

Definition 3: **AI-based PHPA Model (Ω)**: We define Ω as a deep neural network model for PHPA tasks. Ω contains two key attributes: *network architecture* and *hyperparameter configuration*, where the selection of the two attributes of Ω will directly affect the its performance as shown in Figure 1.

Definition 4: **Network Architecture Search Space ($\boldsymbol{A}$)**: In our paper, we define $\boldsymbol{A} = \{A_1, A_2, ..., A_P\}$ as the set of different network architectures from a network architecture search space for Ω. A_p is a network architecture in $\boldsymbol{A}$, and P is the number of different network architectures in $\boldsymbol{A}$. In this paper, the network architectures we focus on include the types of the convolutional block (residual block or dense block), the number of convolutional layers per block, the width of convolutional block, the growth rate, and the size of input features [6].

Definition 5: **Hyperparameter Configuration Space ($\boldsymbol{B}$)**: We define $\boldsymbol{B} = \{B_1, B_2, ..., B_Q\}$ as the set of different hyperparameter configurations from a hyperparameter configuration space for Ω. B_q represents a hyperparameter configuration in $\boldsymbol{B}$, and Q represents the number of different hyperparameter configurations in $\boldsymbol{B}$. In this paper, the hyperparameter configurations we study include learning rate, optimizer, weight decay, and conditional parameters of the optimizer (e.g., RMSprop alpha, SDG momentum, Adam beta1, Adam beta2) [9].

Definition 6: **AI model design Instance ($\Phi_{a,b}$)**: We define $\Phi_{p,q}$ as an AI model design instance that includes a particular network architecture A_p and a specific hyperparameter configuration B_q, which is used to configure Ω for PHPA tasks. Our CrowdDesign is developed to identify the optimal AI model design instance Φ^* for Ω that generates the estimated PHPA labels for all studied images with the highest accuracy among all candidate AI model design instances.

Definition 7: **PHPA Label Estimated by an AI model design Instance ($\widehat{\boldsymbol{L}^{\boldsymbol{\Phi}_{a,b}}}$)**: We define $\widehat{\boldsymbol{L}^{\boldsymbol{\Phi}_{a,b}}}$ to be the PHPA labels estimated by a particular AI model design instance $\Phi_{a,b}$ for all images in the studied PHPA application. In particular, $\widehat{\boldsymbol{L}_i^{\boldsymbol{\Phi}_{a,b}}}$ represents the estimated PHPA label for D_i by $\Phi_{a,b}$.

This paper focuses on utilizing the crowdsourced human intelligence to infer the likelihood of each AI model design instance in providing accurate labels in PHPA applications. Therefore, we further present a key definition of the crowdsourcing tasks, which are designed to acquire human intelligence from crowdsourcing platforms.

Definition 8: **Crowdsourcing Query (Ψ)**: We define Ψ as a crowdsourcing task to collect the human intelligence from freelance workers in crowdsourcing platforms. We note that labeling all studied images through crowd workers is impractical due to budget and time limitations, particularly given the large number of social media images in PHPA applications. In the crowdsourcing tasks, each image from the selected subset is labeled by a group of C crowd workers $\boldsymbol{W} = \{W_1, W_2, ..., W_C\}$, where W_c represents the c^{th} crowd worker and C represents the number of participating crowd workers. We define $\widehat{L_i^{W_c}}$ to be the PHPA label from a crowd worker W_c for an image D_i if D_i is selected in Ψ. The selection process of images for Ψ will be discussed in the solution section.

The goal of our optimal AI model design problem in PHPA applications is to utilize the collaborative intelligence of AI and humans to identify the optimal AI model design instance that achieves the best PHPA performance as follows:

$$\arg\max_{\Phi^*} \left(\Pr(\widehat{L_i^{\Phi^*}} = L_i \mid \boldsymbol{D}, \boldsymbol{A}, \boldsymbol{B}, \Psi)\right), \forall\ 1 < i < I \quad (1)$$

This problem is challenging because it is not a trivial task to translate the highly complex AI model design problem to a simplified task for crowd workers given the interdependence between network architecture and hyperparameter configuration in the AI model design. We present our solution CrowdDesign to address such a challenge in the next section.

IV. Solution

Our CrowdDesign is a crowdsourcing-driven AI model design framework that utilizes the complementary AI and human intelligence to address the QAMD problem in PHPA applications. An overview of CrowdDesign is shown in Figure 2. The CrowdDesign includes two key modules:

- *Joint Architecture-Hyperparameter Space Reduction (JASR)*: it designs a joint network architecture and hyperparameter space reduction scheme that explicitly reduces the AI model design search space while maintaining a high likelihood of including the optimal AI model design instance in the reduced search space via a principled budget-constraint multi-armed bandit learning framework.
- *Probabilistic Reasoning-driven Optimal Model Design (POMD)*: it designs a probabilistic reasoning-driven AI model design framework that models the AI model design instances from the reduced search space obtained by JASR module and the crowd workers as "estimators" to collaboratively infer the probability of each AI model design instance in providing correct PHPA labels. The inferred probability is used to identify the optimal AI model design instance for the studied PHPA application.

A. Joint Architecture-Hyperparameter Space Reduction

In the first subsection, we discuss our joint architecture hyperparameter space reduction design. Our objective is to

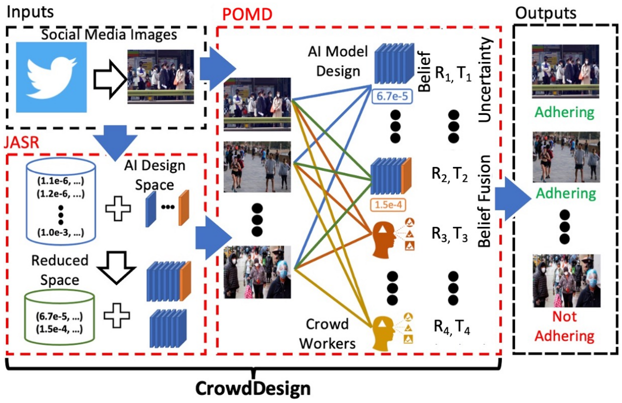


Figure 2. Overview of CrowdDesign Framework

explicitly reduce the AI model design search space and ensure the reduced search space maintains a high likelihood of including the optimal AI model design instance. The JASR achieves this objective by optimizing the search space using PHPA images with ground truth labels from the training dataset, which were labeled by public health professionals prior to the AI model design search process. We first formally define the reduced AI design search space as follows.

Definition 9: **Reduced AI Model Design Search Space** (Λ): We define Λ to be a reduced AI model search space which is significantly smaller compared to the original AI model design search space $\mathbf{\Phi}$ while maintaining a high likelihood of including the optimal AI model design instance Φ^*. In particular, we have:

$$\Lambda \in \mathbf{\Phi}, J = |\mathbf{\Phi}|, J << P \times Q, \underset{\Phi^*}{\arg\max} \left(Pr(\Phi^* \in \Lambda)\right) \quad (2)$$

where J represents the number of AI model design instances in Λ and P represents the size of the network architecture search space $\boldsymbol{A}$ and Q represents the size of the hyperparameter configuration space $\boldsymbol{B}$. Our JASR is designed to learn such a reduced AI design space.

To learn the reduced AI design space Λ, our JASR module explicitly models the AI design space reduction problem as a budget-constant non-stochastic multi-armed bandit (BNMB) problem. In particular, in the BNMB problem, a player plays a large number of different bandit machines under a limited budget. The objective is to identify a reduced set of bandit machines that have the highest winning rates. On the one hand, the player tries to keep playing the bandit machines that yield a good winning rate (i.e., *exploitation*) to maximize the overall reward. On the other hand, the player also tries to explore other bandit machines (i.e., *exploration*) in order to avoid missing the bandit machines with higher winning rates. Similarly, in our JASR framework, our goal is to identify a set of best-performing AI model design instances given the limited budget on computational time. On the one hand, our JASR framework focuses on keeping allocating computational time to further train the AI model with designs that yield a good PHPA performance. On the other hand, our JASR framework also keeps exploring the AI model search space by allocating computational time to other alternative AI model design instances that have not yielded the best PHPA performance. In particular, we first introduce several key definitions in our BNMB problem to learn the reduced AI model design search space.

Definition 10: **Budget**: We define the budget in JASR module as the computational time required to identify the optimal AI model design instance. The JASR module can explore all possible AI model design instances from the entire search space if the computational time is unlimited. However, the amount of computational time available in a real-world PHPA application is always finite.

Definition 11: **Action**: Similar to the BNMB problem where a player can take actions to play different bandit machines, the action in our JASR module is to allocate computational time to train the AI model with different designs.

Definition 12: **Reward**: The player in the BNMB problem can get different rewards after playing different bandit machines, which help the player identify the bandit machine with the highest winning rate. We define the reward in the JASR module as the performance of the AI model design instance on the validation set of the studied PHPA application, which is used to learn the probability of each AI model design instance in providing correct PHPA labels.

Definition 13: **Goal**: Like the BNMB problem that aims at identifying a reduced set of machines with the highest rewards, we define the goal of JASR to be identifying a set of best-performing AI model design instances as the reduced design search space Λ.

We achieve the goal of our JASR module by effectively balancing the trade-off between exploration and exploitation in the BNMB problem [4]. Specifically, following the discussion in [37], this trade-off is managed using an Upper Confidence Bound (UCB) method. The UCB for each AI model design Φ_t at time t is given by:

$$UCB_t(\Phi) = \hat{r}_t(\Phi) + \sqrt{\frac{2\log T}{n_t(\Phi)}}, \quad (3)$$

where $\hat{r}_t(\Phi)$ is the average reward and $n_t(\Phi)$ is the number of times Φ has been selected. This approach allows us to exploit model designs with high rewards while simultaneously encouraging exploration of lesser-known designs, thus optimizing the search for the best model design instance Φ^*. In particular, our JASR module exploits the AI model design instances that yield high PHPA performance on the validation set by keeping allocating computational time to further train the AI model with those designs. In addition, our JASR module also consistently explores other alternative AI model design instances to prevent our JASR module from keeping allocating all computational time to train the AI model with a suboptimal design. The joint design of exploitation and exploration makes JASR capable of exploring the entire AI model design search space and identifying the reduced AI model search space effectively. We also observe that the AI model design instances in the reduced search spaces could be overfitted to the validation set and resulting in non-negligible performance loss in the testing set. In the next subsection, we will discuss how to effectively leverage the crowdsourced human intelligence to identify the best AI model design instance on the testing set.

B. Probabilistic Reasoning-driven Optimal Model Design

In the POMD module, we present a probabilistic reasoning-driven optimal model design framework to leverage the crowd-sourced human intelligence (i.e., queried annotations from crowd workers through the crowdsourcing query defined in Definition 8) to identify the optimal AI model design instance from the reduced AI model design search space generated by the JASR module.

We first discuss how to perform crowdsourcing query Ψ in the POMD module. We note that it is not practical to query the crowd workers to label all studied images given the budget and time constraints in the presence of a significant amount of social media images in PHPA applications. Therefore, our POMD module samples a subset of images for Ψ where different AI model design instances in the reduced search space cannot reach a consensus. In particular, our POMD module first calculates the *consensus* of the PHPA labels estimated by all AI model instances in Λ for each image D_i using Shannon entropy. The consensus indicates the degree of agreement among different AI model design instances on the estimated PHPA labels for D_i. We then select the images with $\delta \times I$ lowest consensus scores for Ψ. Here, δ represents the percentage of studied PHPA images that are sampled for Ψ. The value of δ is determined by the trade-off between the accuracy performance and the crowdsourcing cost in the studied PHPA application. I represents the total number of studied PHPA images. Our next step is to leverage the crowd labels returned by Ψ to identify the optimal AI module design. In particular, we first introduce a key definition in our POMD module.

Definition 14: **Hybrid Estimation Committee ($\boldsymbol{H}$)**: We define $\boldsymbol{H} = \{H_1, H_2, ..., H_N\}$ as a hybrid estimation committee which includes all J different AI model design instances in Λ generated by the JASR module and all C different crowd workers who participate in the crowdsourcing query Ψ. A unit H_n can be a PHPA model design instance Φ_j or a crowd worker W_c, which estimates the PHPA label for a studied PHPA image.

Our CrowdDesign framework models different units in $\boldsymbol{H}$ as *estimators* that can collaboratively estimate the labels for all studied PHPA images. Our goal is to infer the probability of each estimator in generating correct PHPA labels of all studied images, which is used to identify the optimal AI model design instance that achieves the best PHPA performance. To that end, we first model the probability of each unit of H in generating correct PHPA labels for all studied images through probabilistic reasoning, which models the reliability and epistemic uncertainty of multiple sources when fusing the inputs from different sources. In particular, we define R_n^k as the belief of an estimator H_n on the PHPA label of a studied image to be k. In particular, R_n^k represents the degree of confidence that an estimator H_n estimates the PHPA label of a studied image to be k. We further define T_n^k as the uncertainty of H_n on the PHPA label of a studied image to be k. In particular, T_n^k represents the inability of an estimator H_n to determine whether the PHPA label of a studied image is k or not.

Our next step is to fuse the belief and uncertainty of all estimators in $\boldsymbol{H}$ to jointly identify the optimal AI model design instance of the studied PHPA application. To that end, we integrate the belief and uncertainty of all estimators in an iterative manner through the belief fusion function, which is a key function in probabilistic reasoning that is used to determine the combined belief and uncertainty of two sources by fusing the individual belief and uncertainty of each source. In particular, we have:

$$\begin{aligned} R_{n_1,n_2}^k &= \mathcal{B}(H_{n_1}, H_{n_2}) = \frac{R_{n_1}^k \times T_{n_2}^k + R_{n_2}^k \times T_{n_1}^k}{T_{n_1}^k + T_{n_2}^k - T_{n_1}^k \times T_{n_2}^k} \\ T_{n_1,n_2}^k &= \mathcal{F}(H_{n_1}, H_{n_2}) = \frac{T_{n_1}^k \times T_{n_2}^k}{T_{n_1}^k + T_{n_2}^k - T_{n_1}^k \times T_{n_2}^k} \end{aligned} \tag{4}$$

where H_{n_1} and H_{n_2} are any two estimators in $\boldsymbol{H}$. R_{n_1,n_2}^k is the combined belief for H_{n_1} and H_{n_2}, and T_{n_1,n_2}^k is the combined uncertainty for H_{n_1} and H_{n_2}. We then fuse the belief and uncertainty from all estimators in $\boldsymbol{H}$ by iteratively utilizing the above belief fusion function to generate the combined belief $R_{\boldsymbol{H}}^k$ and combined uncertainty $T_{\boldsymbol{H}}^k$ for all estimators in $\boldsymbol{H}$ as follows:

$$\begin{aligned} R_{\boldsymbol{H}}^k &= \mathcal{B}(H_N, \mathcal{B}(H_{N-1}, ..., \mathcal{B}(H_3, \mathcal{B}(H_2, H_1)))) \\ T_{\boldsymbol{H}}^k &= \mathcal{F}(H_N, \mathcal{F}(H_{N-1}, ..., \mathcal{F}(H_3, \mathcal{F}(H_2, H_1)))) \end{aligned} \tag{5}$$

Given the combined belief $R_{\boldsymbol{H}}^k$ and uncertainty $T_{\boldsymbol{H}}^k$, we then set the class label estimated by our CrowdDesign framework to be the one that has the highest belief value $R_{\boldsymbol{H}}^k$ among all possible class labels k for each studied image D_i.

We observe that one key challenge to derive the optimal AI model design instance using the above approach is that the belief and uncertainty of each estimator is unknown *a priori*. To address this challenge, we design a novel iterative learning scheme to infer the belief R_n^k and uncertainty T_n^k of each estimator H_n in $\boldsymbol{H}$. In particular, we define $\boldsymbol{U} = \{U_1, U_2, ..., U_N\}$ as the set of PHPA estimation reliability for all estimators in $\boldsymbol{H}$, where each $U_n = \{U_n^1, U_n^2, ..., U_n^K\}$. In particular, U_n^k is the probability that an estimator H_n correctly infers the PHPA label of a studied image when its ground-truth label is k. We further define $\boldsymbol{V} = \{V_1, V_2, ..., V_I\}$ as PHPA label probability for all studied images, where each $V_i = \{V_i^1, V_i^2, ..., V_i^K\}$. In particular, we define V_i^k to be the probability that the PHPA label of a studied image to be k.

Given the two definitions above, we first optimize the PHPA label probability $\boldsymbol{V}$ using the PHPA estimation reliability $\boldsymbol{U}$. We observe that if different estimators in H with high reliability agree with each other on the estimated PHPA label of an image to be k, then the PHPA label of the image is likely to be k. Therefore, We calculate the PHPA label probability $\boldsymbol{V}$ as follows:

$$\begin{aligned} {V_i^k}^{**} = &\sum_{H_{n_1}, H_{n_2} \in \boldsymbol{H}^{i,k}} {U_{n_1}^k}^{*} \times {U_{n_2}^k}^{*} \times \frac{|\alpha_{n_1}^k \cap \alpha_{n_2}^k|}{|\alpha_{n_1}^k \cup \alpha_{n_2}^k|} \\ &\Big/ \sum_{H_{n_1}, H_{n_2} \in \boldsymbol{H}^{i,k}} {U_{n_1}^k}^{*} \times {U_{n_2}^k}^{*} \end{aligned} \tag{6}$$

where $*$ and $**$ represent two consecutive iterations. $\boldsymbol{H}^{i,k}$ represents the set of estimators in $\boldsymbol{H}$ that infer the PHPA label for a studied image D_i to be k. $\alpha_{n_1}^k$ and $\alpha_{n_2}^k$ represent the set of images whose PHPA labels are estimated as k by H_{n_1} and H_{n_2}, respectively. In addition, the value of V_i^k is 0 when no estimators share an agreement on the PHPA label of D_i to be k.

Next, we infer the PHPA estimation reliability $\boldsymbol{U}$ using the updated PHPA label probability $\boldsymbol{V}$. We observe that an estimator H_n has a high value of U_n^k when the estimator can correctly estimate the PHPA label of an image with a high probability of being k. We then update the PHPA estimation reliability $\boldsymbol{U}$ as follows:

$$U_n^{k^{**}} = \sum_{D_i \in \boldsymbol{D^{n,k}}} \sum_{n_1 \neq n}^{N} V_i^{k^{**}} \times \frac{|\alpha_{n_1}^k \cap \alpha_n^k|}{|\alpha_{n_1}^k \cup \alpha_n^k|} / \sum_{D_i \in \boldsymbol{D^{n,k}}} V_i^{k^{**}} \tag{7}$$

where $\boldsymbol{D^{n,k}}$ is the set of images where an estimator H_n infers the PHPA label to be k.

Using the above two equations, we iteratively compute $\boldsymbol{U}$ and $\boldsymbol{V}$ until the values of $\boldsymbol{U}$ and $\boldsymbol{V}$ remain the same between two consecutive iterations. We refer to the converged values of $\boldsymbol{U}$ and $\boldsymbol{V}$ as $\boldsymbol{U^o}$ and $\boldsymbol{V^o}$. We then leverage the $\boldsymbol{U^o}$ and $\boldsymbol{V^o}$ to derive the belief $R_n^{k^o}$ and uncertainty $T_n^{k^o}$ as follows:

$$\begin{aligned} R_n^{k^o} &= \Omega\Big(\sum_{\forall i \in \boldsymbol{\Delta_k^{H_n}}} V_i^{k^o}\Big) \times U_n^{k^o} \\ T_n^{k^o} &= 1 - \Omega\Big(\sum_{\forall i \in \boldsymbol{\Delta_k^{H_n}}} V_i^{k^o}\Big) \end{aligned} \tag{8}$$

where $\Omega(\cdot)$ is a normalization function to normalize the input between 0 and 1. $\sum_{\forall i \in \boldsymbol{\Delta_k^{H_n}}} V_i^{k^o}$ indicates the likelihood that G_n is certain about estimated labels for the images of class k.

Finally, the optimized belief $R_n^{k^o}$ and uncertainty $T_n^{k^o}$ for each estimator in $\boldsymbol{H}$ is plugged into Equation (5) to derive the combined belief $R_n^{\boldsymbol{H}^o}$ and identify the optimal AI model design instance. In addition, we summarize the CrowdDesign framework in Algorithm 1. We also provide an example to illustrate the usage of subjective logic in CrowdDesign. In particular, Subjective logic serves as a foundational element in the CrowdDesign framework, providing a robust mathematical approach to managing the uncertainty and variability inherent in crowdsourced data. This probabilistic reasoning framework extends beyond traditional binary assessments by incorporating degrees of belief, disbelief, and uncertainty, thus allowing for more nuanced decision-making in complex AI model design scenarios. In CrowdDesign, each participant's feedback is converted into a structured opinion represented as a triplet: belief (b) that quantifies the supportive evidence for a hypothesis; disbelief (d) that quantifies the counter-evidence; and uncertainty (u) that captures the ambiguity or lack of evidence. For instance, a participant confident in a particular AI model's efficacy might provide feedback represented as $b = 0.8, d = 0.1, u = 0.1$, indicating strong belief and minimal uncertainty. These individual opinions are aggregated using Dempster-Shafer theory, a robust method for synthesizing evidence from diverse sources. This aggregation process results in a collective opinion that includes consolidated belief, disbelief, and uncertainty scores for each AI model under consideration, emphasizing contributions that are more certain and minimizing the impact of those with significant doubt. For example, as shown in Figure 3, consider a scenario where 3 participants evaluate a new AI model: Worker 1 is highly confident ($b = 0.9, d = 0.05, u = 0.05$), Worker 2 is uncertain ($b = 0.3, d = 0.1, u = 0.6$), and Worker 3 is moderately skeptical ($b = 0.4, d = 0.4, u = 0.2$). The aggregated opinion might show moderate belief ($b = 0.53$), some disbelief ($d = 0.18$), and notable uncertainty ($u = 0.29$), indicating that while there is some optimism about the model, further investigation or evidence is necessary due to the mixed beliefs and high uncertainty. The decision-making process in CrowdDesign leverages this aggregated opinion to guide the final decision on adopting, modifying, or rejecting a model design. This ensures that decisions are informed by a comprehensive synthesis of crowd intelligence, balancing optimism with skepticism and reducing uncertainty where possible. The use of subjective logic not only democratizes the decision-making process but also ensures that it is robust, well-considered, and substantiated by detailed evidence, thus enhancing the reliability and effectiveness of crowd-AI model in our design.

Algorithm 1 CrowdDesign

```
initialize $\Omega$ (Definition 3)
obtain $\boldsymbol{A}$ (Definition 4)
obtain $\boldsymbol{B}$ (Definition 5)
generate $\Lambda$ (Definition 9) using JASR (Section IV-A)
generate $\boldsymbol{H}$ (Definition 14)
for each $H_n$ in $\boldsymbol{H}$ do
    for each $k$ in $[1, 2, ..., K]$ do
        initialize each $R_n^k$
        initialize each $T_n^k$
    end for
end for
while any $U_n^k$ and $V_i^k$ not coverage do
    for each $H_n$, $k$, $i$ do
        update $U_n^k$ (Equation (6))
        update $V_i^k$ (Equation (7))
    end for
end while
for each $H_n$, $k$, $i$ do
    generate $U_n^{k^o}$ and $V_i^{k^o}$
end for
for each $H_n$, $k$, $i$ do
    generate $R_n^{k^o}$ and $T_n^{k^o}$ (Equation (8))
end for
generate $Y_{\boldsymbol{H}}^k$ (Equation (5))
output $k^*$ for each $D_i$
```

V. Evaluation

A. Datasets and Crowdsourcing Settings

1) Two Real-world PHPA Applications: In our experiments, we evaluate CrowdDesign through two real-world PHPA applications: 1) *Mask Wearing Policy Adherence (MWPA)* application and *Social Distancing Policy Adherence (SDPA)* application. Both MWPA and SDPA are crucial for understanding their effectiveness in mitigating disease spread, informing tailored strategies, identifying high-risk areas, and

This article has been accepted for publication in IEEE Transactions on Emerging Topics in Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TETC.2024.3496835

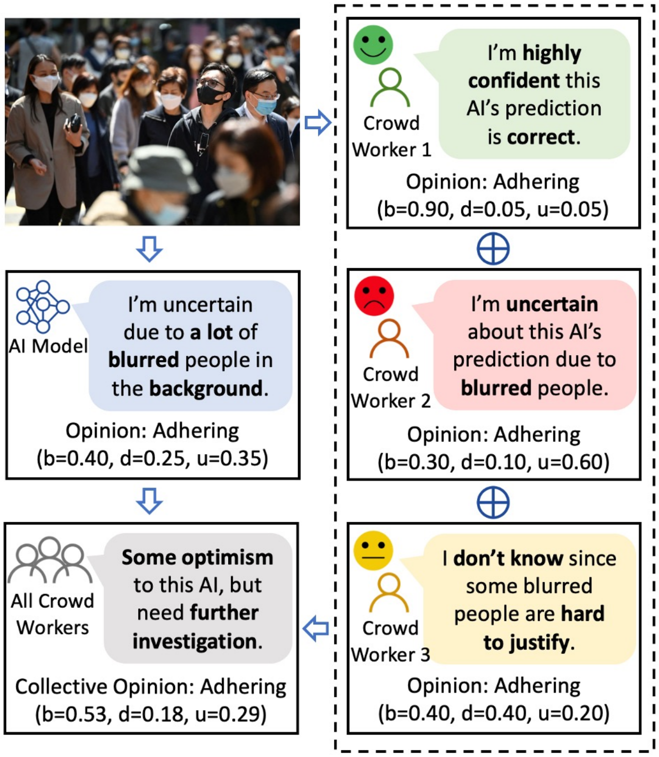

Crowd Worker 1
I'm highly confident this AI's prediction is correct.
Opinion: Adhering (b=0.90, d=0.05, u=0.05)
AI Model
I'm uncertain due to a lot of blurred people in the background.
Opinion: Adhering (b=0.40, d=0.25, u=0.35)
Crowd Worker 2
I'm uncertain about this AI's prediction due to blurred people.
Opinion: Adhering (b=0.30, d=0.10, u=0.60)
All Crowd Workers
Some optimism to this AI, but need further investigation.
Collective Opinion: Adhering (b=0.53, d=0.18, u=0.29)
Crowd Worker 3
I don't know since some blurred people are hard to justify.
Opinion: Adhering (b=0.40, d=0.40, u=0.20)

Figure 3. Illustration of Subjective Logic in CrowdDesign

refining public messaging during the emergent health crisis like COVID-19 [2]. In particular, both applications contain a set of public health-crisis related images collected from a widely used online social media platform (i.e., Twitter) during the COVID-19 pandemic. Following the standard practice in PHPA [2], we categorize the PHPA label in MWPA application into two classes: *adhering* (i.e., all people in the image wear the face mask properly based on the CDC's mask wearing guidelines [1]) and *not adhering* (i.e., not all people in the image wear the face mask properly). Similarly, in the SDPA application, we also categorize the PHPA label into two classes: *adhering* (i.e., all people in the image practice social distance based on the CDC's social distancing guidelines [2]) and *not adhering* (i.e., not all people in the image practice social distance). In both MWPA and SDPA, we ask three domain experts to annotate the PHPA label of each studied image and apply the majority voting to obtain the ground truth label of the image. The summary of the two datasets is presented in Table I. In addition, each dataset is split into training, validation, and testing sets using a ratio of 6:2:2. Specifically, we utilize the training and validation sets to train CrowdDesign and compared baselines for PHPA tasks. We use the testing set to evaluate the PHPA performance for CrowdDesign and compared baselines. Note that we used only 1,299 labeled images for the MWPA application and 616 labeled images for the SDPA application to train our model, which represents a data-scarce scenario compared to the vast amount of social media images generated during public health crises.

Table I
STATISTICS OF TWO PHPA APPLICATIONS

Application	MWPA	SDPA
Data Collection Time	May, 2020	May, 2020
# of Images	2,165	1,027
percentage of *Adhering*	75.4%	41.0%
percentage of *Not Adhering*	24.6%	59.0%

2) Crowdsourcing Settings: We utilize the widely used crowdsourcing platform Amazon Mechanical Turk (AMT) [3] to recruit crowd workers in our experiments. To ensure the annotation quality, we set two requirements for the crowd workers to participate in our task: 1) the crowd worker has finished at least 1000 approved tasks, and 2) the overall approval rate of the worker is greater than 95%. In our experiment, we recruit 479 crowd workers for MWPA application and 232 crowd workers for the SDPA application. The average annotation accuracy of the recruited crowd workers are 92.8% and 80.9% on the MWPA and SDPA applications, respectively. We pay a crowd worker $0.05 for each annotation task of an image. The IRB approval has been granted for all protocols in this project.

B. Baselines and Experimental Settings

We compare CrowdDesign with a rich set of state-of-the-art baselines for PHPA tasks, which include:

1) *Deep Neural Network (DNN)*:

- **ResNet** [38]: a convolutional network framework that imposes residual block architecture to improve the overall performance of image classification.
- **DenseNet** [39]: a densely connected convolutional neural network that leverages cross-layer connections to boost visual feature representation in image classification.
- **VGG** [40]: a popular neural network that leverages stacked deep convolutional network layers to ensure desirable network complexity for image classification tasks.
- **P>M>F** [33]: a transformer-based few-shot learning pipeline that achieves accurate and robust classification by incorporating pre-training on external data, meta-training with few-shot tasks, and task-specific fine-tuning.
- **Robust CLIP** [34]: an unsupervised adversarial fine-tuning approach to enhance the robustness of the CLIP vision encoder, which yields robustness on various zero-shot vision classification tasks.

2) *Crowd-AI Hybrid Learning*:

- **Deep Active** [14]: an active learning method that selects a subset of data samples for crowd labeling and leverages the crowd labels to retrain the AI model to improve the overall PHPA classification performance.
- **CrowdLearn** [15]: a crowd-AI hybrid framework that leverages crowdsourced human intelligence to troubleshoot AI models and improve the overall application performance.

[1] https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/masks.html)

[2] https://www.cdc.gov/coronavirus/2019-ncov/community/tribal/social-distancing.html

[3] https://www.mturk.com/

- **LL4AL** [16]: a crowd-AI collaborative method that utilizes a crowdsourcing uncertainty-aware deep estimation model to identify and fix AI failure cases in image classification.
- **CrowdOptim** [10]: A crowd-AI hybrid framework leverages crowdsourced human intelligence to optimize the hyperparameter configuration of a pre-selected network architecture.
- **CrowdNAS** [8]: A crowd-guided neural architecture searching framework utilizes crowd intelligence to identify the optimal neural architecture with pre-configured hyperparameters through the maximum likelihood estimation.

3) *AI Model Design*:

- **DEHB** [22]: a widely used AI model design framework that introduces a non-stochastic infinite-armed bandit-based strategy to improve the convergence speed of the AI model optimization process.
- **BOHB** [41]: a representative AI model design scheme that leverages the Bayesian optimization scheme to optimize AI model design.
- **MnasNet** [26]: a lightweight AI model design approach that introduces a factorized hierarchical AI model design searching process through a multi-objective reinforcement learning scheme.

To ensure a fair comparison, we use the same input to all compared baselines: 1) the images collected from the PHPA applications; 2) the ground-truth labels for the PHPA images in the training and validation sets; 3) the labeled PHPA images from crowdsourcing query. Specifically, we use the queried crowd labels to fine-tune the DNN and AI model design baselines so that all compared schemes have the same inputs and the performance of compared baselines is optimized. In the experiments, our CrowdDesign framework and compared baselines are implemented using PyTorch 1.1.0 libraries and trained on NVIDIA Quadro RTX 6000 GPUs. For DNN and crowd-AI baselines, we leverage widely used pre-configured AI model designs and further optimize the parameters of each compared scheme on the training and validation datasets. For AI model design baselines, we follow the standard practice in AI model optimization to optimize the AI model design using the training and validation datasets [4].

Following a standard AI model design process [6], [9], we set network architecture search space in our experiments to include 1) the types of the convolutional block to be residual block or dense block, 2) the number of convolutional layers per block to be between 1 and 36, 3) the width of the convolutional block to be between 2^1 and 2^7, 4) the growth rate to be between 32 and 48, and 5) the size of input features to be between 64 and 96. We also consider the hyperparameter configuration search space in the experiments to include 1) the learning rate to be between 10^{-6} and 10^{-3}, 2) the weight decay to be between 0 and 10^{-3}, 3) three candidate optimizers inclining SGD, RMSprop, and ADAM, 4) the conditional parameters of SDG momentum, RMSprop alpha, Adam beta1, and Adam beta2 to be between 0.8 and 1.0, and 5) the number of epochs to be between 30 and 150 in our experiments.

In our evaluation, we leverage four evaluation metrics that are widely used in image classification tasks with imbalanced data: 1) *Accuracy (Acc)*, 2) *F1-Score*, 3) *Kappa Score ($\mathcal{K}$-Score)*, and 4) *Matthews Correlation Coefficient (MCC)*. We include $\mathcal{K}$-Score and MCC because both MWPA and SDPA datasets are imbalanced, and $\mathcal{K}$-Score and MCC are shown to be reliable in evaluating the imbalanced classification performance. Higher values on these metrics represent better PHPA classification results.

C. Experimental Results

1) Classification Performance on Mask Wearing Policy Adherence: We first evaluate the classification performance of all compared schemes on assessing the mask wearing policy adherence. In particular, we set the query ratio of human intelligence to be 10% and the number of crowd workers per task in the crowdsourcing query to be 5. We will investigate the impact of different crowdsourcing query ratios and the numbers of crowd workers of our CrowdDesign in Section V-C3. We summarize the evaluation results in Table II. We observe that CrowdDesign clearly outperforms all compared baselines on all evaluation metrics. For example, the performance gains of CrowdDesign compared to the best-performing baseline (i.e., CrowdNAS) on Accuracy, F1-Score, $\mathcal{K}$-Score, and MCC are 3.93%, 2.78%, 9.14%, and 9.26%, respectively. Such performance improvements are mainly achieved by our principled subject logic-based AI model design framework that jointly utilizes the AI and human intelligence to identify the optimal AI model design instance that can accurately assess mask wearing policy adherence in PHPA.

Table II
PERFORMANCE ON MASK WEARING POLICY ADHERENCE

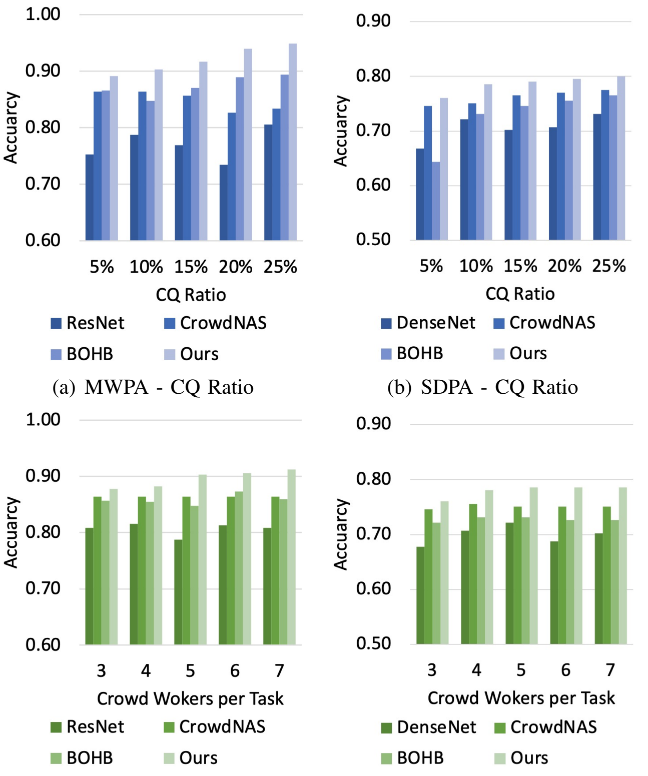
	Evaluation Metrics			
Algorithm	Acc	F1-Score	$\mathcal{K}$-Score	MCC
ResNet	0.7875	0.8563	0.4497	0.4509
DenseNet	0.7737	0.8444	0.4310	0.4346
VGG	0.7552	0.8127	0.4820	0.5311
P>M>F	0.8199	0.8863	0.4580	0.4714
Robust CLIP	0.8037	0.8522	0.5751	0.6231
Deep Active	0.8106	0.8735	0.4974	0.4975
CrowdLearn	0.8083	0.8693	0.5109	0.5135
LL4AL	0.7945	0.8548	0.5079	0.5201
CrowdNAS	0.8637	0.9088	0.6395	0.6398
CrowdOptim	0.8591	0.9051	0.6318	0.6327
DEHB	0.8406	0.8885	0.6119	0.6228
BOHB	0.8476	0.8918	0.6381	0.6557
MnasNet	0.7829	0.8454	0.4874	0.5020
CrowdDesign	**0.9030**	**0.9366**	**0.7309**	**0.7324**

2) Classification Performance on Social Distancing Policy Adherence: Secondly, we evaluate the classification performance of all compared methods in assessing social distancing policy adherence. Compared to the MWPA application, we have a different application objective in SDPA that focuses on identifying whether people in an image maintain social distance or not in public areas. In particular, we also set the crowdsourcing query ratio to be 10% and the number of crowd workers per task in the crowdsourcing query to be 5. We summarize the evaluation results in Table III. We note that CrowdDesign consistently surpasses all compared methods in terms of the performance on accurately assessing the social distancing policy adherence. The consistent performance gains achieved by CrowdDesign over two diversified PHPA applications (i.e., MDPA and SDPA) demonstrate the adaptability of CrowdDesign to identify optimal AI model design instance in PHPA applications with different objectives.

Table III
PERFORMANCE ON SOCIAL DISTANCING POLICY ADHERENCE

	Evaluation Metrics			
Algorithm	Acc	F1-Score	$\mathcal{K}$-Score	MCC
ResNet	0.6829	0.6524	0.3664	0.3729
DenseNet	0.7220	0.6545	0.4220	0.4222
VGG	0.6439	0.6605	0.3218	0.3577
P>M>F	0.6829	0.6632	0.3730	0.3844
Robust CLIP	0.7317	0.7027	0.4620	0.4685
Deep Active	0.7122	0.6424	0.4017	0.4019
CrowdLearn	0.6878	0.6559	0.3751	0.3810
LL4AL	0.6780	0.6916	0.3858	0.4269
CrowdNAS	0.7512	0.6222	0.4528	0.4842
CrowdOptim	0.7115	0.6015	0.4291	0.4633
DEHB	0.6683	0.6822	0.3672	0.4063
BOHB	0.7317	0.6746	0.4464	0.4464
MnasNet	0.6927	0.6631	0.3859	0.3927
CrowdDesign	**0.7854**	**0.7660**	**0.5719**	**0.5830**

3) Robustness of CrowdDesign on Different Crowdsourcing Settings: Third, we perform the robustness study to evaluate the performance of CrowdDesign on both MWPA and SDPA applications across different crowdsourcing settings. In particular, we change the crowdsourcing query (CQ) ratio δ in both applications from 5% to 25% and change the number of crowd workers per task C from 3 to 7. We compare the performance of CrowdDesign with the best-performing baselines in each category of compared baselines (i.e., ResNet, CrowdNAS, and BOHB for MWPA application, DenseNet, CrowdNAS, and BOHB for SDPA application). We summarize the evaluation results in Figure 4. We note that the performance of CrowdDesign is stable and clearly outperforms the best-performing baselines in both applications when the crowdsourcing query ratio δ and the number of crowd workers per task C change. The results show the robustness of our CrowdDesign scheme in addressing the optimal AI model design instance problem in PHPA applications over different crowdsourcing settings, especially when dealing with data-scarce scenarios with only a limited amount of crowd annotation data.



(a) MWPA - CQ Ratio

(b) SDPA - CQ Ratio

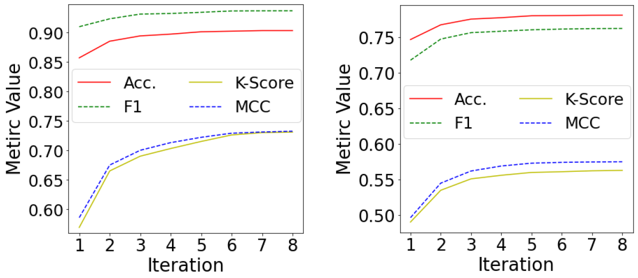
(c) MWPA - Crowd Workers per Task

(d) SDPA - Crowd Workers per Task

Figure 4. Robustness of CrowdDesign

4) Convergence of CrowdDesign: Fourth, we study the convergence of CrowdDesign by tracking its performance over different iterations during the iterative subjective logic learning process (discussed in Section IV-B). The results are shown in Figure 5. We observe that the performance of CrowdDesign, measured by various evaluation metrics, converges to optimized values quickly and remains stable afterward for both MWPA and SDPA. Such performance indicates that our principled subjective logic framework is effective in terms of obtaining optimized AI model designs. Additionally, the fast convergence rate demonstrates the efficiency and scalability of our CrowdDesign scheme in real-world PHPA applications.

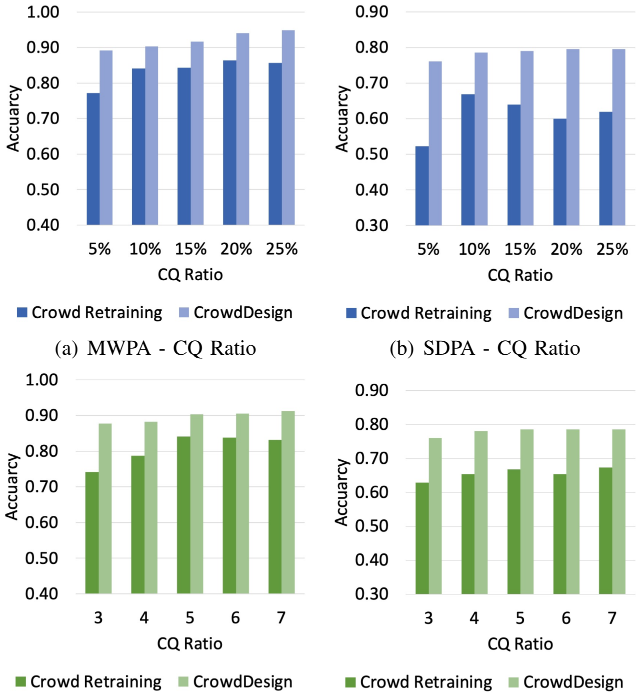
5) Performance Comparisons between OAMD Optimization and AI Model Retraining: Finally, we conducted experiments to compare the performance between our CrowdDesign that leverages crowd-labeled data for OAMD optimization and the version that leverages the crowd input data for model retraining (we refer to it as *Crowd Retraining*). In particular, we compared the performance between CrowdDesign and Crowd Retraining using the same set of crowd-labeled data from crowdsourcing queries. In addition, to ensure a comprehensive evaluation, we varied the crowdsourced query (CQ) ratio in both applications from 5% to 25% and changed the number of crowd workers per task (C) from 3 to 7, as we did in the



(a) MWPA (b) SDPA

Figure 5. Convergence of CrowdDesign

other evaluation comparisons. We show the evaluation results in Figure 6. We note that our CrowdDesign, which effectively utilizes the crowd labels to optimize the AI model, consistently outputs the Crowd Retraining scheme that leverages the crowd labels for retraining when the CQ ratio or crowd worker per task varies. The evaluation results validate the assumption of our paper that efficiently leveraging the crowd-labeled data for OAMD optimization ensures desirable PHPA performance.



(a) MWPA - CQ Ratio (b) SDPA - CQ Ratio

(c) MWPA - Crowd Workers per Task (d) SDPA - Crowd Workers per Task

Figure 6. Comparison between Crowd Retraining and Model Optimization

VI. Discussions

A. Limitations and Challenges of CrowdDesign

In this subsection, we discuss a few limitations and challenges of CrowdDesign that could offer insights for future research directions. First, the reliance on crowdsourced human intelligence within CrowdDesign introduces inherent challenges due to the variability in data quality from the crowd, which can significantly affect the model's accuracy and reliability. The crowdsourced data quality can be influenced by factors like the clarity of task descriptions, participant motivation, and the design of incentive structures. To mitigate these risks, the framework includes a comprehensive onboarding process for new contributors that features interactive tutorials and example tasks, which clarify expectations and demonstrate successful task completion strategies. Continuous engagement strategies are also deployed, including a tiered reward system where contributors can achieve different levels of certification and bonuses for consistent high-quality submissions. Furthermore, CrowdDesign incorporates a robust anomaly detection system that identifies and investigates any sudden changes in data quality, which could indicate issues with task understanding or engagement. This proactive approach not only maintains the integrity of the input data but also enhances contributor performance over time by aligning their efforts more closely with the framework's objectives.

Scalability is a critical aspect of CrowdDesign, especially given the potential for exponential growth in data volume and the complexity of AI model design spaces. CrowdDesign can be engineered on a modular cloud-based architecture that can elastically scale up or down according to real-time demands, ensuring efficient handling of large-scale data without sacrificing processing speed or system stability. To enhance scalability, CrowdDesign can be integrated with Amazon EC2 for dynamic resource allocation and Google Kubernetes Engine (GKE) to orchestrate containerized AI models, promoting seamless scalability and efficient resource management across distributed environments. For managing the complexity of the AI model design space, CrowdDesign employs a decompositional approach where the overall problem is segmented into smaller and discrete components that can be solved independently in parallel pipelines. In particular, the JASR module in CrowdDesign can be further scaled up by utilizing parallelized budget-constant non-stochastic multi-armed bandit (BNMB) model. This ensures that the search space can be segmented into a set of subspaces for parallel search of the AI model design search space, enhancing the scalability of CrowdDesign to large-scale datasets while ensuring that individual segments of search spaces collectively contribute to globally optimized AI model optimization results. This approach not only facilitates more manageable processing loads but also allows for specialized optimization techniques to be applied to different segments of the design space. Each segment incorporates machine learning algorithms optimized for specific types of data and tasks, ensuring that the system remains robust even as the complexity and size of the data grow.

We are working on sequentially stacked convolutional blocks because they are well-suited for analyzing social media images in public health policy adherence assessments [42]. These blocks efficiently extract image features that capture key patterns such as mask-wearing or social distancing behaviors in PHPA applications. Their low computational cost and efficient implementation make them ideal for processing diversified and complex image datasets from social media. However, we also recognize the potential for improving model performance by extending our approach to more advanced architectures, such as transformers, which are particularly effective at capturing long-range dependencies and contextual information in complex visual data. The CrowdDesign framework, while initially focused on convolutional blocks, can be extended to explore transformer architectures through recent advances in neural architecture search (NAS). Current research in NAS for transformers, such as the work on AutoFormer and Evolved Transformer [43], demonstrates how automated search methods can optimize the configuration of multi-headed attention mechanisms, the number of layers, and tokenization strategies in transformer models. By integrating the transformer architecture search space introduced these NAS techniques to our JASR module, we can build upon the flexibility of the CrowdDesign framework to automatically discover optimal transformer architectures tailored to the public health policy adherence tasks. This would allow us to explore the space of attention-based architectures systematically, while maintaining computational efficiency through our budget-constrained multi-armed bandit learning framework, enabling us to incorporate state-of-the-art transformers and further enhance model performance for our target applications.

B. CrowdDesign's Applicability in AI-driven Applications

The CrowdDesign framework, primarily developed for public health policy adherence in our paper, has a broad applicability across diverse application domains. This versatile approach integrates crowdsourced human intelligence with an advanced AI model design framework, making it ideal for domains requiring AI model optimization based on diversified input data and user feedback. Below, we illustrate its potential through applications in recommendation systems and intelligent transportation systems (ITS), showcasing its adaptability to other AI-driven applications. For example, in the domain of recommendation systems, the CrowdDesign framework integrates a sophisticated design that captures both quantitative metrics such as purchase rate, viewing duration, and user engagement level (e.g., likes, shares, and comments) as well as qualitative user feedback to continuously refine AI models. This process is vital across various platforms, such as e-commerce or media streaming services, where user satisfaction directly influences business outcomes. Data scientists employ the framework to analyze algorithmic outputs using objective performance metrics, including click-through rates, conversion rates, and user retention statistics. Simultaneously, end-users contribute subjective feedback reflecting their personal experiences, satisfaction, and the relevance of the recommendations they receive. Crowd input plays a critical role in adjusting the AI models within the framework, ensuring that the models evolve in response to user feedback, which directly impacts the accuracy and relevancy of the models' outputs. By integrating this real-time, user-generated data, CrowdDesign effectively tailors the AI's learning process, optimizing recommendation systems to better align with current consumer preferences and behaviors. This dual-input approach enables CrowdDesign to transform disparate data points into a structured format using its subjective logic mechanism: belief (derived from positive user feedback and/or enhanced performance in quantitative metrics), disbelief (primarily stemming from suboptimal results in quantitative metrics and/or negative use feedback), and uncertainty (highlighting areas with insufficient data or conflicting feedback). For instance, an e-commerce platform might use CrowdDesign to fine-tune its product recommendation algorithms by assigning greater weight to user feedback on product relevance, especially during peak shopping seasons. This strategic integration of crowd feedback allows the AI models to adapt more dynamically to changing market conditions and user preferences, enhancing both the personalization and effectiveness of the recommendations. Such dynamic adjustment ensures that the recommendation algorithms are not only adaptive but also responsive to real-time user behavior and feedback, thus enhancing personalization and improving user satisfaction.

In the domain of ITS, the CrowdDesign framework can be applied in developing AI models for accurately predicting traffic accident rates. This application benefits from the CrowdDesign approach that integrates crowdsourced human intelligence and data-driven AI insights to optimize the AI model, enhancing the accuracy and reliability of traffic accident predictions, which are crucial for public safety and urban planning. In particular, using CrowdDesign, traffic data scientists can leverage the training data collected from quantitative traffic metrics (e.g., vehicle speeds, traffic volume, and weather conditions) from traffic sensors to effectively train the initial AI model for traffic predictions. Meanwhile, the onsite qualitative inputs from local authorities and commuters (e.g., reports of hazardous road conditions, experiences of traffic incidents, and feedback on traffic signal efficacy) can be further leveraged by CrowdDesign to continuously optimize the AI models, ensuring desirable prediction accuracy. These combined insights are particularly valuable in dynamic environments like urban traffic systems, where standalone sensor data may not fully capture the subtleties and complexities of real-world conditions. For example, while sensors can report rain, crowdsourced feedback from drivers via mobile apps can provide additional context about the actual road conditions, such as the effectiveness of drainage systems and the presence of unexpected obstacles like fallen branches. Such information allows the model to dynamically adjust the architecture and hyperparameters of the AI model for more accurate predictions, incorporating both sensor data and human observations. This significantly enhances the model's adaptability to nuanced situational changes with an updated network architecture.

C. Feasibility and Implications of Using Social Media Imagery for PHPA

The utilization of social media imagery in our paper provides an innovative approach to PHPA by leveraging the immediacy and vast scale of such data sources. Social media platforms offer real-time, pervasive, and large-scale visual data that reflect public behaviors and social trends [3]. This immediacy and breadth are crucial in rapidly evolving situations, such as public health emergencies, where timely data are essential for responsive decision-making. The feasibility of using social media data for PHPA is supported by existing literature that demonstrates the effective use of these platforms for gathering real-time public health-related data. For example, studies have shown that social media can be a valuable source for tracking disease outbreaks, public sentiment towards health campaigns, and compliance with health guidelines [18], [19]. These studies provide a foundation for believing that social media imagery, which often contains visual evidence of compliance such as mask-wearing or social distancing, can similarly serve as a reliable indicator for PHPA. In terms of the implications, the use of social media allows for rapid, cost-effective data collection across diverse geographic locations and communities, potentially leading to more agile and informed public health responses. Moreover, it offers a means to assess policy adherence passively without necessitating intrusive methods or extensive manual data collection. This approach not only enhances the speed and breadth of data collection but also supports more dynamic and immediate adjustments to public health policies based on observed behaviors. Traditional PHPA techniques typically include direct observation by health officials or trained volunteers who manually monitor and record adherence behaviors in public or healthcare settings, and surveys or interviews where individuals report their own or others' adherence behaviors [44]. These methods, while providing high accuracy and reliability, suffer from several limitations. Direct observation is labor-intensive and can be intrusive, often limited to smaller, more controlled population samples. Surveys and interviews are prone to biases such as self-reporting bias and the social desirability bias, where respondents might not always provide accurate information [2]. Those approaches are also time-consuming and costly, making them less effective and scalable for rapid response scenarios during health crises. In contrast, our CrowdDesign approach harnesses the widespread availability of social media data to overcome these challenges by enabling broader and more agile PHPA assessments. It effectively identifies the optimized AI model design to provide desirable PHPA label estimations for the vast amount of real-time social media imagery data. This allows for real-time monitoring and assessment of public health policy adherence, providing public health officials with timely insights that are crucial during rapidly evolving health crises.

However, utilizing social media imagery for PHPA via the CrowdDesign framework introduces specific technical challenges and ethical concerns. One of the primary technical challenges is ensuring the quality and reliability of the data extracted from social media. Social media platforms often contain a high volume of unstructured and heterogeneous data, which can vary significantly in quality and context. To address this, CrowdDesign develops advanced image processing and machine learning algorithms to filter and preprocess imagery data effectively. These algorithms enhance data quality by identifying and discarding irrelevant or low-quality images and by normalizing the data to reduce variability introduced by different user-generated content. Moreover, the representativeness of social media samples is another critical issue, as the demographic distribution of social media users does not typically mirror that of the general population. This discrepancy can introduce sample bias [45], potentially skewing the PHPA outcomes. In CrowdDesign, we mitigate this bias through a hybrid model that integrates data from multiple social media platforms e.g., Twitter/X, Instagram, and Facebook/Meta, and, when possible, combines this with traditional data sources (e.g., healthcare surveys, official public health records, and in-person observations at public venues). This integrated approach broadens the demographic coverage and enhances the generalizability of the findings.

Ethical considerations in CrowdDesign are important, particularly regarding privacy and consent. Social media images often contain personally identifiable information, posing significant privacy concerns. To navigate these ethical waters, CrowdDesign applies rigorous data anonymization techniques (e.g., pixelation, blurring, and hashing of direct identifiers) and advanced computer vision algorithms that include facial recognition tools to detect and obscure faces or other identifiable markers in images before they are processed. Additionally, the framework adheres to strict data governance protocols (e.g., the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA)) that comply with both ethical standards and legal requirements concerning user data. CrowdDesign also addresses the ethical issue of consent by deploying mechanisms that ensure only publicly available images, where users have consented to broader visibility, are harvested and utilized for analysis. Furthermore, the framework features an ethical oversight module that continuously reviews adherence to privacy standards and consent norms, adjusting data collection and processing practices as needed to align with evolving ethical guidelines and user preferences.

D. The potential of Using Labeled data for both OAMD Optimization and Retraining

The potential of using labeled data for both OAMD optimization and AI model retraining is vast and can be harnessed to enhance the effectiveness and efficiency of AI systems. OAMD optimization focuses on refining the configuration of network architecture and hyperparameters. This approach allows for quick adaptation of models without the need for extensive retraining, thereby conserving computational resources and time. In contrast, retraining AI models with new, crowdsourced labeled data addresses different challenges, such as adapting to concept drift in dynamic environments. This continuous retraining helps improve the model's accuracy and robustness, as it learns from the most recent and diverse data, which may introduce previously unseen features or classes. It also helps in reducing overfitting by ensuring that the model does not overly specialize to the noise or biases of a specific dataset. Integrating these two approaches can synergistically enhance AI system performance. For example, one could employ a sequential integration where OAMD optimization is used for immediate model refinement, followed by periodic retraining to adapt to new data trends over time. Alternatively, parallel integration could be applied where optimization and retraining are conducted concurrently to ensure that the model architecture and parameters are continuously fine-tuned while its weights are adapted to the changing data landscape. Additionally, a feedback mechanism could be beneficial, where insights from retraining sessions inform future OAMD processes, creating a cycle that continuously refines and adapts the model. However, integrating both strategies requires careful consideration of challenges such as ensuring data quality and diversity, managing increased computational demands, and balancing the model's specificity and generalization capabilities. The integration of OAMD optimization and model retraining, if properly implemented, can lead to the development of robust, efficient, and adaptable AI systems that are better suited to meet the evolving demands and complexities of real-world applications. The key lies in finding the right balance and integration strategy that satisfies the specific needs and constraints of the operational environment.

VII. Conclusion

The paper introduces a CrowdDesign framework to solve the optimal AI model design problem in PHPA applications. In particular, our CrowdDesign develops a crowdsourcing-driven AI model design scheme that leverages crowdsourced human intelligence to identify the optimal AI model design instance for a PHPA application. Our CrowdDesign effectively translates the highly complex AI model design problem into a simplified problem that can be solved by crowd workers while considering the interdependence between the network architecture and hyperparameter configuration in the AI model design. Our CrowdDesign is shown to achieve the highest PHPA accuracy compared to a broad set of baselines in two real-world PHPA applications. We believe CrowdDesgin provides useful insights to develop novel crowdsourcing-driven AI model design frameworks in addressing the optimal AI model design problem in AI-driven applications beyond PHPA.

Acknowledgement

This research is supported in part by the National Science Foundation under Grant No. CNS-2427070, IIS-2331069, IIS-2202481, IIS-2130263, CNS-2131622, CNS-2140999. 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 U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] J. Budd, B. S. Miller, E. M. Manning, V. Lampos, M. Zhuang, M. Edelstein, G. Rees, V. C. Emery, M. M. Stevens, N. Keegan *et al.*, "Digital technologies in the public-health response to covid-19," *Nature medicine*, vol. 26, no. 8, pp. 1183–1192, 2020.

[2] V. Negri, D. Scuratti, S. Agresti, D. Rooein, G. Scalia, A. R. Shankar, J. L. F. Marquez, M. J. Carman, and B. Pernici, "Image-based social sensing: combining ai and the crowd to mine policy-adherence indicators from twitter," in *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*. IEEE, 2021, pp. 92–101.

[3] S.-F. Tsao, H. Chen, T. Tisseverasinghe, Y. Yang, L. Li, and Z. A. Butt, "What social media told us in the time of covid-19: a scoping review," *The Lancet Digital Health*, vol. 3, no. 3, pp. e175–e194, 2021.

[4] M. Feurer and F. Hutter, "Hyperparameter optimization," in *Automated machine learning*. Springer, Cham, 2019, pp. 3–33.

[5] X. Hu, L. Chu, J. Pei, W. Liu, and J. Bian, "Model complexity of deep learning: A survey," *Knowledge and Information Systems*, vol. 63, no. 10, pp. 2585–2619, 2021.

[6] A. Zela, A. Klein, S. Falkner, and F. Hutter, "Towards automated deep learning: Efficient joint neural architecture and hyperparameter search," *arXiv preprint arXiv:1807.06906*, 2018.

[7] M.-H. Tsou, "Research challenges and opportunities in mapping social media and big data," *Cartography and Geographic Information Science*, vol. 42, no. sup1, pp. 70–74, 2015.

[8] Y. Zhang, R. Zong, Z. Kou, L. Shang, and D. Wang, "Crowdnas: A crowd-guided neural architecture searching approach to disaster damage assessment," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW2, pp. 1–29, 2022.

[9] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, "Hyperband: A novel bandit-based approach to hyperparameter optimization," *Journal of Machine Learning Research*, vol. 18, no. 185, pp. 1–52, 2018. [Online]. Available: http://jmlr.org/papers/v18/16-558.html

[10] Y. Zhang, R. Zong, L. Shang, Z. Kou, H. Zeng, and D. Wang, "Crowdoptim: A crowd-driven neural network hyperparameter optimization approach to ai-based smart urban sensing," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW2, pp. 1–27, 2022.

[11] R. Egele, P. Balaprakash, I. Guyon, V. Vishwanath, F. Xia, R. Stevens, and Z. Liu, "Agebo-tabular: joint neural architecture and hyperparameter search with autotuned data-parallel training for tabular data," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2021, pp. 1–14.

[12] J. Guerrero-Viu, S. Hauns, S. Izquierdo, G. Miotto, S. Schrodi, A. Biedenkapp, T. Elsken, D. Deng, M. Lindauer, and F. Hutter, "Bag of baselines for multi-objective joint neural architecture search and hyperparameter optimization," *arXiv preprint arXiv:2105.01015*, 2021.

[13] F. Alam, F. Ofli, M. Imran, and M. Aupetit, "A twitter tale of three hurricanes: Harvey, irma, and maria," *arXiv preprint arXiv:1805.05144*, 2018.

[14] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in *International Conference on Learning Representations*, 2018.

[15] D. Zhang, Y. Zhang, Q. Li, T. Plummer, and D. Wang, "Crowdlearn: A crowd-ai hybrid system for deep learning-based damage assessment applications," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2019, pp. 1221–1232.

[16] M. Shukla and S. Ahmed, "A mathematical analysis of learning loss for active learning in regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3320–3328.

[17] R. M. Merchant, E. C. South, and N. Lurie, "Public health messaging in an era of social media," *Jama*, vol. 325, no. 3, pp. 223–224, 2021.

[18] J. Xue, J. Chen, R. Hu, C. Chen, C. Zheng, Y. Su, and T. Zhu, "Twitter discussions and emotions about the covid-19 pandemic: Machine learning approach," *Journal of medical Internet research*, vol. 22, no. 11, p. e20550, 2020.

[19] E. Bonnevie, S. D. Rosenberg, C. Kummeth, J. Goldbarg, E. Wartella, and J. Smyser, "Using social media influencers to increase knowledge and positive attitudes toward the flu vaccine," *Plos one*, vol. 15, no. 10, p. e0240828, 2020.

[20] E. Afful-Dadzie, A. Afful-Dadzie, and S. B. Egala, "Social media in health communication: A literature review of information quality," *Health Information Management Journal*, vol. 52, no. 1, pp. 3–17, 2023.

[21] J. L. Bender, A. B. Cyr, L. Arbuckle, and L. E. Ferris, "Ethics and privacy implications of using the internet and social media to recruit participants for health research: A privacy-by-design framework for online recruitment," *Journal of medical Internet research*, vol. 19, no. 4, p. e104, 2017.

[22] N. Awad, N. Mallik, and F. Hutter, "Dehb: Evolutionary hyperband for scalable, robust and efficient hyperparameter optimization," *arXiv preprint arXiv:2105.09821*, 2021.

[23] Y. Liu, Y. Sun, B. Xue, M. Zhang, G. G. Yen, and K. C. Tan, "A survey on evolutionary neural architecture search," *IEEE transactions on neural networks and learning systems*, vol. 34, no. 2, pp. 550–570, 2021.

[24] Y. Chen, X. Song, C. Lee, Z. Wang, R. Zhang, D. Dohan, K. Kawakami, G. Kochanski, A. Doucet, M. Ranzato *et al.*, "Towards learning universal hyperparameter optimizers with transformers," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32 053–32 068, 2022.

[25] Y. Hirose, N. Yoshinari, and S. Shirakawa, "Nas-hpo-bench-ii: A benchmark dataset on joint optimization of convolutional neural network architecture and training hyperparameters," in *Asian Conference on Machine Learning*. PMLR, 2021, pp. 1349–1364.

[26] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le, "Mnasnet: Platform-aware neural architecture search for mobile," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2820–2828.

[27] J. Hong, K. Lee, J. Xu, and H. Kacorri, "Crowdsourcing the perception of machine teaching," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–14.

[28] A. Guo, A. Jain, S. Ghose, G. Laput, C. Harrison, and J. P. Bigham, "Crowd-ai camera sensing in the real world," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, pp. 1–20, 2018.

[29] L. Lin, D. Bermejo-Peláez, D. Capellán-Martín, D. Cuadrado, C. Rodríguez, L. García, N. Díez, R. Tomé, M. Postigo, M. J. Ledesma-Carbayo *et al.*, "Combining collective and artificial intelligence for global health diseases diagnosis using crowdsourced annotated medical images," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 3344–3348.

[30] D. Yoo and I. S. Kweon, "Learning loss for active learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 93–102.

[31] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM computing surveys (csur)*, vol. 53, no. 3, pp. 1–34, 2020.

[32] F. Pourpanah, M. Abdar, Y. Luo, X. Zhou, R. Wang, C. P. Lim, X.-Z. Wang, and Q. J. Wu, "A review of generalized zero-shot learning methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 4, pp. 4051–4070, 2022.

[33] S. X. Hu, D. Li, J. Stühmer, M. Kim, and T. M. Hospedales, "Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9068–9077.

[34] C. Schlarmann, N. D. Singh, F. Croce, and M. Hein, "Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models," *arXiv preprint arXiv:2402.12336*, 2024.

[35] Y. Chen, Z. Liu, H. Xu, T. Darrell, and X. Wang, "Meta-baseline: Exploring simple meta-learning for few-shot learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9062–9071.

[36] A. Zhao, M. Ding, Z. Lu, T. Xiang, Y. Niu, J. Guan, and J.-R. Wen, "Domain-adaptive few-shot learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1390–1399.

[37] L. Li and A. Talwalkar, "Random search and reproducibility for neural architecture search," *Proceedings of Machine Learning Research*, vol. 119, pp. 3671–3680, 2020.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[39] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks." in *CVPR*, vol. 1, no. 2, 2017, p. 3.

[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

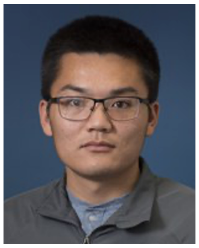
[41] S. Falkner, A. Klein, and F. Hutter, "Bohb: Robust and efficient hyperparameter optimization at scale," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1437–1446.

[42] S. D. Young *et al.*, "Social media images as an emerging tool to monitor adherence to covid-19 public health guidelines: Content analysis," *Journal of Medical Internet Research*, vol. 24, no. 3, p. e24787, 2022.

[43] M. Chen, H. Peng, J. Fu, and H. Ling, "Autoformer: Searching transformers for visual recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 12 270–12 280.

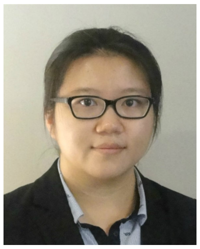
[44] S. Kwon, A. D. Joshi, C.-H. Lo, D. A. Drew, L. H. Nguyen, C.-G. Guo, W. Ma, R. S. Mehta, F. M. Shebl, E. T. Warner *et al.*, "Association of social distancing and face mask use with risk of covid-19," *Nature Communications*, vol. 12, no. 1, pp. 1–10, 2021.

[45] L.-E. Casper Ferm and P. Thaichon, "Successful covid-19 prevention factors and their effect on the economy: A comparison between thailand, vietnam and australia," *COVID-19, Technology and Marketing: Moving Forward and the New Normal*, pp. 59–83, 2021.

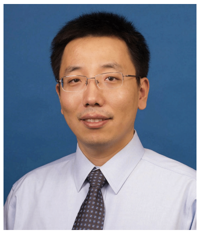


Yang Zhang received his Ph.D. in Computer Science and Engineering from the University of Notre Dame in 2022. He is now a teaching assistant professor in the School of Information Sciences at the University of Illinois Urbana-Champaign. His research interests lie in human-centered artificial intelligence (AI) and social (human-centric) sensing, intelligence, and computing. He is the recipient of the Outstanding Graduate Research Award at the University of Notre Dame and the W. J. Cody Research Associates at Argonne National Laboratory. He is a member of IEEE.

Ruohan Zong is a Ph.D. student in the School of Information Sciences at the University of Illinois Urbana-Champaign (UIUC). She received an M.S. in Computer Science from Columbia University and a B.E. in Computer Science and Technology from Sichuan University. Her primary research interests are human-centered AI and AI for social good. Ruohan is the recipient of the Illinois Distinguished Fellowships at UIUC. She is a student member of IEEE.



Lanyu Shang is a Ph.D. student in the School of Information Sciences at the University of Illinois Urbana-Champaign (UIUC). She received an M.S. in Data Science from New York University and a B.S. in Applied Mathematics from the University of California - Los Angeles (UCLA). Her research interest primarily lies in online misinformation detection using social media data. Lanyu is also the recipient of the Outstanding Graduate Teaching Award at the University of Notre Dame, the Best Paper Award at ACM/IEEE ASONAM 2022, and the Best Paper Honorable Mention at IEEE SmartComp 2022. She is a student member of IEEE.



Dong Wang received his Ph.D. in Computer Science from University of Illinois Urbana-Champaign (UIUC) in 2012. He is now an associate professor in the School of Information Sciences at the University of Illinois Urbana-Champaign. Dr. Wang's research interests lie in the area of social (human-centric) sensing, intelligence and computing, human-centered AI, AI for social good, data quality, and big data analytics. He is the recipient of NSF CAREER Award, Google Faculty Research Award, Young Investigator Program (YIP) Award from the ARO, Wing Kai Cheng Fellowship from the University of Illinois, the Best Paper Award of 2022 ACM/IEEE International Conference on Advances in Social Networks Analysis and Mining (ASONAM), the Best Paper Award of 16th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS) and the Best Paper Honorable Mention of 8th IEEE SmartComp. He is a senior member of IEEE.