JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 30 No. 2, pp. 1-11, February 2025
https://doi.org/10.9708/jksci.2025.30.02.001

Machine Learning-Based Analysis of Semiconductor Production Schedule Logs: A Hybrid Approach Using Clustering and Decision Tree

Raekyung Ahn*, Seok-Won Lee**

*Student, Dept. of Artificial Intelligence, Ajou University, Suwon, Korea
**Professor, Dept. of Software and Computer Engineering, Dept. of Artificial Intelligence, Ajou University, Suwon, Korea

[Abstract]

Semiconductor production management becomes increasingly challenging due to the growing complexity of products, raising issues with the inefficiency of on-site scheduling. The complexity of the production environment and the multitude of variables make it difficult for traditional scheduling methods to effectively respond to real-time production changes, limiting the ability to achieve optimal productivity and quality. In this context, production experts are demanding solutions that provide insights into production direction and operational status. Based on the requirements, I conducted research on a solution that uses machine learning to analyze actual semiconductor Fab log data, identifying and visualizing decision-making factors that impact scheduling. This study is significant in that it suggests a direction for future schedulers and provides a foundation for building an Autonomous Fab.

▸ **Key words**: Decision making, unsupervised learning, decision tree, correlation analysis, hybrid learning, scheduling, semiconductor production, manufacturing scheduling

[요 약]

반도체 생산관리는 제품 복잡도의 증가에 따라 어려워지고 있지만, 현장 스케줄링은 효율적이지 않다는 문제가 제기되고 있다. 생산 환경의 복잡성과 다양한 변수들로 인해 기존의 스케줄링 방식은 실시간으로 변화하는 생산 조건에 효과적으로 대응하기 어렵고, 이로 인해 최적의 생산성과 품질을 달성하는 데 한계가 있다. 이에 따라 생산 전문가들은 생산 방향과 운영 상태를 파악할 수 있는 솔루션을 요구하고 있다. 필자는 현업의 요구사항을 바탕으로 실제 반도체 Fab의 로그 데이터를 머신 러닝을 통해 분석하여 스케줄에 영향을 미치는 의사결정 요소를 파악하고 시각화 할 수 있는 솔루션에 대해 연구하였다. 이는 향후 스케줄러의 방향성을 제시하여 Autonomous Fab을 구축하는데 기반을 제공한다는 점에서 의의가 있다.

▸ **주제어**: 의사결정, 비지도 학습, 의사결정 트리, 상관관계 분석, 하이브리드 학습, 스케줄링, 반도체 생산, 생산 스케줄링

• First Author: Raekyung Ahn, Corresponding Author: Seok-Won Lee
*Raekyung Ahn (rmoress@ajou.ac.kr), Dept. of Artificial Intelligence, Ajou University
**Seok-Won Lee (leesw@ajou.ac.kr), Dept. of Software and Computer Engineering, Dept. of Artificial Intelligence, Ajou University
• Received: 2024. 12. 09, Revised: 2025. 02. 03, Accepted: 2025. 02. 03.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

4차 산업혁명의 시작과 더불어 생산관리 자동화의 중요성은 더욱 대두되고 있다[1]. 또한, 빅데이터를 기반으로 한 4차 산업혁명의 기술들은 생산관리의 다양한 영역에 적용되고 있다. 그 결과 생산 효율,제품 품질, 공정 최적화 등의 분야에서 괄목한 만한 성과를 보이고 있다[2][3]. 인공지능은 빅데이터를 활용하여 공정의 비정상적인 패턴을 식별하고, 최적의 생산 계획을 자동으로 생성하는 것을 목표로 하고 있다. 또한 기계의 센서를 분석하여 이상을 탐지하는 IoT는 실시간 데이터를 수집하여 공정을 모니터링하고, 향후 발생할 고장을 예측하여 유지보수와 자원관리를 가능하게 한다[4].

MSS(Manufacturing Scheduling System)은 생산계획에 따라 생산 물량을 공정별로 할당하고 작업순서를 관리하는 스케줄러와 실시간으로 최적의 작업대상을 선정하는 디스패처로 구성된다. 특히 반도체 생산은 각기 다른 공정 및 설비 특성을 보이는 프로세스의 집합으로 이루어져 복잡도가 높다고 알려져 있다. 그렇기 때문에 특정 교육을 받은 소수의 전문가만의 스케줄러의 영향도를 분석할 수 있고, 데이터의 양이 방대하고 공정 특성이 각기 다르기 때문에 하나의 생산라인을 분석하는 데는 수시간이 소요되고 기존 데이터분석 툴인 엑셀이나 데이터 시각화 툴인 Spotfire로 라인 전체를 보는 것은 불가능에 가깝다.

이러한 한계를 극복하기 위해 본 연구에서는 실제 반도체 생산라인의 복잡한 스케줄을 분석하고, 그 주요 결정 인자들을 기계학습 기법을 통해 추출하여 의사결정 트리 모델을 통해 가시화하는 것이다.

하지만 최근 연구(강화학습 기반의 스케줄[8], 딥러닝 및 강화학습에 의한 스케줄링[9])에서는 기계학습을 통한 의사결정 모델이 블랙박스로 작용하고 있다. 딥러닝과 같은 고급 기계학습 기술이 의사결정을 내리지만, 그 이유나 근거를 사용자가 명확히 이해하기 어렵다. 즉, 기계학습 모델의 해석 가능성 부족이 산업 현장에서의 신뢰성을 낮추고, 실제 적용을 방해하는 주요 요인으로 작용할 수 있다.

본 연구는 이러한 문제를 해결하기 위해, 반도체 생산라인에서 발생하는 스케줄 로그 데이터를 비지도 학습으로 분석하고, 의사결정 트리 모델을 사용하여 의사결정 과정을 명확하게 시각화하고, 비전문가도 이해할 수 있도록 기계학습 결과를 도출하는 데 중점을 두었다. 구체적으로, 본 연구는 다음과 같은 목적을 가진다:

- 반도체Fab 스케줄을 결정하는 주요 인자 파악
- 자동화된 의사결정의 규칙과 특징변수 값의 Threshold 및 휴리스틱과의 차이점 파악
- 고객 Demand의 변화가 스케줄링에 끼치는 영향 분석
- 비전문가도 이해할 수 있는 기계학습 결과 산출

본 연구를 통해 복잡한 반도체 생산 스케줄의 주된 결정 인자를 의사결정 트리를 통해 나타냄으로써 NP-hard 문제로 불리는 Job Shop 방식의 생산라인도 몇 가지 주성분으로 가시화 할 수 있다는 부분에서 생산관리 업무를 좀 더 편리하게 할 수 있을 것으로 기대한다. 또한, 사람이 휴리스틱으로 조절하는 스케줄 팩터의 작동방식과 자동화된 의사결정의 연결성을 보여주어 스케줄러가 반도체 생산 전문가의 의지대로 움직이고 있다는 부분을 증명하여 향후 무인 자동화된 생산라인을 구축하는 데 기여할 수 있다. 이를 통해 현재 단순 스케줄 중심의 생산라인을 향후 시뮬레이션 및 강화학습 기반의 Autonomous Fab을 구축을 위한 초석을 다지는 중요한 기회를 제공할 것이다.

II. Preliminaries

1. Related works

1.1 Scheduling

스케줄링은 제한된 리소스를 효율적으로 할당하여 작업을 수행하는 과정을 계획, 조정, 최적화하는 것을 말한다. 스케줄링은 제조업, 프로젝트 관리, 교통, 컴퓨터 운영체제 등의 다양한 분야에서 사용되며, 그 목적과 방법은 적용 분야와 상황에 따라 다양화 된다. 특히, 제조업 에서의 스케줄링은 공장의 복잡한 문제를 해결하는 데 사용 되고 있으며, 다양한 알고리즘과 기법이 활발히 연구되고 있는 분야다. 또한 생산 제품의 복잡성이 점점 더 증가함에 따라 스케줄링은 학문적인 분야와 실제 생산이 이루어지는 공장에서도 중요한 문제이다. 제조업의 스케줄링은 여러 작업에 대한 일련의 프로세스를 기계에 순차적으로 할당하는 최적의 순서를 결정하는 것을 목표로 한다[6]. 이 때 스케줄러는 작업의 제약 조건을 지켜야 하는 것 뿐만 아니라, 설비의 효율성 증대, 제품 품질 향상, 예기치 않은 설비 문제가 발생했을 때 빠르게 대응할 수 있는 유연성을 갖추어야 한다.

1.2 Job Shop Scheduling Problem

Job Shop 스케줄링 문제는 제조업의 복잡도와 난이도가 올라감에 따라 많은 연구 분야에서 많은 관심을 받고 있다. 이는 제조 생산시스템에 대한 효과적인 스케줄링 전략을 도출하는 프레임워크로 활용할 수 있기 때문이며, NP-Hard 조합의 최적화 문제에 대한 솔루션으로 사용되어 최적화 알고리즘을 개발하는 데 있어 중요한 역할을 한다[7]. 반도체 제조산업에서 스케줄러 및 디스패쳐는 다양한 방식으로 쓰이고 있다. Job Shop 문제의 스케줄링과 디스패칭에 많은연구가 있지만, 대부분의 연구는 반도체 제조공장인 Fab의 구조와 웨이퍼의 생산 납기에 기반하여 다음 순서의 설비를 선택하는 것에 중점을 두고 있다. 최근 강화학습의 등장으로 MDP(Markov Decision Process)를 기반으로 한 자동차 제조산업에서의 실시간 의사결정[8], 딥러닝 및 강화학습의 결과인 보상(Reward)을 최대화하는 스케줄 의사결정[9][10] 등의 연구가 제안되고 있다.

하지만 딥러닝 기반 의사결정 모델은 생산라인에서 각 제품 프로세스의 순서가 왜 선택되었는지 사용자가 이해하기 어려운 블랙박스 형태로 되어 있다. 즉, 기계학습, 특히 딥러닝을 통해 도출된 의사결정의 결과는 이해하기 어려운 부분이 많다는 문제점이 있다. 따라서, 기계학습을 통한 의사결정을 수학적으로 이해하는 방법론이 필요하며[11], 이를 통해 기계학습으로 이루어진 의사결정의 이해 가능성을 높이고, 현장에 적합한 실용적인 적용이 가능하도록 해야 한다.

본 연구는 이러한 문제를 해결하기 위해 실제 생산라인에서 발생한 스케줄 로그를 기계학습으로 분석하여 의사결정 과정에 대한 설명 가능성을 보여주는 데 초점을 맞추었다. 이는 기계학습 모델을 기업에 적용할 때 발생하는 이해 부족 문제를 해결하고, 기술적 신뢰성을 높여 현업에서의 실질적인 활용 가능성을 극대화 하는데 기여할 것이다.

III. The Proposed Scheme

반도체 생산 자동화는 MES(Manufacturing Execution System)을 기반으로 작동하며, 제조 공정의 효율성을 높이기 위해 실시간으로 데이터를 생성, 추적하고 제어한다[12]. MES 로그는 실시간으로 생성되며 각각의 시스템 모듈에서 서로 다른 헤더로 그 성질이 구분된다. 스케줄링 및 디스패칭 결과도 MES에서 로그 형태로 실시간 통신하며, 주기적으로 로그의 일부를 파싱(Parsing)하여 데이터베이스에 저장하는 형태로 지나간 과거 이력을 추적할 수 있는 Traceability를 보장한다[13].

이러한 Full Automation 기반 생산 방식에도 불구하고 생산환경은 끊임없이 변화한다. 반도체 산업은 국제 경쟁이 점점 더 강해지고있는 분야로써 기업들은 높은품질의 제품을 더 빠르고 저렴하게 생산하기 위하여 많은 노력을 기울이고있다. 더욱이 급변하는 시장 상황과고객의 요구, 기술 변화에 빠르고 유연하게 대응하는 것이 기업의 핵심 역량이다.

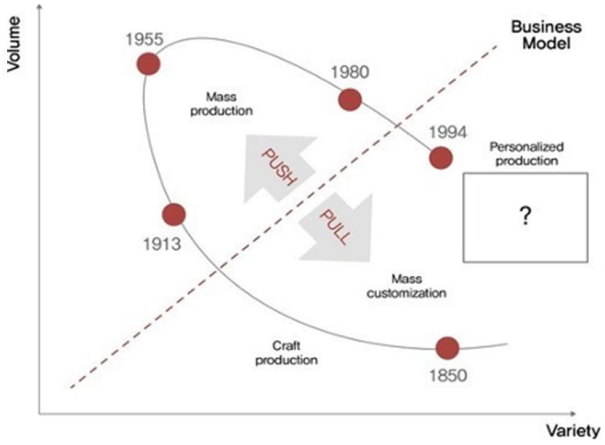

Fig. 1. Production paradigms according to production variety and volume.[10]

Fig. 1, Fig. 2는 생산 다양성과 생산량에 따른 생산 모델의 진화를 보여주는 그래프이다. 이 그래프는 Mass Production에서 Mass Customization으로, 그리고 Personalized Production으로의 전환을 나타내며, Push와 Pull의 생산 방식 변화를 시각적으로 나타낸다. 이러한 변화는 시장의 요구와 기업의 생산 전략에 맞춰 조정되며, 반도체 생산의 유연성을 더욱 높이고자 하는 목표와 일치한다. 시장 변화에 맞춘 생산 계획 조정은 Fab의 스케줄 방향성을 결정짓는 중요한 요소로 작용한다.

Fig. 3은 반도체 생산 관리 흐름을 나타내고 있으며, 시장의 변화에 따라 생산 계획이 끊임없이 변화하는 과정을 보여준다. 생산 계획의 변화는 Fab의 스케줄 방향성에 영향을 미친다. 특히 Push와 Pull의 생산 모델 변화에 따라 고객 맞춤형 생산과 대량 생산의 조화가 중요한 요소로 작용한다. 이처럼 빠르게 변화하는 시장과 기술에 맞춰 Fab의 스케줄이 유연하게 조정되어야 한다.

반도체 생산 자동화는 MES를 기반으로 작동하며, 제조 공정의 효율성을 높이기 위해 실시간으로 데이터를 생성, 추적하고 제어한다[12]. MES 로그는 실시간으로 생성되며 각각의 시스템 모듈에서 서로 다른 헤더로 그 성질이 구분

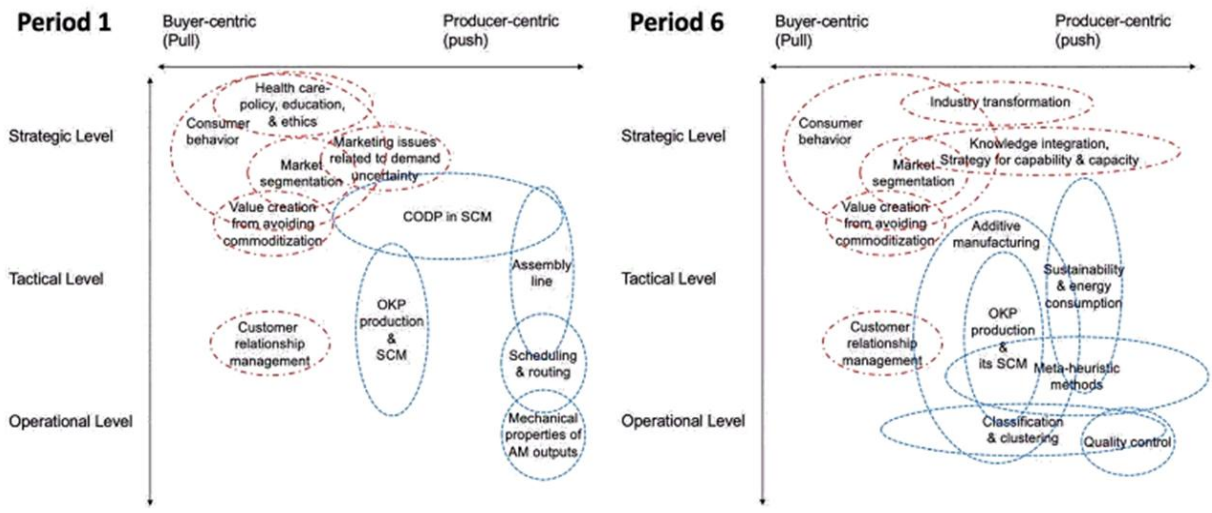

Fig. 2. Paradigm shift in the mass customisation research area between Period 1 (1992 ~ 1996) and Period 6 (2017 ~ 2019).[10]

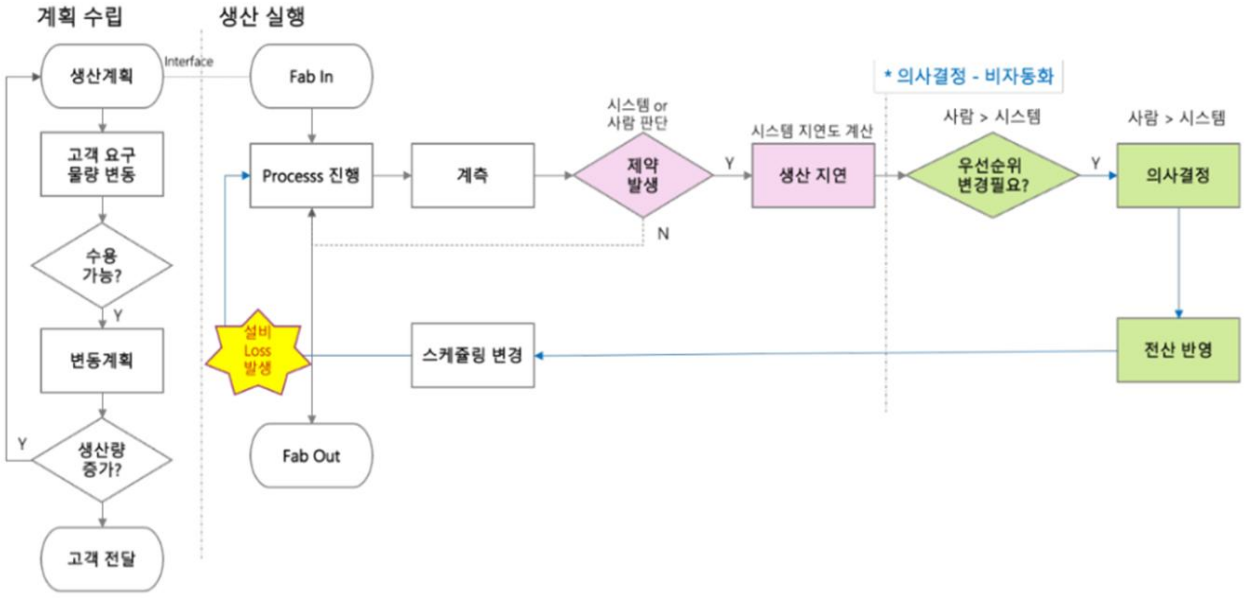

Fig. 3. Semiconductor Production Management Flowchart

된다. 스케줄링 및 디스패칭 결과도 MES에서 로그 형태로 실시간 통신하며, 주기적으로 로그의 일부를 파싱(Parsing)하여 데이터베이스에 저장하는 형태로 지나간 과거 이력력을 추적할 수 있는 Traceability를 보장한다[13].

이러한 Full Automation 기반 생산방식에도 불구하고 생산환경은 끊임없이 변화한다. 반도체 산업은 국제 경쟁이 점점 더 강해지고 있는 분야로써 기업들은 높은 품질의 제품을 더 빠르고 저렴하게 생산하기 위하여 많은 노력을 기울이고 있다. 더욱이 급변하는 시장 상황과 고객의 요구, 기술 변화에 빠르고 유연하게 대응하는 것이 기업의 핵심 역량이다. Fig. 3은 반도체 생산관리 흐름을 나타내고 있다. 시장의 변화에 맞춰서 생산 계획은 끊임없이 변화하여 이는 결과적으로 Fab의 스케줄 방향성에 영향을 끼친다.

Domain Experts인 생산관리 실무자는 생산계획의 변동에 대해 제품의 프로세스별 긴급도를 조절하여 시스템에 입력하는 방법으로 일을 하는데 그들이 Input으로 넣는 파라메터가 생산라인에 어떻게, 얼만큼의 영향을 끼치는지 주기적으로 모니터링 하면서 값을 조절하고 있다. 하나의 생산라인의 스케줄링을 한번에 분석하는 것은 공정 및 설비 특성이 각기 다르기 때문에 굉장히 어려운 작업이다. 게다가 스케줄링을 담당하는 시스템 전문가가 분석 툴을 만들어 확인하고 있지만 데이터의 양이 방대하고 설비별 설정된 파라메터의 값이 서로 다르기 때문에 전체 생산

라인을 한눈에 보기는 불가능에 가깝다. 본 연구는 반도체 생산라인의 이러한 한계점을 극복하기 위해 기계학습을 활용하여 생산 현장에서 발생하는 실제 데이터를 기반으로 생산 스케줄을 좌우하는 주요 인자를 파악하고자 한다. 본 연구에서 생산 스케줄링을 결정하는 주요 영향 인자가 무엇인지 분석하고, 각 인자 별 영향도에 대해 파악할 수 있는 솔루션을 마련하고자 하였다. 이러한 연구 목적을 달성하기 위한 연구 문제는 다음과 같다.

RQ1. 현업 전문가의 휴리스틱(경험)에 기반한 의사결정이 실제 자동화된 생산라인에 어떻게 반영이 되는가?

Job Shop방식 생산 스케줄링에 있어 휴리스틱은 다양한 방식이 있다[14]. 아래는 대표적인 5가지 생산 방식에 대해 나열하였다.

(1) FIFO(First In First Out) : 먼저 도착한 Job이 먼저 스케줄이 되는 원칙이다.

(2) EDD(Earliest Due Date) : 공급 납기가 빠른Job부터 스케줄이 되는 방식이다.

(3) NJF(Nearest Job First) : Job이 저장된 위치를 보고 가장 가까이 위치한 Job을 우선 할당하는 방식이다.

(4) SMED(Single-Minute Exchange of Die): 생산설비의 전환 시간을 최소화하여, 작업 간의 전환을 빠르게 하여 생산성을 향상시키는 스케줄 방식이다[14].

(5) FLNQ(Fewest Lots in Next Queue) : 설비 사용율에 초점을 맞춘 방식으로, 다음 순서의 설비의 부하가 적은 순서대로 스케줄하는 방식이다.

반도체 생산라인에서는 위와 같이 나열된 방식이 혼합되어 사용되고 있으며, 생산경험이 많은 반도체 생산 전문가가 설정한 우선순위로 적용되고 있다. 관리자에 따라 다른 방식으로 우선순위가 선정되기 때문에 생산하는 제품과 그 구성에 따라 운영하는 방식에 차이가 있어 운영자의 관점에 따라 주관적인 개입이 발생하고 있다. 본 연구에서는 생산전문가가 생각하는 우선순위와 실제반도체 생산라인의 스케줄링 결과를 기계학습을 통해 분석하는 솔루션을 통해 사람과 시스템 작동에서의 차이를 제시하고자 한다. 이를 통해 스케줄러가 관리자가 원하는 방향대로 생산을 지원하고있음을 가시화하여, 생산관리에 있어서 사람의 개입을 최소화하여 자동화 공장의 완성도를 높이고자 한다.

RQ2. 사람이 입력한 의사결정 파라메터가 최종 고객 Demand 만족에 기여하는 바는 얼마나 되는가? 만약 이를 머신 러닝으로 분석하게 된다면 어떤 방법이 적합하겠는가?

생산관리의 최종 목적은 고객이 원하는 제품을 원하는 시기에 생산하여 납품하는 것이다. 생산진도 전문가는 변동성이 많은 고객의 요구에 맞춰 Fig. 3과 같이 제품별 생산량과 제품납기를 조절하는 활동을 한다. 하지만, 이러한 생산계획에 대한 변동이 잦을수록 스케줄의 방향성이 제대로 가고 있는가에 대해 의구심을 품는 경향이 있다. 그러므로 본 연구는 설비에 디스패칭 된 Job에 사용된 파라메터의 영향도를 비지도학습으로 분석하여 각 파라메터별 영향도를 파악하고자 한다. 분석하는 방법으로는 영향을 주는 인자를 분석하는 주성분 분석(PCA), 클러스터링을 기반으로 예측할 수 있는 K-Means 군집화 알고리즘과 비지도학습의 결과에 Label을 붙여 스케줄러의 동작 방식에 대해 인사이트를 제공할 의사결정 트리를구축함으로써 생산라인이 어떠한 규칙에 의해 의사결정 되는지 보여주고자 한다.

RQ3. Job shop 방식의 반도체 생산의 복잡성을 반영한 스케줄러는 어떠한 요소에 의해 좌우되는가? 수백 개의 파라메터 중에서 Demand에 의한 진도, 설비 효율 최대화, 품질 향상 등 생산 라인을 움직이는 주요 파라메터는 무엇이 되는가? 또한 그들 간의 상관관계는 어떻게 정의할 수 있는가?

Job Shop 방식은 생산 과정에서 각 작업이 서로 다른 공정을 거쳐야 하고, 각 공정은 독립적으로 일정과 자원을 할당받는 생산 방식이다. 즉, 작업마다 필요한 기계와 진행 시간이 상이하며, 작업 간에 유연한 스케쥴링이 필요하다. 이 방식은 다양한 제품 흐름을 소화할 수 있는 유연한 생산 방식이며, 생산 라인에 다양한 작업이 동시에 존재하므로 효율적인 자원 배분과 스케쥴 관리가 중요하다.

본 연구에서 분석한 스케줄러는 파라메터가 300개 이상되고, 모델 라인에서 실제 사용되는 파라메터는 77개이다. 이 중 모든 Job에 해당되는 파라메터가 있기도 하고, 스케줄링에 주요한 영향을 끼치는 파라메터가 있기도 하다. 첫 번째로, 데이터 전처리를 통해 수집한 데이터를 파싱(Parsing)하여 파라메터 코드와 값으로 분류하고, 분류된 데이터에 모든 Job에 해당되는 값을 표준편차를 비교하여 노이즈와 영향력 있는 데이터를 분류하였다. 두번째로, 여러 정규화 방법을 통해 데이터 스케일링을 진행하여 데이터의 특성을 강조 시키는 방법을 통해 실험을 진행 하였다. 세번째로, 선정된 파라메터별 상관관계 분석을 통해 각각의 파라메터의 독립성을 확인하였다.

IV. Data Engineering

4.1 Data Parsing and Feature Selection

1. Data Parsing

RDBMS에 저장된 스케줄 로그는 기계와 JobID, Transaction Time, Job에 해당되는 제품과 프로세스, 스케줄된 이유를 나타내는 reason description으로 구성되어 있다. Reason description은 dispatching된 job 에 대한 분류 코드와 값 또는 단순히 이력을 표기하기 위한 상태값으로 구성되어 있다. 각 코드별로 내포하고 있는 값의 성질이 다르므로 다른 구분자로 구분되어 있어 파이썬의 re(Regular Expression) 라이브러리를 활용하여 데이터프레임을 구성하였다.

2. Feature selection

스케줄 로그 파싱(Parsing) 결과 총 77개의 파라메터가 추출되었다. 특징변수 선택에는 생산관리 현장 전문가가 지적한 11개의 특징변수와 LTV(Law of Total Variance)[15]에 따라 통계적으로 분산 σ^2(1)이 0에 가까운 특징변수 13개를 제거하였다.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \qquad (1)$$

이후에는 반도체 생산 전문가가 바라보았을 때 변별력이 떨어진다고 판단한 파라메터를 제거하여 유효한 파라메터를 추출하였다. 생산관리 전문가의 휴리스틱에 근거한 변별력의 설명은 특정 공정 영역에 특화된 변수와 공정 개선 평가에 사용되는 특수한 경우에 해당하는 변수를 의미한다.

최종 12개의 특징변수가 선택 되었으며 파라메터는 제 3장 1절 연구문제 1의 제품구성에 따른 생산진도, Job의 대기시간(FIFO, EDD), 설비효율(SMED, FLNQ) 및 제품의 품질의 4가지 카테고리로 구분할수 있으며, 각 파라메터의 영향성은 제 5장 실험 및 결과 분석에서 다루도록 한다.

V. Experiment Results

5.1 Data Correlation

스케줄 파라메터 간에는 연관이 있는 인자가 있기도 하고, 각각 독립적으로 작용하는 인자가 있기도 하다. 이를 알아보기 위해서 피어슨 상관계수 $\rho_{X,Y}$(2)를 이용하여 상관관계 분석을 진행하였다.

$$\rho_{X,Y} = \frac{\sum_{n}^{i}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{n}^{i}(X_i - \overline{X})^2}\sqrt{\sum_{n}^{i}(Y_i - \overline{Y})^2}} \qquad (2)$$

12개의 파라메터의 상관계수를 비교한 결과 Table 1의 상관계수의 절대값이 0.3 이하로 일반적으로 모든 특징변수는 독립적으로 작용하고 있다고 해석할 수 있으나, 상대적으로 W1-W4의 상관계수가 0.1 이상으로 다른 특징변수들보다 상관 관계가 있다고 해석 할 수 있다.

실제로, 생산관리 현장 전문가의 휴리스틱 관점에서 보면, W1과 W2-4 사이에서 음의 상관관계가 상대적으로 높은 것이 주목할 만하다. 왜냐하면, 생산 전문가의 의견에 따르면 이 두 파라메터(W1, W2-W4)는 실제로 생산 현장에서도 상호 배타적인 영향을 끼쳐 의사결정을 할 때 반드시 고려하는 요소이기 때문이다. 이는 Fig. 3에서 나타낸 바와 같이 생산 계획 또는 공정 제약 발생으로 인해 생산 지연이 발생되어 Job의 우선순위가 변경되면 설비 생산성에 영향을 끼친다는 Domain Knowledge와 일맥상통한다. 결론적으로, 제 2절-2.특징변수 선택에서 언급한 바와 같이 생산 스케줄에 영향을 주는 독립적인 인자가 잘 선정되었음을 상관관계 분석을 통해 확인 할 수 있었다.

5.2 Principal Component Analysis(PCA)

주성분 분석을 통해 특징변수 선택을 통해 선발된 12개의 파라메터 중 몇 개의 파라메터가 생산라인에 영향도가 큰 지 분석을 진행하였다. 특히, W1(Chamber별 효율), W2(Step별 생산 목표), W3(Blcok별 생산 목표)와 같은 파라메터들이 주성분 분석에서 중요한 영향을 미치는 것으로 나타났다. PCA 분석 결과, 이들 파라메터는 생산 계획과 공정 제약 발생에 중요한 역할을 하며, W1과 W2-4 사이의 상관관계가 상대적으로 높다는 상관계수 분석 결과와 일치한다. 이러한 파라메터들은 기계학습 모델에서 중요한 결정 요소로 작용하며, 주성분 분석에 따라 우선순위가 정해졌다. 각 파라메터의 정의는 Table 2와 같다.

주성분 분석 전에 원본 데이터를 정규화하여 파라메터별 스케일이 다른 이슈를 줄이고 이상치의 영향을 최소화하고자 하였다. Fig.4는 PCA 분석후 주성분 수를 선정하기 위해 분산 변화를 나타내는Scree Plot을 실험한 결과이다.

원본 데이터셋과 Robust Scaler는 2번째 특징변수에서 변곡점을 나타내어 현장의 의사결정 인자를 반영하기에

Table 1. Correlation between schedule weight parameters

Feature	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12
W1	1.00	-0.19	-0.25	-0.17	-0.07	-0.06	-0.03	-0.06	-0.05	-0.06	-0.02	-0.01
W2	-0.19	1.00	-0.21	-0.14	-0.06	-0.05	-0.03	-0.05	-0.04	-0.05	-0.02	-0.01
W3	-0.25	-0.21	1.00	-0.20	-0.08	-0.07	-0.04	-0.07	-0.06	-0.07	-0.03	-0.02
W4	-0.17	-0.14	-0.20	1.00	-0.06	-0.05	-0.03	-0.05	-0.04	-0.05	-0.02	-0.01
W5	-0.07	-0.06	-0.08	-0.06	1.00	-0.02	-0.01	-0.02	-0.02	-0.02	-0.01	0.00
W6	-0.06	-0.05	-0.07	-0.05	-0.02	1.00	-0.01	-0.02	-0.01	-0.02	-0.01	0.00
W7	-0.03	-0.03	-0.04	-0.03	-0.01	-0.01	1.00	-0.01	-0.01	-0.01	0.00	0.00
W8	-0.06	-0.05	-0.07	-0.05	-0.02	-0.02	-0.01	1.00	-0.01	-0.02	-0.01	0.00
W9	-0.05	-0.04	-0.06	-0.04	-0.02	-0.01	-0.01	-0.01	1.00	-0.01	-0.01	0.00
W10	-0.06	-0.05	-0.07	-0.05	-0.02	-0.02	-0.01	-0.02	-0.01	1.00	-0.01	0.00
W11	-0.02	-0.02	-0.03	-0.02	-0.01	-0.01	0.00	-0.01	-0.01	-0.01	1.00	0.00
W12	-0.01	-0.01	-0.02	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Table 2. Definition of selected features

Feature	Category	Description
W1	Efficiency	Consider the production efficiency of each chamber.
W2	Progress	Adjust the production target progress for each step.
W3	Progress	Control the production progress for each block(a group of continuous steps).
W4	Progress	Assign a weighted priority based on product characteristics.
W5	Job	Evaluate and assign weights to each lot based on priority.
W6	Efficiency	Reward cases where the recipe matches the current equipment operation.
W7	Efficiency	Evaluate overall chamber efficiency.
W8	Progress	Determine product priorities based on rule-based decision-making.
W9	Quality	Ensure the next process starts within a designated time to prevent quality degradation.
W10	Efficiency	Assign a score when the recipe group matches the equipment process.
W11	TAT (Turn Around Time)	Adjust the turnaround time for each lot.
W12	FIFO (First In, First Out)	Assign priority to jobs exceeding a certain waiting time.

적합하지 않으며, Standard Scaler 또한 모든 특징변수의 값이 평준화 되어 스케줄링의 특징을 나타내는 데 적합하지 않다. 반면에, minimax, max abs, uni-quantile, gs-quantile scaler와 L2 norm scaler는 PCA Component 4에서 변곡점을 가졌으므로, 이 모델의 주성분인자는 4개로 판단할 수 있다.

5.3 K-Means Clustering

"K-Means Clustering은 단순하지만 의사결정트리 학습을 수행하기에 효과적이다"[16]. 비지도 학습의 레이블을 정하기 위해 Table3, Table4와 같이 K값의 변화에 따른 클러스터의 군집성을 나타내는 실루엣 스코어와 클러스터 간의 분리도를 나타내는 DBI(Davies-Bouldin Index)를 측정 하였다. 일반적으로 실루엣 스코어가 0.5이상인 경우 군집화가 잘 되었다고 하고, DBI가 작을수록 클러스터간 분리가 잘 되었다고 평가한다.

Table4의 실루엣 스코어를 보면 robust scaling 외의 standard, minmax,max abs, uni-quantile, gs-quantile, L2 norm scaler에서 실루엣 스코어가 0.5 이상으로 K-Means Clustering에 적합한 정규화 방법으로 판단 할 수 있다. 하지만, DBI는 K값에 따라큰 편차를 보이지 않고 있어 주어진 데이터셋은 DBI로 K값을 결정하기는 어렵다고 볼 수 있다. 최적의 Cluster 값을 찾기 위해 Distortion score를 elbow 방법을 이용해 분석한 결과 Table5와 같은 결과가 나왔다. 원본 데이터셋과 표준 정규화 방법에서는 최적의 K값이 산출되지 않았으며, 다른 정규화 방법에서는 최적의 K값=4,5로 제 2절 비지도학습에 의한 주성분 분석의 Scree Plot과 비교하여 최적의 K는 4가 적절한 것으로 판단하였다. 이를 증명하기 위해 생산관리 전문가와 최적의 K값이 4와 5일경우를 비교하여 의사결정 트리를 구성하였다.

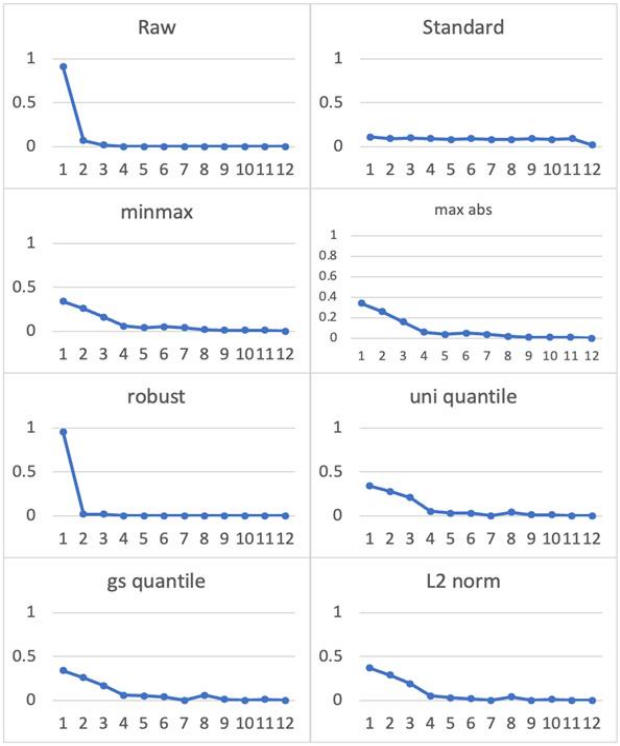

Fig. 4. Scree Plot by Normalization Methodologies

Table 3. K-Means Clustering, **Silhouette coefficient**

K value	2	3	4	5	6	7	8	9	10	11	12
raw	0.79	0.77	0.85	0.73	0.73	0.75	0.77	0.77	0.76	0.85	0.85
sc	0.27	0.31	0.42	0.47	0.51	0.53	0.57	0.59	0.6	0.64	0.65
minmax	0.52	0.58	0.6	0.63	0.66	0.59	0.48	0.49	0.5	0.73	0.49
maxabs	0.52	0.58	0.47	0.58	0.63	0.43	0.63	0.49	0.5	0.78	0.49
robust	0.02	0.8	-0.27	-0.24	-0.2	-0.17	-0.16	0.07	0.08	0.1	0.1
uni_quantile	0.36	0.38	0.4	0.42	0.62	0.67	0.56	0.57	0.7	0.72	0.72
gs_quantile	0.3	0.33	0.35	0.58	0.6	0.51	0.42	0.54	0.54	0.56	0.56
L2 norm	0.37	0.39	0.4	0.43	0.45	0.47	0.74	0.58	0.74	0.76	0.76

Table 4. K-Means Clustering, DBI

K value	2	3	4	5	6	7	8	9	10	11	12
raw	0.17	0.24	0.22	0.33	0.33	0.28	0.25	0.38	0.39	0.32	0.37
sc	0.79	0.6	0.66	0.55	0.47	0.4	0.35	0.41	0.35	0.29	0.48
minmax	0.46	0.54	0.48	0.44	0.4	0.53	0.48	0.45	0.42	0.36	0.45
maxabs	0.46	0.54	0.57	0.42	0.42	0.57	0.41	0.45	0.42	0.25	0.45
robust	1.23	0.21	1.26	1.46	1.16	1.3	1.4	1.4	0.28	1.09	1.03
uni_quantile	0.76	0.73	0.71	0.69	0.57	0.6	0.56	0.54	0.43	0.4	0.39
gs_quantile	0.86	0.87	0.77	0.68	0.6	0.6	0.61	0.53	0.51	0.49	0.49
L2 norm	0.75	0.74	0.74	0.72	0.7	0.7	0.57	0.57	0.44	0.41	0.4

Table 5. K-Means Clustering, optimal K(distortion elbow method)

Scaler	**raw**	**sc**	**minmax**	**maxabs**
Optimal K	no optimal	no optimal	4	4
Scaler	**robust**	**uni_quantile**	**gs_quantile**	**L2 norm**
Optimal K	5	4	4	4

5.4 Decision Tree

제 2절의 비지도 학습 결과를 통해 주어진 데이터셋에 K-Means Clustering 결과를 레이블로 반영하여 의사결정 트리를 이용해 스케줄러의 의사결정 규칙을 산출하였다. 이 때, K값은 4, 5로 K-Means 군집화에서 구한 값을 근거로 실험을 진행하였고, 반도체 생산 전문가의 해석 용이성을 위해 정규화 이전 원본 데이터셋으로 의사결정 트리의 임계 값(threshold)을 산출할 수 있도록 하였다. 그 결과 max abs, uni-quantile, gs-quantile, L2-norm 정규화 방법에서 K값에 상관없이 Fig. 5와 같이 의사결정 트리가 수렴하였다.

산출된 의사결정 트리는 Fig. 3. 반도체 생산관리 흐름도에서 볼 수 있듯이 사람 또는 시스템이 변경한 생산 우선순위 변경이 가장 최우선으로 스케줄러의 Job 선택에 이용되고, 차순위로는 설비 효율을 고려하여 스케줄링 한다는 것을 알수 있었다. Fig. 5 의사결정 트리에서 새롭게 발견한 점은 W4(우선순위2)의 특징변수 값이 W1(효율)의 특징변수 값보다 높은데도 불구하고 최종 Job의 결정에서는 W1을 우선해서 고려했다는 점이다. 이는 반도체 생산 전문가가 휴리스틱으로 추정하고 있던 현상으로 기존에 현업 담당자가 사용하는 데이터 시각화 분석툴인 Spotfire를 통해 공정군별로 데이터 분석을 진행 하였을 때는 눈에 띄지 않았던 부분이다.

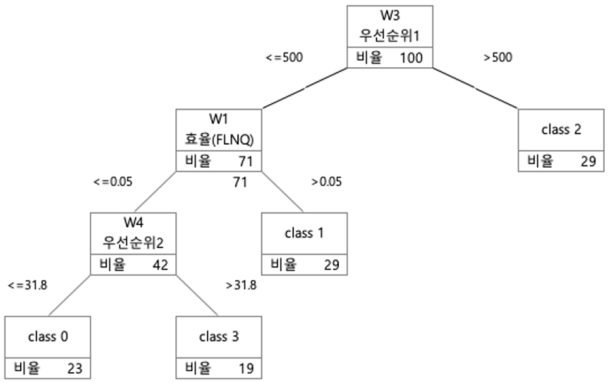

Fig. 5. Decision Tree

산출된 의사결정 트리를 클래스별로 정리하면다음과 같다.

1) Class0 : W3(생산 우선순위1)의 중요도가 낮으면 설비에서 진행중이니 Job이 다음공정군의 Job 공급에 문제가 없는지 확인한 후 W4(우선순위2)의 값이 31.8 이하면 스케줄한다. 이 때의 비율은 23% 생산관리 담당자가 설정한 우선순위가 낮은 제품의 Job이다. 이는 스케줄된 Job의 23%는 고객의 요구사항 변동이나 시장상황 변동에 관계 없이 스케줄 된다는 의미로 연구문제 1의 휴리스틱중 어떠한 것도 스

케줄 결과에 영향을 준 것은없다고 해석할 수 있다.

2) Class1 : W3이 낮고 설비 효율에 영향을 끼칠 만큼 설비에 Job 분배가 안되었으면(>0.05) 설비 효율을 우선으로 스케줄한다 .즉, 이 때 스케줄러의 의사결정 우선순위는 생산의 긴급도보다 설비 효율이 우선라고 할수 있다. 이 비율은 29%로 생산 우선순위 변동이 설비 효율에 영향을 주는 비율이라고 해석할 수있다. 즉 설비의 사용율을 고려하여 FLNQ가 의사결정의 키로 작용하였다.

3) Class2 : W3(생산 우선순위1)이 높으면(>500) 설비 효율과 상관없이 스케줄한다. 이 때 비율은 29%로 고객 요구사항 변동이 Fab 전체 스케줄 결과에 영향을 끼치는 비율이 29%라는 것을 의미한다. 또한, 사람의 우선순위 변동이 그만큼 시장상황의 변동에 의해 자주 일어난다고 할 수 있다. Class2에서 decision-making factor로 작용한 휴리스틱은 EDD로 제품 납기를 맞추기 위한 SCM(Supply Chain Management) 관리가 이루어지고 있다는 것을 의미한다.

4) Class3 : W3의 긴급도가 낮지만 효율에 영향을 최소한으로 끼치는 범위 내에서(<=0.05) W3(우선순위2)을 고려하여 스케줄 한다. 즉, 이는 다음순서의 Job 스케줄링으로 효율에 끼치는 영향을 복구 할 수 있다는 의미이다. 다시 말하면, W3이 기준치 이하(<=500)이면 효율이 다음 Job 스케줄에서 방어되는 선에서 W4를 고려하여 스케줄하는 비율은 19%라는 것이다. Class3에서는 EDD(W3)가 FLNQ보다 의사결정에 먼저 사용되지만 설비의 사용율 저하를 방어하는 동시에 EDD(W4)를 추가 결정인자로 사용함으로써 고객 납기 달성을 위해 스케줄러가 움직이고 있다고 해석 할 수 있다.

이번 연구를 통해 생산진도 전문가와 공통된 합의점이 있었던 부분은 Fig. 3과 같이 고객의요구사항 변동을 반영하여 생산 우선순위를 변경하면 스케줄링 시스템을 통해 현장에 반영되고 있었다는 점이다. 또한, 생산의 [우선순위1]이 변경되면 [효율]에 영향을 끼쳐 [우선순위2]의 값에 상관없이 효율에 관련된 특징변수가 스케줄러의 Job 선택에 영향을 준다는 점에서 설비 이용율(FLNQ)에 대해 자동화된 시스템이 의사결정을 하고 있다는것을 확인할 수 있었다.

V. Conclusions

본 연구는 반도체 생산 관리에 있어서 스케줄링에 영향을 미치는 주요 인자를 추출하는 것에 목적을 두었다. 반도체 스케줄링에 영향을 미치는 주요 인자를 추출하기 위해 데이터 클리닝 및 반도체 생산 전문가와의 논의를 통해 특징변수를 선택하였다. 선택된 12개의 특징변수로 피어슨 상관계수 분석, PCA, K-Means Clustering을 진행하였다. K-Means Clustering의 실루엣 스코어 실험 결과K값이 증가함에 따라 실루엣 스코어가 증가하여 각각의 특징변수는 독립적으로 동작한다는 것으로 증명되었고, 이는 4개의 파라메터에서만 약한 상관관계를 보이는 상관계수 분석에서도 확인 할 수 있었다. PCA 분석 결과 4개의 특징변수가 생산 라인의 스케줄링에 중요하게 영향을 끼친다는 것을 확인하였으나, K-Means 군집화와 의사결정 트리를 이용해 스케줄러의 의사결정 파라메터를 분석한 결과 3개의 파라메터가 실제 Job이 선택되는 의사결정에 있어서 주된 요인으로 작용하고 있음을 확인하였다. 이는 반도체 생산 전문가와 논의 했을 때 생산라인을 움직이는 주요 인자와도 일치한다는 것을 확인 할 수 있었다.

본 연구를 바탕으로 다음과 같이 결론을 도출할 수 있다.

첫 번째, 생산 계획의 변동이 자주 발생할수록 생산 우선순위가 빈번히 바뀌기 때문에 설비 효율에는 부정적인 영향을 끼치는 것이 스케줄러 측면에서도 증명되었다. 또한 설비 효율을 증가시키기 위해서는 생산 진도를 우선순위로 스케줄링 하는 방식에서 설비 효율(FLNQ, SMED) 위주의 스케줄 파라메터 설정이 필요하다는 공감대를 반도체 생산 전문가와 이룰 수 있었다.

두 번째, 스케줄링은 각기 다른 파라메터로 인해 작동된다. 피어슨 상관계수 분석 결과 각각의 특징변수는 상관관계가 아주 작거나 없었다. 다만 상대적으로 생산진도 우선순위와 효율 간의 음의 상관관계가 약하게 나타났고 이는 의사결정 트리의 결정 노드가 분리되면서 다른 의사결정 포인트로 작용하였다. 그러므로 스케줄 파라메터는 독립적으로 작용하고 이는 생산 전문가의 전략에 따라 Fab 스케줄 방향을 쉽게 바꿀 수 있다는 것을 의미한다.

세 번째, 연구문제에서 제시된 휴리스틱 중FIFO, EDD, NJF, SMED는 실제 생산 현장의 스케줄 의사결정에 끼치는 영향은 미미하였으며, 생산 진도 관리자가 조절하는 우선순위대로 생산을 지원함과 동시에 설비 이용율 저하에 대한 방어 또한 의사결정에 이용된다는 점이 증명되었다.

본 연구를 통해 Job Shop 방식의 반도체 생산 자체는 복잡하더라도, 기계학습을 통해 하나의 반도체 Fab을 움직이는 특징변수를 의사결정 트리로 분석한 결과 Fab이 움직이는 방향을 한 눈에 확인 할 수 있었다. 또한 의사결정 트리는 기계학습 전문가가 아니더라도 스케줄 결과에 대한 해석이 쉬워 생산 진도 및 스케줄러 전문가가 스케줄 파라메터 관리의 효율성을 향상시키는 데 기여할 수 있다.

결론적으로 본 연구는 실제 생산 현장에서 반도체 생산 스케줄의 의사결정 포인트를 파라메터화하여 의사결정 트

리를 구축한 것으로 스케줄러의 현 수준을 진단하고 향후 이를 활용한 연구 방향을 제시할 수 있다는 점에서 의의가 있다. 본 연구 결과를 토대로 스케줄러의 목적함수(Objective function)을 재정의 하여 시뮬레이션을 통해 실제 반도체 현장에 적합한 스케줄 솔루션을 찾는 것은 생산관리 분야에서 혁신을 이끌 수 있는 중요한 단계이다. 또한, 기존 스케줄러와 새로운 스케줄러의 비교를 통해 반도체 생산의 특성을 거시적 및 미시적 특성을 더욱 깊이 이해하고, 기존 시뮬레이션의 한계점인 속도 문제를 해결함으로써 생산 현장의 효율성을 크게 향상시킬 수 있을 것으로 기대된다. 이러한 접근은 반도체 산업의 지속 가능한 성장과 경쟁력 강화에 기여할 뿐만 아니라, 복잡한 제조 공정의 최적화를 위한 새로운 방향을 제시하는 데에도 중요한 역할을 할 것으로 예상한다.

하지만 본 연구에는 몇 가지 한계점이 있다. 첫째, 다양한 반도체 공장에 대한 케이스 스터디가 필요하다. 본 연구는 특정 공장의 데이터를 기반으로 진행되었기 때문에, 다양한 공장의 데이터를 활용하여 모델을 일반화할 수 있는 작업이 추가적으로 필요하다. 둘째, 데이터가 방대해질 경우 딥러닝 모델과 본 연구에서 제시한 방법을 비교하여 성능 분석이 필요하다. 데이터가 커질수록 딥러닝 모델의 성능이 기존 방법론보다 우수할 수 있으므로, 이를 비교하는 실험을 통해 최적의 방법론을 도출해야 한다.

이러한 한계점을 보완하기 위한 추후 연구에서는 다양한 반도체 공장의 데이터를 수집하고, 딥러닝을 포함한 최신 기법을 적용하여 보다 일반화된 모델을 개발하는 것이 중요하다. 또한, 스케줄링 시스템의 성능을 더욱 고도화하고, 딥러닝 모델을 통한 효율적인 의사결정을 구현하는 연구가 필요하다. 이를 통해 생산 관리의 혁신을 이루고, 효율적인 스케줄링 시스템을 구축할 수 있을 것이다.

ACKNOWLEDGEMENT

Seok-Won Lee's work was supported by the BK21 FOUR program of the National Research Foundation of Korea funded by the Ministry of Education(NRF5199991014091) and the Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2024-RS-2023-00255968) grant funded by the Korea government(MSIT)

REFERENCES

[1] A. Attajer, S. Darmoul, S. Chaabane, Y. Sallez, F. Riane, "An analytic hierarchy process augmented with expert rules for product driven control in cyber-physical manufacturing systems," Computers in Industry, Vol. 143, 2022.

[2] M. Sesana and G. Tavola, "Resilient manufacturing systems enabled by AI support to AR equipped operator," IEEE International Conference on, 2021.

[3] S. Bag, A. Telukdarie, J.H.C. Pretorius, S. Gupta, "Industry 4.0 and supply chain sustainability: framework and future research directions," Benchmarking: An International Journal, DOI: 10.1108/BIJ-03-2018-0056, 2018.

[4] S. Ayvaz, K. Alpay, "Predictive maintenance system for production lines in manufacturing: A machine learning approach using IoT data in real-time," Expert Systems with Applications, Vol. 173, 2021.

[5] J. Park, J. Chun, S.H. Kim, Y.K. Kim, J. Park, "Learning to schedule Job-Shop problems: Representation and policy learning using graph neural network and reinforcement learning," International Journal of Production Research, Vol. 59, Issue 11, pp. 3360-3377, 2021.

[6] M.R. Garey, D.S. Johnson, R. Sethi, "Complexity of flowshop and jobshop scheduling," Mathematics of Operations Research, Vol. 1, No. 2, pp. 117-129, 1976.

[7] T.P. Gros, J. Gros, V. Wolf, "Real-time decision making for a car manufacturing process using deep reinforcement learning," Winter Simulation Conference (WSC), pp. 3032-3044, 2020.

[8] R. Liu, R. Piplani, C. Toro, "Deep reinforcement learning for dynamic scheduling of a flexible job shop," International Journal of Production Research, Vol. 60, No. 13, pp. 4049-4069, 2022.

[9] A. Kuhnle, M.C. May, L. Schäfer, G. Lanza, "Explainable reinforcement learning in production control of job shop manufacturing system," International Journal of Production Research, Vol. 60, Issue 19, pp. 5812-5834, Oct 2022.

[10] S. Kim, K. Lee, "The paradigm shift of mass customization research," International Journal of Production Research, Vol. 61, No. 10, pp. 3350-3376, 2023.

[11] S. Arora, "Opening the black box of deep learning: Some lessons and take-aways," Keynote Talk, Abstract Proceedings of the 2021 ACM SIGMETRICS/International Conference on Measurement and Modeling of Computer Systems, pp. 1-1, 2021.

[12] T. Kim, "An empirical study on manufacturing process mining of smart factory," Journal of the Korea Safety Management & Science, Vol. 24, No. 4, pp. 149-156, 2022, DOI: 10.9723/JKSIIS.2020.25.2.073.

[13] S. Shingo, "A Revolution in Manufacturing: The SMED System," Routledge, 1985, DOI: 10.4324/9781315136479.

[14] S. Thomassey, A. Fiordaliso, "A hybrid sales forecasting system

based on clustering and decision trees," Decision Support Systems, Vol. 42, No. 1, pp. 408-421, Oct 2005, DOI: 10.1016/j.dss.2005.01.008.

[15] N.A. Mustapa, A. Senawi, C.Z. Liang, "Feature selection using law of total variance with fast correlation-based filter," IEEE 8th International Conference on Software Engineering and Computer Systems (ICSECS), 2023.

Authors

Raekyung Ahn received the B.S. Manufactu -ring System Engineering in Northumbria University, UK in 2008, M.S. degrees in Intelligent software of Information and Com -munication Technology at Ajou University, Korea, in 2024 and Ph.D student of Dept. of Applied Artificial Intelligence at Ajou University, Korea, respectively. She specializes in manufacturing scheduling, machine learning, and artificial intelligence applications in semiconductor production. Her research focuses on optimizing production scheduling using explainable AI techniques and developing automated decision-making models for smart manufacturing systems. She has experience in industrial engineering and intelligent software engineering, with a strong interest in integrating machine learning into real-world manufacturing environments.

Seok-Won Lee is currently a Full Professorand Chair in the Dept. of Software & Computer Engineering and Dept. of Applied Artificial Intelligence, and Vice President for International Affair at Ajou University, Republic of Korea since 2012. His areas of specialization include software engineering with specific expertise in ontological requirements engineering and domain modeling, and knowledge engineering with specific expertise in knowledge acquisition, machine learning and knowledge-based systems, and information assurance with specific expertise in software security & privacy. He haspublished more than 180 peer reviewed articles. He is a professional senior member of IEEE, ACM and AAAI.