# Computers & Industrial Engineering

## Knowledge Distillation for Cost-Effective Fault Prediction in Manufacturing Process

--Manuscript Draft--

| | |
|---|---|
| **Manuscript Number:** | |
| **Article Type:** | Research Paper |
| **Keywords:** | fault prediction; inspection cost; knowledge distillation; active inspection; semiconductor manufacturing |
| **Corresponding Author:** | Jaewoong Shim<br>Seoul National University of Science & Technology<br>KOREA, REPUBLIC OF |
| **First Author:** | Junbong Heo |
| **Order of Authors:** | Junbong Heo |
| | Minhyeok Son |
| | Jaewoong Shim |
| **Abstract:** | Manufacturing processes involve various inspections aimed at identifying faulty products before they reach the final stages. These inspections are typically categorized into basic, conducted for all products, and advanced, conducted only for selected sampled products due to cost constraints. Recent advancements have leveraged inspection data to train machine learning models that predict potential faults in manufactured products. However, models using only basic inspection results, referred to as the basic model, often underperform compared to those that also use advanced inspection results, referred to as the advanced model, due to limited information. In this study, we propose a novel approach to train a basic model using knowledge distillation from an advanced model, achieving high prediction accuracy with reduced inspection costs. Additionally, we integrate this distilled basic model into an active inspection framework during the inference phase to further improve the cost-effectiveness of inspections. Within this framework, the distilled basic model and the advanced model are selectively used, optimizing the usage of inspection costs. The effectiveness of our approach is demonstrated through a case study involving real-world data from a semiconductor manufacturer. |
| **Suggested Reviewers:** | Eunji Kim<br>eunjikim@cau.ac.kr |
| | Dongil Kim<br>d.kim@ewha.ac.kr |

Cover Letter

SEOULTECH SEOUL NATIONAL UNIVERSITY OF
SCIENCE & TECHNOLOGY

July 17, 2024

Dear Editor,

Please find the enclosed manuscript, “**Knowledge Distillation for Cost-Effective Fault Prediction in Manufacturing Process**”, which we would like to submit for consideration in *Computers & Industrial Engineering* Journal.

Manufacturing processes involve various inspections aimed at identifying faulty products before they reach the final stages. These inspections are typically categorized into basic, conducted for all products, and advanced, conducted only for selected sampled products due to cost constraints. Recent advancements have leveraged inspection data to train machine learning models that predict potential faults in manufactured products. However, models using only basic inspection results, referred to as the basic model, often underperform compared to those that also use advanced inspection results, referred to as the advanced model, due to limited information. In this study, we propose a novel approach to train a basic model using knowledge distillation from an advanced model, achieving high prediction accuracy with reduced inspection costs. Additionally, we integrate this distilled basic model into an active inspection framework during the inference phase to further improve the cost-effectiveness of inspections. Within this framework, the distilled basic model and the advanced model are selectively used, optimizing the usage of inspection costs. The effectiveness of our approach is demonstrated through a case study involving real-world data from a semiconductor manufacturer.

This manuscript is the authors' original work and has not been published nor has it been submitted simultaneously elsewhere. All authors have checked the manuscript and have agreed to the submission.

We hope that you will consider this paper as suitable publication in this journal.

Sincerely,

Jaewoong Shim, Ph.D.
Assistant Professor of Industrial Engineering
Seoul National University of Science & Technology

Tel: +82 2 970 6485
E-mail: jaewoong@seoultech.ac.kr

# Knowledge Distillation for Cost-Effective Fault Prediction in Manufacturing Process

Junbong Heo[a], Minhyeok Son[b], Jaewoong Shim[a,b,*]

[a] *Department of Data Science, Seoul National University of Science & Technology, 232 Gongneung-ro, Nowon-gu, Seoul 01811, Republic of Korea*

[b] *Department of Industrial Engineering, Seoul National University of Science & Technology, 232 Gongneung-ro, Nowon-gu, Seoul 01811, Republic of Korea*

## Abstract

Manufacturing processes involve various inspections aimed at identifying faulty products before they reach the final stages. These inspections are typically categorized into basic, conducted for all products, and advanced, conducted only for selected sampled products due to cost constraints. Recent advancements have leveraged inspection data to train machine learning models that predict potential faults in manufactured products. However, models using only basic inspection results, referred to as the basic model, often underperform compared to those that also use advanced inspection results, referred to as the advanced model, due to limited information. In this study, we propose a novel approach to train a basic model using knowledge distillation from an advanced model, achieving high prediction accuracy with reduced inspection costs. Additionally, we integrate this distilled basic model into an active inspection framework during the inference phase to further improve the cost-effectiveness of inspections. Within this framework, the distilled basic model and the advanced model are selectively used, optimizing the usage of inspection costs. The effectiveness of our approach is demonstrated through a case study involving real-world data from a semiconductor manufacturer.

*Keywords:* fault prediction, inspection cost, knowledge distillation, active inspection, semiconductor manufacturing

## Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT; Ministry of Science and ICT) (No. RS-2022-00165783).

## References

[*] Corresponding author. Tel.: +82 2 970 6485

*Email addresses:* gjwnsqhd@g.seoultech.ac.kr (Junbong Heo), shawn22587@g.seoultech.ac.kr (Minhyeok Son), jaewoong@seoultech.ac.kr (Jaewoong Shim)

Highlights (for review)

- Knowledge distillation is employed for training fault prediction model
- Fault prediction accuracy is improved without additional inspection cost.
- Integrating with active inspection framework further improve the cost-effectiveness
- Case study on real-world data demonstrate the effectiveness of our method

# Knowledge Distillation for Cost-Effective Fault Prediction in Manufacturing Process

## Abstract

Manufacturing processes involve various inspections aimed at identifying faulty products before they reach the final stages. These inspections are typically categorized into basic, conducted for all products, and advanced, conducted only for selected sampled products due to cost constraints. Recent advancements have leveraged inspection data to train machine learning models that predict potential faults in manufactured products. However, models using only basic inspection results, referred to as the basic model, often underperform compared to those that also use advanced inspection results, referred to as the advanced model, due to limited information. In this study, we propose a novel approach to train a basic model using knowledge distillation from an advanced model, achieving high prediction accuracy with reduced inspection costs. Additionally, we integrate this distilled basic model into an active inspection framework during the inference phase to further improve the cost-effectiveness of inspections. Within this framework, the distilled basic model and the advanced model are selectively used, optimizing the usage of inspection costs. The effectiveness of our approach is demonstrated through a case study involving real-world data from a semiconductor manufacturer.

*Keywords:* fault prediction, inspection cost, knowledge distillation, active inspection, semiconductor manufacturing

## 1. Introduction

Fault prediction refers to the task of preemptively identifying potential faulty products in the manufacturing process. Despite numerous inspection steps designed to filter out faulty products, some faults may not be detected until the final shipment inspection, or worse, they may be discovered by customers after delivery. Late detection of faulty products leads to unnecessary continuation of the manufacturing process, incurring additional costs and potentially damaging the manufacturer's reputation if faulty products reach customers. If potential faults could be detected early, it would be possible to rectify or discard the defective products before they cause further issues. To this end, there have been efforts to develop machine learning-based fault prediction models (Kang, 2020; Dogan & Birant, 2021; Bai et al., 2017; Lee et al., 2017; Qian et al., 2022; Hsu & Liu, 2021; Meyes et al., 2019). These models aim to use inspection results to identify defects early in the manufacturing process, thereby preventing costly outcomes.

Although ideally every product would undergo all types of inspections, the need for high productivity and reduced production costs often makes this impractical. Consequently, inspections are categorized into two types: basic and advanced. Basic inspections are essential and conducted on all products to ensure that no product is overlooked. In contrast, due to their higher costs and time requirements, advanced inspections are not feasible for every product and are thus performed only on a selected subset of products.

Fault prediction models utilize these two types of inspection data as input variables. Depending on the range of inspection data used as input variables, there are two model options: a basic model and an advanced model. The basic model utilizes only the results of basic inspections as input variables, whereas the advanced model incorporates results from both basic and advanced inspections. The selection between the basic and advanced models is contingent on whether the product has undergone advanced inspections. Typically, the advanced model offers higher prediction accuracy since it leverages a broader range of input variables.

Given trained basic and advanced models, a product must undergo both basic and advanced inspections to be evaluated by the advanced model, which results in higher inspection costs. However, this also leads to potentially higher performance. In contrast, predictions made by the basic model only require results from the basic inspection, resulting in relatively lower inspection costs and, likely, lower performance. The aim of this research is to develop a fault prediction model that achieves high performance while incurring low inspection costs. Specifically, the goal is to achieve high prediction accuracy using only basic inspection results as input variables, thereby improving the cost-effectiveness of the inspections.

In this study, we propose a method to train an improved basic model utilizing knowledge distillation. Traditionally, knowledge distillation is a technique that transfers knowledge from a model with more complex architecture to a simpler one, primarily aiming to achieve high performance with reduced computational costs (Gou et al., 2021; Hinton et al., 2015). However, in this study, we employ knowledge distillation to achieve high performance with reduced inspection costs. Specifically, after the advanced model is trained, knowledge is transferred to the basic model during its training phase. Consequently, the trained basic model maintains low inspection costs required for inference while enhancing prediction performance. Furthermore, the basic model, improved through knowledge distillation, can be integrated into an active inspection framework (Shim et al., 2021). In this framework, it is used selectively alongside the advanced model, further enhancing the cost-effectiveness of inspections for fault prediction accuracy. We investigate the effectiveness of the proposed method through a case study using a real-world dataset provided by a semiconductor manufacturer.

This paper is structured as follows. In section 2, we introduce related research. In section 3, we explain the proposed method to training basic model leveraging knowledge distillation. In section 4, we present the experimental results of a case study. In section 5, we provide the conclusions of the paper.

## 2. Related work

### *2.1. Machine learning for fault prediction*

Fault prediction within the manufacturing industry has significantly evolved, increasingly leveraging advanced machine learning techniques to enhance accuracy. Early methods primarily utilized statistical techniques and simple predictive models (Kumar et al., 2009; Ma et al., 2009; Psarommatis et al., 2020). However, as computational resources expanded, these approaches have gradually evolved into more sophisticated machine learning techniques (Saadat et al., 2022; Frumosu et al., 2020; Kang et al., 2018). These modern approaches leverage a wide range of product-related information as input variables, such as process parameters, equipment sensor values, and inspection results. The output variable typically determines if a product will be identified as faulty at later stages, including during final shipment inspections or customer inspections (Tercan & Meisen, 2022).

A broad spectrum of learning algorithms has been developed for fault prediction, ranging from basic classifiers such as logistic regression, decision trees, Bayesian classifiers, and k-nearest neighbors, to more robust ensemble models like random forests and gradient boosting (Dogan & Birant, 2021; Kang et al., 2018; Chien et al., 2007; Bai et al., 2019). Additionally, with the rise of deep learning, neural network models have increasingly been adopted for fault prediction (Kang, 2020; Bai et al., 2017; Meyes et al., 2019; Hsu & Liu, 2021; Qian et al., 2022). These models excel at handling complex patterns and processing large datasets efficiently, making them particularly suitable for the intricate demands of modern manufacturing fault detection.

Despite their advantages, all machine learning algorithms necessitate data for both training and inference phases. The training phase requires securing both input and output variable values, while inference requires only input variable values. Data acquisition, especially obtaining inspection results, incurs substantial financial and temporal costs within the manufacturing system (Farooq et al., 2017). Practical constraints often limit the budget available for inspection costs, thereby posing challenges to the widespread application of machine-learning-based fault prediction.

To overcome these challenges, the active inspection framework has been proposed by Shim et al. (2021). Given trained models, this framework improves the cost-effectiveness of inspection costs by selectively using basic and advanced models during the model's inference phase. Herein, we propose a method that applies during the model training phase, utilizing knowledge distillation techniques. The fault prediction model trained via our method can be integrated within active inspection framework during the model inference phase. This integration of methodologies will be described in subsection 3.3.

### 2.2. Knowledge distillation

Knowledge distillation is a machine learning technique originally developed to address the challenges of deploying large, complex models on devices with limited computational power or environments requiring rapid inference (Gou et al., 2021; Hinton et al., 2015). The core principle of knowledge distillation involves transferring the knowledge from a large, sophisticated teacher model to a smaller, simpler student model. This method not only reduces the model size but also aims to retain the performance capabilities of the larger model, thus enabling the smaller model to deliver high accuracy predictions with reduced computational costs.

Various methods of knowledge distillation have been developed over time. The most widely used approach is the soft target method (Hinton et al., 2015), where the outputs probabilities of the teacher model are employed as soft targets for the student model, in addition to hard labels. Another technique is feature-based knowledge distillation, which utilizes the intermediate features or representations learned by the teacher to guide the training of the student model (Wu et al., 2021; Ji et al., 2021). Additionally, relation-based distillation has emerged, focusing on replicating the relational aspects between data points captured by the teacher model (Park et al., 2019).

In the manufacturing industry, knowledge distillation has been applied to streamline the implementation of predictive models across various production aspects. These applications include predictive maintenance (Gong et al., 2023; Xu et al., 2021), defect detection (Tong et al., 2023; Liu et al., 2023), fault prediction (Ji et al., 2022; Zhang et al., 2020), and energy consumption prediction (Li et al., 2023), where quick and efficient processing directly impacts production efficiency and operational costs. By employing distilled models, manufacturers can deploy machine learning model directly onto the production floor, allowing for real-time decision-making.

In this study, we employ knowledge distillation to enhance the practicality of fault prediction in manufacturing system. Unlike traditional applications that primarily focus on computational efficiency, we leverage knowledge distillation to improve the efficiency of inspection costs involved in acquiring input variable values. Our objective is to achieve high prediction accuracy with reduced inspection costs, thereby boosting the practical utility of fault prediction in real-world settings.

## 3. Proposed method

### *3.1. Problem statement*

In this study, we develop a fault prediction model that predicts future faults of each product, utilizing a variety of inspection results as input variables. The inspection process in manufacturing can be divided into two categories: basic inspection and subsequent advanced inspection. Through the basic inspection, we acquire $n$ basic inspection items $X_1^{\text{basic}}, X_2^{\text{basic}}, \ldots, X_n^{\text{basic}}$ for all products. On the contrary, we obtain $m$ advanced inspection items $X_1^{\text{adv}}, X_2^{\text{adv}}, \ldots, X_m^{\text{adv}}$, which are only available for a select number of products due to the high cost of these inspections.

The fault prediction model outputs a binary variable indicating whether a product will ultimately be deemed faulty. Thus, the model essentially functions as a binary classifier. This binary classifier, for any given input variable value, outputs a probability estimate ranging from 0 to 1. This estimate is used as a criterion for determining whether a new product is normal or faulty, and is denoted as a fault score $p$.

Given that advanced inspection items are only available for a select number of products, two types of fault prediction models can be trained: the basic model $F^{\text{basic}}$ and the advanced model $F^{\text{adv}}$ (Shim et al., 2021). The basic model $F^{\text{basic}}$ predicts faults using only the basic inspection items as input variables. It can be represented as follows:

$$p^{\text{basic}} = F^{\text{basic}}(X_1^{\text{basic}}, X_2^{\text{basic}}, \ldots, X_n^{\text{basic}}) \tag{1}$$

The advanced model $F^{\text{adv}}$ predicts faults using both basic and advanced inspection items. It can be represented as follows:

$$p^{\text{adv}} = F^{\text{adv}}(X_1^{\text{basic}}, X_2^{\text{basic}}, \ldots, X_n^{\text{basic}}, X_1^{\text{adv}}, X_2^{\text{adv}}, \ldots, X_m^{\text{adv}}) \tag{2}$$

Naturally, the advanced model $F^{\text{adv}}$ is expected to outperform the basic model $F^{\text{basic}}$, as advanced inspection items can contain significant information crucial for fault prediction. The problem is that advanced model is not always available due to the cost of advanced inspection items.

The objective of this study is to develop a high-performance basic model, based on the assumption that a complete training set, containing both basic and advanced inspection items, is available during the model training phase. As the basic model does not require advanced inspection items as input variables, it reduces the inspection costs necessary for model inference. Consequently, this approach allows us to achieve high fault prediction performance even with lower inspection costs. To accomplish this, we apply knowledge distillation from the advanced model to the basic model. The notations used to describe the proposed method are summarized in Table 1.

| Notation | Description |
|---|---|
| $F^{\text{basic}}$ | The basic model |
| $F^{\text{adv}}$ | The advanced model |
| $X_1^{\text{basic}}, X_2^{\text{basic}}, \ldots, X_n^{\text{basic}}$ | Basic inspection items |
| $X_1^{\text{adv}}, X_2^{\text{adv}}, \ldots, X_m^{\text{adv}}$ | Advanced inspection items |
| $m$ | The number of the advanced inspection items |
| $n$ | The number of the basic inspection items |
| $p^{\text{basic}}$ | Fault score estimated by the basic model. |
| $p^{\text{adv}}$ | Fault score estimated by the advanced model. |
| $y$ | Target variable ; 0 for normal product, 1 for faulty product |

**Table. 1.** Notations used in this paper.

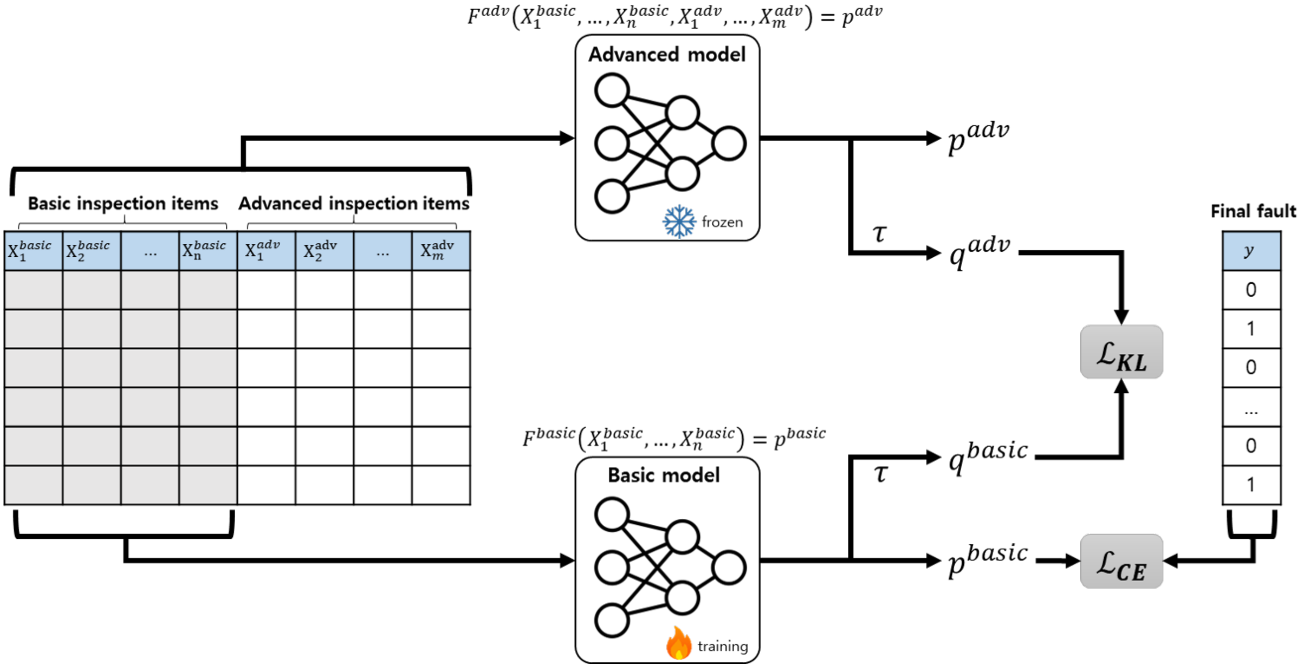

**Fig. 1.** Illustrative description of the proposed method for training a basic model.

### *3.2. Knowledge distillation for training cost-effective basic model*

To create a high-performance basic model, we employ a knowledge distillation technique. While the original purpose of knowledge distillation is to transfer knowledge to a simpler model in terms of architecture, our goal is to transfer knowledge from a model with more input variables to one with fewer, specifically from the advanced model $F^{\text{adv}}$ to the basic model $F^{\text{basic}}$. Ultimately, the distilled basic model can achieve high performance without advanced inspection items.

When using neural networks as fault prediction models, the basic model $F^{\text{basic}}$ is typically trained with the binary cross-entropy loss $\mathcal{L}_{CE}$:

$$\mathcal{L}_{CE} = -y \log(p^{\text{basic}}) - (1-y) \log(1-p^{\text{basic}}) \tag{3}$$

where $p^{\text{basic}}$ is the fault score predicted by the basic model $F^{\text{basic}}$. Specifically, $p^{\text{basic}}$ is calculated by applying a sigmoid function to the logit $z^{\text{basic}}$ from the last layer of the fault prediction model $F^{\text{basic}}$, as follows:

$$p^{\text{basic}} = \frac{1}{1 + \exp(-z^{\text{basic}})} \tag{4}$$

By minimizing $\mathcal{L}_{CE}$, the output of basic model $p^{\text{basic}}$ can be closer to the ground-truth label $y$.

Given a trained advanced model $F^{\text{adv}}$, the Kullback-Leibler divergence loss $\mathcal{L}_{KL}$ is introduced to utilize the knowledge of the advanced model in training the basic model $F^{\text{basic}}$. Here, the distillation temperature $\tau$ is introduced to derive the smoothed probabilities, $q^{\text{basic}}$ and $q^{\text{adv}}$ as follows:

$$q^{\text{basic}} = \frac{1}{1 + \exp(-z^{\text{basic}}/\tau)} \tag{5}$$

$$q^{\text{adv}} = \frac{1}{1 + \exp(-z^{\text{adv}}/\tau)} \tag{6}$$

where $\tau$ serves to adjust the level of smoothness. As $\tau$ increases, $q$ becomes closer to 0.5. Then, $\mathcal{L}_{KL}$ is calculated as follows:

$$\mathcal{L}_{KL} = \tau^2 \left\{-q^{adv} \log(q^{basic}) - (1 - q^{adv}) \log(1 - q^{basic})\right\} \tag{7}$$

Minimizing this loss reduces the difference between the outputs of the basic and advanced models, enabling the basic model to acquire knowledge by mimicking the advanced model. The total loss for training the basic model is expressed as a weighted sum of these two losses:

$$\mathcal{L}_{Total} = (1 - \alpha)\mathcal{L}_{CE} + \alpha\mathcal{L}_{KL} \tag{8}$$

where $\alpha$ balances $\mathcal{L}_{CE}$ and $\mathcal{L}_{KL}$. By minimizing $\mathcal{L}_{Total}$, we obtain the distilled basic model that achieve high performance using only the basic inspection items.

Additionally, we consider the use of non-neural network models as fault prediction models, $F^{\text{basic}}$ and $F^{\text{adv}}$. The concept of knowledge distillation is primarily applied to neural network models. However, for the tabular data we are currently dealing with, other machine learning models, such as random forests, still exhibit high performance (Shwartz-Ziv & Armon, 2022). Therefore, we also apply the concept of knowledge distillation to non-neural network models.

For the implementation of knowledge distillation in non-neural network models, the basic model is constructed as a regression model rather than a binary classification model (Fukui et al., 2019). This regression-based basic model is explicitly trained to estimate the output of the trained advanced model, $F^{\text{adv}}$, representing $p^{\text{adv}}$. The basic model is trained to minimize the mean squared error, $(p^{\text{adv}} - p^{\text{basic}})^2$, thereby aligning its outputs closely with the fault score from the advanced model.

During the inference phase, the output of the basic model, $p^{basic}$, is utilized as the fault score to determine whether a product is normal or faulty.

### *3.3. Integrating into active inspection framework*

The active inspection framework (Shim et al., 2021) aims to improve the accuracy of fault prediction cost-effectively by selectively applying basic and advanced models during the inference phase. Given trained basic model

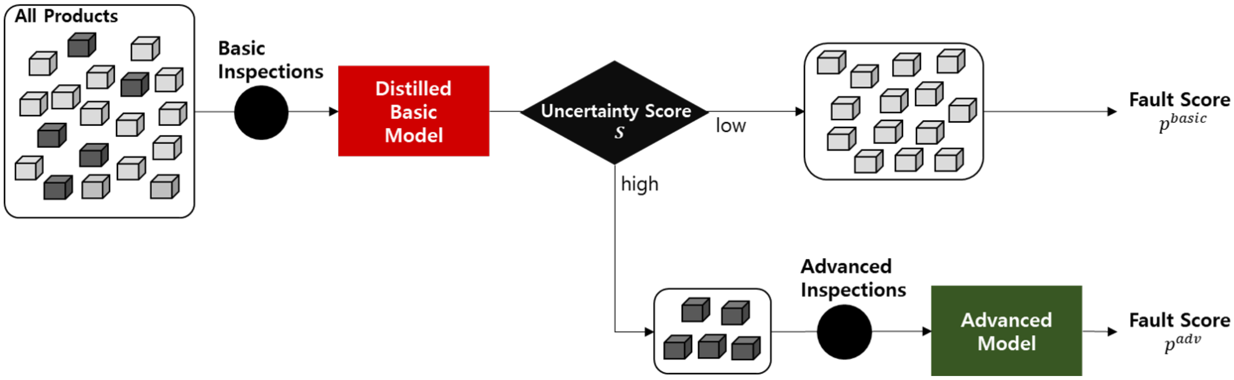

**Fig. 2.** Active inspection with the proposed distilled basic model.

and advanced model, the framework operates as follows: Initially, every product is subjected to basic inspections to collect values on basic inspection items, $X_1^{\text{basic}}, X_2^{\text{basic}}, \ldots, X_n^{\text{basic}}$. These values are then used to compute the uncertainty score $S$, reflecting the predictive uncertainty of the basic model $F^{\text{basic}}$ for each product.

Products exhibiting a high uncertainty score then proceed to advanced inspections to obtain the values of advanced inspection items, $X_1^{\text{adv}}, X_2^{\text{adv}}, \ldots, X_m^{\text{adv}}$. Finally, products that only underwent basic inspections are assessed using the basic model $F^{\text{basic}}$ to determine $p^{\text{basic}}$, while products that underwent both inspections are evaluated using the advanced model $F^{\text{adv}}$ to determine $p^{\text{adv}}$. This framework ensures that only products with a high necessity for the advanced model, as indicated by uncertainty scores, undergo advanced inspections. This selective process enhances the cost-effectiveness of the inspections.

Once the distilled basic model is trained using the method proposed in subsection 3.2, it can be integrated with the active inspection framework during the model inference phase to further enhance the cost-effectiveness of inspections. This integration is illustrated in Fig. 2. Within this framework, both the uncertainty score $S$ and the fault score $p^{\text{basic}}$ are calculated based on the distilled basic model. By selectively utilizing the distilled basic model alongside the advanced model, we can further enhance fault prediction accuracy cost-effectively.

## 4. Case study

### *4.1. Data description*

| Inspection recipe | Number of instances | Number of input variables | |
|---|---|---|---|
| | | Basic inspection items | Advanced inspection items |
| `Recipe1` | 18,471 | 8 | 91 |
| `Recipe2` | 245,006 | 8 | 80 |

**Table. 2.** Dataset description.

To demonstrate the effectiveness of the proposed method, a case study was conducted using a real-world dataset provided by a semiconductor manufacturer in the Republic of Korea. The semiconductor manufacturing process consists of sequential phases: wafer fabrication, wafer test, assembly, and final test. After wafer fabrication, each die on the wafer undergoes electrical characteristic inspections during the wafer test phase. Subsequently, in the assembly phase, the dies are manufactured into final products, which are then subjected to a final test to determine whether they pass or fail based on their functional integrity. Our dataset includes inspection results from the wafer test phase and the status of each die, pass or fail, from the final test phase. Among the various inspections conducted during the wafer test phase, some are performed under severe conditions, making them time-consuming and costly. Therefore, these inspections are carried out only on a sampled subset of dies and are therefore categorized as advanced inspections. The remaining inspections are categorized as basic inspections. The objective of this case study is to predict the failures in the final test based on the inspection results from the wafer test.

We utilized two datasets corresponding to different inspection recipes of wafer test, `Recipe1` and `Recipe2`. Both recipes contain identical basic inspection items, but `Recipe1` includes a greater number of advanced inspection items than `Recipe2`. Each product corresponds to an instance, and the total number of instances is higher in `Recipe2`. Each instance contains information about whether it is faulty or normal in the final test, represented as 1 or 0, with a fault rate of less than 1%. Inspection items serve as the input variables for the model, and the final test fault status is the output variable. A detailed description of the dataset is shown in Table 2.

### *4.2. Experimental setting*

We randomly divided the dataset into two parts: 50% for training and 50% for testing. The training set was used to train both the basic and advanced models, while the test set was utilized to evaluate the fault prediction performances.

For the fault prediction models, we employed two learning algorithms: neural network (NN) and random forest (RF). The NN model was configured with a single hidden layer comprising 10 neurons and utilized the ReLU activation function. The Adam optimizer was employed to update the parameters. We set aside a randomly selected 20% of the training set as a validation set, and the model that achieved the highest performance on this validation set was selected, with a cap at a maximum of 200 epochs. For knowledge distillation (KD) applied to the basic model, we set the $\tau$ at 2 and the $\alpha$ at 0.5. Regarding the RF models, each consisted of 500 trees, with each tree requiring a minimum of 10 instances for a split decision. For the advanced model and the basic model without KD, the RF classifier was employed using the Gini impurity as the splitting criterion. For the basic model trained with KD, an RF regressor was used, employing mean squared error as the splitting criterion.

In simulating the active inspection framework, we employed two different methods for calculating the uncertainty score $S$: margin and biased margin.

- margin: This method calculates the inverse of the absolute difference between the fault score $p^{\text{basic}}$ and the normal score $1 - p^{\text{basic}}$ (Scheffer et al., 2001):

$$S = \frac{1}{|p^{\text{basic}} - 0.5|} \tag{9}$$

- biased margin: This method modifies the margin technique to account for class imbalance. It assigns higher uncertainty to instances near the fault rate $FR$, which represents the percentage of faulty products in the training dataset (Attenberg & Ertekin, 2013). It is calculated as:

$$S = \begin{cases} \dfrac{p^{\text{basic}}}{FR}, & \text{if } p^{\text{basic}} \leq FR; \\ \dfrac{1 - p^{\text{basic}}}{1 - FR}, & \text{otherwise,} \end{cases} \tag{10}$$

Additionally, random sampling served as a baseline for comparison.

We defined advanced inspection rates as the proportion of products subjected to advanced inspections relative to the total number of products in the test set, reflecting the incurred inspection costs. We analyzed performance variations across different advanced inspection rates, which varied from zero to one in increments of 0.1. A rate of zero means that only basic inspections are performed on all products, whereas a rate of one indicates that every product undergo advanced inspections.

To summarize, the following approaches were compared in the experiments:

- **Random**: Uses a basic model trained without KD, and selects instances for advanced inspection through random sampling without utilizing uncertainty scoring.
- **Margin**: Uses a basic model trained without KD, applying the active inspection framework with margin as the uncertainty score.
- **BiasedMargin**: Uses a basic model trained without KD, applying the active inspection framework with biased margin as the uncertainty score.
- **KD_Random**: Uses a basic model trained with KD, and selects instances for advanced inspection through random sampling without utilizing uncertainty scoring.
- **KD_Margin**: Uses a basic model trained with KD, applying the active inspection framework with margin as the uncertainty score.
- **KD_BiasedMargin**: Uses a basic model trained with KD, applying the active inspection framework with biased margin as the uncertainty score.

The effectiveness of each method was assessed using the area under the receiver operating characteristic curve (AUROC). Each experiment was independently repeated 30 times, and the average results are presented.

### *4.3. Results and discussion*

The comparative results of the fault prediction models are displayed in Table 3. As shown in the table, for both `Recipe1` and `Recipe2` datasets, and employing NN and RF as the underlying models, the basic model trained with KD consistently outperformed the conventional basic model trained without KD. Additionally, the models trained with KD exhibited smaller standard deviations, suggesting more stable training outcomes. Notably, in the `Recipe1` dataset, the basic model trained with KD even surpassed the performance of the advanced model. In the relatively smaller `Recipe1` dataset, it is conjectured that KD acted as an effective regularizer, contributing to these results (Yuan et al., 2020).

| | | NN | RF |
|---|---|---|---|
| `Recipe1` | Basic model without KD | 0.7097 ± 0.0574 | 0.7173 ± 0.0272 |
| | Basic model with KD | **0.7346 ± 0.0527** | **0.7542 ± 0.0220** |
| | Advanced model | 0.7327 ± 0.0485 | 0.7477 ± 0.0287 |
| `Recipe2` | Basic model without KD | 0.7282 ± 0.0101 | 0.6955 ± 0.0127 |
| | Basic model with KD | **0.7297 ± 0.0085** | **0.7177 ± 0.0115** |
| | Advanced model | 0.7374 ± 0.0106 | 0.7324 ± 0.0117 |

**Table. 3.** Comparison results of the fault prediction models in AUROC. (mean ± standard deviation)

The comparative results of the fault prediction models integrated within the active inspection framework are depicted in Fig. 3. The x-axis represents the advanced inspection rate, which also reflects the associated inspection costs, while the y-axis denotes the fault prediction performance derived from the framework. An x-axis value of 0 means that only the basic model was used for all products, whereas a value of 1 indicates that only the advanced model was employed for all products. Points closer to the top left indicate higher performance with lower inspection costs, thus demonstrating better cost-effectiveness.

The performances of the active inspection framework utilizing the conventional basic model are represented by dashed lines. For the **Random**, the fault prediction performance increased linearly with the advanced inspection rate. Meanwhile, **Margin** and **BiasedMargin**, which utilize the uncertainty score, demonstrated greater effectiveness in terms of inspection costs compared to **Random**, aligning with the results in the original paper (Shim et al., 2021).

The performances of the active inspection framework integrated with the basic model trained with KD are shown with solid lines. The performance at an advanced inspection rate of 0 reflected the enhanced performance of the basic model due to KD, as previously confirmed in Table 3. This integration allowed **KD_Random**, **KD_Margin**, and **KD_BiasedMargin** to outperform **Random**, **Margin**, and **BiasedMargin** across all advanced inspection rates respectively. **KD_Random** exhibited a performance trend that interpolates between the basic and advanced model performances, similar to **Random**. Meanwhile, **KD_Margin** and **KD_BiasedMargin** displayed superior cost-effectiveness compared to **Random**. Specifically, in `Recipe1` using NN (Fig. 3(a)), **KD_BiasedMargin** outperformed other baseline methods. Performance was better at advanced inspection rates ranging from 0.1 to 0.9 than at the extremes of 0 or 1, implying that selectively utilizing the basic model trained with KD and the advanced model can yield better results than exclusively using one model. In `Recipe2` using NN (Fig. 3(c)), **KD_Margin** outperformed other methods. In `Recipe1` using RF (Fig. 3(b)), **Margin** at an advanced inspection rate of 0.1 demonstrated the best performance, and in `Recipe2` using RF (Fig. 3(d)), **BiasedMargin** showed superior performance on the whole.

Additionally, some results showed the active inspection framework performing worse in terms of cost-effectiveness than the random baselines, such as with **KD_BiasedMargin** in `Recipe2` using NN. This indicates that the appropriate uncertainty score for using the distilled basic model should be investigated further, considering the characteristics of each dataset and learning algorithm.

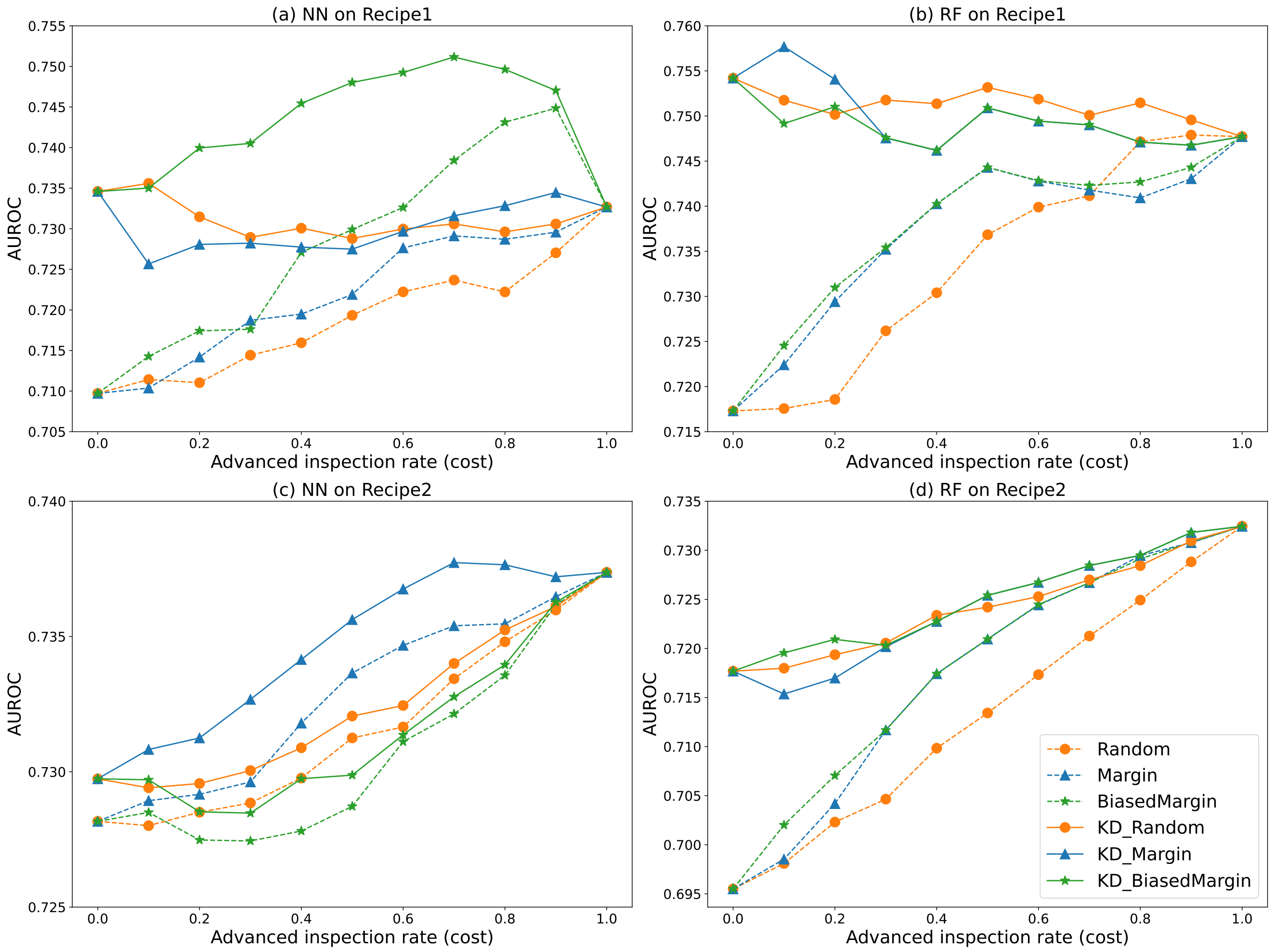

**Fig. 3.** Comparison of average performance with active inspection framework

In summary, employing proposed KD during the training phase of the basic model has improved fault prediction performance without additional inspection costs. Further integrating this enhanced model within the active inspection framework has shown improved cost-effectiveness in inspection.

## 5. Conclusion

To accurately predict faulty products through a machine learning model, it is essential to conduct inspections to obtain values for input variables. Due to cost constraints, advanced inspections are typically limited to a subset of sampled products. In this study, we proposed a knowledge distillation-based method to train an enhanced basic model that does not require additional results from advanced inspections. By transferring knowledge from the advanced model to the basic model during the training process, we were able to improve the prediction performance of the basic model while maintaining low inspection costs. Integrating this basic model into the active inspection framework further enhanced the cost-effectiveness of inspections for fault prediction accuracy. The effectiveness of this methodology was demonstrated through a case study using a real-world dataset provided by a semiconductor manufacturer.

For future work, we plan to explore various knowledge distillation techniques to train more cost-effective fault prediction models. Specifically, we plan to employ a teaching assistant model to bridge the information gap between advanced and basic inspections. Additionally, rather than a one-way transfer, we aim to facilitate mutual knowledge sharing between the basic and advanced models to leverage their combined strengths. Lastly, we intend to refine the sampling process of the active inspection framework, ensuring that the most appropriate models are used for each product considering the inspection budget.

## References

Attenberg, J., & Ertekin, S. (2013). Class imbalance and active learning. *Imbalanced Learning: Foundations, Algorithms, and Applications*, (pp. 101–149).

Bai, Y., Li, C., Sun, Z., & Chen, H. (2017). Deep neural network for manufacturing quality prediction. In *Proceedings of the Prognostics and System Health Management Conference* (pp. 1–5). IEEE.

Bai, Y., Sun, Z., Zeng, B., Long, J., Li, L., de Oliveira, J. V., & Li, C. (2019). A comparison of dimension reduction techniques for support vector machine modeling of multi-parameter manufacturing quality prediction. *Journal of Intelligent Manufacturing*, *30*, 2245–2256.

Chien, C.-F., Wang, W.-C., & Cheng, J.-C. (2007). Data mining for yield enhancement in semiconductor manufacturing and an empirical study. *Expert Systems with Applications*, *33*, 192–198.

Dogan, A., & Birant, D. (2021). Machine learning and data mining in manufacturing. *Expert Systems with Applications*, *166*, 114060.

Farooq, M. A., Kirchain, R., Novoa, H., & Araujo, A. (2017). Cost of quality: Evaluating cost-quality trade-offs for inspection strategies of manufacturing processes. *International Journal of Production Economics*, *188*, 156–166.

Frumosu, F. D., Khan, A. R., Schiøler, H., Kulahci, M., Zaki, M., & Westermann-Rasmussen, P. (2020). Cost-sensitive learning classification strategy for predicting product failures. *Expert Systems with Applications*, *161*, 113653.

Fukui, S., Yu, J., & Hashimoto, M. (2019). Distilling knowledge for non-neural networks. In *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* (pp. 1411–1416). IEEE.

Gong, R., Wang, C., Li, J., & Xu, Y. (2023). Lightweight fault diagnosis method in embedded system based on knowledge distillation. *Journal of Mechanical Science and Technology*, *37*, 5649–5660.

Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision*, *129*, 1789–1819.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, .

Hsu, C.-Y., & Liu, W.-C. (2021). Multiple time-series convolutional neural network for fault detection and diagnosis and empirical study in semiconductor manufacturing. *Journal of Intelligent Manufacturing*, *32*, 823–836.

Ji, M., Heo, B., & Park, S. (2021). Show, attend and distill: Knowledge distillation via attention-based feature matching. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 7945–7952). volume 35.

Ji, M., Peng, G., Li, S., Cheng, F., Chen, Z., Li, Z., & Du, H. (2022). A neural network compression method based on knowledge-distillation and parameter quantization for the bearing fault diagnosis. *Applied Soft Computing*, *127*, 109331.

Kang, S. (2020). Joint modeling of classification and regression for improving faulty wafer detection in semiconductor manufacturing. *Journal of Intelligent Manufacturing*, *31*, 319–326.

Kang, S., Kim, E., Shim, J., Chang, W., & Cho, S. (2018). Product failure prediction with missing data. *International Journal of Production Research*, *56*, 4849–4859.

Kumar, S., Chow, T. W., & Pecht, M. (2009). Approach to fault identification for electronic products using mahalanobis distance. *IEEE Transactions on Instrumentation and Measurement*, *59*, 2055–2064.

Lee, K. B., Cheon, S., & Kim, C. O. (2017). A convolutional neural network for fault classification and diagnosis in semiconductor manufacturing processes. *IEEE Transactions on Semiconductor Manufacturing*, *30*, 135–142.

Li, Y., Hu, F., Liu, Y., Ryan, M., & Wang, R. (2023). A hybrid model compression approach via knowledge distillation for predicting energy consumption in additive manufacturing. *International Journal of Production Research*, *61*, 4525–4547.

Liu, J., Li, H., Zuo, F., Zhao, Z., & Lu, S. (2023). Kd-lightnet: A lightweight network based on knowledge distillation for industrial defect detection. *IEEE Transactions on Instrumentation and Measurement*, .

Ma, M.-D., Wong, D. S.-H., Jang, S.-S., & Tseng, S.-T. (2009). Fault detection based on statistical multivariate analysis and microarray visualization. *IEEE Transactions on industrial informatics*, *6*, 18–24.

Meyes, R., Donauer, J., Schmeing, A., & Meisen, T. (2019). A recurrent neural network architecture for failure prediction in deep drawing sensory time series data. *Procedia Manufacturing*, *34*, 789–797.

Park, W., Kim, D., Lu, Y., & Cho, M. (2019). Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3967–3976).

Psarommatis, F., May, G., Dreyfus, P.-A., & Kiritsis, D. (2020). Zero defect manufacturing: state-of-the-art review, shortcomings and future directions in research. *International journal of production research*, *58*, 1–17.

Qian, H., Sun, B., Guo, Y., Yang, Z., Ling, J., & Feng, W. (2022). A parallel deep learning algorithm with applications in process monitoring and fault prediction. *Computers and Electrical Engineering*, *99*, 107724.

Saadat, R., Syed-Mohamad, S. M., Azmi, A., & Keikhosrokiani, P. (2022). Enhancing manufacturing process by predicting component failures using machine learning. *Neural Computing and Applications*, *34*, 18155–18169.

Scheffer, T., Decomain, C., & Wrobel, S. (2001). Active hidden Markov models for information extraction. In *Proceedings of the International Symposium on Intelligent Data Analysis* (pp. 309–318).

Shim, J., Kang, S., & Cho, S. (2021). Active inspection for cost-effective fault prediction in manufacturing process. *Journal of Process Control*, *105*, 250–258.

Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, *81*, 84–90.

Tercan, H., & Meisen, T. (2022). Machine learning and deep learning based predictive quality in manufacturing: a systematic review. *Journal of Intelligent Manufacturing*, *33*, 1879–1905.

Tong, G., Li, Q., & Song, Y. (2023). Two-stage reverse knowledge distillation incorporated and self-supervised masking strategy for industrial anomaly detection. *Knowledge-Based Systems*, *273*, 110611.

Wu, Y., Rezagholizadeh, M., Ghaddar, A., Haidar, M. A., & Ghodsi, A. (2021). Universal-kd: Attention-based output-grounded intermediate layer knowledge distillation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 7649–7661).

Xu, Q., Chen, Z., Wu, K., Wang, C., Wu, M., & Li, X. (2021). Kdnet-rul: A knowledge distillation framework to compress deep neural networks for machine remaining useful life prediction. *IEEE Transactions on Industrial Electronics*, *69*, 2022–2032.

Yuan, L., Tay, F. E., Li, G., Wang, T., & Feng, J. (2020). Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3903–3911).

Zhang, W., Biswas, G., Zhao, Q., Zhao, H., & Feng, W. (2020). Knowledge distilling based model compression and feature learning in fault diagnosis. *Applied soft computing*, *88*, 105958.