Sensors and Materials, Vol. 36, No. 11 (2024) 4781–4799
MYU Tokyo

S & M 3831

Applying the Generative Model Integrated with the Diffusion Technique to Improve Virtual Sample Generation in Environmental Sound Classification

Yao-San Lin and Mei-Ling Huang*

Department of Industrial Engineering and Management, National Chin-Yi University of Technology,
No. 57, Sec. 2, Zhongshan Rd., Taiping, Taichung 411, Taiwan (R.O.C.)

(Received June 24, 2024; accepted October 23, 2024)

Keywords: GAN, generative model, diffusion technique, ESC, virtual sample generation

We propose a novel framework for environmental sound classification (ESC) to address the challenge of insufficient training samples in sound recognition systems for manufacturing environments. Because of sample scarcity, traditional systems often perform poorly, so in this research, we utilize generative adversarial networks (GANs) to generate virtual sound samples and augment existing datasets. The proposed method integrates a robust Bayesian inference approach with a modified GAN architecture to generate high-quality synthetic samples, particularly for rare events and emergencies on production lines. The framework aims to enhance the stability and performance of ESC systems by expanding training data in a controlled manner. Experimental results demonstrate the potential of this approach to reduce sample collection costs and improve the practical application of ESC technology in manufacturing systems. Key aspects discussed include technological innovation, cost-effectiveness, implementation challenges, and ethical considerations related to synthetic audio data generation. The results of this research will advance ESC's real-time monitoring and anomaly detection capabilities in diverse manufacturing environments.

1. Introduction

To achieve a smart factory in the fourth industrial revolution manufacturing process, it is necessary to transform all equipment into innovative equipment and connect to a centralized system for real-time information exchange. Sound can be an effective means of making a device smart because of its ability to contain status information from various devices and its ease of recording using only a microphone. The sound emitted by the machine on the production line often directly reflects its operating status. This reflection may include a specific sound pattern produced by the regular operation of the machine, or it may reveal a precursor to an imminent malfunction or malfunction of the machine. For example, a particular machine may emit a uniform and stable sound when operating normally, but once it encounters a malfunction, the frequency or amplitude of the sound may change abnormally. Through in-depth analysis of these

*Corresponding author: e-mail: huangml@ncut.edu.tw
https://doi.org/10.18494/SAM5202

ISSN 0914-4935 © MYU K.K.
https://myukk.org/

subtle sound changes, machine failures can be predicted, and maintenance can be performed in advance, thereby saving repair time and reducing production losses. Over the past few decades, the monitoring system used on industrial production lines has constantly evolved from initial manual control to more advanced automated monitoring systems. However, even with modern automated surveillance systems, some limitations remain. Traditional dashboards and monitoring systems focus on machine operation data and the production line's primary status. Still, they require a complete understanding of the actual health status of the machine and the working status of employees to be effective. As technology rapidly advances, the operation of production lines constantly evolves. Traditionally, dashboards and monitoring systems have been used to supervise production lines. While these conventional systems currently support most production lines among small and medium enterprises and provide basic operational information, they need real-time status monitoring. In the current era of information explosion, emerging technologies have increased the demand for monitoring production lines in multiple industries. Against this setting, production line sound has become a signal source with tremendous potential, which can be analyzed using advanced technology for more comprehensive and real-time control.

This study was motivated by the potential value of sound on production lines. Traditionally, production line management has mainly relied on dashboards and monitoring systems. Although these systems can provide some basic operational information, they need to better reflect the real-time status of machines and employees. We can capture more subtle and hidden intelligence through unconventional signal sources like sound. The sound emitted by the machine contains the state of mechanical movement. Every rattling and turning sound is how the machine expresses its operating state. By analyzing these sounds, managers can accurately establish models of regular machine operation and abnormal situations, thereby achieving precise control of production line operations.

In this study, we investigated the practical uses of production line sounds, specifically in detecting and responding to malfunctions or emergencies. The sounds emitted by machines on the production line can provide valuable information about their operating status. These sounds may contain distinct patterns of regular operation or indicate signs of machine malfunction or impending failure. By carefully analyzing these sounds, we can identify potential problems the machines may encounter and perform preventive maintenance to minimize production losses caused by machine failure. In modern manufacturing environments, sound monitoring systems are increasingly used to monitor equipment and predict failures in real time. For instance, in automotive production lines, where precision and uptime are critical, sound monitoring systems have been deployed to track the operational sounds of machinery continuously. These systems can detect early signs of mechanical wear, misalignment, or malfunction before they lead to equipment failure by identifying subtle changes in sound frequency, pitch, or amplitude. For example, if a machine's bearings begin to degrade, the system can recognize deviations from the typical sound profile and alert the maintenance team to address the issue proactively. Studies have shown that such predictive maintenance strategies can reduce unplanned downtime by as much as 25% and improve production efficiency by 15%.

Another prominent application is in aerospace manufacturing, where high-precision machining is essential. Sound monitoring systems are integrated into the assembly and machining processes to detect even the most minor irregularities in machine sounds. These systems enable manufacturers to perform predictive maintenance, thus preventing costly delays and ensuring the production of high-quality components. For instance, the system can detect when a cutting tool is nearing the end of its useful life by recognizing shifts in the sound profile caused by tool wear. It allows operators to replace the tool before it fails, avoiding damage to the workpiece and costly rework. In electronics manufacturing, where production environments are typically noisy and distinguishing specific sound patterns is challenging, sound monitoring systems play a key role in maintaining high product quality. By employing advanced environmental sound classification (ESC) techniques to filter out background noise and focus on the critical operational sounds of machines, these systems enable real-time monitoring of equipment performance and help reduce defect rates by the early identification of potential issues. Their ability to analyze sound anomalies in real time ensures continuous production flow and high product quality.

By applying sound monitoring systems in these diverse manufacturing scenarios, industries can optimize production processes, minimize equipment failure rates, and reduce maintenance costs. Not only is the overall production efficiency improved, but the safety and reliability of manufacturing operations are also enhanced.

On a technical level, the development of artificial intelligence and machine learning in recent years has provided new possibilities for production line sound analysis. Through deep learning technology, we can train models to identify abnormal patterns in machine sounds and quickly determine whether there are potential faults. At the same time, emotional analysis of employee voices can help us understand the atmosphere of the work site more comprehensively so as to take targeted improvement measures. Applying these technologies improves the accuracy of supervision and enhances the ability to predict and respond to abnormal situations. In addition, sound, as a noninvasive data collection method, is more convenient and cost-effective than other sensors. Without additional equipment installation and maintenance, sound monitoring systems can be more easily integrated into existing production environments, reducing implementation costs.

Advances in machine learning and deep learning technologies provide new possibilities for sound analysis on production lines. By adopting big data and deep neural networks, models can be trained to identify anomalies in different sound patterns. This brilliant analysis system can improve the accuracy of machine sounds and distinguish the emotions and context in employees' speech, thereby more comprehensively assessing the status of the production line. These technological advances make instant monitoring and rapid response more feasible. For example, once the system detects an abnormal sound coming from a particular machine, it can immediately send an alarm to notify the appropriate personnel for maintenance, thereby minimizing production interruption time. The system can also identify unusual language patterns in communication between employees, alert managers to possible problems or improvement points, and promote more effective team collaboration.

The convenience and cost-effectiveness of the sound monitoring system are also crucial advantages of this study. Compared with other sensors, such as temperature and pressure sensors, sound sensors do not require additional hardware installation, reducing implementation costs. This is a relatively low-risk investment for companies upgrading existing production lines or introducing new technologies. In addition, the nonintrusive nature of the sound sensing system also facilitates employee acceptance. Compared with some monitoring equipment that may require large-scale modifications on the production line, deploying sound monitoring systems is more straightforward, does not have much impact on existing workflows, and helps improve the scalability and sustainability of the system.

Typically, environmental sounds encompass the auditory elements present in a given setting, including natural sounds like wind, water, and birdsong, as well as artificial sounds such as traffic, machinery, and urban noise. In this study, we utilize deep learning to analyze production line sounds, focusing on identifying and addressing failures or emergencies. The sounds emitted by machines on the production line and the vocal cues and emotional expressions among workers might hold vital information regarding the state of production operations. Such information is rarely readily available, and its direct analysis may be challenging, especially if management requires insights provided by sound. In this regard, investigating infrequent emergencies and utilizing small-sample machine learning would aid in achieving a more comprehensive understanding of extreme situations that may arise on the production line.

When the amount of collected data is limited, the generative model (GM) can be used to create a virtual sample. This approach aims to address the problems that arise owing to the small size of the dataset. With the help of a GM, one can successfully overcome the obstacle of limited data. The GM generates a synthetic sample that mirrors the original data, resulting in more precise analysis and informed decision-making.

This study aims to develop a monitoring system that can effectively respond to abnormal events on the production line to improve operational efficiency and safety. To overcome the challenge of insufficient data on abnormal events, we will advance our development of small-sample machine learning technology. We focus on constructing a multi-objective environmental sound recognition model for the production line. We will begin by reconstructing the transfer learning (TL) framework with robust Bayesian inference to develop the model structure. Then, we will attempt to establish a small sample data amplification model using a generative adversarial network (GAN) and knowledge transfer. The proposed model will recognize manufacturing process sounds based on deep TL technology.

2. Literature Reviews

2.1 Application of sound recognition in production lines

Sound recognition technology is being increasingly applied in production lines to enhance automation and efficiency. By utilizing sound information, systems can detect faults in audio signaling devices, analyze equipment vibrations and audio data to identify defects, and convert analog product data to digital format for recognition. Identifying anomalies within noisy industrial settings also reflects the difficulties encountered in categorizing environmental sounds with diverse and unpredictable background noises. Approaches like signal processing techniques and machine-learning-based methodologies can be directly employed to elevate the precision of ESC by enhancing noise resilience and feature extraction.(1) It enables the development of autonomous production systems that can quickly adapt to changes in production environments and product lots, thereby improving the overall approach to maintenance and ultimately minimizing periods of inactivity in production due to unexpected equipment malfunctions, such as in the automotive and machinery industries.(2)

ESC can be improved with signal processing and machine learning techniques that enhance noise resilience and feature extraction. Deep learning techniques, such as long short-term memory (LSTM) models and GANs, can be applied to classify environmental sounds, especially in processing intricate and temporal sound sequences.(3) Diverse and inclusive datasets that accurately represent the nuances and complexities of environmental sound data play a significant role in the research on ESC. To ensure that models in this field can generalize effectively, it is absolutely crucial to employ creative data augmentation techniques. With the proper implementation of these techniques, models can be trained to recognize and adapt to a wider range of scenarios, leading to more accurate and effective results overall.

The advancement and improvement of highly reliable tool condition monitoring (TCM) systems play a vital role in enhancing the longevity of tools used in various industrial processes, thereby significantly contributing to the optimization of tool life span, and in turn, leads to the assurance of high-quality workpiece output, meeting manufacturers' stringent standards and requirements across different sectors. Moreover, implementing such advanced TCM systems also plays a crucial role in effectively managing costs associated with tool maintenance and replacements, ultimately aiding manufacturers in maintaining a competitive edge in the market. Emphasizing the importance of integrating sustainable manufacturing technology into production processes demonstrates a commitment to reducing environmental impact. This aligns with the collective worldwide initiatives to improve environmental conservation and mitigate the effects of energy and resource scarcity. By prioritizing sustainable manufacturing practices, companies contribute to advancing a more ecologically balanced and socially responsible global economy, thereby fostering long-term sustainability and resilience.(4) The exploration of sound signal processing for TCM in high-end manufacturing industries, particularly in sectors like aerospace, showcases the promising potential for enhancing the overall quality of machining processes involved in producing high-value components. This utilization of sound signal processing allows for real-time monitoring and analysis of tool conditions. It opens avenues for the further implementation of advanced techniques to optimize manufacturing operations and ensure the final products' precision and reliability.(4)

With artificial neural networks and machine learning techniques, sound recognition systems can accurately classify faulty devices, improve detection accuracy, and optimize production processes. Overall, the integration of sound recognition in production lines offers a reliable and efficient method for quality control and process optimization.

2.2 Challenges of insufficient data in ESC

ESC faces a significant challenge when dealing with insufficient labeled data, which is common in many real-world scenarios. In manufacturing environments, it is often challenging to capture comprehensive datasets because many critical sound events, such as machine breakdowns or rare failures, occur infrequently. Traditional ESC systems rely heavily on large, labeled datasets for high accuracy. However, obtaining a balanced dataset that covers all potential sound categories in production lines is costly and time-consuming. As a result, these systems often need help with generalization and exhibit poor performance when detecting rare but critical sound events.

Several approaches have been proposed to address the data scarcity problem in ESC. One traditional method is data augmentation, which involves artificially increasing the size of the training dataset through techniques such as noise addition, time stretching, pitch shifting, and other transformations. These methods can help mitigate the lack of training data but are limited when generating novel sound patterns, especially for those of rare sound events crucial in industrial environments.(3) In recent research, TL, where pretrained models on similar tasks are fine-tuned for ESC, has also been explored. This approach leverages knowledge from larger, related datasets, allowing ESC models to perform better with limited data.(2)

More advanced techniques for addressing insufficient data in ESC involve using GMs. For example, GANs have been successfully applied to generate synthetic sound data that can augment real datasets. GANs can create realistic, high-quality sound samples that mimic rare events, enabling ESC systems to improve their performance even when the available real-world data is limited.(1) These synthetic samples expand the training dataset and enhance the model's ability to generalize across different sound categories, especially in highly variable environments such as industrial production lines.

Despite these advancements, insufficient data in ESC remains a critical issue that requires further exploration. Traditional methods of addressing data scarcity have limitations, and the integration of GMs like GANs presents a promising alternative that could revolutionize how ESC systems handle sparse and unbalanced datasets. In the following sections, we build on this recent work by integrating GANs with Bayesian inference to generate virtual samples for sound classification, further advancing the field of ESC in manufacturing environments.

2.3 Virtual sample generation in manufacturing

Virtual sample generation (VSG) provides an alternative solution to addressing the challenges of small datasets and insufficient training samples. It enhances machine learning models by generating additional training samples, addressing the challenge of small datasets, and improving model performance.(5–7) Synthetic data is artificially generated data designed to closely resemble real-world data and enhances the precision and reliability of predictive models. Creating virtual datasets involves sophisticated methods such as manipulating sample distributions and integrating background elements to replicate real-life manufacturing settings. This meticulous approach is a robust foundation for training machine learning models.(1,8)

The core concept of VSG suggests that integrating information diffusion and fuzzy theory with VSG methods offers robust solutions for pattern recognition in scenarios with limited training samples. However, the quality of samples generated through VSG is affected by complicated learning tasks and the growing size of the virtual samples. Advanced GAN architectures play the role of gatekeeper in monitoring the sample qualities. Integrating VSG techniques, including advanced GAN architectures, into manufacturing processes supports rapid decision-making and addresses the limitations posed by small sample sizes. Successive application of deep learning tools is now feasible owing to significant innovations in VSG, particularly GANs with refinements like WGANMTD. These aim to produce high-quality virtual samples for numerical datasets.(9) These innovations in VSG and GANs have potential applications in manufacturing, where they can support rapid decision-making and address the small-sample problem, enhancing decision-support systems.

Performance improvements in monitoring systems using virtual datasets demonstrate the effectiveness of VSG, with significant accuracy gains in detecting operational states of manufacturing equipment. New approaches in video-style transfer, such as the dual-VSG method, have shown promising results by combining nonlinear interpolation with a self-supervised learning framework. Additionally, the use of advanced GAN architectures has demonstrated the potential of video-style transfer to revolutionize decision-making processes in manufacturing by addressing the limitations posed by small sample sizes.(10,11)

Various methods of VSGs, such as the Newton-VSG and GANs, generate virtual samples for small datasets with nonlinear and asymmetric distributions.(12) These techniques help improve learning performance, especially when traditional models struggle to generalize complex datasets or the assumed distribution is not elastic enough for small datasets. By leveraging VSG, manufacturing processes can benefit from enhanced training processes, improved prediction accuracy, and better decision-making based on limited data.

3. Methodology

3.1 TL framework with robust Bayesian inference

TL improves the model's performance by applying knowledge learned on one task to another related task. The core concept is the improvement process for new learning tasks using the learned data characteristics and knowledge. Thus, the learning model can be more generalized and become adaptable. In mathematics, TL means that the relationship of a mapping function, $X_s \rightarrow Y_s$, can explain another mapping function $X_t \rightarrow Y_t$. In other words, $Y_s = f_T\left(X_s\right)$ learned on the basis of $\{X_s, Y_s\}$ will still be valid for explaining $\{X_t, Y_t\}$ so that the mapping relationship, $Y_t = f_T\left(X_t\right)$, can be established, where $\{X_s, Y_s\}$ can be regarded as known data characteristics and f_T learned knowledge. The inference-like operation reflects the purpose that TL leverages the given data to estimate the unknown relationship. Therefore, TL can efficiently reduce the need for a large amount of training data and solve the problem of insufficient training data when handling small-sample learning.

Let $\mathcal{D}$ denote the data population, where X_s and X_t are subsets of $\mathcal{D}$. For a learning task $\mathcal{T}$, which is defined by the user and not necessarily related to $\mathcal{D}$, the purpose of model establishment is to find a functional relationship, $Y = f_T\left(X\right)$, from the existing dataset, $\{X, Y\}$, to approximate $Y = f_{\mathcal{T}}\left(X\right)$, where $Y = f_{\mathcal{T}}\left(X\right)$ is the unknown and ideal function, which does not really exist and can only be found by estimation, $Y = f_T\left(X\right)$. A deep TL model develops f_T, a deep neural network with heavy computing of a nonlinear function for approximating $f_{\mathcal{T}}$.

In terms of sound recognition, X_s can be regarded as the data collected on site, such as voiceprint, sound intensity, temperature, humidity, time and other observed values, $X_s = \{x_1, x_2, \dots x_n\}$, that is, the environmental dataset X that can be described by n features. Y_s is the sound category, such as the state of each machine (e.g., starting sound, running sound, deceleration sound, stopping sound, emergency stop sound, fault buzzer of machine No. 1, and starting sound of machine No. 2), the voice of the staff (smooth or impatient), the sound of feeding, workpiece placement, and transport on the conveyor belt. That is, $Y_s = \{y_1, y_2, \dots y_m\}$, where y_i represents the ith defined sound among m categories. The user can expand $\{X, Y\}$ to describe the surrounding environment in detail, and $\{X, Y\}$ is subject to user definition. When the entire production condition changes, such as the addition of new machines to the operation, replanning of production lines, and power outages, the feature set X and the category set Y also vary their coverages. That is, X would be expanded to $\left\{X_s, X_t\right\} = \left\{x_1, \cdots x_n, x_{n+1}, \cdots x_{n+k}\right\}$, and Y stored as sound categories could also be expanded to $\left\{Y_s, Y_t\right\} = \left\{y_1, \cdots y_m, y_{m+1}, \cdots y_{m+l}\right\}$. The TL mechanism establishes $Y_t = f_T\left(X_t\right)$ based on $Y = f_T\left(X\right)$ and its corresponding $\{X_s, Y_s\}$ to estimate $Y_t = f_{\mathcal{T}}\left(X_t\right)$ when $\{X_t, Y_t\}$ is appended to the present dataset.

Since f_T is the estimator of $f_{\mathcal{T}}$, we considered an operable parameter θ to be introduced into f_T for better approximating the unknown $f_{\mathcal{T}}$, that is, f_T can be expressed in the form of $f_T(\cdot|\theta)$. The operable parameter θ mentioned above does not refer to the population parameters, nor is it derived from the original dataset. θ, which is operable, can generally refer to the adjustment parameters involved in conducting learning tasks. For example, in nonlinear functions of deep network architectures (CNN, GAN), θ could be the weights, learning rates, or other parameters in the network architecture. However, the unknown parameters θ can be estimated using the collected data. Therefore, in this study, we integrate the concept of robust Bayesian inference into the TL framework and propose a deep TL framework based on robust Bayesian inference. As Eq. (1) shows, we proposed $f_T(X_t|\theta)$ to approximate $f_{\mathcal{T}}$, where θ can be formulated as a function of X_s, that is, $\theta = \mathcal{P}\left(\theta|X_s\right)$.

$$Y_t = f_T\left(X_t|\theta\right) \tag{1}$$

3.2 Membership function family of robust Bayesian inference

Emergency situations on the production line may not arise often, but they may cause heavy losses each time because "no one knows it will happen." It can also be attributed to the fact that underlying characteristics inside the coming $\left\{x_{n+1}, \cdots x_{n+k}\right\}$ have not been observed, resulting in the inability to propose corresponding and reference measures. In the context of limited

$y = f_T(x)$, we often fail to obtain robust learning results. TL has limitations when applying the learned knowledge to a new and independent dataset. Therefore, in addition to developing robust Bayesian inference as a deep TL framework as mentioned above, we further integrate a small-sample-data augmentation model into it.

In the previous MTD (Mega-Trend-Diffusion) research,(13) the architecture for integrating robust Bayesian inference with GAN was developed, and a membership function based on robust Bayesian inference was proposed for data augmentation. In this study, we further improve and refine the membership function for summarizing a membership function family and for developing an algorithm to build a small dataset augmentation model based on the deep TL framework.

When developing priori distribution parameter estimation for nonlinear distribution, we considered the Weibull distribution density function as a flexible function family for modeling the fuzzy distribution of small sample data. The shape of Weibull density is elastic and suitable for describing the spread of nonlinear and asymmetric distribution data. It has good adaptability for fitting the priori distribution of robust Bayesian inference. The general form of Weibull density has three parameters, which determine the appearance shape parameter (β), the scale parameter (λ), and the data threshold parameter (γ) of the distribution function. When the threshold parameter is zero ($\gamma = 0$), it degenerates into a two-parameter Weibull distribution, that is, the Weibull distribution for nonnegative random variables. The probability density function of the general Weibull distribution is defined as

$$f(x,\beta,\lambda,\gamma) = (\beta/\lambda)\left((x-\beta)/\lambda\right)^{\beta-1} e^{-\left((x-\beta)/\lambda\right)^{\beta}}, x \geq \gamma, \beta > 0, \lambda > 0, \tag{2}$$

where data x must be greater than or equal to the threshold parameter γ, and the shape parameter β and scale parameter λ are real numbers greater than 0.

For the augmentation of virtual samples and based on the previously proposed WGAN_MTD2, we introduced the priori distribution function q into the three-parameter Weibull density, considering a contaminated parameter ε, where ε is between 0 and 1. The contaminated parameter ε can be regarded as the interference term used to control the preset q proportion of influence on the entire membership function. Therefore, the membership function family that defines robust Bayesian inference has the following complete form:

$$MF = (1-\varepsilon)MF_0(\theta \mid \lambda) + \varepsilon q, \text{ where } q \in Q = \left\{ f \mid f(x,\gamma,\lambda,\beta) \right\}. \tag{3}$$

The results of the previous research(9) suggest a symmetric Gaussian priori distribution for unimodal and symmetric data distribution situations. However, in a small dataset, it is difficult to determine its kurtosis and whether it is symmetrical from the data dispersion. In view of this, we introduced the density function of the three-parameter Weibull distribution, $f(x, \gamma, \lambda, \beta)$, which can cover more types of data, including multimodal, asymmetry, and discretized data distributions. For the univariate case with consideration of a single variable x, the membership function of its robust Bayesian inference can be presented as

$$MF = \begin{cases} (1-\varepsilon)\dfrac{x-LB}{CL-LB} + \varepsilon f\left(x,\gamma,\lambda,\beta\right), & x \le CL \\ (1-\varepsilon)\dfrac{UB-x}{UB-CL} + \varepsilon f\left(x,\gamma,\lambda,\beta\right), & \text{otherwise} \end{cases} \tag{4}$$

In Eq. (4), *LB* and *UB* are, respectively, the possible lower and upper bounds of the data distribution of a single variable *x*, where *CL* respresents the center points of the data distribution. Therefore, the membership function of robust Bayesian inference is a linear combination of two functions, MF_0 and *f*, as shown in Fig. 1. Then the Weibull density would be more suitable and flexible for modeling various data distributions than the Gaussian-based function. The user only needs to adjust the weights of MF_0 and *f* separately for different data patterns. Through ε parameter intervention, the calculation of *MF* does not fully rely on MF_0 computed from the patterns presented by small datasets, but also considers the subjective information partially from *f*.

3.3 VSG through the modified GANs with cooperating robust Bayesian inference (RBI)

In this study, we adopted GANs to control the quality of the generated virtual sample to confirm that the generation will not output a dataset that differs too much from the original data. GAN is used for comparison to make the generated virtual samples closer to the real data. We adopted GANs to control the quality of the generated virtual sample and confirmed that the generation does not output a dataset that differs too much from the original data. GAN is used for comparison to make the generated virtual samples closer to the real data. In the past research,[14] GAN simulation generation has achieved adequate performance when producing new samples following the target dataset. However, when inputting the small-size training

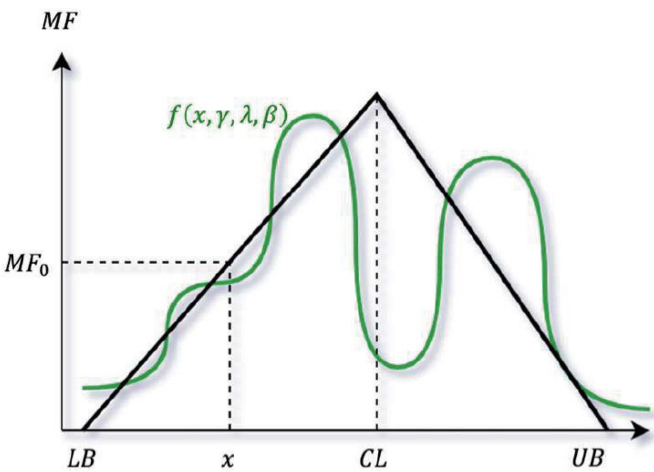

Fig. 1. (Color online) Membership function of robust Bayesian inference. *LB* and *UB* are, respectively, the possible lower and upper bounds of the data distribution of a single variable *x*, where *CL* represents the center points of the data distribution. The membership of robust Bayesian inference is a linear combination of two functions, MF_0 and *f*.

datasets to GAN, the population knowledge base of the given small dataset is not sufficient to generate virtual samples, and the phenomenon of overfitting would occur, as anticipated.

In this study, we considered the framework of WGAN-GP[15] as the guide for generating virtual samples with several modifications. The difference is that the original WGAN-GP samples from Gaussian distribution or uniform distribution were the input of the generator *G*, while for the model proposed in our study, $X\,|_{MF}$, developed as mentioned above, was adopted as the input of *G*. As shown in Fig. 2, the existing small sample dataset X_s is used as the input of the discriminator *D* to compare the virtual sample generated by the generator *G*, and then through the gradient penalty (GP) weight update, the final generated virtual samples possessed acceptable quality similar to that of the original data. The modified WGAN-GP algorithm mainly replaces the input of the generator *G* with $X\,|_{MF}$. In the GP section, adding the loss function, $L = D_\omega(\tilde{x}) - D_\omega(x) + \delta\left(\nabla_{\hat{x}} D_\omega(\tilde{x})_2 - 1\right)^2$, as a penalty term would help accelerate meeting the convergence conditions. Therefore, the concept of the modified algorithm can be described by the pseudocode shown in Fig. 3.

Before implementing the algorithm, its initial settings require a given gradient penalty coefficient (δ), the number of iterations needed to identify that a virtual dataset has been generated (d_{cirtic}), and the batch size (b) of virtual samples to be generated. The network training process adopts *Adam* as the optimization function for updating weights iteratively. It requires the initialization of the *Adam* function and network parameters, including starting weights α_0 (for the *Adam* function), ω_0 (for the *D* network), and υ_0 (for the *G* network).

To address the challenge of insufficient data in ESC, our approach integrates GANs with Bayesian inference. This combination allows us to generate high-quality synthetic sound samples that effectively augment the limited real-world datasets used for training ESC models. Below, we outline the step-by-step operational mechanics of this method and how it addresses the core challenges of sound sample generation. The process begins with Bayesian inference, which provides a statistical framework for estimating the priori distribution of different sound

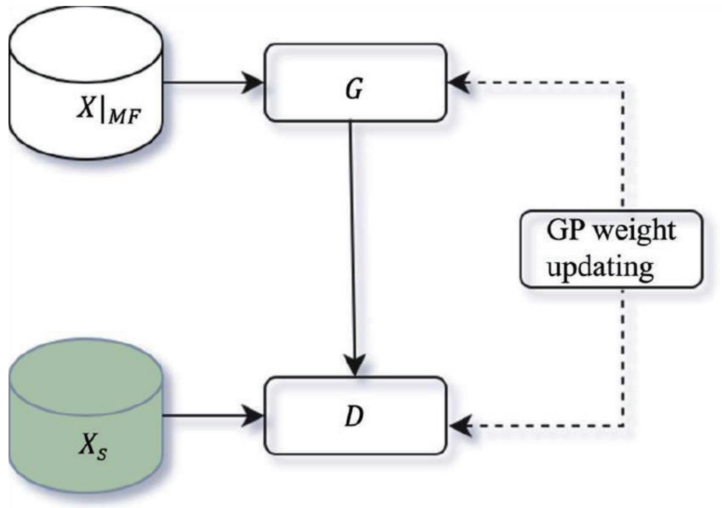

Fig. 2. (Color online) We considered the framework of WGAN-GP[14] as the guide for generating virtual samples with several modifications. The main difference is that our proposed model adopted $X|_{MF}$ as the input of *G*.

While υ hasn't converged

For $t = 1, \cdots, d_{critic}$

For $i = 1, \cdots, b$

x drawn from X_s, z drawn from $X|_{MF}$, i.e. $z \sim MF(z)$, and a random ϵ drawn from $U[0,1]$

$\tilde{x} \longleftarrow G_\upsilon(z)$

$\hat{x} \longleftarrow \epsilon x + (1-\epsilon)\tilde{x}$

$L^{(i)} \longleftarrow D_\omega(\tilde{x}) - D_\omega(x) + \delta(\nabla_{\hat{x}} \| D_\omega(\hat{x}) \|_2 - 1)^2$

End for

$\omega \longleftarrow Adam(\nabla_\omega \frac{1}{b} \sum_{i=1}^{b} L^{(i)})$

End for

draw m samples to form $\{z^{(i)}\} \sim MF(z)$

$\upsilon \longleftarrow Adam(\nabla_\upsilon \frac{1}{b} \sum_{i=1}^{b} -D_\omega(G_\upsilon(z)))$

End while

Fig. 3. Pseudocode of WGAN-GP with $X|_{MF}$ input for its generator *G*.

categories. In a manufacturing environment, where certain sounds (e.g., machine breakdown) are rare, Bayesian inference enables us to model these events using available data to establish reasonable probability distributions. These priori distributions are crucial in guiding the generation of synthetic samples and ensuring that the generated sounds are realistic and reflect the likelihood of rare events that are under-represented in the original dataset. Once the priori distributions are established, these are fed into the GAN architecture. The GAN consists of two neural networks: the generator and the discriminator. The generator takes the priori distributions and uses them as input to create synthetic sound samples. During this process, the generator iteratively produces new sound samples that the discriminator then evaluates. The discriminator's role is to distinguish between accurate sound data and synthetic samples generated by the GAN. Through this adversarial training process, the generator learns to produce increasingly realistic sound samples that mimic the characteristics of real-world data.

Integrating Bayesian inference into the GAN framework is crucial in ensuring that the generated samples are relevant to the specific sound classification task. By using Bayesian priors, we avoid the issue of generating random or irrelevant samples. Instead, the synthetic data is tailored to the sounds most critical to the ESC system's performance, such as sounds that indicate malfunctions or unusual events in a production line. This focused approach improves the quality of the generated samples and ensures that the augmented dataset is diverse enough to cover rare and essential events. Additionally, a gradient penalty (WGAN-GP) is applied during the training process to prevent the GAN from generating overfitted or low-quality samples. This regularization technique stabilizes the learning process by constraining the model's gradients, ensuring that the synthetic samples remain within a realistic range and that the generator does not overfit the training data. This combined approach of using Bayesian inference with GANs allows for the creation of a robust dataset that addresses the problem of insufficient data in ESC.

The ESC system can better recognize rare or complex sound events as a result of generating synthetic samples that closely resemble real-world sounds, thus improving the overall classification accuracy and reliability.

In real-world industrial applications, handling large datasets efficiently is crucial for any sound classification framework. The proposed GM integrated with the diffusion technique is designed with scalability in mind. One of the framework's key features is its ability to scale seamlessly to accommodate increasing amounts of data without significant computational bottlenecks. The GAN-based VSG method is computationally efficient and capable of generating large datasets from small amounts of initial sound data. It allows the system to augment datasets dynamically, improving model performance without requiring extensive real-world data collection.

The framework can be adapted for parallel and distributed computing architectures to enhance scalability further. By distributing the computational load across multiple processors or machines, the system can handle large volumes of sound data, such as those collected in large-scale production environments or across multiple manufacturing sites. Additionally, the model's architecture allows for batch processing of sound data, ensuring that the generation and classification tasks can be performed simultaneously with reduced processing times and improved real-time application feasibility. Another factor that ensures scalability is using CNNs for the classification task. CNNs are well known for their ability to process high-dimensional data efficiently. In our system, CNNs handle the input spectrograms generated from the sound data, leveraging their shared-weight architecture to minimize computational costs while maximizing feature extraction performance. As a result, the framework can maintain high classification accuracy even as the dataset size increases.

4. Experimental Studies

Figure 4 shows the flowchart of model verification with the production line scenario.

4.1 Audio data collection and data labeling

On-site information is collected individually within each of the four main production line layouts: functional, product, fixed, and modular. Different production lines represent different manufacturing industries, so the training materials must be extended to as many small- and medium-sized enterprises as possible. In addition to helping managers understand the current situations of the production line, on-site sound monitoring also helps managers make quicker decisions on emergency response. For infrequent emergencies and more common emergencies, it is necessary to discuss with on-site personnel to clarify and define normal and abnormal situations.

The sounds of the factory environment, operators, machines, and pieces of equipment can each be subdivided into different types that represent different physical meanings. Therefore, sound data labeling must be carried out after the audio data is collected, especially for labeling audio data in emergencies. Since various on-site operators or management cadres can provide

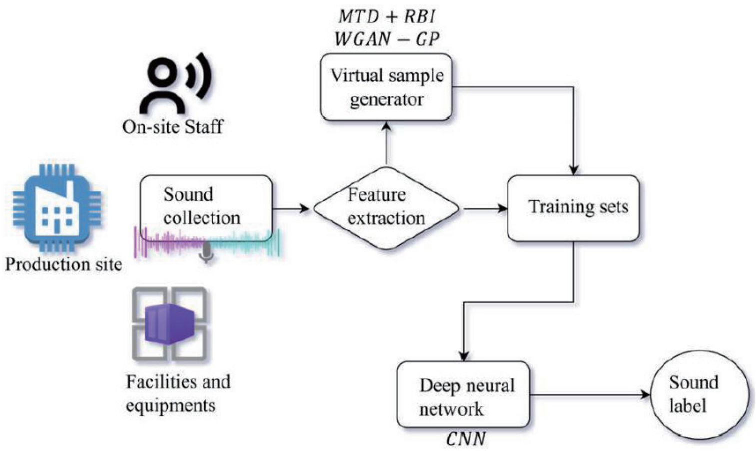

Fig. 4. (Color online) Flowchart of model verification with the production line scenario.

relevant information and the emergency definition varies from factory to factory and for each machine, discussions with on-site personnel are needed to establish clear label definitions.

In the experimental study, we collected sound datasets from a production line consisting of two machines and two operators. To easily identify them, we labeled the datasets as Machine 1, Machine 2, Operator 1, and Operator 2. For example, "Machine 1" refers to the sound dataset collected from the first machine, and so on.

4.2 VSG

In establishing a small sample data augmentation model based on the aforementioned deep TL framework, the test data is used for interactive verification in advance, and the quality of the amplified samples is compared to grasp the optimal range of each parameter setting before entering the actual production line. When facing various unbalanced datasets, the gap in the amount of data between categories is more serious. This is due to differences in the data collection time range and the frequency of events. Therefore, through on-site inspections, parameters are further adjusted, including robust Bayesian inference parameters, (γ, λ, β), as well as the relevant parameters δ of the WGAN-GP algorithm shown in Fig. 3, such as d_{critic} and b, making the virtual sample generator more general.

Concerning the differences among each dataset, we increased the number of virtual samples generated incrementally to observe the impact on accuracy. Our aim was to carefully investigate how the incremental increase in virtual samples influenced the overall accuracy of the dataset. We systematically appended five virtual samples to join the training set each time for the same ESC models to investigate the changing accuracies.

As shown in Table 1, the first column lists the sizes of various training sets and the first row indicates the learning accuracies of the original (training) datasets comprising five data values each for four datasets: Machine 1, Machine 2, Operator 1, and Operator 2. It shows differences among various datasets since the accuracies vary from 47.8% to 91.2%. With each increase of five virtual samples, starting from 15 virtual samples, we made the training set with sizes from 20 to 65, which means the virtual sample size increased from 15 to 60.

Table 1
Averaged learning accuracies for five sound datasets with different sizes of training sets: Machine 1, Machine 2, Machine 3, Operator 1, and Operator 2. The training set size increases from 5 to 65; the number within the parentheses indicates the virtual sample size.

Training set size (virtual sample size)	Machine 1 (%)	Machine 2 (%)	Operator 1 (%)	Operator 2 (%)	Machine 3 (%)
5 (0)	47.8	89.4	91.2	63.8	50.7
20 (15)	56.2	87.6	85	70.6	51.9
25 (20)	63.4	87.8	88.2	61.4	56.7
30 (25)	64.2	88.6	92	72.4	63.4
35 (30)	46.4	90	91.2	51.8	66.7
40 (35)	53.8	93	97.2	61.2	68.3
45 (40)	63	91.8	88	71.8	75.1
50 (45)	51.4	87.4	86.8	69.4	74.8
55 (50)	59.6	94.6	92	68.8	70.9
60 (55)	61.8	92.8	91.6	67.6	69.3
65 (60)	47.4	88.4	87.8	40.6	65.5

4.3 Deep network training and tuning

We adopted CNNs as a learning tool for sound recognition in production environments. It was tested in production lines to better grasp the optimal values of starting parameter settings to improve recognition accuracy and practical usability.

Table 1 shows the averaged learning accuracies for five sound datasets with different sizes of training sets: Machine 1, Machine 2, Machine 3, Operator 1, and Operator 2. The training set size increases from 5 to 65, and the number within the parentheses indicates the virtual sample size. The training data set from Machine 1 increased by 16.4%, the training set of Machine 2 increased by 5.2%, the training set of Operator 1 increased by 6%, and the training set of Operator 2 increased by 8.6%. For each dataset, we marked the highest learning accuracy in red. The highest learning accuracies were obtained with different training sets. For Machine 1, the highest accuracy occurs with a training set of size 30, while the highest accuracy occurs with a training set of size 55 for Machine 2. In the cases of Operator 1 and Operator 2, the highest accuracies occur with the sizes of 40 and 30, respectively. The overall improvement is greatly improved compared with the five original samples. According to our verification and analysis, we found that using GAN to generate samples is of significant help to the ESC model with low accuracy due to an insufficient sample number, and the effect will be more obvious for cases with lower original accuracy. The size of the virtual sample is a key factor in the accuracy of the ESC model. Blindly increasing the number of GANs does not cause a continuous rise in accuracy.

A critical factor in evaluating the effectiveness of any ESC system is its ability to generalize across different manufacturing environments. We tested the system on various industrial soundscape datasets to evaluate the model's generalizability. For example, the system was exposed to various sounds in automotive production lines, including machine operations, conveyor belts, and welding sounds. The system successfully identified early signs of equipment wear and malfunctions in these environments, demonstrating its ability to adapt to the high-noise, high-variability environment typical of the automotive sector. The system was also tested on the more delicate, less noisy environments, such as assembly lines for small components, of the electronics manufacturing industry. In these scenarios, the system was able to filter out background noise and focus on the critical operational sounds of machines, such as soldering stations and robotic arms. These results demonstrated the model's flexibility in handling both high-noise and low-noise environments, confirming its generalizability.

Integrating GANs and Bayesian inference is vital in enhancing the model's generalizability. The GANs allow the system to generate synthetic sound samples tailored to each specific environment, thereby augmenting the dataset with relevant training data. This ensures that the model can handle varying sound conditions, even when the availability of real-world data from a particular environment is limited. By leveraging the virtual data, the system can improve its classification accuracy and reliability across different industrial settings.

5. Conclusions

In this study, we proposed a new framework of the ESC model that can be applied to handle the model training challenges of sound recognition systems when insufficient samples are available. Traditional systems often perform poorly owing to the scarcity of samples, so we used GAN to generate virtual sound samples to improve model performance. The research methods included using GANs to generate artificial sound samples and integrating them with existing datasets to expand training materials, with the aim of enhancing the stability and performance of ESC systems. The research results demonstrated the potential of GAN in expanding training data, which may reduce sample collection costs and improve the practical application of ESC technology in manufacturing systems.

The application of GAN in ESC represents a novel technological application. This method can generate high-quality virtual samples (synthetic sound data) and simulate the operating environment of various production lines. This diversity can make the training process and results of the ESC model more robust and general. For example, it can help the system better identify different types of machines or sounds in various noisy environments, thus greatly improving its practicality and applicability. The traditional voiceprint data collection process is often time-consuming, laborious, and costly, especially when many diverse samples are required. GAN technology can significantly reduce these costs. Enterprises can use limited real data to generate many high-quality virtual samples, thereby greatly saving human and financial resources. This would accelerate the product development cycle and allow small- and medium-sized enterprises to develop high-performance ESC systems, thus promoting innovation and competition in the entire industry.

Although GAN technology has broad prospects, its practical application still faces challenges. The primary issue is ensuring the quality and diversity of the generated samples. Samples that are too similar or of poor quality can cause model overfitting or degraded performance. Another challenge is balancing the proportion of real data and virtual samples. Over-reliance on virtual samples can lead to poorly performing models with overfitting issues in real-world scenarios. In addition, effective methods need to be developed to evaluate and verify the reliability of synthetic samples to ensure that they improve model performance rather than introducing bias.

Using GANs to generate virtual samples of sound data involves complicated ethical issues. The first is the issue of privacy: in addition to environmental sounds such as production line machines, the sources of virtual sound generated in this experiment also include the operators' voices. If the virtual samples generated are too close to the voices of real individuals, they may be abused for fraud or identity purposes. Secondly, there is the issue of informed consent: the original sound samples used to train GAN must be used with full authorization. Potential discrimination issues must be considered, such as ensuring that the generated sample does not reinforce or amplify existing gender, racial, or other biases. Therefore, researchers and businesses need to develop strict ethical guidelines and legal frameworks to regulate the use of this technology, protect individual privacy, and prevent possible misuse.

The decision to use CNNs as the primary learning strategy for this study was based on their proven effectiveness in analyzing audio data represented as spectrograms, which can be viewed as two-dimensional images. CNNs have demonstrated strong performance in pattern recognition tasks, particularly when spatial hierarchies of features need to be extracted, which is highly applicable to sound data. In the context of ESC, sound signals are typically transformed into spectrograms before being fed into the model. The spatial arrangement of frequencies and their intensity over time is well suited to the ability of CNNs to capture local and global patterns in the data. CNNs excel at learning from visual representations of sound, such as Mel-spectrograms, where the structure of the sound is preserved in a form similar to an image. It allows CNNs to detect critical features like frequency shifts, harmonic patterns, and transient sound changes—elements crucial for identifying environmental sounds in manufacturing processes. Additionally, the shared-weight architecture and convolutional layers of CNNs make them computationally efficient, which is essential when processing large datasets or when real-time classification is required, as is the case in many industrial settings.

Other learning strategies, such as recurrent neural networks (RNNs) or LSTM networks, were considered but ultimately not chosen for several reasons. While RNNs and LSTMs are often used for sequential data like audio, they are generally more suited to tasks where temporal dependences are paramount, such as speech recognition or natural language processing. In ESC, where the goal is to classify distinct sounds rather than to process sequential speech data, CNNs have proven more efficient and accurate in recognizing spatial patterns across spectrograms. Furthermore, RNNs and LSTMs typically require more computational resources and training time, making them less suitable for real-time monitoring in manufacturing environments. Moreover, recent studies have shown that CNNs outperform RNN-based models in tasks involving environmental sound classification owing to their ability to handle the high-dimensionality of sound spectrograms.[3] The ability of CNNs to perform automatic feature extraction without the need for extensive preprocessing is another advantage, as it reduces the complexity of the model training process and makes CNNs more robust for generalization to unseen sound data. On the basis of these considerations, CNNs were selected as the most appropriate learning strategy in this study.

The research results can help further chip development of deep learning networks. A system-on-chip (SoC) is a microcomputer that integrates a central processing unit, memory, timer/counter, and various input and output interfaces on an integrated circuit chip. Compared with the general-purpose CPU used in personal computers, the biggest advantage is that it is small and saves costs. A SoC can be placed inside an instrument with limited storage. The input and output interfaces are simple, but the functions that can be implemented are also more specific. In recent years, with the evolution of artificial intelligence and the popularization of deep learning, the design of AI single chips for various field applications has begun to be widely discussed. Regardless of the application, all AI chips can be defined as specifically designed to run machine learning work. Integrated circuits are designed for loads that are processed in a manner similar to how the human brain functions, and they process decisions and tasks in a complex and rapidly changing world. A single chip has a small storage space and a simple input and output interface. These limitations may be insufficient when dealing with large amounts of data and complex calculations. The calculation of WGAN-GP must consume power, and the large amounts of data generated requires storage space. Low power consumption and low-calculation-intensive chip operation become important in this case. This may involve optimizing the chip architecture to ensure the efficient management of energy and computing resources when handling complex calculations. At the same time, since WGAN-GP may require large amounts of data and storage space, this is another aspect that needs to be considered in chip design.

In this study, the above issues present the challenges of balancing SoC microprocessors and WGAN-GP, which requires trade-offs between energy conservation, computational performance, and storage requirements. The design has two important features: low power consumption and low-computing-intensive chip operation. However, storage space is another challenge for real-time calculation and identification of live sound data. In the future study, we will cooperate with experts and scholars from electrical engineering and information engineering to engage in SoC development of deep learning networks with low power consumption, low computing intensity, and high response speed. The deep learning network, shown in Fig. 4, could be an architectural blueprint for proposing a new type of deep learning chip design in the manufacturing process sound recognition model based on TL technology. Synthetic sound data generated from virtual samples enhances sensor performance by providing additional training samples for machine learning models, which improves their ability to recognize and classify various environmental sounds. In our research, we leverage advanced techniques like GMs and diffusion methods to create virtual sound samples that closely resemble real-world data. This process improves the accuracy and reliability of predictive models, thereby improving sensor performance. By generating diverse and inclusive datasets through synthetic data, users can train sensors to recognize a broader range of sound scenarios. This leads to more effective and accurate results in sound classification tasks, making the technology more reliable and applicable in various real-world situations. Integrating synthetic sound data with existing datasets expands the training materials available for sensors, ultimately enhancing their stability and overall performance in sound recognition systems.

Acknowledgments

This study is partially supported by the National Science and Technology Council, Taiwan (grant contract no. NSTC 112-2221-E-167-032-MY2).

References

1 Y. Tagawa, R. Maskeliūnas, and R. Damaševičius: Electronics **10** (2021) 2329. https://doi.org/10.3390/electronics10192329
2 R. Espinosa, H. Ponce, and S. Gutiérrez: Appl. Soft Comput. **108** (2021) 107465. https://doi.org/10.1016/j.asoc.2021.107465
3 P. Becker, C. Roth, A. Roennau, and R. Dillmann: 2020 IEEE 7th Int. Conf. Industrial Engineering and Applications (ICIEA, 2020) 921. https://doi.org/10.1109/ICIEA49774.2020.9102002
4 K. Guo and J. Sun: Mech. Syst. Signal Process. **157** (2021) 107738. https://doi.org/10.1016/j.ymssp.2021.107738
5 D.-C. Li, S.-C. Chen, Y.-S. Lin, and K.-C. Huang: Appl. Sci. **11** (2021) 10823. https://doi.org/10.3390/app112210823
6 L. Shen and Q. Qian: Comput. Mater. Sci. **211** (2022) 111475. https://doi.org/10.1016/j.commatsci.2022.111475
7 X. Zhong and H. Ban: Ann. Nucl. Energy **175** (2022) 109201. https://doi.org/https://doi.org/10.1016/j.anucene.2022.109201
8 J. Kim, H. Lee, S. Jeong, and S.-H. Ahn: J. Manuf. Syst. **58** (2021) 431. https://doi.org/10.1016/j.jmsy.2020.12.020
9 Y.-S. Lin, L.-S. Lin, and C.-C. Chen: Symmetry **14** (2022) 339. https://doi.org/10.3390/sym14020339
10 L. S. Lin, Y. S. Lin, D. C. Li, and Y. H. Liu: Decis. Support Syst. **172** (2023) 10. https://doi.org/https://doi.org/10.1016/j.dss.2023.113996
11 L. S. Lin, Y. S. Lin, D. C. Li, and Y. T. Chen: Appl. Soft Comput. **143** (2023) 20. https://doi.org/10.1016/j.asoc.2023.110406
12 L.-S. Lin, Y.-S. Lin, and D.-C. Li: Neurocomputing **548** (2023) 126408. https://doi.org/10.1016/j.neucom.2023.126408
13 X. Yu, Y. He, Y. Xu, and Q. Zhu: J. Phys. Conf. Ser. **1325** (2019) 012079. https://doi.org/10.1088/1742-6596/1325/1/012079
14 Y.-S. Lin, M.-L. Huang, D.-C. Li, and J.-Y. Yang: Sens. Mater. **36** (2024) 2439. https://doi.org/10.18494/SAM4780
15 I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville: Adv. Neural Inf. Process. Syst. **30** (2017). https://doi.org/10.48550/arXiv.1704.00028

About the Authors

Yao-San Lin is an associate professor of the Department of Industrial Engineering and Management at National Chin-Yi University of Technology, Taiwan, and has taught statistical analysis, information management, and machine learning for many years. In his past ten years of research, he has won the IEEE Best Paper Award (2019 and 2013), and his research results have been published in international academic journals: Applied Soft Computing, Decision Support Systems, Neurocomputing, European Journal of Operational Research, International Journal of Production Research, Computers & Operations Research, and Expert Systems with Applications. His current research interests are mainly in the fields of manufacturing and natural language processing. (yslin@ncut.edu.tw)

Mei-Ling Huang received her M.S. and Ph.D. degrees in industrial engineering from the University of Wisconsin–Madison and National Chiao Tung University, respectively. Currently, she is affiliated with the Department of Industrial Engineering and Management at National Chin-Yi University of Technology. Her research interests include quality management, quality engineering, data mining, and medical diagnosis. (huangml@ncut.edu.tw)