Improving Quality Control with Industrial AIoT at HP Factories: Experiences and Learned Lessons

Joy Qiping Yang* Siyuan Zhou* Duc Van Le* Daren Ho† Rui Tan‡
*HP-NTU Digital Manufacturing Corporate Lab, Nanyang Technological University †HP Inc.
‡School of Computer Science and Engineering, Nanyang Technological University

***Abstract*—Enabled by the increasingly available embedded hardware accelerators, the capability of executing advanced machine learning models at the edge of the Internet of Things (IoT) triggers wide interest of applying the resulting Artificial Intelligence of Things (AIoT) systems in industrial applications. The *in situ* inference and decision made based on the sensor data containing patterns with certain sophistication allow the industrial system to address a variety of heterogeneous, local-area non-trivial problems in the last hop of the IoT networks, avoiding the wireless bandwidth bottleneck and unreliability issues and also the cumbersome cloud. However, the literature still lacks presentations of industrial AIoT system developments that provide insights into the challenges and offer important lessons for the relevant research and engineering communities, no matter the development is successful or not. In light of this, we present the design, deployment, and evaluation of an industrial AIoT system for improving the quality control of Hewlett-Packard's ink cartridge manufacturing lines. While our development has obtained promising results, we also discuss the lessons learned from the whole course of the effort, which could be useful to the developments of other industrial AIoT systems.**

***Index Terms*—Industrial AIoT, quality control**

I. Introduction

The recent advances of machine learning (ML) techniques in dealing with sophisticated industrial data patterns and the increasingly available embedded hardware for accelerating ML trigger the interest of studying and implementing industrial Artificial Intelligence of Things (AIoT) [1] that integrates artificial intelligence (AI) with the Internet of Things (IoT) edge. The AIoT systems will have distributed, *in situ* inference and decision capabilities to avoid the handicaps encountered when transmitting data to remote central servers for decision making. As such, AIoT is promising for addressing a variety of local-area, non-trivial problems in the industrial processes.

However, there is no one-size-fits-all AIoT system that can be used for all industrial applications. The designs and implementations of the AIoT systems in general need to be highly customized based on the specific objectives, operational procedures, and practical constraints of the industrial processes. Due to the application-oriented nature and the high/prohibitive cost of the system design from scratch, it is wise to integrate commercial off-the-shelf (COTS) hardware modules, follow academically proven approaches, and use available software components to implement the desired functionalities. However, many task-specific designs such as the configuration and training of the used ML models still require substantial work to achieve the objectives. The main challenges often come from the deviations of the real-world conditions from the assumptions made by the relevant research artifacts. This is expected, because the relevant research in general needs a set of clearly defined assumptions to render a satisfactory level of academic rigor in addressing a specific problem while isolate other problems, but real-world tasks in industrial practices face many coupled problems. Therefore, the design of a working industrial AIoT system requires holistic considerations with many inputs from the domain experts and technicians.

Despite the heterogeneity of industrial AIoT systems, the systematic description of an effort that designs and implements an AIoT system for a specific industrial application can provide insights into understanding the potential challenges that would be faced by other AIoT system designs. However, so far, such a systematic presentation is still lacking. As such, in this paper, we present our recent effort of developing an industrial AIoT system for improving the quality control (QC) of the ink cartridge manufacturing lines at the factories of Hewlett-Packard (HP) Inc. This development includes the key elements of AIoT, including sensing, data processing, design and deployment of embedded ML models at the IoT edge. We present the motivation, the details of our system design, and more importantly, the experiences and lessons learned from this effort that can be useful to the design and implementation of other industrial AIoT systems.

In this paper, the target application is HP's ink extraction testing (IET), which is a destructive and accelerated testing on randomly selected samples of the manufactured ink cartridges. It is the final QC procedure which aims at detecting any defective batch in which the ink cartridges' performance deviates from the specification. Specifically, the IET machine (referred to as *tester* for short in this paper) extracts the ink from the tested cartridge at a prescribed rate, which is much faster than those on printers, and records the liquid pressure of the ink throughout the course. The *profile curve* of the liquid pressure versus the volume of the extracted ink provides rich information regarding the performance of the tested ink cartridge. Thus, the match between the recorded profile and a preset template profile is the main criterion to pass the test. The alarms due to detected mismatch will be further classified manually by well trained technicians. Depending on

This research was conducted in collaboration with HP Inc. and supported by the Singapore Government through the Industry Alignment Fund-Industry Collaboration Projects Grant.

978-1-6654-4108-7/21/$31.00

the manual classification results, further QC actions will be taken. However, the factories' current IET procedure faces two main challenges as follows.

First, it is desirable to solidify the technicians' experience-based approach of manually classifying alarms as a computable classifier for the purpose of operation consistency and knowledge transfer. However, the pressure profiles exhibit a significant degree of variability and the technicians' manual classification incorporates extensive domain knowledge regarding the internals of the ink cartridges, which may be descriptive and not quantifiable. The attempt of converting the manual classification approach into a computable rule-based classifier results in many questions of how to properly define the features, configure the rules, and set the thresholds.

Second, the operations of the tester inevitably introduce uncertainties that result in false alarms. For example, from the technicians' experiences, formation of air bubbles in the tester's ink tubes is one of the major factors causing false alarms, because a bubble with a sufficiently large volume affects the liquid pressure measurement. Performing a tube flush before each test can largely resolve the issue, but it significantly reduces the testing throughput. From the historical records, the overall alarm rate of the deployed testers is about 30 times of the defect rate of the manufactured ink cartridges, suggesting most alarms are false. For quality assurance, upon any alarm, the factories' current practice is to flush the tester's tube and perform the destructive test on an additional ink cartridge sample to reconfirm the technician's manual classification result. Thus, it is desirable to have an approach that can reliably identify the false alarms and avoid the unnecessary additional destructive tests.

To address the above two challenges, we have designed and implemented an AIoT system that classifies the tester's alarms into product-induced (i.e., true alarms) and tester-induced (i.e., false alarms). The primary design goal is to achieve high recall and precision in identifying the product-induced and tester-induced alarms. Specifically, our AIoT system has three components. First, the *ML-based profile classifier* captures the product engineers' experiences in classifying the alarms. Second, we develop a heuristic-based anomaly detection (AD) approach that classifies the pressure profiles based on domain knowledge on the patterns contained in the profiles. Third, based on a key observation that the air bubbles are often formed at the joint of the tester's ink tubes, we deploy a *smart camera* at the joint and design convolutional neural network (CNN) and computer vision algorithms that run on the camera to detect and estimate the presence and volume of air bubbles. We have deployed our AIoT system in HP's manufacturing lines. Through controlled experiments, our heuristic-based AD approach achieves a recall of 95.2% in detecting the defective ink cartridges. Moreover, the smart camera can correctly detect the presence of air bubbles in 94% of the testing images. This paper presents the design and evaluation processes of the AIoT system. We also discuss the key experiences and lessons learned from the whole course of the effort, which could be useful to the developments of other industrial AIoT systems.

The remainder of this paper is organized as follows. §II reviews related work. §III presents the background about IET and overviews our AIoT system. §IV, §V, and §VI present the designs of ML-based profile classifiers, heuristic-based AD approach, and smart camera, respectively. §VII presents deployment of our system and evaluation results. §VIII discusses the experiences and learned lessons. §IX concludes this paper.

II. Related Work

Challenges in deploying ML and AIoT in industries: Industrial AIoT is the combination of AI and industrial IoT to improve the level of automation in analyzing and creating useful insights from the industrial sensor data [2]. Deploying an industrial AIoT system often faces challenges of making decision on the design and implementation of IoT hardware infrastructures (e.g., edge, fog, and cloud) and software components (e.g., ML models) based on the specific objectives and practical constraints of the industrial processes. A number of studies [3]–[7] have investigated practical challenges and provided some insights on deploying industrial AIoT systems. Alkhabbas *et al.* [3] conduct a survey that distributes a questionnaire containing 14 questions about the deployment decisions of IoT systems. Their findings based on the responses of 66 IoT system designers from 18 countries show that the reliability, performance, security, and cost are the four main factors affecting the designer's decisions on deploying IoT systems. The studies [4]–[7] discuss practical challenges and lessons learned from deploying ML algorithms for various applications. For instance, with experiences from building data analytics platforms at Twitter, Lin and Ryaboy [4] observe that at the first step, the data scientists often spend a lot of efforts in understanding and cleansing the collected data before they can develop ML algorithms. Budd *et al.* [5] identify that the lacking of training data labels is a key challenge of developing ML algorithms for medical image analysis. As presented in [6], practical ML systems often employ simple ML models such as random forests, decision trees, and shallow neural networks to shorten the deployment time and gain better interpretability. For instance, Haldar *et al.* [6] report that in the process of applying deep ML models for AirBnB search, after several failed attempts with complex neural networks, they finally deployed a simple neural network that simplifies the deployment process while providing reasonably good performance. In addition, Hazelwood *et al.* [7] discuss several key factors that drive the decisions on designing ML models for data center infrastructures at Facebook. Similar to the above studies, in this paper, we present our experiences and lessons learned from the design and implementation of an industrial AIoT system. As our work considers different specific objectives, operational procedures, and practical constraints, this paper will provide new insights.

QC in production processes: QC is a set of procedures for determining whether a product meets a predefined set of quality criteria or the customer's requirements [8]. It also provides the information to determine the need for corrective actions in the manufacturing process. AIoT technologies have been

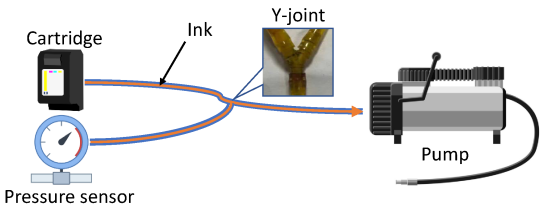


Fig. 1. Illustration of testing an ink cartridge in IET machines.

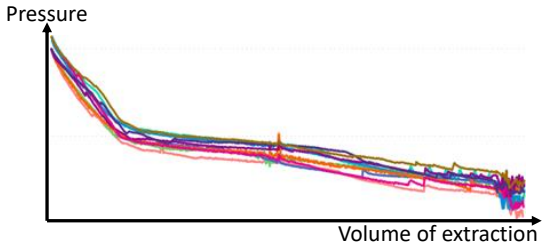


Fig. 2. Measured profiles of an ink cartridge model.

adopted to improve QC of manufacturing lines. For instance, at Siemens' electronics plant in Amberg, Germany [9], various ML models and edge computing are used to design a predictive model-based QC framework for testing the quality of printed circuit boards (PCBs). The framework helps improve the recall in detecting defective PCBs and reduce testing overheads. In this paper, we present our effort in developing an industrial AIoT system for improving the QC of the ink cartridge manufacturing lines at the HP's factories.

III. Background, Motivation, & System Overview

In this section, we present the background of the ink extraction testing (IET) and discuss its current problems in practice. Then, we overview the design of our AIoT system for improving the IET process.

A. IET Background and Problem Statement

As discussed in §I, the IET is the final QC process of the ink cartridge manufacturing. Specifically, a number of randomly selected ink cartridge samples are tested using the tester. The tester can run six ink cartridges simultaneously. Fig. 1 illustrates how the tubes connect a tested ink cartridge, a stepper motor pump, and a pressure sensor. A transparent plastic Y-joint is used to join the tubes. A workstation computer of the tester controls the stepper motor pump to extract ink from the ink cartridge at a steady volume rate for a certain time duration. Meanwhile, a liquid pressure sensor continuously measures the pressure in the tube and reports the readings to the workstation computer. The resulting curve of the measured liquid pressure versus the volume of the extracted ink is a profile of the tested ink cartridge. The ink cartridges of different models have distinct profiles. Fig. 2 shows profile samples of an ink cartridge model.

The tester adopts a *bound-based detector* to assess a measured profile against a *template profile* with an upper bound and a lower bound The template profile is defined based on the specification of the ink cartridge. The bound-based detector classifies a profile *normal* if the profile completely lies within the belt area between the two bounds; otherwise, the tester classifies the profile *abnormal*. To achieve high recall in capturing defective cartridges, the factories' current practice is to impose stringent bounds. As a result, the tester generates alarms frequently. As mentioned in §I, many alarms are actually false. This is because that the liquid pressure measurements can be noisy and biased.

Specifically, the pressure sensing is subject to both endogenous and exogenous noises. Endogenous noises are mainly from the thermal noises of the pressure sensor and the random control errors of the stepper motor pump. Exogenous noises are mainly caused by vibrations and blockage of the ink tubes. The vibration is caused by the movements of nearby human operators and bulky manufacturing machines, while the blockage is caused by the hardening ink residue trap within the tube. In addition, the tester is subject to the following biases. An improper manual insertion of the tested ink cartridge onto the tester may cause loss of back pressure of the cartridge and deviation from the template profile. An air bubble formed in the tester's ink tubes with a sufficiently large volume can also affect the pressure sensing.

In the current protocol of the factories, the alarm-triggering profiles will be further classified manually by the technicians into false positives (i.e., tester-induced) and true positives (i.e., product-induced). The manual classifications are based on the technicians' knowledge received during training and also their own experiences. As such, the classification results may lack high confidence and consistency. To ensure that there is no doubt regarding the QC result of a tested batch, the technicians may need to perform maintenance of the tester and conduct destructive tests with additional samples. A common maintenance performed is to flush the tubes with water to purge out ink and air bubbles at the end of every test. However, the frequent maintenance reduces the IET throughput significantly; the additional destructive tests increase the cost. Therefore, it is desirable to develop a system that can reliably and consistently classify the alarms generated by the bound-based detector, such that all or part of the unnecessary tester maintenance and additional destructive tests can be avoided.

B. AIoT System Overview

In this work, we follow the *progressive system development methodology* to design and implement an AIoT system to replace the factories' current practice of manually classifying the alarm-triggering profiles into normal and abnormal profiles. During the whole course of designing our AIoT system, we have developed three main components as follows.

(1) ML-based profile classifiers: We design and train several ML-based classifiers to classify the profiles. The training processes are based on historical profiles labeled by the product engineers. Specifically, we design multiple classifiers based on supervised, semi-supervised, and unsupervised ML models. Each classifier takes different features as input to classify a profile. Ensemble methods are also used to integrate the results of the multiple classifiers.

(2) Heuristic-based anomaly detection: The ML-based classifiers face challenges of limited and imbalanced training dataset. Thus, we also develop a heuristic approach which considers the profile classification as an anomaly detection (AD) problem. The profiles of good ink cartridges, albeit measured in the presence of noises and biases, should be detected normal; the profiles of defective cartridges should be detected abnormal.

(3) Smart camera: From the technicians' experiences, formation of an air bubble at the Y-joint of the ink tubes can affect the pressure measurement, which likely leads to false alarms. We design a smart camera system to monitor the Y-joint. It runs a CNN to detect air bubble and a computer vision algorithm to estimate the volume of the bubbles. The results are used to assist the profile classifier or the AD algorithm in deciding the nature of any alarm generated by the tester.

All computing for the profile classification and bubble detection is executed on a Raspberry Pi single-board computer which is deployed close to the sensors generating data. Specifically, the Pi is connected directly with the camera and tester to receive the captured images and measured pressure profiles. The designs of the above three system components are presented in §IV, §V, and §VI, respectively. Their performance will be evaluated via field experiments as presented in §VII.

IV. ML-based Pressure Profile Classifiers

A. Preparation of Design Data

We receive a dataset containing 550,508 pressure profiles of 723 ink cartridge models collected from the testers deployed in HP's factories in 18 months. The dataset includes the profile labels which are generated by the tester using the bound-based detector. Specifically, the bound-based detector classifies about 2% of profiles abnormal. However, the actual defect rate of the manufactured ink cartridges is about 0.07% only. This result suggests that most abnormal profile labels generated by the bound-based detector are inaccurate. We work with HP's product engineers and domain experts to manually relabel the abnormal profiles in the dataset. However, the relabeling is tedious and extremely time-consuming. We can only confirm 134 abnormal profiles. Eventually, we have a dataset consisting of about 530,000 profiles with reliable "normal" labels, merely 134 profiles with reliable "abnormal" labels, and about 110,000 profiles that were classified abnormal by the bound-based detector but unlabeled after the relabeling process. This renders the training dataset imbalanced with limited data with abnormal labels. The difficulty of the labeling process will be further discussed in §VIII.

B. Design of ML-based Classifiers

As discussed in §I, each ML approach addresses a specific problem based on a set of assumptions, but real-world tasks often face mixes of many problems. In practice, it is often more efficient to try multiple ML approaches than relying on a single approach unless we clearly know that the conditions of the task well match the assumptions of the single approach. As such, we have tried two unsupervised, one semi-supervised, and one unsupervised ML approaches, which are presented below. In addition, as an ensemble of multiple ML-based classifiers is more accurate than any of the single classifier [10], in the performance evaluation, we also try the ensembles of the four classifiers with distinct combination rules.

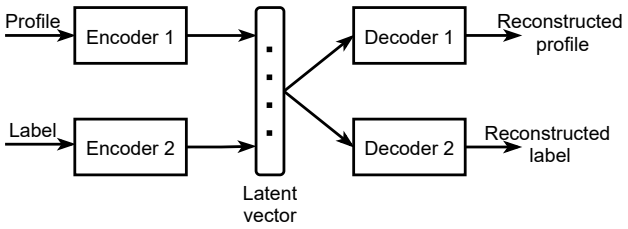
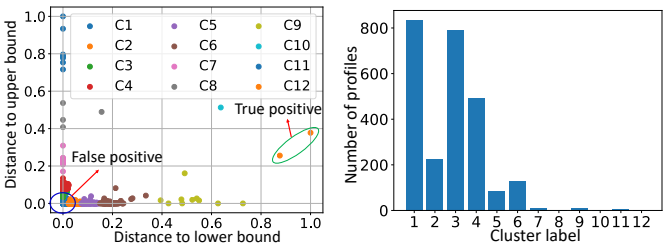


Fig. 3. Architecture of the MVAE-based profile classifier.

Supervised ML-based classifiers: Inspired by the success of the CNNs in many classification tasks, we first build a CNN-based profile classifier. We normalize the pressure measurements of the profile into $[-1, 1]$, and feed them into a CNN consisting of an input layer, a convolution layer, four fully-connected (FC) layers, and an output layer. The rectified linear units (ReLUs) are used as the activation function for convolution and FC layers, while the softmax activation is used at the output layer. Additionally, we build a decision tree (DT)-based profile classifier which is widely adopted in manufacturing applications due to its good interpretability [11]. From the prior domain knowledge, the normal profile curve can be divided into three phases: early, middle, and last phases according to the volume of extraction. In the early phase, the pressure often sharply drops. Then, the pressure remains flat and stable in the middle phase. At last, the pressure further drops. Thus, we considered variability of the pressure measurement in these three phases as inputs for the DT-based classifier. Specifically, we implement a change point detection algorithm [12] to divide the profile curve into three phases. The change point, minimum, maximum, median, mean, and slope rate of the pressure measurements in the three phases are fed to the DT to predict the profile label. The CNN and DT are trained using the relabeled training dataset.

Semi-supervised ML-based classifier: To use both unlabeled and labeled profiles, we develop a profile classifier based on a semi-supervised learning ML model, called multimodal variational autoencoder (MVAE) [13]. Fig. 3 shows the architecture of our MVAE-based classifier which contains two encoders and two decoders. The MVAE is trained using labeled and unlabeled profiles as follows. For the labeled profiles, Encoder 1 and Encoder 2 take profile and its label, respectively, to jointly generate a latent vector. Then, Decoder 1 and Decoder 2 use the latent vector to reconstruct the original profile and label, respectively. The classifier is trained to minimize the differences between the original and reconstructed profiles/labels. For the unlabelled profiles, Encoder 1 and Decoder 1 are trained using the pressure profiles only. Upon a new profile, the trained MVAE takes the profile as input to predict the profile label at the output of Decoder 2.

Unsupervised ML-based classifier: We develop an unsupervised classifier that can leverage a large number of normal



(a) Scatter plot of distances (b) Distribution of profiles

Fig. 4. k-means clustering results.

profiles classified by the bound-based detector. In particular, we develop an algorithm to determine the lower and upper bounds that the normal profiles lie within. For each abnormal profile classified by the bound-based detector, we compute pairwise distances between itself and the upper and lower bounds. The profile distances are fed into the k-means algorithm to group the abnormal profiles into k clusters. Fig. 4(a) shows the scatter plot of the distances of 3,159 randomly selected abnormal profiles to the upper/lower bounds, which are divided into $k = 12$ clusters by the k-means algorithm. Fig. 4(b) presents the distribution of the abnormal profiles among the 12 clusters. The clusters are ordered based on the ℓ_2 norm of the two distances to the upper and lower bounds. Then, we classify the profiles in the clusters with the cluster label $k \leq k_{th}$ as normal and those with $k > k_{th}$ as abnormal.

V. Heuristic-based Anomaly Detection

From the experiences of evaluating the ML-based classifiers (cf. §VII), the imbalanced training dataset and limited training samples pose substantial challenges for the classifiers to achieve high accuracy. In general, ML techniques such as resampling [14] and few-shot learning [15] can be used to mitigate these problems. However, such techniques cannot completely address our issues with the ML-based classifiers. For instance, the resampling can be used to create a more balanced dataset. However, it cannot help expand the training data distribution to cover unobserved/unlabelled abnormal profile samples. Moreover, the few-shot learning can build accurate ML models with limited training samples based on prior knowledge about the data structure and learning process. Meanwhile, we have limited knowledge about dynamics of the pressure-volume profiles. Thus, we develop a heuristic approach which treats the profile classification as an AD problem. Specifically, our approach considers the abnormal profiles as outliers which do not follow the expected pattern of the normal profiles. Upon a new profile, a distance-based similarity score between itself and the normal profiles is calculated. The profile is considered abnormal if the score is lower than a threshold. This AD approach provides good interpretability in that it gives information for understanding the classification results. In this section, we present four categories of false alarms and then describe the AD approach.

A. Categories of Alarm-Triggering Normal Profiles

As mentioned in §III, the liquid pressure measurements are subject to various biases due to the human operators and the tester deviations. The biases can cause different patterns of the normal profiles that trigger the bound-based detector. From the product engineers' domain knowledge and experiences, the normal profiles can be divided into four categories as follows.

Miss-configuration profiles are caused by setting a wrong reference point by the human operator at the beginning of the test. With the wrong reference point, the measured profiles have a similar pattern to the profiles of good ink cartridges. However, they are shifted beyond the belt area between the two bounds of the template profile which is used by the tester to classify the profiles into normal and abnormal. As a result, these miss-configuration profiles trigger false alarms.

Miss-calibration profiles are caused by configuring a wrong gain to scale the sensor's raw readings to the pressure unit in the calibration process of the pressure sensor.

No-cartridge profiles are measured when the ink cartridges are not inserted properly onto the tester. Without the ink from the cartridge, the motor pump of the tester pulls the air through the tube only. Under this condition, the measured pressure profile is nearly a flat line.

Tube-blocking profiles are measured when the ink tubes are blocked by air bubbles or ink residue. Specifically, the tube-blocking profiles have a liquid pressure drop in the early stage of the extraction due to presence of the air bubbles inside the tube. Then, they quickly increase and recover to the pattern which is similar to a shift-up variation of the normal profile.

B. Anomaly Detection (AD)

From the technician's experiences, the last phase of the profiles often includes the pressure measurement fluctuations caused by over-extraction in which the tester's motor pump still operates when the internal valve of the ink cartridge is already closed. The air gaps traveling through the tube introduce measurement fluctuations that can trigger the bound-based detector. First, our AD algorithm excludes such fluctuations from the input profile. Our experiments in §VII show that the over-extraction has a strong correlation with the presence of air bubble in the tube. Thus, we use air bubble as an indicator to determine whether the measurement fluctuations are caused by over-extraction. Then, we apply data analytics methods to extract the features of the normal profiles that are used to distinguish the abnormal profiles as outliers. Specifically, we check whether a testing profile belongs to any of the four categories presented in §V-A. If yes, it is normal; otherwise, it is abnormal. The checking is as follows.

For the miss-configuration, no-cartridge, and tube-blocking categories, we used the mean subtraction method to normalize the original profile by subtracting its pressure measurements from its average. Dynamic time warping (DTW) distances [16] between all pairs of normalized training profiles in the normal profile category i are calculated. We define γ_i as the detection threshold for category i and $\gamma_i = \mu + 3\sigma$, where μ and σ are mean and variance of calculated DTW distances. Upon

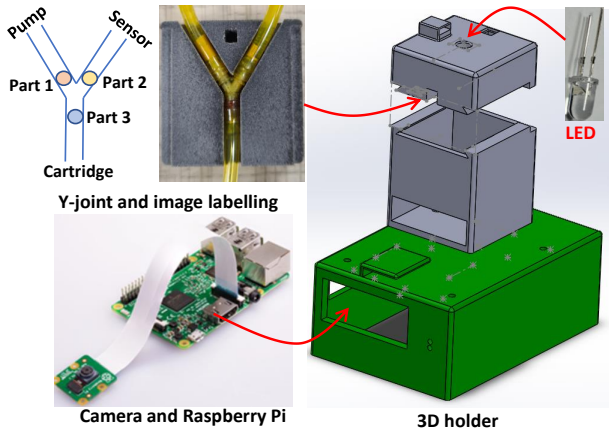


Fig. 5. Smart camera fixed into a 3D-printed holder for Y-joint monitoring.

a new profile, we first calculate the DTW distance between itself with all training profiles of the category i. If the mean of calculated distances is less than γ_i, the profile is considered normal in the category i.

For the miss-calibration category, we use a scale-matching method to extract profile features. Each training profile is equally divided into 10 segments and the maximum among the pressure measurements of each segment is determined. The mean and variance of the maximum over the same segment across all training profiles are calculated. For a new profile, we first determine the maximum of its 10 segments, and then compute their scale with respect to the mean and variance obtained from the training profiles. The profile is considered normal if all scales of its 10 segments fall within a suitable range between each other. If the profile is considered normal by the above scale-matching approach, we additionally perform the DTW distance-based AD process to confirm whether the profile is normal.

VI. Smart Camera System

As mentioned earlier, the presence of the air bubbles inside the tester's tube can affect the pressure sensing and is indicative of over-extraction. Thus, we design and deploy a smart camera with an embedded image processing pipeline to monitor the air bubbles during the ink extraction.

A. Hardware Components

Fig. 5 illustrates our camera system that consists of three main components: the low-cost camera, the edge node, and the light source. For the camera, we select the Raspberry Pi camera module which can capture up to 90 images per second. The captured images are transferred to a Raspberry Pi 4 edge node that runs the CNN and traditional computer vision (CV) algorithms. An external light source is used to illuminate the ink tube for the camera. To reduce the impact of the tube's vibration on the camera's image sensing, all hardware components and Y-joint are fixed into a custom 3D-printed holder as shown in Fig. 5. We deploy the camera system to monitor the air bubbles at the Y-joint of the tube since the air bubbles are often trapped by the Y-joint.

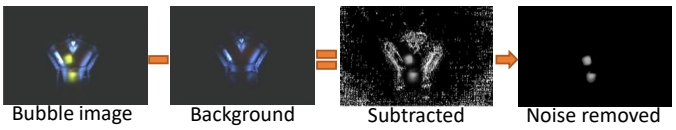


Fig. 6. Workflow of CV-based air bubble size measuring.

B. Image Processing

We implement a two-step image processing pipeline to process the captured images at the Raspberry Pi. First, the image is fed to a CNN to detect the air bubbles in the Y-joint. Specifically, each image is characterized by three labels that indicate the presence of the bubbles in three tube channels of the Y-joint as shown in Fig. 5. To train the designed CNN, we collected and manually labeled an dataset of 1,494 and 1,455 images with and without the bubbles, respectively. Different from relabeling the pressure profiles, this labeling process is easy because human can easily recognize the bubbles.

Second, we develop a CV-based framework to determine the size of detected bubbles as shown in Fig. 6. In particular, a previously captured image without air bubble is used as the background. Upon a new image with bubbles, a background subtraction method is used to extract the bubble areas by subtracting the image from the background. Then, the morphological processing is adopted to remove the noises from the extracted bubble areas. Finally, the number of points with the pixel value greater than zero is yielded as the size of the air bubble. The background is updated once a new image without the bubbles is captured.

C. Usages of the Smart Camera

We use the camera system to reduce the maintenance overheads and improve the classification of the ML-based classifiers or the heuristic-based AD. First, it provides an indicator to determine whether the bubbles are completely removed after performing a water flushing round. As mentioned earlier, the current protocol of the factories performs water flushing to purge out ink and bubbles at the end of every test. This process is labor intensive and usually requires a number of attempts. Thus, to reduce the flushing overheads, the camera system can be used to check whether the air bubbles are completely removed from the tube. Once the tube is clear without bubbles, the flushing process can be stopped. Second, the bubble detection and size measurement functions can be used to avoid the measurement fluctuations during the over-extraction period. Specifically, in the last phase of the tests, we stop the pressure measurement when a bubble with a certain size is detected. The presence of the bubble is also used as an indicator to determine and exclude the over-extraction period.

VII. Deployment and Evaluation Experiments

A. Deployment

We deploy our AIoT system to an operational tester in an HP factory. Specifically, we use Python and several ML libraries including PyTorch, TensorFlow Lite, and Scikit-Learn to implement the ML-based classifiers and AD module running

TABLE I
ACCURACY (IN %) OF ML-BASED CLASSIFIERS OVER 88 PROFILES COLLECTED FROM CONTROLLED EXPERIMENTS.

Metrics	Classifiers				Ensemble	
	CNN	DT	MVAE	k-means	Veto	Majority
Accuracy	34	51.1	42	39.7	64.7	40.9
Abnormal recall	7.9	44.4	38	15.8	82.5	17.4
Abnormal precision	100	77.7	66.6	100	72.2	100
Normal recall	100	68	52	100	20	100
Normal precision	30.1	32.6	25	32	31.2	32.4

on one Raspberry Pi 4. At the end of each testing round, the tester reports the measured profiles of six tested cartridges to the workstation computer. The profiles triggering alarms are then transferred to the Pi for further classification into normal (i.e., the tester-induced alarm) or abnormal (i.e., the product-induced alarm) profiles. We also deploy six units of the smart cameras to monitor the bubbles at the Y-joints of six tubes connected to six testing modules. The camera periodically captures an image of the Y-joint and transfers it to the Pi at every two seconds during the testing period.

We first evaluate the performance of our system in controlled experiments, in which we induce biases and noises to the tester to generate normal profiles, and defects to good ink cartridges to create abnormal profiles. This section presents the results of the controlled experiments. We are now working with our industry partner to run long-term evaluation on the cartridge manufacturing lines. We plan to present the long-term evaluation results in a longer version of this paper that will be made publicly available.

B. Accuracy of Profile Classification

We perform a set of controlled experiments to evaluate the accuracy of our ML-based classifiers and AD module. We intentionally induce the tester's biases and noises to generate the normal profiles of four categories (cf. §V). Specifically, we create seven miss-configuration profiles by setting an arbitrary reference point in the beginning of tests for seven good ink cartridges. Eight miss-calibration profiles are created by setting a wrong gain parameter to scale the pressure sensor's raw readings to the pressure unit. We also generate six no-cartridge profiles by inserting the ink cartridges improperly such that no ink is extracted under the pressure from the pumps. Moreover, we induce bubbles and ink residue inside the tubes to create four tube-blocking profiles. In summary, we create 25 normal profiles that trigger false alarms. Additionally, we manually induce defects to good ink cartridges by damaging the vent of the cartridges or releasing the pressure into the cartridge to create 15 abnormal profiles. In addition, we run tests for 48 defective cartridges and generate abnormal profiles. As a result, we have 63 abnormal profiles. In summary, our controlled experiments generate a total of 88 pressure profiles whose labels are also confirmed by the domain experts.

We use the overall classification accuracy, recall, and precision in detecting the normal and abnormal profiles as the evaluation metrics. Table I shows the evaluation metrics of four ML-based classifiers over 88 profiles. For the k-means-based classifier, we adopt the settings of $k = 12$ and $k_{th} = 3$. We also evaluate two ensemble approaches including *veto* and *majority*, which combine the results of four ML-based classifiers to yield the final result. Specifically, with a primary focus on achieving high recall in capturing defective cartridges, the veto approach considers the profile as abnormal if any of four classifiers outputs abnormal. The majority approach yields the majority of the classifiers' results as the final result. From Table I, the four classifiers (i.e., CNN, DT, MVAE, and k-means) show best performance in different metrics. For instance, DT has the highest accuracy and abnormal recall, while CNN and k-means exhibit the best abnormal precision, and normal recall. Moreover, two ensemble approaches mostly show better accuracy performance. The veto approach has the highest accuracy and abnormal recall.

Table II shows the performance of the AD module. The columns headed by miss-configuration, miss-calibration, no-cartridge, and tube-blocking present evaluation metrics of the AD module in detecting 88 profiles by comparing its similarity score with the normal profiles in each of four category only. The overall column shows the performance results when the scores between the testing profile and the normal profiles in all four categories are used. The AD approach achieves an overall accuracy of 96.5% in classifying the testing profiles. Moreover, it always has better accuracy performance, compared with that of the best-performing ML-based classifier, i.e., the veto.

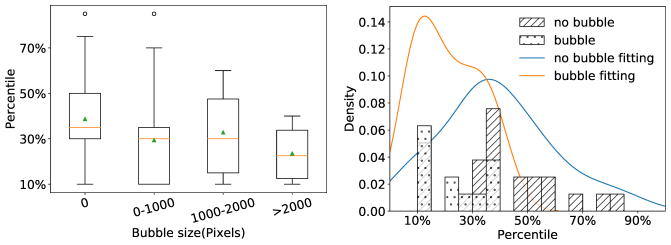
C. Performance of Camera System

1) Accuracy of bubble detection and size measurement: We use 450 captured images in the controlled experiments to evaluate the accuracy of bubble detection by the camera system. The CNN can detect the air bubbles in 450 testing images with an accuracy of 94%. It cannot detect small air bubble in co-presence of the diluted ink inside the Y-joint. However, the small air bubbles generate little/no impact on the pressure measurements. Moreover, we use 49 images with the air bubbles to evaluate the accuracy of the size measurement by the CV method. We adopt the intersection over union (IoU) as the evaluation metric. In particular, for each image, we calculate the IoU between the detected bubble areas and the ground truth of the bubble areas. The bubble size measurement is considered correct if the calculated IoU is higher than 0.5. Our CV method achieves an accuracy of 79.5% in measuring the sizes of the air bubbles in 49 testing images.

2) Impact of air bubble on pressure measurement: We use our camera system to capture the top view of the Y-joint at the beginning of the ink extraction for 81 ink cartridges of 6 models over a 7-day operation period of the tester. We perform an analysis on the captured images and the corresponding profiles to study how the bubbles affect pressure measurements. Specifically, we cannot directly compare the 81 pressure profiles with and without bubbles since the profiles of different cartridge models fall in different measurement ranges. Thus, we compare the average of testing profiles with that of profiles of the same cartridge model in our historical dataset.

TABLE II
ACCURACY OF HEURISTIC-BASED AD OVER 88 PROFILES COLLECTED FROM CONTROLLED EXPERIMENTS.

Metrics	Anomaly Detection				Overall	Veto
	Miss-configuration	Miss-calibration	No-cartridge	Tube-blocking		
Accuracy	95.7%	95.7%	100%	100%	96.5%	64.7%
Abnormal recall	95.2%	95.2%	100%	100%	95.2%	82.5%
Abnormal precision	100%	100%	100%	100%	100%	72.2%
Normal/Category recall	100%	100%	100%	100%	100%	20%
Normal/Category precision	70%	72.7%	100%	100%	89.2%	31.2%



(a) Distribution of percentiles. (b) Fitted distributions.

Fig. 7. Impact of air bubbles on pressure measurements. The percentile represents the percentage of historical profiles whose average is lower than the average of the testing profile. In (a), the box, line, triangle, upper and lower whiskers represent middle 50%, median, average, ranges for the bottom 25% and the top 25% of the samples, respectively.

We use the percentage (i.e., percentile) of historical profiles whose average over time is lower than that of the testing profile to characterize the testing profile. Fig. 7(a) shows the box plots of the percentiles of 81 testing profiles which are divided into three groups based the measured bubble size. The percentiles of the profiles with the bubble size lower than 2,000 pixels have similar average and median. Meanwhile, when the bubble size is greater than 2,000 pixels, the profile percentiles fluctuate in narrower ranges and have lower average. To further investigate the impact of the bubble size on the distribution of the profile percentile, we fit two probabilistic distributions to model the percentiles of the profiles without the bubbles and with the bubble size greater than 2,000 pixels. Fig. 7(b) shows the histograms of the percentiles and the fitted density functions. We can see that the mean percentile of profiles with bubbles is lower than that of the profiles without bubbles. We also conduct a one-sided Kolmogorov–Smirnov test using testing profiles to check the null hypothesis that the percentile of profile with the bubbles of the size greater than 2,000 pixels is higher than that of the profiles without the bubbles. We obtain a p-value of 0.0273. Thus, the null hypothesis can be rejected. This result implies that the bubbles with a large size make the pressure measurements statistically lower.

3) Correlation between the presences of air bubble and over-extraction pressure fluctuation: As mentioned in §V, the measured pressure often has fluctuations during the over-extraction period. These fluctuations should be excluded from the profiles for better classification performance. However, it is non-trivial to determine the starting point of the fluctuations in the presence of measurement noises. From prior observations, the over-extraction often coincides with bubbles in the tubes. Now, we analyze the Pearson correlation between the bubble presences and the over-extraction fluctuations. We collect a dataset consisting of 17 profiles and five profiles with and without over-extraction, respectively. An image of Y-joint is captured for each profile. The Pearson correlation is 0.7483 over 22 collected data points. This result implies the strong correlation between presences of the bubbles and the over-extraction fluctuation. Therefore, our AIoT system uses bubble presence to assist the determination of the presence of over-extraction fluctuation.

VIII. EXPERIENCES AND LEARNED LESSONS

As a systematic attempt of developing an industrial AIoT system for improving the QC of ink cartridge manufacturing, our research has generated experiences and learned lessons that the future industrial practices can consider. The experiences and lessons are summarized as follows.

(1) Classifiers vs. heuristics: In the early stage of our system development, we considered the problem of dividing the profiles into normal and abnormal classes as a classification problem. However, the four ML-based classifiers cannot achieve a high accuracy in the deployment. A main reason is the limited and imbalanced training dataset, which is also related to the second challenge that we will discuss shortly. Then, we investigated the characteristics of the normal and abnormal profiles. Specifically, the tester often induces stable biases and noises to the pressure measurement of all tested ink cartridges over a certain period of time. The profiles of defective ink cartridges are rare ones which do not follow the pattern of the profile of good cartridges under the tester-induced noises and biases. Thus, we further designed a heuristic approach that considers the profile classification as an AD problem. Our evaluation results based on the controlled tests shows that the AD approach outperforms the ML-based profile classifiers. From our experience, the quality of the training data is crucial to the development of effective ML classifiers. It is often very difficult to achieve satisfactory performance if the data is limited or include high-variance noises and biases. In such cases, simpler, heuristic solutions (e.g., AD approach in our case) can be more effective.

(2) Curse from data labeling: ML classifier's attractive advances recently are mainly owing to availability of big labeled training data and standardized hardware acceleration. For the tasks that humans are good at, creating big labeled training datasets is feasible. Manual labeling services (e.g.,

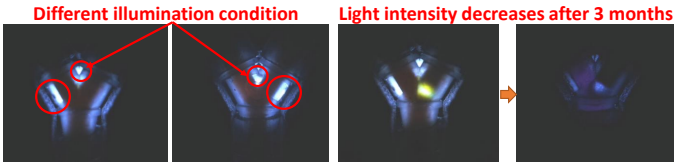


Fig. 8. Impact of sensor condition on data quality.

Google's [17]) are now established. However, data labeling is very challenging for developing an industrial AIoT system. Such labeling processes cannot be performed by normal persons based on their instinct and/or basic knowledge. Differently, they require experts' experience and prior knowledge. In our work, relabeling the pressure profiles is highly non-trivial and requires a collaboration with the tester domain experts. In particular, the experts sometimes lack high confidence and consistency for assigning labels for high-variance profiles. This can be solved if they can access meta information about the internals of the tested ink cartridges and tester's parameters. However, this meta information was not collected in the historical database. Even if the meta information is available, frequently referring to the detailed meta information inevitably adds overhead to the relabeling process. Eventually, we can only relabel a limited number of profile samples, which lead to the poor performance of our ML-based profile classifiers. The use of ML classifier in our AIoT system is limited to the bubble detection, which is a task that a normal human can complete after receiving some simple guidance. From this experience, it is reasonable to argue that the success of applying ML classification to an industrial task highly depends on the availability of sufficient labeled data.

(3) System challenges: Sensor inconsistency and deviation pose challenges for the deployment of industrial AIoT systems in practices. In our system, we use a camera to capture images to train the CNN for detecting the air bubbles. A light source was used to provide a stable and sufficient illumination for the camera to capture the training images. Then, the trained CNN was deployed to six sets of cameras. However, the trained CNN did not show the same performance on them. This is because the quality of captured images across six cameras are different due to the deviation in installation and working condition of the cameras and light sources. Fig. 8(a) shows two images captured by two camera sets. We can see that they have different illumination conditions, which affect the performance of the CNN. Moreover, the illumination condition of a certain camera can drift over time due to wear and tear of the light source. Fig. 8(b) presents two images captured by the same camera set at the beginning of the deployment and three months later. The light intensity of the light source is weakened. As a result, the CNN cannot correctly detect the air bubbles in the images captured with weakened lighting conditions. Although the dimming was caused by that the light was kept on all the time, which was then replaced with on-demand switch-on, the long-term wear and tear are inevitable. This calls for new research to obviate negative impacts of sensor inconsistency and deviation on performance of AIoT systems. The method proposed in [18] may be promising to address the issues. Specifically, we can model the relationship between the images captured by different cameras or under different controlled illumination levels. Then, we can use the modeled relationship to augment the training dataset. As such, the trained CNN can have the capability to deal with different cameras and illumination levels.

IX. Conclusion

This paper presented the design, deployment, and evaluation of an industrial AIoT system for improving the quality control of Hewlett-Packard's ink cartridge manufacturing lines. Specifically, the evaluation results showed that our AIoT system can help improve the accuracy of the HP's testers in detecting defective ink cartridges. The lessons learned and experiences discussed in this paper can be useful to the developments of other industrial AIoT systems.

Acknowledgment

We thank Chou Po-Yi, HP product engineer, for providing expert knowledge on ink cartridge and the operation management team for supporting us to conduct tests and evaluation at an HP manufacturing facility.

References

[1] I. Gosh, "AIoT: When artificial intelligence meets the internet of things," August 12, 2020. [Online]. Available: https://bit.ly/3aMVmRb
[2] J. MSV, "Why AIoT is emerging as the future of industry 4.0," August 12, 2019,. [Online]. Available: https://bit.ly/3rJLtuB
[3] F. Alkhabbas, R. Spalazzese, M. Cerioli, M. Leotta, and G. Reggio, "On the deployment of IoT systems: An industrial survey," in *ICSA-C*, 2020.
[4] J. Lin and D. Ryaboy, "Scaling big data mining infrastructure: the twitter experience," *ACM SIGKDD Explorations Newsletter*, vol. 14, 2013.
[5] S. Budd, E. C. Robinson, and B. Kainz, "A survey on active learning and human-in-the-loop deep learning for medical image analysis," *arXiv:1910.02923*, 2019.
[6] M. Haldar, M. Abdool, P. Ramanathan, T. Xu, S. Yang, H. Duan, Q. Zhang, N. Barrow-Williams, B. C. Turnbull, B. M. Collins *et al.*, "Applying deep learning to airbnb search," in *ACM KDD*, 2019.
[7] K. Hazelwood, S. Bird, D. Brooks, S. Chintala, U. Diril, D. Dzhulgakov, M. Fawzy, B. Jia, Y. Jia, A. Kalro *et al.*, "Applied machine learning at facebook: A datacenter infrastructure perspective," in *IEEE HPCA*, 2018.
[8] G. P. Staff, "Quality Control In Manufacturing," January 01, 2021. [Online]. Available: https://bit.ly/2MjlY2J
[9] J. Schmitt, J. Bönig, T. Borggräfe, G. Beitinger, and J. Deuse, "Predictive model-based quality inspection using machine learning and edge cloud computing," *Advanced Engineering Informatics*, vol. 45, 2020.
[10] D. Opitz and R. Maclin, "Popular ensemble methods: An empirical study," *J. Artif. Intell. Res.*, vol. 11, pp. 169–198, 1999.
[11] E. Coopersmith, G. Dean, J. McVean, and E. Storaune, "Making decisions in the oil and gas industry," *Oilfield review*, 2000.
[12] G. J. Van Den Burg and C. K. Williams, "An evaluation of change point detection algorithms," *arXiv:2003.06222*, 2020.
[13] M. Wu and N. Goodman, "Multimodal generative models for scalable weakly-supervised learning," in *32nd NIPS*, 2018.
[14] A. More, "Survey of resampling techniques for improving classification performance in unbalanced datasets," arXiv:1608.06048, 2016.
[15] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Comput. Surv.*, 2020.
[16] X. Cai, T. Xu, J. Yi, J. Huang, and S. Rajasekaran, "DTWNet: a dynamic time warping network," *NeurIPS*, vol. 32, 2019.
[17] https://cloud.google.com/vision/automl/docs/human-labeling.
[18] W. Luo, Z. Yan, Q. Song, and R. Tan, "Phyaug: Physics-directed data augmentation for deep sensing model transfer in cyber-physical systems," in *IPSN*, 2021.